REVERSAL OF FORTUNE: GEOGRAPHY AND INSTITUTIONS IN THE MAKING OF THE MODERN WORLD INCOME DISTRIBUTION*

DARON ACEMOGLU
SIMON JOHNSON
JAMES A. ROBINSON

Among countries colonized by European powers during the past 500 years, those that were relatively rich in 1500 are now relatively poor. We document this reversal using data on urbanization patterns and population density, which, we argue, proxy for economic prosperity. This reversal weighs against a view that links economic development to geographic factors. Instead, we argue that the reversal reflects changes in the institutions resulting from European colonialism. The European intervention appears to have created an "institutional reversal" among these societies, meaning that Europeans were more likely to introduce institutions encouraging investment in regions that were previously poor. This institutional reversal accounts for the reversal in relative incomes. We provide further support for this view by documenting that the reversal in relative incomes took place during the late eighteenth and early nineteenth centuries, and resulted from societies with good institutions taking advantage of the opportunity to industrialize.

I. INTRODUCTION

This paper documents a reversal in relative incomes among the former European colonies. For example, the Mughals in India and the Aztecs and Incas in the Americas were among the richest civilizations in 1500, while the civilizations in North America, New Zealand, and Australia were less developed. Today the United States, Canada, New Zealand, and Australia are an order of magnitude richer than the countries now occupying the territories of the Mughal, Aztec, and Inca Empires.

* We thank Joshua Angrist, Abhijit Banerjee, Olivier Blanchard, Alessandra Cassella, Jan de Vries, Ronald Findlay, Jeffry Frieden, Edward Glaeser, Herschel Grossman, Lawrence Katz, Peter Lange, Jeffrey Sachs, Andrei Shleifer, Fabrizio Zilibotti, three anonymous referees, and seminar participants at the All-Universities of California History Conference at Berkeley, the conference on "Globalization and Marginalization" in Bergen, The Canadian Institute of Advanced Research, Brown University, the University of Chicago, Columbia University, the University of Houston, Indiana University, Massachusetts Institute of Technology, National Bureau of Economic Research summer institute, Stanford University, the Wharton School of the University of Pennsylvania, and Yale University for useful comments. Acemoglu gratefully acknowledges financial help from The Canadian Institute for Advanced Research and the National Science Foundation Grant SES-0095253. Johnson thanks the Massachusetts Institute of Technology Entrepreneurship Center for support.

The Quarterly Journal of Economics, November 2002

Our main measure of economic prosperity in 1500 is urbanization. Bairoch [1988, Ch. 1] and de Vries [1976, p. 164] argue that only areas with high agricultural productivity and a developed transportation network can support large urban populations. In addition, we present evidence that both in the time series and the cross section there is a close association between urbanization and income per capita.[1] As an additional proxy for prosperity we use population density, for which there are relatively more extensive data. Although the theoretical relationship between population density and prosperity is more complex, it seems clear that during preindustrial periods only relatively prosperous areas could support dense populations.

With either measure, there is a negative association between economic prosperity in 1500 and today. Figure I shows a negative relationship between the percent of the population living in towns with more than 5000 inhabitants in 1500 and income per capita today. Figure II shows the same negative relationship between log population density (number of inhabitants per square kilometer) in 1500 and income per capita today. The relationships shown in Figures I and II are robust—they are unchanged when we control for continent dummies, the identity of the colonial power, religion, distance from the equator, temperature, humidity, resources, and whether the country is landlocked, and when we exclude the "neo-Europes" (the United States, Canada, New Zealand, and Australia) from the sample.

This pattern is interesting, in part, because it provides an opportunity to distinguish between a number of competing theories of the determinants of long-run development. One of the most popular theories, which we refer to as the "geography hypothesis," explains most of the differences in economic prosperity by geographic, climatic, or ecological differences across countries. The list of scholars who have emphasized the importance of geographic factors includes, inter alia, Machiavelli [1519], Mon-

1. By economic prosperity or income per capita in 1500, we do not refer to the economic or social conditions or the welfare of the masses, but to a measure of total production in the economy relative to the number of inhabitants. Although urbanization is likely to have been associated with relatively high output per capita, the majority of urban dwellers lived in poverty and died young because of poor sanitary conditions (see, for example, Bairoch [1988, Ch. 12]).

It is also important to note that the Reversal of Fortune refers to changes in relative incomes across different areas, and does not imply that the initial inhabitants of, for example, New Zealand or North America themselves became relatively rich. In fact, much of the native population of these areas did not survive European colonialism.

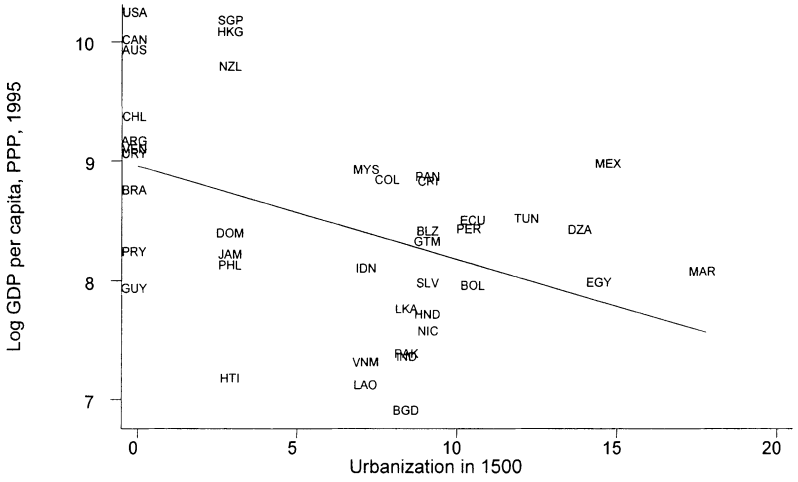


FIGURE I

Log GDP per Capita (PPP) in 1995 against Urbanization Rate in 1500

Note. GDP per capita is from the World Bank [1999]; urbanization in 1500 is people living in towns with more than 5000 inhabitants divided by total population, from Bairoch [1988] and Eggimann [1999]. Details are in Appendices 1 and 2.

tesquieu [1748], Toynbee [1934–1961], Marshall [1890], and Myrdal [1968], and more recently, Diamond [1997] and Sachs [2000, 2001]. The simplest version of the geography hypothesis emphasizes the time-invariant effects of geographic variables, such as climate and disease, on work effort and productivity, and therefore predicts that nations and areas that were relatively rich in 1500 should also be relatively prosperous today. The reversal in relative incomes weighs against this simple version of the geography hypothesis.

More sophisticated versions of this hypothesis focus on the time-varying effects of geography. Certain geographic characteristics that were not useful, or even harmful, for successful economic performance in 1500 may turn out to be beneficial later on. A possible example, which we call "the temperate drift hypothesis," argues that areas in the tropics had an early advantage, but later agricultural technologies, such as the heavy plow, crop rotation systems, domesticated animals, and high-yield crops, have favored countries in the temperate areas (see Bloch [1966], Lewis [1978], and White [1962]; also see Sachs [2001]). Although plausible, the temperate drift hypothesis cannot account for the

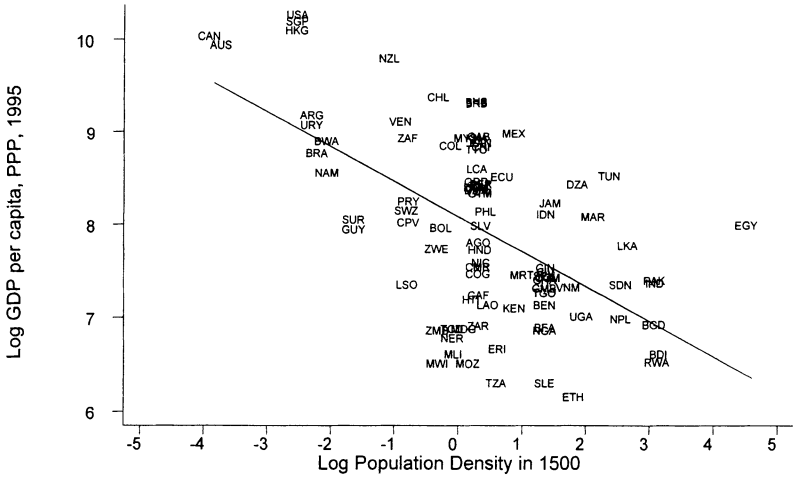


FIGURE II

Log GDP per Capita (PPP) against Log Population Density in 1500

Note. GDP per capita from the World Bank [1999]; log population density in 1500 from McEvedy and Jones [1978]. Details are in Appendix 2.

reversal. First, the reversal in relative incomes seems to be related to population density and prosperity before Europeans arrived, *not* to any inherent geographic characteristics of the area. Furthermore, according to the temperate drift hypothesis, the reversal should have occurred when European agricultural technology spread to the colonies. Yet, while the introduction of European agricultural techniques, at least in North America, took place earlier, the reversal occurred during the late eighteenth and early nineteenth centuries, and is closely related to industrialization. Another version of the sophisticated geography hypothesis could be that certain geographic characteristics, such as the presence of coal reserves or easy access to the sea, facilitated industrialization (e.g., Pomeranz [2000] and Wrigley [1988]). But we do not find any evidence that these geographic factors caused industrialization. Our reading of the evidence therefore provides little support to various sophisticated geography hypotheses either.

An alternative view, which we believe provides the best explanation for the patterns we document, is the "institutions hypothesis," relating differences in economic performance to the organization of society. Societies that provide incentives and opportunities for investment will be richer than those that fail to do so (e.g., North and Thomas [1973], North and Weingast [1989],

and Olson [2000]). As we discuss in more detail below, we hypothesize that a cluster of institutions ensuring secure property rights for a broad cross section of society, which we refer to as *institutions of private property,* are essential for investment incentives and successful economic performance. In contrast, *extractive institutions,* which concentrate power in the hands of a small elite and create a high risk of expropriation for the majority of the population, are likely to discourage investment and economic development. Extractive institutions, despite their adverse effects on aggregate performance, may emerge as equilibrium institutions because they increase the rents captured by the groups that hold political power.

How does the institutions hypothesis explain the reversal in relative incomes among the former colonies? The basic idea is that the expansion of European overseas empires starting at the end of the fifteenth century caused major changes in the organization of many of these societies. In fact, historical and econometric evidence suggests that European colonialism caused an "institutional reversal": European colonialism led to the development of institutions of private property in previously poor areas, while introducing extractive institutions or maintaining existing extractive institutions in previously prosperous places.[2] The main reason for the institutional reversal is that relatively poor regions were sparsely populated, and this enabled or induced Europeans to settle in large numbers and develop institutions encouraging investment. In contrast, a large population and relative prosperity made extractive institutions more profitable for the colonizers; for example, the native population could be forced to work in mines and plantations, or taxed by taking over existing tax and tribute systems. The expansion of European overseas empires, combined with the institutional reversal, is consistent with the reversal in relative incomes since 1500.

Is the reversal related to institutions? We document that the reversal in relative incomes from 1500 to today can be explained,

2. By the term "institutional reversal," we do not imply that it was societies with good institutions that ended up with extractive institutions after European colonialism. First, there is no presumption that relatively prosperous societies in 1500 had anything resembling institutions of private property. In fact, their relative prosperity most likely reflected other factors, and even perhaps geographic factors. Second, the institutional reversal may have resulted more from the emergence of institutions of private property in previously poor areas than from a deterioration in the institutions of previously rich areas.

at least statistically, by differences in institutions across countries. The institutions hypothesis also suggests that institutional differences should matter more when new technologies that require investments from a broad cross section of the society become available. We therefore expect societies with good institutions to take advantage of the opportunity to industrialize, while societies with extractive institutions fail to do so. The data support this prediction.

We are unaware of any other work that has noticed or documented this change in the distribution of economic prosperity. Nevertheless, many historians emphasize that in 1500 the Mughal, Ottoman, and Chinese Empires were highly prosperous, but grew slowly during the next 500 years (see the discussion and references in Section III).

Our overall interpretation of comparative development in the former colonies is closely related to Coatsworth [1993] and Engerman and Sokoloff [1997, 2000], who emphasize the adverse effects of the plantation complex in the Caribbean and Central America working through political and economic inequality,[3] and to our previous paper, Acemoglu, Johnson, and Robinson [2001a]. In that paper we proposed the disease environment at the time Europeans arrived as an instrument for European settlements and the subsequent institutional development of the former colonies, and used this to estimate the causal effect of institutional differences on economic performance. Our thesis in the current paper is related, but emphasizes the influence of population density and prosperity on the policies pursued by the Europeans (see also Engerman and Sokoloff [1997]). In addition, here we document the reversal in relative incomes among the former colonies, show that it was related to industrialization, and provide evidence that the interaction between institutions and the opportunity to industrialize during the nineteenth century played a central role in the long-run development of the former colonies.[4]

3. In this context, see also Frank [1978], Rodney [1972], Wallerstein [1974–1980], and Williams [1944].

4. Our results are also relevant to the literature on the relationship between population and growth. The recent consensus is that population density encourages the discovery and exchange of ideas, and contributes to growth (e.g., Boserup [1965], Jones [1997], Kremer [1993], Kuznets [1968], Romer [1986], and Simon [1977]). Our evidence points to a major historical episode of 500 years where high population density was detrimental to economic development, and therefore sheds doubt on the general applicability of this recent consensus.

The rest of the paper is organized as follows. The next section discusses the construction of urbanization and population density data, and provides evidence that these are good proxies for economic prosperity. Section III documents the "Reversal of Fortune"—the negative relationship between economic prosperity in 1500 and income per capita today among the former colonies. Section IV discusses why the simple and sophisticated geography hypotheses cannot explain this pattern, and how the institutions hypothesis explains the reversal. Section V documents that the reversal in relative incomes reflects the institutional reversal caused by European colonialism, and that institutions started playing a more important role during the age of industry. Section VI concludes.

II. Urbanization and Population Density

II.A. Data on Urbanization

Bairoch [1988] provides the best single collection and assessment of urbanization estimates. Our base data for 1500 consist of Bairoch's [1988] urbanization estimates augmented by the work of Eggimann [1999]. Merging the Eggimann and Bairoch series requires us to convert Eggimann's estimates, which are based on a minimum population threshold of 20,000, into Bairoch-equivalent urbanization estimates, which use a minimum population threshold of 5000. We use a number of different methods to convert between the two sets of estimates, all with similar results. Appendix 1 provides details about data sources and construction. Briefly, for our base estimates, we run a regression of Bairoch estimates on Eggimann estimates for all countries where they overlap in 1900 (the year for which we have most Bairoch estimates for non-European countries). This regression yields a constant of 6.6 and a coefficient of 0.67, which we use to generate Bairoch-equivalent urbanization estimates from Eggimann's estimates.

Alternatively, we converted the Eggimann's numbers using a uniform conversion rate of 2 as suggested by Davis' and Zipf's Laws (see Appendix 1 and Bairoch [1988, Ch. 9]), and also tested the robustness of the estimates using conversion ratios at the regional level based on Bairoch's analysis. Finally, we constructed three alternative series without combining estimates from different sources. One of these is based on Bairoch, the

second on Eggimann, and the third on Chandler [1987]. All four alternative series are reported in Appendix 3, and results using these measures are reported in Table IV.

While the data on sub-Saharan Africa are worse than for any other region, it is clear that urbanization in sub-Saharan Africa before 1500 was at a higher level than in North America or Australia. Bairoch, for example, argues that by 1500 urbanization was "well-established" in sub-Saharan Africa.[5] Because there are no detailed urbanization data for sub-Saharan Africa, we leave this region out of the regression analysis when we use urbanization data, although African countries are included in our regressions using population density.

Table I gives descriptive statistics for the key variables of interest, separately for the whole world, for the sample of ex-colonies for which we have urbanization data in 1500, and for the sample of ex-colonies for which we have population density data in 1500. Appendix 2 gives detailed definitions and sources for the variables used in this study.

II.B. Urbanization and Income

There are good reasons to presume that urbanization and income are positively related. Kuznets [1968, p. 1] opens his book on economic growth by stating: "we identify the economic growth of nations as a sustained increase in per-capita or per-worker product, most often accompanied by an increase in population and usually by sweeping structural changes. . . . in the distribution of population between the countryside and the cities, the process of urbanization."

Bairoch [1988] points out that during preindustrial periods a large fraction of the agricultural surplus was likely to be spent on transportation, so both a relatively high agricultural surplus and a developed transport system were necessary for large urban populations (see Bairoch [1988, Ch. 1]). He argues "the existence of true urban centers presupposes not only a surplus of agricul-

5. Sahelian trading cities such as Timbuktu, Gao, and Djenne (all in modern Mali) were very large in the middle ages with populations as high as 80,000. Kano (in modern Nigeria) had a population of 30,000 in the early nineteenth century, and Yorubaland (also in Nigeria) was highly urbanized with a dozen towns with populations of over 20,000 while its capital Ibadan possibly had 70,000 inhabitants. For these numbers and more detail, see Hopkins [1973, Ch. 2].

TABLE I
DESCRIPTIVE STATISTICS

	Whole world (1)	Base sample for urbanization (2)	Base sample for population density (3)	Below median urbanization in 1500 (4)	Above median urbanization in 1500 (5)	Below median population density in 1500 (6)	Above median population density in 1500 (7)
Log GDP per capita (PPP) in 1995	8.3	8.5	7.9	8.8	8.1	8.3	7.5
	(1.1)	(0.9)	(1.0)				
Urbanization in 1995	53.0	57.5	45.4	64.9	49.7	53.5	36.7
	(23.8)	(22.4)	(22.2)				
Urbanization in 1500	7.3	6.4	6.4	2.4	10.5	2.3	9.5
	(5.0)	(5.0)	(5.0)				
Log population density in 1500	1.0	0.2	0.5	−0.9	1.4	−0.6	1.6
	(1.6)	(1.9)	(1.5)				
Population density in 1500	9.2	6.3	4.8	1.2	11.7	0.8	9.1
	(24.3)	(16.4)	(11.7)				
Log population density in 1000	0.6	0.11	0.08	−1.20	1.22	−0.94	1.04
	(1.5)	(2.0)	(1.5)				
Average protection against expropriation, 1985–1995	7.1	6.9	6.5	7.5	6.3	6.8	6.2
	(1.8)	(1.5)	(1.4)				
Constraint on the executive in 1990	3.6	4.9	3.7	5.1	4.6	4.0	3.5
	(2.3)	(2.1)	(2.3)				
Constraint on the executive in first year of independence	3.6	3.3	3.4	3.8	2.8	3.6	3.3
	(2.4)	(2.5)	(2.3)				
European settlements in 1900	29.6	23.2	12.5	30.5	6.0	18.7	4.7
	(41.7)	(28.7)	(22.1)				
Number of observations	162	41	91	21	20	47	44

Standard deviations are in parentheses. Number of observations varies across rows due to missing data. The first three columns report mean values for the sample indicated at the head of the column. The last four columns report mean values for former colonies below and above the median, separately for the base urbanization and population density samples. For detailed sources and descriptions see Appendix 2.

tural produce, but also the possibility of using this surplus in trade" [p. 11].[6] See de Vries [1976, p. 164] for a similar argument.

We supplement this argument by empirically investigating the link between urbanization and income in Table II. Columns (1)–(6) present cross-sectional regressions. Column (1) is for 1900, the earliest date for which we have data on urbanization and income per capita for a large number of countries. The regression coefficient, 0.038, is highly significant, with a standard error of 0.006. It implies that a country with 10 percentage points higher urbanization has, on average, 46 percent (38 log points) greater income per capita (throughout the paper, all urbanization rates are expressed in percentage points, e.g., 10 rather than 0.1—see Table I). Column (2) reports a similar result using data for 1950. Column (3) uses current data and shows that even today there is a strong relationship between income per capita and urbanization for a large sample of countries. The coefficient is similar, 0.036, and precisely estimated, with a standard error of 0.002. This relationship is shown diagrammatically in Figure III.

Below, we draw a distinction between countries colonized by Europeans and those never colonized (i.e., Europe and non-European countries not colonized by Western Europe). Columns (4) and (5) report the same regression separately for these two samples. The estimates are very similar: 0.037 for the former colonies sample, and 0.033 for the rest of the countries. Finally, in column (6) we add continent dummies to the same regression. This leads to only a slightly smaller coefficient of 0.030, with a standard error of 0.002.

Finally, we use estimates from Bairoch [1978, 1988] to construct a small unbalanced panel data set of urbanization and income per capita from 1750 to 1913. Column (7) reports a re-

6. The view that urbanization and income (productivity) are closely related is shared by many other scholars. See Ades and Glaeser [1999], De Long and Shleifer [1993], Tilly and Blockmans [1994], and Tilly [1990]. De Long and Shleifer, for example, write "The larger preindustrial cities were nodes of information, industry, and exchange in areas where the growth of agricultural productivity and economic specialization had advanced far enough to support them. They could not exist without a productive countryside and a flourishing trade network. The population of Europe's preindustrial cities is a rough indicator of economic prosperity" [p. 675].

A large history literature also documents how urbanization accelerated in Europe during periods of economic expansion (e.g., Duby [1974], Pirenne [1956], and Postan and Rich [1966]). For example, the period between the beginning of the eleventh and mid-fourteenth centuries is an era of rapid increase in agricultural productivity and industrial output. The same period also witnessed a proliferation of cities. Bairoch [1988], for example, estimates that the number of cities with more than 20,000 inhabitants increased from around 43 in 1000 to 107 in 1500 [Table 10.2, p. 159].

TABLE II

URBANIZATION AND PER CAPITA INCOME

	Cross-sectional regression in 1913, all countries (1)	Cross-sectional regression in 1950, all countries (2)	Cross-sectional regression in 1995, all countries (3)	Cross-sectional regression in 1995, only for ex-colonies (4)	Cross-sectional regression in 1995, never colonized countries only (5)	Cross-sectional regression in 1995, all countries, with continent dummies (6)	Panel data set through 1913 (7)
	Dependent variable is log GDP per capita						
Urbanization	0.038	0.026	0.036	0.037	0.033	0.030	0.026
	(0.006)	(0.002)	(0.002)	(0.003)	(0.007)	(0.002)	(0.004)
R^2	0.69	0.57	0.63	0.69	0.34	0.68	0.93
Number of observations	22	128	162	93	51	162	55

Standard errors are in parentheses. Log GDP per capita through 1913 is from Bairoch [1978]. Urbanization is percent of population living in towns with at least 5000 people, from Bairoch [1988] through 1900 with supplementary sources as described in Appendix 1. Log GDP per capita in 1950 is from Maddison [1995]; this regression uses urbanization in 1960 from the World Bank's *World Development Indicators* [1999]. Log GDP per capita (PPP) and Urbanization data for 1995 are from the World Bank's *World Development Indicators* [1999]. Population density is total population divided by arable land area, both from McEvedy and Jones [1978]. For detailed sources and descriptions see Appendix 2.

The countries and approximate years for which we have data (used in the unbalanced panel regression in column (7)) are Australia (1830, 1860, and 1913), Austria (1830, 1860, 1913), Belgium (1830, 1860, 1913), Britain (1750, 1830, 1860, 1913), Bulgaria (1860, 1913), Canada (1830, 1860, 1913), China (1830, 1860), Denmark (1830, 1860, 1913), Finland (1830, 1860, 1913), France (1750, 1830, 1860, 1913), Germany (1830, 1860, 1913), Greece (1860, 1913), India (1830, 1913), Italy (1830, 1860, 1913), Jamaica (1830, 1913), Japan (1750, 1830, 1913), Netherlands (1830, 1860, 1913), Norway (1830, 1860, 1913), Portugal (1830, 1860, 1913), Romania (1830, 1860, 1913), Russia (1750, 1830, 1860, 1913), Spain (1830, 1860, 1913), Sweden (1830, 1860, 1913), Switzerland (1830, 1860, 1913), United States (1750, 1830, 1860, 1913), and Yugoslavia (1830, 1860, 1913).

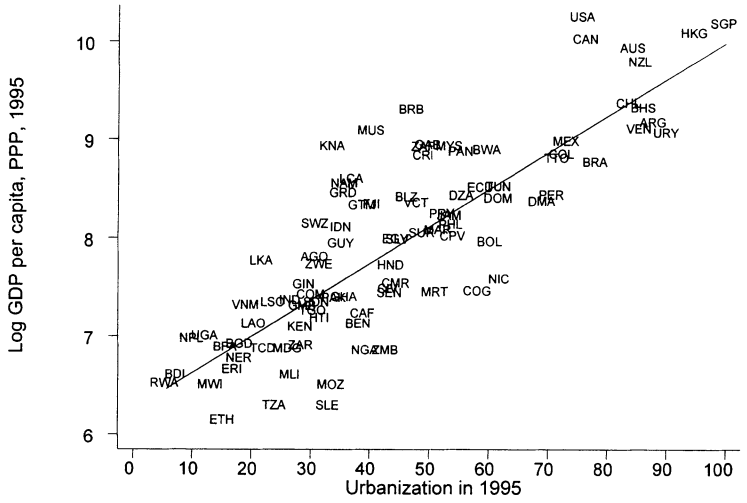


FIGURE III

Log GDP per Capita (PPP) in 1995 against the Urbanization Rate in 1995

Note. GDP per capita and urbanization are from the World Bank [1999]. Urbanization is percent of population living in urban areas. The definition of urban areas differs between countries, but the usual minimum size is 2000–5000 inhabitants. For details of definitions and sources for urban population in 1995, see the United Nations [1998].

gression of income per capita on urbanization using this panel data set and controlling for country and period dummies. The estimate is again similar: 0.026 (s.e. = 0.004). Overall, we conclude that urbanization is a good proxy for income.

II.C. Population Density and Income

The most comprehensive data on population since 1 A.D. come from McEvedy and Jones [1978]. They provide estimates based on censuses and published secondary sources. While some individual country numbers have since been revised and others remain contentious (particularly for pre-Columbian Meso-America), their estimates are consistent with more recent research (see, for example, the recent assessment by the Bureau of the Census, www.census.gov/ipc/www/worldhis.html). We use McEvedy and Jones [1978] for our baseline estimates, and test the effect of using alternative assumptions (e.g., lower or higher population estimates for Mexico and its neighbors before the arrival of Cortes).

We calculate population density by dividing total population by arable land (also estimated by McEvedy and Jones). This excludes primarily desert, inland water, and tundra. As much as possible, we use the land area of a country at the date we are considering.

The theoretical relationship between population density and income is more nuanced than that between urbanization and income. With a similar reasoning, it seems natural to think that only relatively rich areas could afford dense populations (see Bairoch [1988, Ch. 1]). This is also in line with Malthus' classic work. Malthus [1798] argued that high productivity increases population by raising birthrates and lowering death rates. However, the main thrust of Malthus' work was how a higher than equilibrium level of population increases death rates and reduces birthrates to correct itself.[7] A high population could therefore be reflecting an "excess" of population, causing low income per capita. So caution is required in interpreting population density as a proxy for income per capita.

The empirical evidence regarding the relationship between population density and income is also less clear-cut than the relationship between urbanization and income. In Acemoglu, Johnson, and Robinson [2001b] we documented that population density and income per capita increased concurrently in many instances. Nevertheless, there is no similar cross-sectional relationship in recent data, most likely because of the demographic transition—it is no longer true that high population density is associated with high income per capita because the relationship between income and the number of children has changed (e.g., Notestein [1945] or Livi-Bacci [2001]).

Despite these reservations, we present results using population density, as well as urbanization, as a proxy for income per capita. This is motivated by three considerations. First, population density data are more extensive, so the use of population density data is a useful check on our results using urbanization data. Second, as argued by Bairoch, population density is closely

7. A common interpretation of Malthus' argument is that these population dynamics will force all countries down to the subsistence level of income. In that case, population density would be a measure of total income, but not necessarily of income per capita, and in fact, there would be no systematic (long-run) differences in income per capita across countries. We view this interpretation as extreme, and existing historical evidence suggests that there were systematic differences in income per capita between different regions even before the modern period (see the references below).

related to urbanization, and in fact, our measures are highly correlated. Third, variation in population density will play an important role not only in documenting the reversal, but also in explaining it.

III. The Reversal of Fortune

III.A. Results with Urbanization

This section presents our main results. Figure I in the introduction depicts the relationship between urbanization 1500 and income per capita today. Table III reports regressions documenting the same relationship. Column (1) is our most parsimonious specification, regressing log income per capita in 1995 (PPP basis) on urbanization rates in 1500 for our sample of former colonies. The coefficient is -0.078 with a standard error of 0.026.[8] This coefficient implies that a 10 percentage point lower urbanization in 1500 is associated with approximately twice as high GDP per capita today (78 log points $\approx$ 108 percent). It is important to note that this is not simply mean reversion—i.e., richer than average countries reverting back to the mean. It is a *reversal.* To illustrate this, let us compare Uruguay and Guatemala. The native population in Uruguay had no urbanization, while, according to our baseline estimates Guatemala had an urbanization rate of 9.2 percent. The estimate in column (1) of Table II, 0.038, for the relationship between income and urbanization implies that Guatemala at the time was approximately 42 percent richer than Uruguay ($\exp(0.038 \times 9.2) - 1 \approx 0.42$). According to our estimate in column (1) of Table III, we expect Uruguay today to be 105 percent richer than Guatemala ($\exp(0.078 \times 9.2) - 1 \approx 1.05$), which is approximately the current difference in income per capita between these two countries.[9]

The second column of Table III excludes North African countries for which data quality may be lower. The result is un-

8. Because China was never a formal colony, we do not include it in our sample of ex-colonies. Adding China does not affect our results. For example, with China, the baseline estimate changes from -0.078 (s.e. = 0.026) to -0.079 (s.e. = 0.025). Furthermore, our sample excludes countries that were colonized by European powers briefly during the twentieth century, such as Iran, Saudi Arabia, and Syria. If we include these observations, the results are essentially unchanged. For example, the baseline estimate changes to -0.072 (s.e. = 0.024).

9. Interestingly, these calculations suggest that not only have relative rankings reversed since 1500, but income differences are now much larger than in 1500.

changed, with a coefficient of −0.101 and standard error of 0.032. Column (3) drops the Americas, which increases both the coefficient and the standard error, but the estimate remains highly significant. Column (4) reports the results just for the Americas, where the relationship is somewhat weaker but still significant at the 8 percent level. Column (5) adds continent dummies to check whether the relationship is being driven by differences across continents. Although continent dummies are jointly significant, the coefficient on urbanization in 1500 is unaffected—it is −0.083 with a standard error of 0.030.

One might also be concerned that the relationship is being driven mainly by the neo-Europes: United States, Canada, New Zealand, and Australia. These countries are settler colonies built on lands that were inhabited by relatively undeveloped civilizations. Although the contrast between the development experiences of these areas and the relatively advanced civilizations of India or Central America is of central importance to the reversal and to our story, one would like to know whether there is anything more than this contrast in the results of Table III. In column (6) we drop these observations. The relationship is now weaker, but still negative and statistically significant at the 7 percent level.

In column (7) we control for distance from the equator (the absolute value of latitude), which does not affect the pattern of the reversal—the coefficient on urbanization in 1500 is now −0.072 instead of −0.078 in our baseline specification. Distance from the equator is itself insignificant. Column (8), in turn, controls for a variety of geography variables that represent the effect of climate, such as measures of temperature, humidity, and soil type, with little effect on the relationship between urbanization in 1500 and income per capita today. The R^2 of the regression increases substantially, but this simply reflects the addition of sixteen new variables to this regression (the adjusted R^2 increases only slightly, to 0.27).

In column (9) we control for a variety of "resources" which may have been important for post-1500 development. These include dummies for being an island, for being landlocked, and for having coal reserves and a variety of other natural resources (see Appendix 2 for detailed definitions and sources). Access to the sea may have become more important with the rise of trade, and availability of coal or other natural resources may have different effects at different points in time. Once again, the addition of these variables has no effect on the pattern of the reversal.

TABLE III
URBANIZATION IN 1500 AND GDP PER CAPITA IN 1995 FOR FORMER EUROPEAN COLONIES

	Dependent variable is log GDP per capita (PPP) in 1995										
	Base sample (1)	Without North Africa (2)	Without the Americas (3)	Just the Americas (4)	With continent dummies (5)	Without neo-Europes (6)	Controlling for latitude (7)	Controlling for climate (8)	Controlling for resources (9)	Controlling for colonial origin (10)	Controlling for religion (11)
Urbanization in 1500	−0.078	−0.101	−0.115	−0.053	−0.083	−0.046	−0.072	−0.088	−0.058	−0.071	−0.060
	(0.026)	(0.032)	(0.051)	(0.029)	(0.030)	(0.026)	(0.025)	(0.030)	(0.029)	(0.028)	(0.033)
Asia dummy					−1.33						
					(0.61)						
Africa dummy					−0.53						
					(0.77)						
America dummy					−0.96						
					(0.57)						
Latitude							1.42				
							(0.92)				
P-value for temperature								[0.51]			
P-value for humidity								[0.40]			
P-value for soil quality								[0.96]			
P-value for resources									[0.16]		

Landlocked									−0.54		
									(0.48)		
Island									0.27		
									(0.33)		
Coal									0.11		
									(0.28)		
Former French colony										−0.59	
										(0.39)	
Former Spanish colony										0.06	
										(0.29)	
P-value for religion											[0.47]
R^2	0.19	0.22	0.26	0.13	0.32	0.09	0.24	0.53	0.45	0.27	0.25
Number of observations	41	37	17	24	41	37	41	41	41	41	41

Standard errors are in parentheses. *P*-values from *F*-tests for joint significance are in square brackets. Dependent variable is log GDP per capita (PPP) in 1995. Base sample is all former colonies for which we have data. Urbanization in 1500 is percent of the population living in towns with 5000 or more inhabitants. The regression that includes continent dummies has Oceania as the base category. The neo-Europes are the United States, Canada, Australia, and New Zealand.

In the "climate" regression we include five measures of temperature, four measures of humidity, and seven measures of soil quality. In the "resources" regression we include reserves of gold, iron, zinc, silver, and oil. Coal is a dummy for the presence of coal, landlocked is a dummy for not having access to the sea, and island is a dummy for being an island. The regression that controls for colonial origin includes dummies for former French colony, Spanish colony, Portuguese colony, Belgian colony, Italian colony, German colony, and Dutch colony. British colonies are the base category. The religion variables are percent of the population who are Muslim, Catholic, and "other"; percent Protestant is the base category. For detailed sources and descriptions see Appendix 2.

Finally, in columns (10) and (11) we add the identity of the colonial power and religion, which also have little effect on our estimate, and are themselves insignificant.

The urbanization variable used in Table III relies on work by Bairoch and Eggimann. In Table IV we use data from Bairoch and Eggimann separately, as well as data from Chandler, who provided the starting point for Bairoch's data. We report a subset of the regressions from Table III using these three different series and an alternative series using the Davis-Zipf adjustment to convert Eggimann's estimates into Bairoch-equivalent numbers (explained in Appendix 1). The results are very similar to the baseline estimates reported in Table III: in all cases, there is a negative relationship between urbanization in 1500 and income per capita today, and in almost all cases, this relationship is statistically significant at the 5 percent level (the full set of results are reported in Acemoglu, Johnson, and Robinson [2001b]).

III.B. Results with Population Density

In Panel A of Table V we regress income per capita today on log population density in 1500, and also include data for sub-Saharan Africa. The results are similar to those in Table IV (also see Figure II). In all specifications we find that countries with higher population density in 1500 are substantially poorer today. The coefficient of -0.38 in column (1) implies that a 10 percent higher population density in 1500 is associated with a 4 percent lower income per capita today. For example, the area now corresponding to Bolivia was seven times more densely settled than the area corresponding to Argentina; so on the basis of this regression, we expect Argentina to be three times as rich as Bolivia, which is more or less the current gap in income between these countries.[10]

The remaining columns perform robustness checks, and show that including a variety of controls for geography and resources, the identity of the colonial power, religion variables, or dropping the Americas, the neo-Europes, or North Africa has very

10. The magnitudes implied by the estimates in this table are similar to those implied by the estimates in Table III. For example, the difference in the urbanization rate between an average high and low urbanization country in 1500 is 8.1 (see columns (4) and (5) in Table I), which using the coefficient of -0.078 from Table III translates into a $0.078 \times 8.1 \approx 0.63$ log points difference in current GDP. The difference in log population density between an average high-density and low-density country in 1500 is 2.2 (see columns (6) and (7) in Table I), which translates into a $0.38 \times 2.2 \approx 0.84$ log points difference in current GDP.

TABLE IV
ALTERNATIVE MEASURES OF URBANIZATION

	Dependent variable is log GDP per capita (PPP) in 1995				
	Base sample (1)	With continent dummies (2)	Without neo-Europes (3)	Controlling for latitude (4)	Controlling for resources (5)
Panel A: Using our base sample measure of urbanization					
Urbanization in 1500	−0.078	−0.083	−0.046	−0.072	−0.058
	(0.026)	(0.030)	(0.026)	(0.025)	(0.029)
R^2	0.19	0.32	0.09	0.24	0.45
Number of observations	41	41	37	41	41
Panel B: Using only Bairoch's estimates					
Urbanization in 1500	−0.126	−0.107	−0.089	−0.116	−0.092
	(0.032)	(0.034)	(0.033)	(0.036)	(0.037)
R^2	0.30	0.37	0.19	0.31	0.49
Number of observations	37	37	33	37	37
Panel C: Using only Eggimann's estimates					
Urbanization in 1500	−0.041	−0.043	−0.022	−0.036	−0.022
	(0.019)	(0.019)	(0.018)	(0.019)	(0.023)
R^2	0.10	0.28	0.04	0.16	0.39
Number of observations	41	41	37	41	41
Panel D: Using only Chandler's estimates					
Urbanization in 1500	−0.057	−0.072	−0.040	−0.054	−0.049
	(0.019)	(0.021)	(0.019)	(0.019)	(0.025)
R^2	0.27	0.43	0.17	0.34	0.66
Number of observations	26	26	23	26	26
Panel E: Using Davis-Zipf Adjustment for Eggimann's series					
Urbanization in 1500	−0.039	−0.048	−0.024	−0.040	−0.031
	(0.015)	(0.020)	(0.014)	(0.015)	(0.017)
R^2	0.14	0.30	0.08	0.23	0.44
Number of observations	41	41	37	41	41

Standard errors are in parentheses. Dependent variable is log GDP per capita (PPP) in 1995. Base sample is all former colonies for which we have data. Urbanization in 1500 is percent of the population living in towns with 5000 or more people. In Panels B, C, D, and E, we use, respectively, Bairoch's estimates, Eggimann's estimates, Chandler's estimates, and a conversion of Eggimann's estimates into Bairoch-equivalent numbers using the Davis-Zipf adjustment. Eggimann's estimates (Panel C) and Chandler's estimates (Panel D) are not converted to Bairoch-equivalent units. The continent dummies, neo-Europes, and resources measures are as described in the note to Table III. For detailed sources and descriptions see Appendix 2. The alternative urbanization series are shown in Appendix 3.

little effect on the results. In all cases, log population density in 1500 is significant at the 1 percent level (although now some of the controls, such as the humidity dummies, are also significant).

TABLE V
POPULATION DENSITY AND GDP PER CAPITA IN FORMER EUROPEAN COLONIES

	Dependent variable is log GDP per capita (PPP) in 1995										
	Base sample (1)	Without Africa (2)	Without the Americas (3)	Just the Americas (4)	With continent dummies (5)	Without neo-Europes (6)	Controlling for latitude (7)	Controlling for climate (8)	Controlling for resources (9)	Controlling for colonial origin (10)	Controlling for religion (11)
Panel A: Log population density in 1500 as independent variable											
Log population density in 1500	−0.38 (0.06)	−0.40 (0.05)	−0.32 (0.07)	−0.25 (0.09)	−0.26 (0.05)	−0.32 (0.06)	−0.33 (0.06)	−0.31 (0.06)	−0.30 (0.06)	−0.32 (0.06)	−0.37 (0.07)
Asia dummy					−0.91 (0.55)						
Africa dummy					−1.67 (0.52)						
America dummy					−0.69 (0.51)						
Latitude							2.09 (0.74)				
P-value for temperature								[0.18]			
P-value for humidity								[0.00]			
P-value for soil quality								[0.10]			
P-value for natural resources									[0.34]		
Landlocked									−0.58 (0.23)		
Island									0.62 (0.23)		

Coal									0.01		
									(0.19)		
Former French colony										−0.48	
										(0.20)	
Former Spanish colony										0.25	
										(0.22)	
P-value for religion											[0.73]
R^2	0.34	0.55	0.27	0.22	0.56	0.24	0.40	0.59	0.54	0.48	0.36
Number of observations	91	47	58	33	91	87	91	90	85	91	85
Panel B: Log population and log land in 1500 as separate independent variables											
Log population in 1500	−0.34	−0.30	−0.32	−0.13	−0.23	−0.27	−0.29	−0.27	−0.27	−0.28	−0.31
	(0.05)	(0.05)	(0.07)	(0.07)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)
Log arable land in 1500	0.26	0.27	0.21	0.16	0.18	0.15	0.20	0.20	0.08	0.21	0.24
	(0.06)	(0.06)	(0.09)	(0.06)	(0.05)	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.07)
R^2	0.35	0.45	0.31	0.17	0.55	0.31	0.41	0.59	0.55	0.47	0.36
Number of observations	91	47	58	33	91	87	91	90	85	91	85
Panel C: Using population density in 1000 A.D. as an instrument for population density in 1500 A.D.											
Log population density	−0.31	−0.4	−0.15	−0.38	−0.18	−0.22	−0.27	−0.26	−0.22	−0.26	−0.25
in 1500	(0.06)	(0.06)	(0.08)	(0.11)	(0.07)	(0.08)	(0.06)	(0.07)	(0.07)	(0.06)	(0.08)
Number of observations	83	43	51	32	83	80	83	83	78	83	77

Standard errors are in parentheses. *P*-values from *F*-tests for joint significance are in square brackets. Dependent variable is log GDP per capita (PPP) in 1995. Base sample is all former colonies for which we have data. Population density in 1500 is total population divided by arable land area. See Table III for an explanation of the sample and covariates in each column. For detailed sources and descriptions see Appendix 2.

The estimates in the top panel of Table V use variation in population density, which reflects two components: differences in population and differences in arable land area. In Panel B we separate the effects of these two components and find that they come in with equal and opposite signs, showing that the specification with population density is appropriate. In Panel C we use population density in 1000 as an instrument for population density in 1500. This is useful since, as discussed in subsection II.C, differences in long-run population density are likely to be better proxies for income per capita. Instrumenting for population density in 1500 with population density in 1000 isolates the long-run component of population density differences across countries (i.e., the component of population density in 1500 that is correlated with population density in 1000). The Two-Stage Least Squares (2SLS) results in Panel C using this instrumental variables strategy are very similar to the OLS results in Panel A.

III.C. Further Results, Robustness Checks, and Discussion

Caution is required in interpreting the results presented in Tables III, IV, and V. Estimates of urbanization and population in 1500 are likely to be error-ridden. Nevertheless, the first effect of measurement error would be to create an attenuation bias toward 0. Therefore, one might think that the negative coefficients in Tables III, IV, and V are, if anything, underestimates. A more serious problem would be if errors in the urbanization and population density estimates were not random, but correlated with current income in some systematic way. We investigate this issue further in Table VI, using a variety of different estimates for urbanization and population density. Columns (1)–(5), for example, show that the results are robust to a variety of modifications to the urbanization data.

Much of the variation in urbanization and population density in 1500 was not at the level of these countries, but at the level of "civilizations." For example, in 1500 there were fewer separate civilizations in the Americas, and even arguably in Asia, than there are countries today. For this reason, in column (6) we repeat our key regressions using variation in urbanization and population density only among fourteen civilizations (based on Toynbee [1934–1961] and McNeill [1999]—see the note to Table VI). The results confirm our basic findings, and show a statistically significant negative relationship between prosperity in 1500 and today. Columns (7) and (8) report robustness checks using variants of

the population density data constructed under different assumptions, again with very similar results.

Is there a similar reversal among the noncolonies? Column (9) reports a regression of log GDP per capita in 1995 on urbanization in 1500 for all noncolonies (including Europe), and column (10) reports the same regression for Europe (including Eastern Europe). In both cases, there is a *positive* relationship between urbanization in 1500 and income today.[11] This suggests that the reversal reflects an unusual event, and is likely to be related to the effect of European colonialism on these societies.

Panel B of Table VI reports results weighted by population in 1500, with very similar results. In Panel C we include urbanization and population density simultaneously in these regressions. In all cases, population density is negative and highly significant, while urbanization is insignificant. This is consistent with the notion, discussed below, that differences in population density played a key role in the reversal in relative incomes among the colonies (although it may also reflect measurement error in the urbanization estimates).

As a final strategy to deal with the measurement error in urbanization, we use log population density as an instrument for urbanization rates in 1500. When both of these are valid proxies for economic prosperity in 1500 and the measurement error is classical, this procedure corrects for the measurement error problem. Not surprisingly, these instrumental-variables estimates reported in the bottom panel of Table VI are considerably larger than the OLS estimates in Table III. For example, the baseline estimate is now -0.18 instead of -0.08 in Table III. The general pattern of reversal in relative incomes is unchanged, however.

Is the reversal shown in Figures I and II and Tables III, IV, and V consistent with other evidence? The literature on the history of civilizations documents that 500 years ago many parts of Asia were highly prosperous (perhaps as prosperous as Western Europe), and civilizations in Meso-America and North Africa were relatively developed (see, e.g., Abu-Lughod [1989], Braudel [1992], Chaudhuri [1990], Hodgson [1993], McNeill [1999], Pomeranz [2000], Reid [1988, 1993], and Townsend [2000]). In con-

11. In Acemoglu, Johnson, and Robinson [2001b] we also provided evidence that urbanization and population density in 1000 are positively correlated with urbanization and population density in 1500, suggesting that before 1500 there was considerable persistence in prosperity both where the Europeans later colonized and where they never colonized.

TABLE VI
ROBUSTNESS CHECKS FOR URBANIZATION AND LOG POPULATION DENSITY

	Dependent variable is log GDP per capita (PPP) in 1995									
	Base sample (1)	Assuming lower urbanization in the Americas (2)	Assuming lower urbanization in North Africa (3)	Assuming lower urbanization in Indian subcontinent (4)	Using least favorable combination of assumptions (5)	Using augmented Toynbee definition of civilization (6)	Using land area in 1995 for population density (7)	Alternative assumptions for log population density (8)	All countries never colonized by Europe (9)	Europe (including Eastern Europe) (10)
	Former colonies								Never colonized	
	Panel A: Unweighted regressions									
Urbanization in 1500	−0.078	−0.089	−0.102	−0.073	−0.105	−0.117			0.068	0.077
	(0.026)	(0.027)	(0.029)	(0.027)	(0.032)	(0.052)			(0.023)	(0.023)
Log population density in 1500							−0.41	−0.32		
							(0.06)	(0.07)		
R^2	0.20	0.22	0.24	0.16	0.21	0.30	0.35	0.21	0.18	0.27
Number of observations	41	41	41	41	41	14	91	91	43	32
	Panel B: Regressions weighted using log population in 1500									
Urbanization in 1500	−0.072	−0.084	−0.097	−0.064	−0.099	−0.118			−0.064	−0.073
	(0.025)	(0.026)	(0.029)	(0.026)	(0.032)	(0.053)			(0.023)	(0.022)
Log population density in 1500							−0.39	−0.29		
							(0.06)	(0.07)		
R^2	0.18	0.22	0.23	0.14	0.20	0.29	0.32	0.19	0.17	0.24
Number of observations	41	41	41	41	41	14	91	91	43	32

Panel C: Including both urbanization and log population density as independent variables

Urbanization in 1500	0.038	0.039	0.017	0.037	0.020	0.072	0.017	0.003	0.028	0.032
	(0.028)	(0.031)	(0.033)	(0.027)	(0.035)	(0.047)	(0.023)	(0.022)	(0.020)	(0.021)
Log population density in 1500	−0.41	−0.41	−0.36	−0.40	−0.37	−0.48	−0.43	−0.41	0.34	0.37
	(0.07)	(0.08)	(0.07)	(0.07)	(0.07)	(0.09)	(0.07)	(0.07)	(0.07)	(0.08)
R^2	0.56	0.56	0.54	0.56	0.54	0.79	0.61	0.60	0.48	0.57
Number of observations	41	41	41	41	41	14	41	41	43	32

Panel D: Instrumenting for urbanization in 1500 using log population density in 1500

Urbanization in 1500	−0.178	−0.181	−0.215	−0.194	−0.242	−0.237	−0.217	−0.239	0.259	0.226
	(0.04)	(0.040)	(0.048)	(0.048)	(0.057)	(0.080)	(0.053)	(0.063)	(0.090)	(0.074)
Number of observations	41	41	41	41	41	14	41	41	43	32

Standard errors are in parentheses. Dependent variable is log GDP per capita (PPP) in 1995. Base sample is all former colonies for which we have data. In our base sample, urbanization in 1500 is percent of the population living in towns with 5000 or more people. Column (2) assumes 9 percent urbanization in the Andes and Central America. Column (3) assumes 10 percent urbanization in North Africa. Column (4) assumes 6 percent urbanization in the Indian subcontinent. Column (5) combines the assumptions of columns (2), (3), (4), and (5) to create the least favorable combination of assumptions for our hypothesis. Column (6) is only civilizations in former European colonies. The augmented Toynbee civilizations, used in column (6), include Andean, Mexic, Yucatec, Arabic (North Africa), Hindu, Polynesian, Eskimo (Canada) North American Indian, South American Indian (Brazil/Argentina/Chile), Australian Aborigine, Malay (Malaysia and Indonesia), Philippines, Vietnam/Cambodia, and Burma. In column (7) population density in 1500 is total population divided by arable land area in 1995. Column (8) halves the population density estimates for Africa. For detailed sources and descriptions see Appendix 2.

trast, there was little agriculture in most of North America and Australia, at most consistent with a population density of 0.1 people per square kilometer. McEvedy and Jones [1978, p. 322] describe the state of Australia at this time as "an unchanging palaeolithic backwater." In fact, because of the relative backwardness of these areas, European powers did not view them as valuable colonies. Voltaire is often quoted as referring to Canada as a "few acres of snow," and the European powers at the time paid little attention to Canada relative to the colonies in the West Indies. In a few parts of North America, along the East Coast and in the Southwest, there was settled agriculture, supporting a population density of approximately 0.4 people per square kilometer, but this was certainly much less than that in the Aztec and Inca Empires, which had fully developed agriculture with a population density of between 1 and 3 people (or even higher) per square kilometer, and also much less than the corresponding numbers in Asia and Africa [McEvedy and Jones 1978, p. 273]. The recent work by Maddison [2001] also confirms our interpretation. He estimates that India, Indonesia, Brazil, and Mexico were richer than the United States in 1500 and 1700 (see, for example, his Table 2-22a).

III.D. The Timing and Nature of the Reversal

The evidence presented so far documents the reversal in relative incomes among the former colonies from 1500 to today. When did this reversal take place? This question is relevant in thinking about the causes of the reversal. For example, if the reversal is related to the extraction of resources from, and the "plunder" of, the former colonies, or to the direct effect of the diseases Europeans brought to the New World, it should have taken place shortly after colonization.

Figure IV shows that the reversal is mostly a late eighteenth- and early nineteenth-century phenomenon, and is closely related to industrialization. Figure IVa compares the evolution of urbanization among two groups of New World ex-colonies, those with low urbanization in 1500 versus those with high urbanization in 1500.[12] We focus on New World colonies since the societies came

12. The initially high urbanization countries for which we have data and are included in the figure are Bolivia, Mexico, Peru, and all of Central America, while the initially low urbanization countries are Argentina, Brazil, Canada, Chile, and the United States.

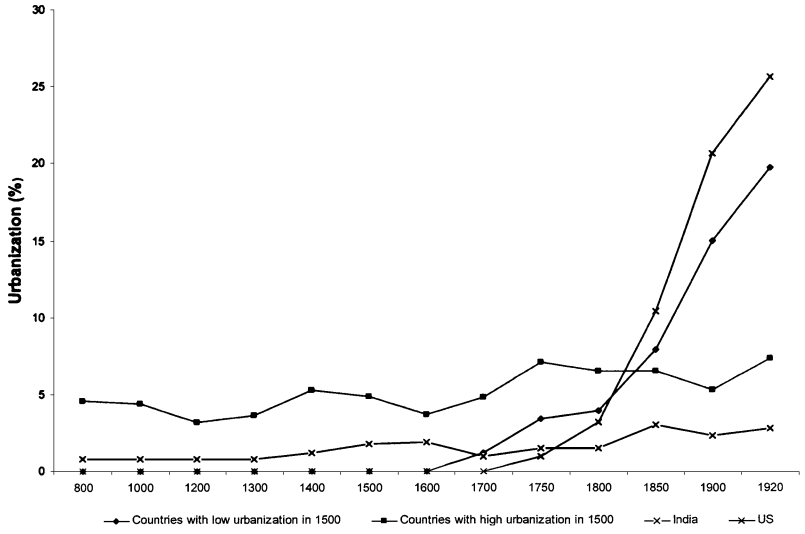


FIGURE IVa
Urbanization Rate in India, the United States, and New World Countries with Low and High Urbanization, 800–1920

Note. Urbanization is population living in urban areas divided by total population. Urban areas have a minimum threshold of 20,000 inhabitants, from Chandler [1987], and Mitchell [1993, 1995]. Low urbanization in 1500 countries are Argentina, Brazil, Canada, Chile, and the United States. High urbanization in 1500 countries are Bolivia, Ecuador, Mexico, Peru, and all of Central America. For details see Appendix 1.

under European dominance very early on. The averages plotted in the figure are weighted by population in 1500. In addition, in the same figure we plot India and the United States separately (as well as including it in the initially low urbanization group). The figure shows that the initially low urbanization group as a whole and the United States by itself overtake India and the initially high urbanization countries sometime between 1750 and 1850.

Figure IVb depicts per capita industrial production for the United States, Canada, New Zealand, Australia, Brazil, Mexico, and India using data from Bairoch [1982]. This figure shows the takeoff in industrial production in the United States, Australia, Canada, and New Zealand relative to Brazil, Mexico, and India. Although the scale makes it difficult to see in the figure, per capita industrial production in 1750 was in fact higher in India, 7, than in the United States, 4 (with U. K. industrial production per capita in 1900 normalized to 100). Bairoch [1982] also reports that in 1750 China had industrial production per capita twice the

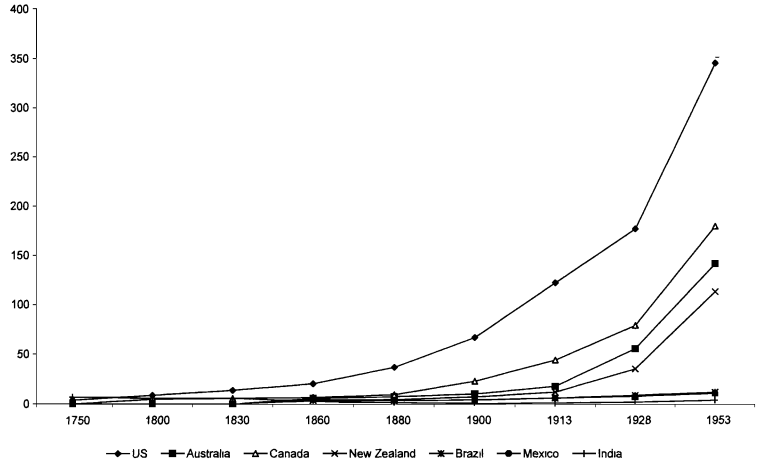


FIGURE IVb
Industrial Production per Capita, 1750–1953

Note. Index of industrial production with U. K. per capita industrialization in 1900 is equal to 100, from Bairoch [1982].

level of the United States. Yet, as Figure IVb shows, over the next 200 years there was a much larger increase in industrial production in the United States than in India (and also than in China).

This general interpretation, that the reversal in relative incomes took place during the late eighteenth and early nineteenth centuries and was linked to industrialization, is also consistent with the fragmentary evidence we have on other measures of income per capita and industrialization. Coatsworth [1993], Eltis [1995], Engerman [1981], and Engerman and Sokoloff [1997] provide evidence that much of Spanish America and the Caribbean were more prosperous (had higher per capita income) than British North America until the eighteenth century. The future United States rose in per capita income during the 1700s relative to the Caribbean and South America, but only really pulled ahead during the late eighteenth and early nineteenth centuries. Maddison's [2001] numbers also show that India, Indonesia, Brazil, and Mexico were richer than the United States in 1700, but had fallen behind by 1820.

U. S. growth during this period also appears to be an industry-based phenomenon. McCusker and Menard [1985] and Galenson [1996] both emphasize that productivity and income growth in North America before the eighteenth century was limited. During the critical period of growth in the United States, between

1840 and 1900, there was modest growth in agricultural output per capita, and very rapid growth in industrial output per capita; the numbers reported by Gallman [2000] imply that between 1840 and 1900 agricultural product per capita increased by about 30 percent, a very small increase relative to the growth in manufacturing output per capita, which increased more than fourfold.

IV. Hypotheses and Explanations

IV.A. The Geography Hypothesis

The geography hypothesis claims that differences in economic performance reflect differences in geographic, climatic, and ecological characteristics across countries. There are many different versions of this hypothesis. Perhaps the most common is the view that climate has a direct effect on income through its influence on work effort. This idea dates back to Machiavelli [1519] and Montesquieu [1748]. Both Toynbee [1934, Vol. 1] and Marshall [1890, p. 195] similarly emphasized the importance of climate, both on work effort and productivity. One of the pioneers of development economics, Myrdal [1968], also placed considerable emphasis on the effect of geography on agricultural productivity. He argued: "serious study of the problems of underdevelopment . . . should take into account the climate and its impacts on soil, vegetation, animals, humans and physical assets—in short, on living conditions in economic development" [Vol. 3, p. 2121].

More recently, Diamond [1997] and Sachs [2000, 2001] have espoused different versions of the geography view. Diamond, for example, argues that the timing of the Neolithic revolution has had a long-lasting effect on economic and social development. Sachs, on the other hand, emphasizes the importance of geography through its effect on the disease environment, transport costs, and technology. He writes: "Certain parts of the world are geographically favored. Geographical advantages might include access to key natural resources, access to the coastline and sea—navigable rivers, proximity to other successful economies, advantageous conditions for agriculture, advantageous conditions for human health" [2000, p. 30]. Also see Myrdal [1968, Vol. 1, pp. 691–695].

This simple version of the geography hypothesis predicts persistence in economic outcomes, since the geographic factors that are the first-order determinants of prosperity are time-in-

variant. The evidence presented so far therefore weighs against the simple geography hypothesis: whatever factors are important in making former colonies rich today are very different from those contributing to prosperity in 1500.

IV.B. The Sophisticated Geography Hypotheses

The reversal in relative incomes does not necessarily reject a more sophisticated geography hypothesis, however. Certain geographic characteristics that were not useful, or that were even harmful, for successful economic performance in 1500 may turn out to be beneficial later on. In this subsection we briefly discuss a number of sophisticated geography hypotheses emphasizing the importance of such time-varying effects of geography.[13]

The first is the "temperate drift hypothesis," emphasizing the temperate (or away from the equator) shift in the center of economic gravity over time. According to this view, geography becomes important when it interacts with the presence of certain technologies. For example, one can argue that tropical areas provided the best environment for early civilizations—after all, humans evolved in the tropics, and the required calorie intake is lower in warmer areas. But with the arrival of "appropriate" technologies, temperate areas became more productive. The technologies that were crucial for progress in temperate areas include the heavy plow, systems of crop rotation, domesticated animals such as cattle and sheep, and some of the high productivity European crops, including wheat and barley. Despite the key role of these technologies for temperate areas, they have had much less of an effect on tropical zones [Lewis 1978]. Sachs [2001, p. 12] also implies this view in his recent paper when he adapts Diamond's argument about the geography of technological diffusion: "Since technologies in the critical areas of agriculture, health, and related areas could diffuse *within* ecological zones, but not across ecological zones, economic development spread through the tem-

13. Put differently, in the simple geography hypothesis, geography has a *main effect* on economic performance, which can be expressed as $Y_{it} = \alpha_0 + \alpha_1 \cdot G_i + \nu_t + \epsilon_{it}$, where Y_{it} is a measure of economic performance in country i at time t, G_i is a measure of geographic characteristics, ν_t is a time effect, and ϵ_{it} measures other country-time-specific factors. In contrast, in the sophisticated geography view, the relationship between income and geography would be $Y_{it} = \alpha_0 + \alpha_1 \cdot G_{it} + \alpha_2 \cdot T_t \cdot G_{it} + \nu_t + \epsilon_{it}$, where T_t is a time-varying characteristic of the world as a whole or of the state of technology. According to this view, the major role that geography plays in history is not through α_1, but through α_2.

perate zones but not through the tropical regions" (italics in the original; also see Myrdal [1968], Ch. 14).

The evidence is not favorable to the view that the reversal reflects the emergence of agricultural technologies favorable to temperate areas, however. First, the regressions in Tables III, IV, and V show little evidence that the reversal was related to geographic characteristics. Second, the temperate drift hypothesis suggests that the reversal should be associated with the spread of European agricultural technologies. Yet in practice, while European agricultural technology spread to the colonies between the sixteenth and eighteenth centuries (e.g., McCusker and Menard [1985], Ch. 3 for North America), the reversal in relative incomes is largely a late eighteenth- and early nineteenth-century, and industry-based phenomenon.

In light of the result that the reversal is related to industrialization, another sophisticated geography hypothesis would be that certain geographic characteristics facilitate or enable industrialization. First, one can imagine that there is more room for specialization in industry, but such specialization requires trade. If countries differ according to their transport costs, it might be those with low transport costs that take off during the age of industry. This argument is not entirely convincing, however, again because there is little evidence that the reversal was related to geographic characteristics (see Tables III, IV, and V). Moreover, many of the previously prosperous colonies that failed to industrialize include islands such as the Caribbean, or countries with natural ports such as those in Central America, India, or Indonesia. Moreover, transport costs appear to have been relatively low in some of the areas that failed to industrialize (e.g., Pomeranz [2000], Appendix A).

Second, countries may lack certain resource endowments, most notably coal, which may have been necessary for industrialization (e.g., Pomeranz [2000] and Wrigley [1988]). But coal is one of the world's most common resources, with proven reserves in 100 countries and production in over 50 countries [World Coal Institute 2000], and our results in Table III and V offer little evidence that either coal or the absence of any other resource was responsible for the reversal. So there appears to be little support for these types of sophisticated geography hypotheses either.[14]

14. Two other related hypotheses are worth mentioning. First, it could be argued that people work less hard in warmer climates and that this matters more

IV.C. The Institutions Hypothesis

According to the institutions hypothesis, societies with a social organization that provides encouragement for investment will prosper. Locke [1980], Smith [1778], and Hayek [1960], among many others, emphasized the importance of property rights for the success of nations. More recently, economists and historians have emphasized the importance of institutions that guarantee property rights. For example, Douglass North starts his 1990 book by stating [p. 3]: "That institutions affect the performance of economies is hardly controversial," and identifies effective protection of property rights as important for the organization of society (see also North and Thomas [1973] and Olson [2000]).

In this context we take a good organization of society to correspond to a cluster of (political, economic, and social) institutions ensuring that a broad cross section of society has effective property rights. We refer to this cluster as *institutions of private property,* and contrast them with *extractive institutions,* where the majority of the population faces a high risk of expropriation and holdup by the government, the ruling elite, or other agents. Two requirements are implicit in this definition of institutions of private property. First, institutions should provide secure property rights, so that those with productive opportunities expect to receive returns from their investments, and are encouraged to undertake such investments. The second requirement is embedded in the emphasis on "a broad cross section of the society." A society in which a very small fraction of the population, for example, a class of landowners, holds all the wealth and political power may not be the ideal environment for investment, even if

for industry than for agriculture, thus explaining the reversal. However, there is no evidence either for the hypothesis that work effort matters more for industry or for the assertion that human energy output depends systematically on temperature (see, e.g., Collins and Roberts [1988]). Moreover, the available evidence on hours worked indicates that people work harder in poorer/warmer countries (e.g., ILO [1995, pp. 36–37]), though of course these high working hours could reflect other factors.

Second, it can be argued that different paths of development reflect the direct influence of Europeans. Places where there are more Europeans have become richer, either because Europeans brought certain values conducive to development (e.g., Landes [1998], and Hall and Jones [1999]), or because having more Europeans confers certain benefits (e.g., through trade with Europe or because Europeans are more productive). In Acemoglu, Johnson, and Robinson [2001b] we presented evidence showing that the reversal and current income levels are not related to the current racial composition of the population or to proxies of whether the colonies were culturally or politically dominated by Europeans.

the property rights of this elite are secure. In such a society, many of the agents with the entrepreneurial human capital and investment opportunities may be those without effective property rights protection. In particular, the concentration of political and social power in the hands of a small elite implies that the majority of the population risks being held up by the powerful elite after they undertake investments. This is also consistent with North and Weingast's [1989, pp. 805–806] emphasis that what matters is: " . . . whether the state produces rules and regulations that benefit a small elite and so provide little prospect for long-run growth, or whether it produces rules that foster long-term growth." Whether political power is broad-based or concentrated in the hands of a small elite is crucial in evaluating the role of institutions in the experiences of the Caribbean or India during colonial times, where the property rights of the elite were well enforced, but the majority of the population had no civil rights or property rights.

It is important to emphasize that "equilibrium institutions" may be extractive, even though such institutions do not encourage economic development. This is because institutions are shaped, at least in part, by politically powerful groups that may obtain fewer rents with institutions of private property (e.g., North [1990]), or fear losing their political power if there is institutional development (e.g., Acemoglu and Robinson [2000, 2001]), or simply may be reluctant to initiate institutional change because they would not be the direct beneficiaries of the resulting economic gains. In the context of the development experience of the former colonies, this implies that equilibrium institutions are likely to have been designed to maximize the rents to European colonists, not to maximize long-run growth.

The organization of society and institutions also persist (see, for example, the evidence presented in Acemoglu, Johnson, and Robinson [2001a]). Therefore, the institutions hypothesis also suggests that societies that are prosperous today should tend to be prosperous in the future. However, if a major shock disrupts the organization of a society, this will affect its economic performance. We argue that European colonialism not only disrupted existing social organizations, but led to the establishment of, or continuation of already existing, extractive institutions in previously prosperous areas and to the development of institutions of private property in previously poor areas. Therefore, European colonialism led to an *institutional reversal,* in the sense that

regions that were *relatively prosperous* before the arrival of Europeans were more likely to end up with extractive institutions under European rule than previously poor areas. The institutions hypothesis, combined with the institutional reversal, predicts a reversal in relative incomes among these countries.

The historical evidence supports the notion that colonization introduced relatively better institutions in previously sparsely settled and less prosperous areas: while in a number of colonies such as the United States, Canada, Australia, New Zealand, Hong Kong, and Singapore, Europeans established institutions of private property, in many others they set up or took over already existing extractive institutions in order to directly extract resources, to develop plantation and mining networks, or to collect taxes.[15] Notice that what is important for our story is not the "plunder" or the direct extraction of resources by the European powers, but the long-run consequences of the institutions that they set up to support extraction. The distinguishing feature of these institutions was a high concentration of political power in the hands of a few who extracted resources from the rest of the population. For example, the main objective of the Spanish and Portuguese colonization was to obtain silver, gold, and other valuables from America, and throughout they monopolized military power to enable the extraction of these resources. The mining network set up for this reason was based on forced labor and the oppression of the native population. Similarly, the British West Indies in the seventeenth and eighteenth centuries were controlled by a small group of planters (e.g., Dunn [1972, Chs. 2–6]). Political power was important to the planters in the West Indies, and to other elites in the colonies specializing in plantation agriculture, because it enabled them to force large masses of natives or African slaves to work for low wages.[16]

What determines whether Europeans pursued an extractive

15. Examples of extraction by Europeans include the transfer of gold and silver from Latin America in the seventeenth and eighteenth centuries and of natural resources from Africa in the nineteenth and twentieth centuries, the Atlantic slave trade, plantation agriculture in the Caribbean, Brazil, and French Indochina, the rule of the British East India Company in India, and the rule of the Dutch East India Company in Indonesia. See Frank [1978], Rodney [1972], Wallerstein [1974–1980], and Williams [1944].

16. In a different vein, Europeans running the Atlantic slave trade, despite their small numbers, also appear to have had a fundamental effect on the evolution of institutions in Africa. The consensus view among historians is that the slave trade fundamentally altered the organization of society in Africa, leading to state centralization and warfare as African polities competed to control the supply of slaves to the Europeans. See, for example, Manning [1990, p. 147], and also

strategy or introduced institutions of private property? And why was extraction more likely in relatively prosperous areas? Two factors appear important.

1. *The economic profitability of alternative policies.* When extractive institutions were more profitable, Europeans were more likely to opt for them. High population density, by providing a supply of labor that could be forced to work in agriculture or mining, made extractive institutions more profitable for the Europeans.[17] For example, the presence of abundant Amerindian labor in Meso-America was conducive to the establishment of forced labor systems, while the relatively high population density in Africa created a profit opportunity for slave traders in supplying labor to American plantations.[18] Other types of extractive institutions were also more profitable in densely settled and prosperous areas where there was more to be extracted by European colonists. Furthermore, in these densely settled areas there was often an existing system of tax administration or tribute; the large population made it profitable for the Europeans to take control of these systems and to continue to levy high taxes (see, e.g.,

Wilks [1975] for Ghana, Law [1977] for Nigeria, Harms [1981]) for the Congo/Zaire, and Miller [1988] on Angola.

17. The Caribbean islands were relatively densely settled in 1500. Much of the population in these islands died soon after the arrival of the Europeans because of the diseases that the Europeans brought (e.g., Crosby [1986] and McNeill [1976]). It is possible that the initial high populations in these islands induced the Europeans to take the "extractive institutions" path, and subsequently, these institutions were developed further with the import of slaves from Africa. An alternative possibility is that the relevant period of institutional development was after the major population decline, but the Caribbean still ended up with extractive institutions because the soil and the climate were suitable for sugar production, which encouraged Europeans to import slaves from Africa and set up labor-oppressive systems (e.g., Dunn [1972] and Engerman and Sokoloff [1997, 2000]).

18. The Spanish conquest around the La Plata River (current day Argentina) during the early sixteenth century provides a nice example of how population density affected European colonization (see Lockhart and Schwartz [1983, pp. 259–260] or Denoon [1983, pp. 23–24]). Early in 1536, a large Spanish expedition arrived in the area, and founded the city of Buenos Aires at the mouth of the river Plata. The area was sparsely inhabited by nonsedentary Indians. The Spaniards could not enslave a sufficient number of Indians for food production. Starvation forced them to abandon Buenos Aires and retreat up the river to a post at Asuncion (current day Paraguay). This area was more densely settled by semisedentary Indians, who were enslaved by the Spaniards; the colony of Paraguay, with relatively extractive institutions, was founded. Argentina was finally colonized later, with a higher proportion of European settlers and little forced labor.

Wiegersma [1988, p. 69], on French policies in Vietnam, or Marshall [1998, pp. 492–497], on British policies in India).

2. *Whether Europeans could settle or not.* Europeans were more likely to develop institutions of private property when they settled in large numbers, for the natural reason that they themselves were affected by these institutions (i.e., their objectives coincided with encouraging good economic performance).[19] Moreover, when a large number of Europeans settled, the lower strata of the settlers demanded rights and protection similar to, or even better than, those in the home country. This made the development of effective property rights for a broad cross section of the society more likely. European settlements, in turn, were affected by population density both directly and indirectly. Population density had a direct effect on settlements, since Europeans could easily settle in large numbers in sparsely inhabited areas. The indirect effect worked through the disease environment, since malaria and yellow fever, to which Europeans lacked immunity, were endemic in many of the densely settled areas [Acemoglu, Johnson, and Robinson 2001a].[20]

Table VII provides econometric evidence on the institutional reversal. It shows the relationship between urbanization or population density in 1500 and subsequent institutions using three different measures of institutions. The first two measures refer to current institutions: protection against expropriation risk between 1985 and 1995 from Political Risk Services, which approximates how secure property rights are, and "constraints on the executive" in 1990 from Gurr's Polity III data set, which can be thought of as a proxy for how concentrated political power is in the hands of ruling groups (see Appendix 2 for detailed sources). Columns (1)–(6) of Table VII show a negative relation-

19. Extraction and European settlement patterns were mutually self-reinforcing. In areas where extractive policies were pursued, the authorities also actively discouraged settlements by Europeans, presumably because this would interfere with the extraction of resources from the locals (e.g., Coatsworth [1982]).

20. European settlements shaped both the type of institutions that developed and the structure of production. For example, while in Potosí (Bolivia) mining employed forced labor [Cole 1985] and in Brazil and the Caribbean sugar was produced by African slaves, in the United States and Australia mining companies employed free migrant labor and sugar was grown by smallholders in Queensland, Australia [Denoon 1983, Chs. 4 and 5]. Consequently, in Bolivia, Brazil, and the Caribbean, political institutions were designed to ensure the control of the laborers and slaves, while in the United States and Australia, the smallholders and the middle class had greater political rights [Cole 1985; Hughes 1988, Ch. 10].

TABLE VII

URBANIZATION, POPULATION DENSITY, AND INSTITUTIONS

	Dependent variable is:								
	Average protection against expropriation risk, 1985–1995			Constraint on executive in 1990			Constraint on executive in first year of independence		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Without additional controls									
Urbanization in 1500	−0.107		−0.001	−0.154		−0.037	−0.132		0.018
	(0.043)		(0.059)	(0.066)		(0.098)	(0.069)		(0.103)
Log population density in 1500		−0.37	−0.37		−0.49	−0.40		−0.33	−0.54
		(0.10)	(0.15)		(0.15)	(0.25)		(0.15)	(0.28)
R^2	0.14	0.16	0.25	0.12	0.12	0.18	0.31	0.16	0.37
Number of observations	42	75	42	41	84	41	42	85	42
Panel B: Controlling for latitude									
Urbanization in 1500	−0.097		−0.001	−0.159		−0.038	−0.128		0.022
	(0.042)		(0.059)	(0.067)		(0.099)	(0.070)		(0.104)
Log population density in 1500		−0.31	−0.34		−0.45	−0.41		−0.30	−0.54
		(0.10)	(0.15)		(0.16)	(0.25)		(0.16)	(0.28)
Latitude	2.87	3.53	2.57	−1.49	2.63	−1.86	1.52	2.68	1.48
	(1.48)	(1.25)	(1.41)	(2.38)	(2.01)	(2.34)	(2.54)	(2.17)	(2.46)
R^2	0.21	0.24	0.31	0.13	0.13	0.19	0.32	0.17	0.38
Number of observations	42	75	42	41	84	41	42	84	42

Standard errors are in parentheses. Regressions use data for all former colonies for which information on urbanization and population density in 1500 is available, as explained in the text. Urbanization in 1500 is percent of the population living in towns with 5000 or more people. Population density in 1500 is total population divided by arable land area from McEvedy and Jones [1978]. Average protection against expropriation risk is an evaluation of the risk that private investments will be expropriated by the government. Constraints on the executive is an assessment of the constitutional limitations on executive power. Regressions with constraints on executive in first year of independence use the earliest available date after independence, and also include the date of independence as an additional regressor. For detailed sources and descriptions see Appendix 2.

ship between our measures of prosperity in 1500 and current institutions.[21]

It is also important to know whether there was an institutional reversal during the colonial times or shortly after independence. Since the Gurr data set does not contain information for nonindependent countries, we can only look at this after independence. Columns (7)–(9) show the relationship between prosperity in 1500 and a measure of early institutions, constraint on the executive in the first year of independence, from the same data set, while also controlling for time since independence as an additional covariate. Finally, the second panel of the table includes (the absolute value of) latitude as an additional control, showing that the institutional reversal does not reflect some simple geographic pattern of institutional change.

The institutions hypothesis, combined with the institutional reversal, predicts that countries in areas that were relatively prosperous and densely settled in 1500 ended up with relatively worse institutions after the European intervention, and therefore should be relatively less prosperous today. The reversal in relative incomes that we have documented so far is consistent with this prediction.

Notice, however, that the institutions hypothesis and the reversal in relative incomes do not rule out an important role for geography during some earlier periods, or working through institutions. They simply suggest that institutional differences are the major source of differences in income per capita *today*. First, differences in economic prosperity in 1500 may be reflecting geographic factors (e.g., that the tropics were more productive than temperate areas) as well as differences in social organization caused by nongeographic influences. Second and more important, as we emphasized in Acemoglu, Johnson, and Robinson [2001a], a major determinant of European settlements, and therefore of institutional development, was the mortality rates faced by Europeans, which is a geographical variable. Similarly, as noted by Engerman and Sokoloff [1997, 2000], whether an area was suitable for sugar production is likely to have been important in

21. When both urbanization and log population density in 1500 are included, it is the population density variable that is significant. This supports the interpretation that it was the differences between densely and sparsely settled areas that was crucial in determining colonial institutions (though, again, this may also reflect the fact that the population density variable is measured with less measurement error).

shaping the type of institutions that Europeans introduced. However, this type of interaction between geography and institutions means that certain regions, say Central America, are poor today not as a result of their geography, but because of their institutions, and that there is not a necessary or universal link between geography and economic development.

V. Institutions and the Making of the Modern World Income Distribution

V.A. Institutions and the Reversal

We next provide evidence suggesting that institutional differences statistically account for the reversal in relative incomes. If the institutional reversal is the reason why there was a reversal in income levels among the former colonies, then once we account for the role of institutions appropriately, the reversal should disappear. That is, according to this view, the reversal documented in Figures I and II and Tables III, IV, V, and VI reflects the correlation between economic prosperity in 1500 and income today working through the intervening variable, institutions.

How do we establish that an intervening variable X is responsible for the correlation between Z and Y? Suppose that the true relationship between Y, and X, and Z is

$$Y = \alpha \cdot X + \beta \cdot Z + \epsilon, \tag{1}$$

where α and β are coefficients and ϵ is a disturbance term. In our case, we can think of Y as income per capita today, X as a measure of institutions, and Z as population density (or urbanization) in 1500. The variable Z is included in equation (1) either because it has a direct effect on Y or because it has an effect through some other variables not included in the analysis. The hypothesis we are interested in is that $\beta = 0$; that is, population density or urbanization in 1500 affects income today *only* via institutions.

This hypothesis obviously requires that there is a statistical relationship between X and Z. So we postulate that $X = \lambda \cdot Z + v$. To start with, suppose that ϵ is independent of X and Z and that v is independent of Z. Now imagine a regression of Y on Z only (in our context, of income today on prosperity in 1500, similar to those we reported in Tables III, IV, V, and VI):

$Y = b \cdot Z + u_1$. As is well-known, the probability limit of the OLS estimate from this regression, $\hat{b}$, is

$$\text{plim}\hat{b} = \beta + \alpha \cdot \lambda.$$

So the results in the regressions of Tables IV, V, VI, and VII are consistent with $\beta = 0$ as long as $\alpha \neq 0$ and $\lambda \neq 0$. In this case, we would be capturing the effect of Z (population density or urbanization) on income working solely through institutions. This is the hypothesis that we are interested in testing. Under the assumptions regarding the independence of Z from ν and ϵ, and of X from ϵ, there is a simple way of testing this hypothesis, which is to run an OLS regression of Y on Z and X:

$$Y = a \cdot X + b \cdot Z + u_2 \tag{2}$$

to obtain the estimates $\hat{a}$ and $\hat{b}$. The fact that ϵ in (1) is independent of both X and Z rules out omitted variable bias, so $\text{plim}\hat{a} = \alpha$ and $\text{plim}\hat{b} = \beta$. Hence, a simple test of whether $\hat{b} = 0$ is all that is required to test our hypothesis that the effect of Z is through X alone.

In practice, there are likely to be problems due to omitted variables, endogeneity bias because Y has an effect on X, and attenuation bias because X is measured with error or corresponds poorly to the real concept that is relevant to development (which is likely to be a broad range of institutions, whereas we only have an index for a particular type of institutions). So the above procedure is not possible. However, the same logic applies as long as we have a valid instrument M for X, such that $X = \gamma \cdot M + \zeta$, and M is independent of ϵ in (1). We can then simply estimate (2) using 2SLS with the first-stage $X = c \cdot M + d \cdot Z + u_3$. Testing our hypothesis that Z has an effect on Y only through its effect on X then amounts to testing that the 2SLS estimate of b, $\tilde{b}$, is equal to 0. Intuitively, the 2SLS procedure ensures a consistent estimate of α, enabling an appropriate test for whether Z has a direct effect.

The key to the success of this strategy is a good instrument for X. In our previous work [Acemoglu, Johnson, and Robinson 2001a] we showed that mortality rates faced by settlers are a good instrument for settlements of Europeans in the colonies and the subsequent institutional development of these countries. These mortality rates are calculated from the mortality of soldiers, bishops, and sailors stationed in the colonies between the seven-

teenth and nineteenth centuries, and are a plausible instrument for the institutional development of the colonies, since in areas with high mortality Europeans did not settle and were more likely to develop extractive institutions. The exclusion restriction implied by this instrumental-variables strategy is that, conditional on the other controls, the mortality rates of European settlers more than 100 years ago have no effect on GDP per capita today, other than their effects through institutional development. This is plausible since these mortality rates were much higher than the mortality rates faced by the native population who had developed a high degree of immunity to the two main killers of Europeans, malaria and yellow fever.

Table VIII reports results from this type of 2SLS test using the log of settler mortality rates as an instrument for institutional development. We look at the same three institutions variables used in Table VII: protection against expropriation risk between 1985 and 1995, and constraint on the executive in 1990 and in the first year of independence. Panel A reports results from regressions that enter urbanization and log population density in 1500 as exogenous regressors in the first and the second stages, while Panel B reports the corresponding first stages. Different columns correspond to different institutions variables, or to different specifications. For comparison, Panel C reports the 2SLS coefficient on institutions with exactly the same sample as the corresponding column, but without including urbanization or population density.

The results are consistent with our hypothesis. In all columns we never reject the hypothesis that urbanization in 1500 or population density in 1500 has *no* direct effect once we control for the effect of institutions on income per capita, and the addition of these variables has little effect on the 2SLS estimate of the effect of institutions on income per capita. This supports our notion that the reversal in economic prosperity reflects the effect of early prosperity and population density working through the institutions and policies introduced by European colonists.

V.B. Institutions and Industrialization

Why did the reversal in relative incomes take place during the nineteenth century? To answer this question, imagine a society like the Caribbean colonies where a small elite controls all the political power. The property rights of this elite are relatively well protected, but the rest of the population has no effective property

TABLE VIII
GDP PER CAPITA AND INSTITUTIONS

	Dependent variable is log GDP per capita (PPP) in 1995					
Institutions as measured by:	Average protection against expropriation risk, 1985–1995		Constraint on executive in 1990		Constraint on executive in first year of independence	
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Second-stage regressions					
Institutions	0.52	0.88	0.84	0.50	0.37	0.46
	(0.10)	(0.21)	(0.47)	(0.11)	(0.12)	(0.16)
Urbanization in 1500	−0.024		0.030		−0.023	
	(0.021)		(0.078)		(0.034)	
Log population density in 1500		−0.08		−0.10		−0.13
		(0.10)		(0.10)		(0.10)
	Panel B: First-stage regressions					
Log settler mortality	−1.21	−0.47	−0.75	−0.88	−1.81	−0.78
	(0.23)	(0.14)	(0.44)	(0.20)	(0.40)	(0.25)
Urbanization in 1500	−0.042		−0.088		−0.043	
	(0.035)		(0.066)		(0.061)	
Log population density in 1500		−0.21		−0.35		−0.24
		(0.11)		(0.15)		(0.17)
R^2	0.53	0.29	0.17	0.37	0.56	0.26
Number of observations	38	64	37	67	38	67
Panel C: Coefficient on institutions without urbanization or population density in 1500						
Institutions	0.56	0.96	0.77	0.54	0.39	0.52
	(0.09)	(0.17)	(0.33)	(0.09)	(0.11)	(0.15)

Standard errors are in parentheses. Dependent variable is log GDP per capita (PPP) in 1995. The measure of institutions used in each regression is indicated at the head of each column. Urbanization in 1500 is percent of the population living in towns with 5000 or more people. Population density is calculated as total population divided by arable land area. Constraint on the executive in 1990, 1900, and the first year of independence are all from the Polity III data set. Regressions with constraint on executive in first year of independence use the earliest available date after independence, and also include the date of independence as an additional regressor.

Panel A reports the second-stage estimates from an IV regression with first-stage shown in Panel B. Panel C reports second-stage estimates from the IV regressions, which do not include urbanization or population density and which instrument for institutions using log settler mortality. Log settler mortality estimates are from Acemoglu, Johnson, and Robinson [2001a]. For detailed sources and descriptions see Appendix 2.

rights. According to our definition, this would not be a society with institutions of private property, since a broad cross section of society does not have effective property rights. Nevertheless, when the major investment opportunities are in agriculture, this may not matter too much, since the elite can invest in the land

and employ the rest of the population, and so will have relatively good incentives to increase output.

Imagine now the arrival of a new technology, for example, the opportunity to industrialize. If the elite could undertake industrial investments without losing its political power, we may expect them to take advantage of these opportunities. However, in practice there are at least three major problems. First, those with the entrepreneurial skills and ideas may not be members of the elite and may not undertake the necessary investments, because they do not have secure property rights and anticipate that they will be held up by political elites once they undertake these investments. Second, the elites may want to *block* investments in new industrial activities, because it may be these outside groups, not the elites themselves, who will benefit from these new activities. Third, they may want to block these new activities, fearing political turbulence and the threat to their political power that new technologies will bring (see Acemoglu and Robinson [2000, 2001]).[22]

This reasoning suggests that whether a society has institutions of private property or extractive institutions may matter much more when new technologies require broad-based economic participation—in other words, extractive institutions may become much more *inappropriate* with the arrival of new technologies. Early industrialization appears to require both investments from a large number of people who were not previously part of the ruling elite and the emergence of new entrepreneurs (see Engerman and Sokoloff [1997], Kahn and Sokoloff [1998], and Rothenberg [1992] for evidence that many middle-class citizens, innovators, and smallholders contributed to the process of early industrialization in the United States). Therefore, there are reasons to expect that institutional differences should matter more during the age of industry.

If this hypothesis is correct, we should expect societies with good institutions to take better advantage of the opportunity to industrialize starting in the late eighteenth century. We can test this idea using data on institutions, industrialization, and GDP from the nineteenth and early twentieth centuries. Bairoch [1982] presents estimates of industrial output for a number of countries at a variety of dates, and Maddison [1995] has esti-

22. In addition, industrialization may have been delayed in some cases because of a comparative advantage in agriculture.

mates of GDP for a larger group of countries. We take Bairoch's estimates of U. K. industrial output as a proxy for the opportunity to industrialize, since during this period the United Kingdom was the world industrial leader. We then run a panel data regression of the following form:

$$y_{it} = \mu_t + \delta_i + \pi \cdot X_{it} + \phi \cdot X_{it} \cdot UKIND_t + \epsilon_{it}, \tag{3}$$

where y_{it} is the outcome variable of interest in country i at date t. We consider industrial output per capita and income per capita as two different measures of economic success during the nineteenth century. In addition, μ_t's are a set of time effects, and δ_i's denote a set of country effects, $UKIND_t$ is industrial output in the United Kingdom at date t, and X_{it} denotes the measure of institutions in country i at date t. Our institutions variable is again constraint on the executive from the Gurr Polity III data set. As noted above, this variable is available from the date of independence for each country. Since colonial rule typically concentrated political power in the hands of a small elite, for the purpose of the regressions in this table, we assign the lowest score to countries still under colonial rule. The coefficient of interest is ϕ, which reflects whether there is an interaction between good institutions and the opportunity to industrialize. A positive and significant ϕ is interpreted as evidence in favor of the view that countries with institutions of private property took better advantage of the opportunity to industrialize. The parameter π measures the direct effect of institutions on industrialization, and is evaluated at the mean value of $UKIND_t$.

The top panel of Table IX reports regressions of equation (3) with industrial output per capita as the left-hand-side variable (see the note to the table for more details). Column (1) reports a regression using only pre-1950 data. The interaction term ϕ is estimated to be 0.132, and is highly significant with a standard error of 0.26. Note that Bairoch's estimate of total U. K. industrialization, which is normalized to 100 in 1900, rose from 16 to 115 between 1800 and 1913. In the meantime, the U. S. per capita production grew from 9 to 126, whereas India's per capita industrial production *fell* from 6 to 2. Since the average difference between the constraint on the executive in the United States and India over this period is approximately 6, the estimate implies that the U. S. industrial output per capita should have increased by 78 points more than India's, which is over half the actual difference.

In column (2) we extend the data through 1980, again with no effect on the coefficient, which stays at 0.132. In columns (3) and (4) we investigate whether independence impacts on industrialization, and whether our procedure of assigning the lowest score to countries still under colonial rule may be driving our results. In column (3) we include a dummy for whether the country is independent, and also interact this dummy with U. K. industrialization. These variables are insignificant, and the coefficient on the interaction between U. K. industrialization and institutions, ϕ, is unchanged (0.145 with standard error 0.035). In column (4) we drop all observations from countries still under colonial rule, and this again has no effect on the results (ϕ is now estimated to be 0.160 with standard error 0.048).

In columns (5) and (6) we use average institutions for each country, $\bar{X}_i$, rather than institutions at date t, so the equation becomes

$$y_{it} = \mu_t + \delta_i + \phi \cdot \bar{X}_i \cdot UKIND_t + \epsilon_{it}.$$

This specification may give more sensible results if either variations in institutions from year to year are endogenous with respect to changes in industrialization or income, or are subject to measurement error. ϕ is now estimated to be larger, suggesting that measurement error is a more important problem than the endogeneity of the changes in institutions.

An advantage of the specification in columns (5) and (6) is that it allows us to instrument for the regressor of interest $\bar{X}_i \cdot UKIND_t$, using the interaction between U. K. industrialization and our instrument for institutions, log settler mortality M_i (so the instrument here is $M_i \cdot UKIND_t$). Once again, institutions might differ across countries because more productive or otherwise different countries have different institutions, and in this case, the interaction between industrialization and institutions could be capturing the direct effects of these characteristics on economic performance. To the extent that log settler mortality is a good instrument for institutions, the interaction between log settler mortality and U. K. industrialization will be a good instrument for the interaction between institutions and U. K. industrialization. The instrumental-variables procedure will then deal with the endogeneity of institutions, the omitted variables bias, and also the attenuation bias due to measurement error. The

TABLE IX
THE INTERACTION OF U. K. INDUSTRIALIZATION AND INSTITUTIONS

	Former colonies, using only pre-1950 data (1)	Former colonies, using data through 1980 (all data) (2)	Former colonies, using only pre-1950 data (3)	Former colonies, using only data pre-1950 and for independent countries (4)	Former colonies, with average institutions for each country, using only pre-1950 data (5)	Former colonies, with average institutions for each country, using only pre-1590 data (6)	Former colonies, with average institutions for each country, instrumenting using settler mortality, only pre-1950 data (7)	Former colonies, with average institutions for each country, instrumenting using settler mortality, only pre-1950 data (8)	Former colonies, with average institutions for each country, instrumenting using settler mortality, only pre-1950 data (9)	Former colonies, with average institutions for each country, instrumenting using settler mortality, only pre-1950 data (10)
Panel A: Dependent variable is industrial production per capita										
U. K. industrialization	0.132	0.132	0.145	0.160	0.202	0.206	0.168	0.169	0.156	0.158
*institutions	(0.026)	(0.027)	(0.035)	(0.048)	(0.019)	(0.022)	(0.030)	(0.032)	(0.065)	(0.065)
Institutions	8.97	−3.36	10.51	7.48						
	(2.30)	(4.46)	(3.50)	(9.51)						
Independence			−14.3			−6.4		1.1		2.0
			(22.9)			(11.4)		(12.6)		(14.2)
U. K. industrialization			−0.12			−0.042		0.046		0.06
*independence			(0.21)			(0.12)		(0.13)		(0.17)
U. K. industrialization									0.13	0.12
*latitude									(0.50)	(0.48)
R^2	0.75	0.74	0.75	0.84	0.89	0.89	0.88	0.88	0.87	0.87
Number of observations	59	75	59	32	59	59	59	59	59	59

Panel B: Dependent variable is log GDP per capita

Log U. K. industrialization	0.078	0.060	0.073	0.079	0.135	0.130	0.159	0.150	0.116	0.111
*institutions	(0.022)	(0.017)	(0.027)	(0.025)	(0.021)	(0.026)	(0.032)	(0.038)	(0.067)	(0.073)
Institutions	−0.027	−0.084	−0.10	−0.11						
	(0.025)	(0.028)	(0.04)	(0.04)						
Independence			0.67			0.12		0.10		0.019
			(0.27)			(0.13)		(0.13)		(0.16)
Log U. K. industrialization			0.035			−0.008		−0.042		0.016
*independence			(0.12)			(0.093)		(0.11)		(0.14)
Log U. K. industrialization									0.42	0.42
*latitude									(0.49)	(0.54)
R^2	0.95	0.92	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96
Number of observations	79	131	79	46	79	79	79	79	79	79

Standard errors are in parentheses. All columns report panel regressions with country and period dummies included. Dependent variable in Panel A is industrial output per capita 1750–1980 from Bairoch [1982]. Dependent variable in Panel B is log GDP per capita 1830–1980 from Maddison [1995]. The institutions variable is "Constraint on the executive," which is an assessment of the constitutional limitations on executive power. The independent variable of interest is total U. K. industrial output interacted with constraint on the executive in each country from the Polity III data set. The main effect of institutions is evaluated at the mean value of U. K. industrialization. Polity III provides information only for independent countries; if a country was a colony at a particular date, we assign the lowest value of constraints on the executive, which is 1. Average institutions are calculated over the values in Polity III for 1750, 1800, 1830, 1860, 1880, 1913, and 1928.

We have an unbalanced panel with the following observations. For industrial output we have data on Australia, Brazil, Canada, India, Mexico, New Zealand, South Africa, and the United States. In the panel regressions for GDP per capita before 1950, we have data on these countries (except South Africa) plus Argentina, Bangladesh, Burma/Myanmar, Chile, Colombia, Egypt, Ghana, India, Indonesia, Pakistan, Peru, and Venezuela. In addition, for the regression using GDP per capita data through 1980, we are also able to include Ethiopia, Ivory Coast, Kenya, Morocco, Nigeria, South Africa, Tanzania, and Zaire. We have data for the following dates: 1750, 1800, 1830, 1860, 1880, 1913, 1928, 1953, and 1980, although not for all countries for all dates. For detailed sources and descriptions see Appendix 2.

2SLS estimates reported in columns (7) and (8) are very similar to the OLS estimates in columns (5) and (6), and are highly significant.[23]

In columns (9) and (10) we add the interaction between latitude and industrialization. This is useful because, if the reason why the United States surged ahead relative to India or South America during the nineteenth century is its geographic advantage, our measures of institutions might be proxying for this, incorrectly assigning the role of geography to institutions. The results give no support to this view: the estimates of ϕ are affected little and remain significant, while the interaction between industrialization and latitude is insignificant. Panel B of Table IX repeats these regressions using log GDP per capita as the left-hand-side variable (the interaction term is now as $M_i \cdot \ln(UKIND_t)$ since the left-hand-side variable is log of GDP per capita). The results are broadly similar to those in Panel A.

Overall, these results provide support for the view that institutions played an important role in the process of economic growth and in the surge of industrialization among the formerly poor colonies, and via this channel, account for a significant fraction of current income differences.

VI. Conclusion

Among the areas colonized by European powers during the past 500 years, those that were relatively rich in 1500 are now relatively poor. Given the crude nature of the proxies for prosperity 500 years ago, some degree of caution is required, but the broad patterns in the data seem uncontroversial. Civilizations in Meso-America, the Andes, India, and Southeast Asia were richer than those located in North America, Australia, New Zealand, or

23. Despite our instrumental-variables strategy, the interaction between institutions and the opportunity to industrialize may capture the possible interaction between industrialization and some country characteristics correlated with our instrument. For example, with an argument along the lines of Nelson and Phelps [1966] or Acemoglu and Zilibotti [2001], one might argue that industrial technologies were appropriate only for societies with sufficient human capital, and that there were systematic cross-country differences in human capital correlated with institutional differences. This interpretation is consistent with our approach, since the correlation between institutions and human capital most likely reflects the fact that in societies with extractive institutions the masses typically did not or could not obtain education. In other words, low levels of human capital may have been a primary mechanism through which extractive institutions delayed industrialization.

the southern cone of Latin America. The intervention of Europe reversed this pattern. This is a first-order fact, both for understanding economic and political development over the past 500 years, and for evaluating various theories of long-run development.

This reversal in relative incomes is inconsistent with the simple geography hypothesis which explains the bulk of the income differences across countries by the direct effect of geographic differences, thus predicting a high degree of persistence in economic outcomes. We also show that the timing and nature of the reversal do not offer support to sophisticated geography views, which emphasize the time-varying effects of geography. Instead, the reversal in relative incomes over the past 500 years appears to reflect the effect of institutions (and the institutional reversal caused by European colonialism) on income today.

Why did European colonialism lead to an institutional reversal? And how did this institutional reversal cause the reversal in relative incomes and the subsequent divergence in income per capita across the various colonies? We argued that the institutional reversal resulted from the differential profitability of alternative colonization strategies in different environments. In prosperous and densely settled areas, Europeans introduced or maintained already-existing extractive institutions to force the local population to work in mines and plantations, and took over existing tax and tribute systems. In contrast, in previously sparsely settled areas, Europeans settled in large numbers and created institutions of private property, providing secure property rights to a broad cross section of the society and encouraging commerce and industry. This institutional reversal laid the seeds of the reversal in relative incomes. But most likely, the scale of the reversal and the subsequent divergence in incomes are due to the emergence of the opportunity to industrialize during the nineteenth century. While societies with extractive institutions or those with highly hierarchical structures could exploit available agricultural technologies relatively effectively, the spread of industrial technology required the participation of a broad cross section of the society—the smallholders, the middle class, and the entrepreneurs. The age of industry, therefore, created a considerable advantage for societies with institutions of private property. Consistent with this view, we documented that these societies took much better advantage of the opportunity to industrialize.

Appendix 1: Urbanization Estimates

This is a shortened version of the Appendix in Acemoglu, Johnson, and Robinson [2001b].

1. Urbanization in 1500

Our base estimates for 1500 consist of Bairoch's [1988] assessment of urbanization augmented by the work of Eggimann [1999]. Merging these two series requires us to convert Eggimann's estimates, based on a minimum population threshold of 20,000, into Bairoch-equivalent urbanization estimates, based on a minimum population threshold of 5000.

To construct our base data, we run a regression of Bairoch estimates on Eggimann estimates for all countries where they overlap in 1900 (the year for which we have the largest number of Bairoch estimates for non-European countries). There are thirteen countries for which we have good overlapping data. This regression yields a constant of 6.6 and a coefficient of 0.67.

We use these results to convert from Eggimann to Bairoch-equivalent urbanization estimates in Colombia, Ecuador, Guatemala (and other parts of Central America), Mexico, and Peru in the Americas. We also use this method for all North African countries and for India (and the rest of the Indian subcontinent), Indonesia, Malaysia, Laos, Burma/Myanmar, and Vietnam in Asia. See Appendix 2 for the precise numbers we use.

There are a number of countries for which Bairoch determines that there was no real urbanization or no pre-European "settled agriculture." In these cases, a reasonable interpretation of Bairoch is that there was no urban population using his definition. In our baseline data we therefore assume zero urbanization for the following countries: Argentina, Brazil, Canada, Chile, Guyana, Paraguay, Uruguay, the United States, and Australia.

For countries where Bairoch determines there was some low level of urbanization, associated with fairly primitive agriculture, he assesses that the urbanization rate was 3 percent. We use this estimate for Cuba, the Dominican Republic, Haiti, and Jamaica in the Americas. We also use this estimate for Hong Kong, the Philippines, and Singapore in Asia and for New Zealand. In the Appendix of Acemoglu, Johnson, and Robinson [2001b], we present qualitative evidence documenting the low levels of urbanization in countries with assigned values of 0 percent or 3 percent urbanization in our baseline data.

While the data on sub-Saharan Africa are worse than for any other region, it is clear that urbanization before 1500 was at a higher level than North America or Australia (see the Appendix of Acemoglu, Johnson, and Robinson [2001b] for detailed discussion and sources). Given the weakness and incompleteness of data for sub-Saharan Africa, we do not include any estimates in our baseline urbanization data set. We do, however, include all of sub-Saharan Africa in our baseline population density data.

We have checked the robustness of our results using alternative methods of converting Eggimann estimates into Bairoch-equivalent numbers. We have calculated conversion ratios at the regional level (e.g., for North Africa and the Andean region separately). We have also constructed an alternative series using a conversion rate of 2, as suggested by Davis' and Zipf's Laws (see Bairoch [1988], Chapter 9.)[24] We have also used Bairoch's overall assessment of urbanization for broad regions, e.g., Asia, without the more detailed information from Eggimann (see the Appendix in Acemoglu, Johnson, and Robinson [2001b] for more detail). We have also used estimates just from Bairoch, just from Eggimann, and just from Chandler. See Table IV for relevant regressions.

Our baseline estimates and the most plausible alternative series are shown in Appendix 2. We have also calculated urbanization rates for all European countries and non-European countries that were never colonized. We have also checked Bairoch's estimates carefully for these countries against the work of Bairoch, Batou, and Chèvre [1988], Chandler and Fox [1974], de Vries [1984], and Hohenberg and Lees [1985]. Our discussion of urbanization in European and never colonized countries is not reported here to conserve space, but it is available from the authors.

2. Urbanization from 1500 to 2000

Eggimann's data only cover countries that are now part of the "Third World." He therefore does not provide any information on the timing of urbanization changes in settler colonies. Bairoch does have some information on urbanization in the United States, Canada, and Australia, but only from 1800 [Bairoch 1988, Table 13.4, p. 221]. For a more complete picture of urbanization from 800 to 1850 across a wide range of countries, we therefore rely

24. We are using a conservative version of Davis' law. See the Appendix in Acemoglu, Johnson, and Robinson [2001b] for a more detailed discussion.

primarily on Chandler's estimates. We should emphasize, however, that wherever there is overlapping information, these estimates are broadly consistent with the findings of Eggimann and Bairoch.[25] As before, we convert urban population numbers into urbanization using population estimates from McEvedy and Jones [1978].

Chandler's data enable us to see changes in urbanization over time across countries, but because his series ends in 1850 (or 1861 for the Americas), we cannot follow the most important trends into the twentieth century. In addition, Chandler's data are reported at 50-year intervals from 1700 (100-year intervals before that), which is only enough to show the broad pattern.

We therefore supplement the analysis with data from two other sources. The UN [1969] provides detailed urbanization data from 1920, focusing on localities with 20,000 or more inhabitants (i.e., the same criterion as Chandler uses outside of Asia). However, this still leaves a gap between 1850 and 1920.

We complete this composite series using data from Mitchell [1993, 1995]. His urbanization data start in 1750, provide information every ten years from 1790 for most countries, and run to 1980. The only disadvantage of this series is the relatively late starting date. The criterion for inclusion in Mitchell's series is also a little different—cities that had at least 200,000 inhabitants around 1970—but this seems to produce broadly consistent estimates for overlapping observations. We use these data both to complete the Chandler series for Mexico, India, and the United States (see Figure IVa) and to provide alternative estimates for the timing of urbanization changes within the Americas.

The data shown in Figure IVa are from Chandler (through 1850), Mitchell (for 1900), and the UN (for 1920 and 1930), converted to Bairoch-equivalent units using the conservative Zipf-Davis adjustment (i.e., multiplying the estimates by 2).

25. The only point of disagreement is whether there was any urbanization in the area now occupied by the United States in 1500. Chandler lists one town (Nanih Waiya) but does not give its population. He also does not indicate any urbanization either before or after this date. Bairoch argues there was no pre-European urbanization and the latest archaeological evidence suggests villages rather than towns [Fagan 2000]. We therefore follow Bairoch in assigning a value of zero. For supportive evidence see Waldman [1985, p. 30].

APPENDIX 2: VARIABLE DEFINITIONS AND SOURCES

Variable	Description	Source
Log GDP per capita (PPP) in 1995	Logarithm of GDP per capita, on Purchasing Power Parity Basis, in 1995.	World Bank, World Development Indicators, CD-Rom, 1999. Data on Suriname is from the 2000 version of this same source.
Log GDP per capita in 1900 and 1950	Logarithm of GDP per capita in 1900 and 1950.	Maddison [1995] for 1950; Bairoch [1978] for 1900.
Industrial output per capita	Index of industrialization with Britain in 1900 equal to 100.	Bairoch [1982].
Total U. K. industrial output	Index equal to 100 in 1900.	Bairoch [1982].
Log population density in 1 A.D., 1000, and 1500 (also log population in 1500 and log arable land in 1500)	Logarithm of population density (total population divided by total arable land) in 1 A.D., 1000, 1500.	McEvedy and Jones [1978].
Urbanization in 1960 and 1995	Percent of population living in urban areas in 1960 and 1995, as defined by the UN (typically 20,000 minimum inhabitants).	World Bank, World Development Indicators, CD-Rom, 1999. For more detail, see p. 159 of the World Bank's World Development Indicators 1999 (hard copy).
Urbanization in 1000, 1500, and 1700	Percent of population living in urban areas with a population of at least 5000 in 1000, 1500, and 1700.	Bairoch and supplemental sources, as described in Appendix 1.
European settlements in 1800 and 1900	Percent of population that was European or of European descent in 1800 and 1900. Ranges from 0 to 0.99 in our base sample.	McEvedy and Jones [1978] and other sources listed in Appendix Table 5 of Acemoglu, Johnson, and Robinson [2000].
Average protection against expropriation risk, 1985–1995	Risk of expropriation of private foreign investment by government, from 0 to 10, where a higher score means less risk. We calculated the mean value for the scores in all years from 1985 to 1995.	Data set obtained directly from Political Risk Services, September 1999. These data were previously used by Knack and Keefer [1995] and were organized in electronic form by the IRIS Center (University of Maryland). The original compilers of these data are Political Risk Services.

APPENDIX 2
(CONTINUED)

Variable	Description	Source
Constraint on executive in 1970, 1990, and first year of independence	A seven-category scale, from 1 to 7, with a higher score indicating more constraints. Score of 1 indicates unlimited authority; score of 3 indicates slight to moderate limitations; score of 5 indicates substantial limitations; score of 7 indicates executive parity or subordination. Scores of 2, 4, and 6 indicate intermediate values.	Polity III data set, downloaded from Inter-University Consortium for Political and Social Research. Variable described in Gurr [1997].
Percent of European descent in 1975 religion variables	Percent of population that was European or of European descent in 1975. Ranges from 0 to 1 in our base sample.	McEvedy and Jones [1978].
	Percentage of the population that belonged in 1980 (or for 1990–1995 for countries formed more recently) to the following religions: Roman Catholic, Protestant, Muslim, and "other."	La Porta et al. [1999].
Colonial dummies	Dummy variable indicating whether country was a British, French, German, Spanish, Italian, Belgian, Dutch, or Portuguese colony.	La Porta et al. [1999].
Temperature variables	Temperature variables are average temperature, minimum monthly high, maximum monthly high, minimum monthly low, and maximum monthly low, all in centigrade.	Parker [1997].

Humidity variables	Humidity variables are morning minimum, morning maximum, afternoon minimum, and afternoon maximum, all in percent	Parker [1997].
Soil quality	Measures of soil quality/climate are steppe (low latitude), desert (low latitude), steppe (middle latitude), desert (middle latitude), dry steppe wasteland, desert dry winter, and highland.	Parker [1997].
Natural resources	Measures of natural resources are percent of world gold reserves today, percent of world iron reserves today, percent of world zinc reserves today, percent of world silver reserves today, and oil resources (thousands of barrels per capita today).	Parker [1997].
Coal	Dummy variable equal to 1 if country has produced coal since 1800.	World Resources Institute [1998] and Etemad and Toutain [1991].
Landlocked	Dummy variable equal to 1 if country does not adjoin the sea.	Parker [1997].
Island	Dummy variable equal to 1 if country is an island.	DK Publishing [1997].
Latitude	Absolute value of the latitude of the country, scaled to take values between 0 and 1, where 0 is the equator.	La Porta et al. [1999].
Log mortality	Log of estimated settler mortality. Settler mortality is calculated from the mortality rates of European-born soldiers, sailors, and bishops when stationed in colonies. It measures the effects of local diseases on people without inherited or acquired immunities.	Acemoglu, Johnson, and Robinson [2001a], based on Curtin [1989] and other sources.

APPENDIX 3

	Base urbanization estimate in 1500	Source of base urbanization estimate in 1500	Urbanization estimate in 1500 using only information from Bairoch	Urbanization estimate in 1500 using only information from Eggimann	Urbanization estimate in 1500 using only information from Chandler	Davis-Zipf adjustment applied to Eggimann series	Population density in 1500		Population density in 1500		Population density in 1500
	Former colonies included in our base sample for urbanization							Former colonies included in base sample for population density but not for urbanization			
Argentina	0.0	Bairoch	0.0	0.0	0.0	0.0	0.11	Angola	1.50	Sudan	14.03
Australia	0.0	Bairoch	0.0	0.0	0.0	0.0	0.03	Bahamas	1.46	Suriname	0.21
Bangladesh	8.5	Eggimann converted to Bairoch	9.0	2.9	.	5.8	23.70	Barbados	1.46	Tanzania	1.98
Belize	9.2	Eggimann (3.8%) converted to Bairoch	7.0	18.0	19.6	7.6	1.54	Benin	4.23	Togo	4.23
Bolivia	10.6	Eggimann (Ecuador and Bolivia) converted to Bairoch	12.0	6.0	.	12.0	0.83	Botswana	0.14	Trinidad and Tobago	1.46
Brazil	0.0	Bairoch	0.0	0.1	.	0.2	0.12	Burkina Faso	4.23	Uganda	7.51
Canada	0.0	Bairoch	0.0	0.0	0.0	0.0	0.02	Burundi	25.00	Zaire	1.50
Chile	0.0	Bairoch	0.0	0.0	0.0	0.0	0.80	Cameroon	1.50	Zambia	0.79
Colombia	7.9	Eggimann converted to Bairoch	7.0	2.0	2.0	4.0	0.96	Cape Verde	0.50	Zimbabwe	0.79
Costa Rica	9.2	Eggimann (3.8%) converted to Bairoch	7.0	18.0	.	7.6	1.54	Central African Republic	1.50		
Dominican Republic	3.0	Bairoch	3.0	0.0	.	0.0	1.46	Chad	1.00		

Algeria	14.0	Eggimann converted to Bairoch	.	11.0	11.0	22.0	7.00	Comoros	4.48
Ecuador	10.6	Eggimann (Ecuador and Bolivia) converted to Bairoch	12.0	6.0	5.0	12.0	2.17	Congo	1.50
Egypt	14.6	Eggimann converted to Bairoch	.	11.9	12.4	23.8	100.46	Cote d'Ivoire	4.23
Guatemala	9.2	Eggimann (3.8%) converted to Bairoch	7.0	18.0	19.6	7.6	1.54	Dominica	1.46
Guyana	0.0	Bairoch	0.0	0.0	.	0.0	0.21	Eritria	2.00
Hong Kong	3.0	Bairoch	3.0	0.0	0.0	0.0	0.09	Ethiopia	6.67
Honduras	9.2	Eggimann (3.8%) converted to Bairoch	7.0	18.0	19.6	7.6	1.54	Gabon	1.50
Haiti	3.0	Bairoch	3.0	0.0	.	0.0	1.32	Gambia	4.23
Indonesia	7.3	Eggimann (Indonesia and Malaysia) converted to Bairoch	9.0	1.0	0.5	2.0	4.28	Ghana	4.23
India	8.5	Eggimann converted to Bairoch	9.0	2.9	1.8	5.8	23.70	Grenada	1.46
Jamaica	3.0	Bairoch	3.0	0.0	.	0.0	4.62	Guinea	4.23
Laos	7.3	Eggimann (Laos and Vietnam) converted to Bairoch	9.0	10.0	10.0	20.0	1.73	Kenya	2.64
Sri Lanka	8.5	Eggimann converted to Bairoch	9.0	2.9	.	5.8	15.47	Lesotho	0.49
Morocco	17.8	Eggimann converted to Bairoch	.	16.7	21.3	33.3	9.08	Madagascar	1.20
Mexico	14.8	Eggimann converted to Bairoch	7.0	12.3	6.5	24.6	2.62	Malawi	0.79

APPENDIX 3
(CONTINUED)

	Base urbanization estimate in 1500	Source of base urbanization estimate in 1500	Urbanization estimate in 1500 using only information from Bairoch	Urbanization estimate in 1500 using only information from Eggimann	Urbanization estimate in 1500 using only information from Chandler	Davis-Zipf adjustment applied to Eggimann series	Population density in 1500		Population density in 1500	Population density in 1500
	Former colonies included in our base sample for urbanization							Former colonies included in base sample for population density but not for urbanization		
Malaysia	7.3	Eggimann (Indonesia and Malaysia) converted to Bairoch	9.0	1.0	0.5	2.0	1.22	Mali	1.00	
Nicaragua	9.2	Eggimann (3.8%) converted to Bairoch	7.0	18.0	19.6	7.6	1.54	Mauritania	3.00	
New Zealand	3.0	Bairoch	3.0	0.0	.	0.0	0.37	Mozambique	1.28	
Pakistan	8.5	Eggimann converted to Bairoch	9.0	2.9	.	5.8	23.70	Namibia	0.14	
Panama	9.2	Eggimann (3.8%) converted to Bairoch	7.0	18.0	19.6	7.6	1.54	Nepal	13.99	
Peru	10.5	Eggimann converted to Bairoch	12.0	5.8	2.5	11.6	1.56	Niger	1.00	
Philippines	3.0	Bairoch	3.0	0.0	.	0.0	1.68	Nigeria	4.23	
Paraguay	0.0	Bairoch	0.0	0.0	.	0.0	0.50	Rwanda	25.00	
Singapore	3.0	Bairoch	3.0	0.0	0.0	0.0	0.09	Swaziland	0.49	

El Salvador	9.2	Eggimann (3.8%) converted to Bairoch	7.0	18.0	19.6	7.6	1.54	Senegal	4.23
Tunisia	12.3	Eggimann converted to Bairoch	.	8.1	11.3	16.3	11.70	Sierra Leone	4.23
Uruguay	0.0	Bairoch	0.0	0.0	.	0.0	0.11	South Africa	0.49
U. S. A.	0.0	Bairoch	0.0	0.0	0.0	0.0	0.09	St. Lucia	1.46
Venezuela	0.0	Bairoch	0.0	0.0	.	0.0	0.44	St. Vincent	1.46
Vietnam	7.3	Eggimann (Laos and Vietnam) converted to Bairoch	9.0	10.0	2.0	20.0	6.14	St. Kitts and Nevis	1.46

Our base urbanization estimates are constructed using information from Bairoch and a conversion from Eggimann's estimates to Bairoch-equivalent estimates (as explained in the text and Appendix 1). Bairoch-only estimates use 9 percent for all Asian countries, 7 percent for Central America and Colombia, 12 percent for Andean countries, 3 percent for countries with minimal urbanization, and 0 percent for all other countries in our base sample. Eggimann-only estimates are not adjusted to Bairoch-equivalent units, and we use zero for countries in his data set without any urban population in 1500. Chandler-only estimates are not adjusted to Bairoch-equivalent units, and we use a value of zero for countries that are in his data set and for which he does not indicate any urban population in 1500. The Davis-Zipf adjustment doubles Eggimann's estimates but uses a low estimate for Central America (details are in the Appendix of Acemoglu, Johnson, and Robinson [2001b]. Population density numbers are calculated from population in McEvedy and Jones [1978]. We divide estimated population in 1500 by land area in 1995 (from World Bank [1999]), adjusted for arable land area using the estimates in McEvedy and Jones [1978]. Where McEvedy and Jones [1978] only provide a regional population estimate, we use their regional land area estimate adjusted for arable land.

In some cases McEvedy and Jones [1978] only provide regional estimates of population in 1500. We therefore use regional averages of population density for: West Africa (Senegal, Gambia, Guinea, Sierra Leone, Ivory Coast, Ghana, Burkina Faso, Togo, Benin, and Nigeria); West-Central Africa (Cameroon, Central African Republic, Gabon, Congo, Zaire, and Angola); Rwanda and Burundi; South-Central Africa (Zambia, Zimbabwe, and Malawi); South Africa, Swaziland, and Lesotho; Namibia and Botswana; the Sahel States (Mauritania, Mali, Niger, and Chad—based on qualitative evidence we assume a slightly higher population density in Mauritania); Eritrea and Ethiopia (based on qualitative evidence we assume a higher population density in Ethiopia); Central America (Guatemala, Belize, El Salvador, Honduras, Nicaragua, Costa Rica, and Panama); Guyana and Suriname are calculated from the average for all the Guyanas; and Pakistan, India, and Bangladesh are calculated from the average for the Indian subcontinent. The population density in Uruguay is assumed to be the same as in Argentina in 1500. Singapore and Hong Kong are assumed to have the same population density as the United States in 1500. Smaller Caribbean islands are assumed to have the same population density as the Dominican Republic in 1500.

A period (.) denotes missing data. For further discussion of sources, see Appendix 1.

DEPARTMENT OF ECONOMICS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SLOAN SCHOOL OF MANAGEMENT, MASSACHUSETTS INSTITUTE OF TECHNOLOGY
DEPARTMENTS OF POLITICAL SCIENCE AND ECONOMICS, UNIVERSITY OF CALIFORNIA, BERKELEY

REFERENCES

Abu-Lughod, Janet L., *Before European Hegemony: The World System A.D. 1250–1350* (Oxford, UK and New York, NY: Oxford University Press, 1989).

Acemoglu, Daron, Simon Johnson, and James A. Robinson, "The Colonial Origins of Comparative Development: An Empirical Investigation," NBER Working Paper No. 7771, 2000.

Acemoglu, Daron, Simon Johnson, and James A. Robinson, "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review,* XCI (2001a), 1369–1401.

Acemoglu, Daron, Simon Johnson, and James A. Robinson, "Reversal of Fortune: Geography and Institutions in the Making of the World Income Distribution," NBER Working Paper No. 8460, 2001b.

Acemoglu, Daron, and James A. Robinson, "Political Losers as a Barrier to Economic Development," *American Economic Review Papers and Proceedings,* XC (2000), 126–130.

Acemoglu, Daron, and James A. Robinson, "Economic Backwardness in Political Perspective," unpublished, 2001.

Acemoglu, Daron, and Fabrizio Zilibotti, "Productivity Differences," *Quarterly Journal of Economics,* CXVI (2001), 563–606.

Ades, Alberto F., and Edward L. Glaeser, "Evidence on Growth, Increasing Returns, and the Extent of the Market," *Quarterly Journal of Economics,* CXIV (1999), 1025–1046.

Bairoch, Paul, "Les grandes tendances des disparités économiques nationales depuis la revolution industrielle," in P. Bairoch and M. Levy Laboyer, eds., *Regional and International Disparities in Economic Development Since the Industrial Revolution,* 7th International Economic History Conference (Edinburgh: 1978).

——, "International Industrialization Levels from 1750 to 1980," *Journal of European Economic History,* XI (1982), 269–333.

——, *Cities and Economic Development: From the Dawn of History to the Present* (Chicago, IL: University of Chicago Press, 1988).

Bairoch, Paul, Jean Batou, and Pierre Chèvre, *La Population des villes Européenes de 800 à 1850: Banque de Données et Analyse Sommaire des Résultats* (Geneva: Centre d'histoire économique Internationale de l'Université de Genève, Libraire Droz, 1988).

Bloch, Marc, *Land and Work in Medieval Europe* (New York: Harper & Row, 1966).

Boserup, Ester, *The Conditions of Agricultural Growth: The Economics of Agrarian Change under Population Pressure* (Chicago, IL: Aldine Publishing Company, 1965).

Braudel, Fernand, *The Structures of Everyday Life: Civilization and Capitalism, Fifteenth–Eighteenth Century* (Berkeley and Los Angeles, CA: University of California Press, 1992).

Chandler, Tertius, *Four Thousand Years of Urban Growth: An Historical Census* (Lewiston, NY: St. David's University Press, 1987).

Chandler, Tertius, and Gerald Fox, *Three Thousand Years of Urban Growth* (New York, NY: Academic Press, 1974).

Chaudhuri, Kirti N., *Asia Before Europe: Economy and Civilization of the Indian Ocean from the Rise of Islam to 1750* (New York, NY: Cambridge University Press, 1990).

Coatsworth, John H., "The Limits of Colonial Absolutism: Mexico in the Eighteenth Century," in Karen Spalding ed., *Essays in the Political, Economic and Social History of Latin America* (Newark, DE: University of Delaware Press, 1982).

——, "Notes on the Comparative Economic History of Latin America and the United States," in Walter L. Bernecker and Hans Werner Tobler, eds., *Development and Underdevelopment in America: Contrasts in Economic Growth in North and Latin America in Historical Perspective* (New York, NY: Walter de Gruyter, 1993).

Cole, Jeffrey A., *The Potosi Mita, 1573–1700: Compulsory Indian Labor in the Andes* (Palo Alto, CA: Stanford University Press, 1985).

Collins, K. J., and D. F. Roberts, eds., *Capacity for Work in the Tropics* (Cambridge, UK: Cambridge University Press, 1988).

Crosby, Alfred, *Ecological Imperialism: The Biological Expansion of Europe 900–1900* (New York, NY: Cambridge University Press, 1986).

Curtin, Philip D., *Death by Migration* (Cambridge, UK: Cambridge University Press, 1989).

De Long, J. Bradford, and Andrei Shleifer, "Princes and Merchants: European City Growth before the Industrial Revolution," *Journal of Law and Economics,* XXXVI (1993), 671–702.

de Vries, Jan, *The Economy of Europe in an Age of Crisis, 1600–1750* (Cambridge, UK: Cambridge University Press, 1976).

——, *European Urbanization, 1500–1800* (Cambridge, MA: Harvard University Press, 1984).

Denoon, Donald, *Settler Capitalism: The Dynamics of Dependent Development in the Southern Hemisphere* (Oxford, UK: Clarendon Press, 1983).

Diamond, Jared M., *Guns, Germs and Steel: The Fate of Human Societies* (New York, NY: W.W. Norton & Co., 1997).

DK Publishing, *World Atlas* (New York, NY: DK Publishing, 1997).

Duby, Georges, *The Early Growth of the European Economy; Warriors and Peasants from the Seventh to the Twelfth Century* (Ithaca, NY: Cornell University Press, 1974).

Dunn, Richard S., *Sugar and Slaves: The Rise of the Planter Class in the English West Indies 1624–1713* (Chapel Hill, NC: University of North Carolina Press, 1972).

Eggimann, Gilbert, *La Population des villes des Tiers-Mondes, 1500–1950* (Geneva: Centre d'histoire économique Internationale de l'Université de Genève, Libraire Droz, 1999).

Eltis, David, "The Total Product of Barbados, 1664–1701," *Journal of Economic History,* LV (1995), 321–336.

Engerman, Stanley L., "Notes on the Patterns of Economic Growth in the British North America Colonies in the Seventeenth, Eighteenth and Nineteenth Centuries" in *Disparities in Economic Development since the Industrial Revolution,* Paul Bairoch and Maurice Levy-Leboyer, eds. (New York, NY: St. Martin's Press, 1981).

Engerman, Stanley L., and Kenneth L. Sokoloff, "Factor Endowments, Institutions, and Differential Paths of Growth among New World Economies," in S. H. Haber, ed., *How Latin America Fell Behind* (Stanford, CA: Stanford University Press, 1997).

Engerman, Stanley L., and Kenneth L. Sokoloff, "Institutions, Factor Endowments, and Paths of Development in the New World," *Journal of Economic Perspectives,* XIV (2000), 217–232.

Etemad, Bouda, and Jean-Claude Toutain, *World Energy Production 1800–1985* (Geneva: Libraire Droz, 1991).

Fagan, Brian M., *Ancient North America: The Archaeology of a Continent,* third edition (London, UK: Thames & Hudson, 2000).

Frank, Andre Gunder, *Dependent Accumulation and Underdevelopment* (London, UK: Macmillan Press, 1978).

Galenson, David W., "The Settlement and Growth of the Colonies: Population, Labor and Economic Development," in Stanley L. Engerman and Robert E. Gallman, eds., *The Cambridge Economic History of the United States,* Volume I (New York, NY: Cambridge University Press, 1996).

Gallman, Robert E., "Economic Growth and Structural Change in the Long Nineteenth Century," in Stanley L. Engerman and Robert E. Gallman, eds., *The Cambridge Economic History of the United States,* Volume II (New York, NY: Cambridge University Press, 2000).

Gurr, Ted Robert, "Polity II: Political Structures and Regime Change, 1800–1986," unpublished paper, Boulder, CO, University of Colorado, 1997.
Hall, Robert E., and Charles I. Jones, "Why Do Some Countries Produce so Much More Output per Worker than Others?" *Quarterly Journal of Economics,* CXIV (1999), 83–116.
Harms, Robert C., *River of Wealth, River of Sorrow: The Central Zaire Basin in the Era of the Slave and Ivory Trade, 1500–1891* (New Haven, CT: Yale University Press, 1981).
Hayek, Freidrich von, *The Constitution of Liberty* (Chicago, IL: University of Chicago Press, 1960).
Hodgson, Marshall G. S., *Essays on Europe, Islam and World History* (Cambridge, UK: Cambridge University Press, 1993).
Hohenberg, Paul M., and Lynn Hollen Lees, *The Making of Urban Europe, 1000–1950* (Cambridge, MA: Harvard University Press, 1985).
Hopkins, Anthony G., *An Economic History of West Africa* (New York, NY: Addison-Wesley Longman, 1973).
Hughes, Robert, *The Fatal Shore* (New York, NY: Vintage Books, 1988).
International Labour Organization (ILO), *Conditions of Work Digest,* XIV (Geneva: ILO, 1995).
Jones, Charles I., "Population and Ideas: A Theory of Endogenous Growth," NBER Working Paper No. 6285, 1997.
Kahn, Zorina, and Kenneth Sokoloff, "Patent Institutions, Industrial Organization and Early Technological Change: Britain and the United States, 1790–1850," in Maxine Berg and Kristine Bruland, eds., *Technological Revolutions in Europe: Historical Perspectives* (Cheltenham, U.K. and Northampton, MA: Elgar, 1998).
Knack, Steven, and Philip Keefer, "Institutions and Economic Performance: Cross-Country Tests Using Alternative Measures," *Economics and Politics,* VII (1995), 207–227.
Kremer, Michael, "Population Growth and Technological Change: One Million B.C. to 1990," *Quarterly Journal of Economics,* CVIII (1993), 681–716.
Kuznets, Simon, *Modern Economic Growth: Rate Structure and Spread* (New Haven, CT: Yale University Press, 1968).
Landes, David S., *The Wealth and Poverty of Nations: Why Some Are So Rich and Some So Poor* (New York, NY: W.W. Norton & Co., 1998).
La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert W. Vishny, unpublished appendix from "The Quality of Government," *Journal of Law, Economics and Organization* XV (1999), 222–279.
Law, Robin C. C., *The Oyo Empire c1600–1836: A West African Imperialism in the Era of the Atlantic Slave Trade* (New York, NY: Oxford University Press, 1977).
Lewis, W. Arthur, *Growth and Fluctuations 1870–1913* (London, UK: George Allen and Unwin, 1978).
Livi-Bacci, Massimo, *A Concise History of World Population,* third edition (Oxford, UK: Blackwell, 2001).
Locke, John, *Two Treatises of Government* (Indianapolis, IN: Hackett, 1690, 1980).
Lockhart, James, and Stuart B. Schwartz, *Early Latin America* (New York, NY: Cambridge University Press, 1983).
Machiavelli, Niccolò, *Discourses on Livy* (New York, NY: Oxford University Press, 1519, 1987).
Maddison, Angus, *Monitoring the World Economy* (Paris: OECD, 1820–1992, 1995).
——, *The World Economy: A Millennial Perspective* (Paris: OECD, Development Centre of the Organization for Economic Cooperation and Development, 2001).
Malthus, Thomas R., *An Essay on the Principle of Population* (Amherst, NY: Prometheus Books, 1798, 1998).
Manning, Patrick, *Slavery and African Life: Occidental, Oriental and African Slave Trades* (New York, NY: Cambridge University Press, 1990).
Marshall, Alfred, *Principles of Economics* (London, UK: Macmillan, 1890).
Marshall, P. J., "The British in Asia: Trade to Dominion, 1700–1765," in P. J.

Marshall ed., *The Oxford History of the British Empire, Volume II The Eighteenth Century* (New York, NY: Oxford University Press, 1998).
McCusker, John J., and Russell R. Menard, *The Economy of British America, 1607–1785* (Chapel Hill, NC: University of North Carolina Press, 1985).
McEvedy, Colin, and Richard Jones, *Atlas of World Population History* (New York, NY: Facts on File, 1978).
McNeill, William H., *Plagues and Peoples* (Garden City, NJ: Anchor Press, 1976).
——, *A World History,* fourth edition (Oxford, UK: Oxford University Press, 1999).
Miller, Joseph C., *Way of Death: Merchant Capitalism and the Angolan Slave Trade, 1730–1830* (Madison, WI: University of Wisconsin Press, 1988).
Mitchell, Brian R., *International Historical Statistics, The Americas 1750–1988,* second edition (New York, NY: Stockton Press, 1993).
Mitchell, Brian R., *International Historical Statistics, Africa, Asia & Oceania 1750–1988,* second edition (New York, NY: Stockton Press, 1995).
Montesquieu, Charles de Secondat, *The Spirit of the Laws* (New York, NY: Cambridge University Press, 1748, 1989).
Myrdal, Gunnar, *Asian Drama: An Inquiry into the Poverty of Nations,* 3 volumes (New York, NY: Twentieth Century Fund, 1968).
Nelson, Richard, and Edmund Phelps, "Investment in Humans, Technological Diffusion and Economic Growth," *American Economic Association Papers and Proceedings,* LVI (1966), 69–75.
North, Douglass C., *Institutions, Institutional Change, and Economic Performance* (New York, NY: Cambridge University Press, 1990).
North, Douglass C., and Robert P. Thomas, *The Rise of the Western World: A New Economic History* (Cambridge, UK: Cambridge University Press, 1973).
North, Douglass C., and Barry R. Weingast, "Constitutions and Commitment: Evolution of Institutions Governing Public Choice in Seventeenth Century England, *Journal of Economic History,* XLIX (1989), 803–832.
Notestein, Frank W., "Population: The Long View," in Theodore W. Schultze ed., *Food for the World* (Chicago, IL: University of Chicago Press, 1945).
Olson, Mancur, *The Rise and Decline of Nations: Economic Growth, Stagflation, and Economic Rigidities* (New Haven, CT: and London, UK: Yale University Press, 1982).
——, *Power and Prosperity: Outgrowing Communist and Capitalist Dictatorships* (New York, NY: Basic Books, 2000).
Parker, Philip M., *National Cultures of the World: A Statistical Reference, Cross-Cultural Statistical Encyclopedia of the World* (Westport, CT: Greenwood Press, 1997).
Pirenne, Henri, *Medieval Cities: Their Origins and the Revival of Trade* (New York, NY: Doubleday, 1956).
Pomeranz, Kenneth, *The Great Divergence: Europe, China, and the Making of the Modern World Economy* (Princeton, NJ: Princeton University Press, 2000).
Postan, M. M., and E. E. Rich, *The Cambridge Economic History of Europe: Volume 2, Trade and Industry in the Middle Ages* (Cambridge, UK: Cambridge University Press, 1966).
Reid, Anthony, *Southeast Asia in the Age of Commerce: Volumes 1 and 2, The Lands Below the Winds* (New Haven, CT: Yale University Press, 1988 and 1993).
Rodney, Walter, *How Europe Underdeveloped Africa* (London, UK: Bogle-L'Ouverture Publications, 1972).
Romer, Paul M., "Increasing Returns and Long-Run Growth," *Journal of Political Economy,* XCIV (1986), 1002–1037.
Rothenberg, Winifred, *The Transformation of Rural Massachusetts, 1750–1850* (Chicago, IL: Chicago University Press, 1992).
Sachs, Jeffrey D., "Notes on a New Sociology of Economic Development," in Lawrence E. Harrison and Samuel P. Huntington, eds., *Culture Matters: How Values Shape Human Progress* (New York, NY: Basic Books, 2000).
——, "Tropical Underdevelopment," NBER Working Paper No. 8119, 2001.
Showers, Victor, *World Facts and Figures* (New York, NY: Wiley, 1979).
Simon, Julian L., *The Economics of Population Growth* (Princeton, NJ: Princeton University Press, 1977).
Smith, Adam, *The Wealth of Nations* (London, UK: Penguin Books, 1778, 1999).

Tilly, Charles, *Coercion, Capital, and European States, A.D. 990–1990* (Cambridge, MA: Basil Blackwell, 1990).
Tilly, Charles, and Wim P. Blockmans eds., *Cities and the Rise of States in Europe, A.D. 1000 to 1800* (Boulder, CO: Westview Press, 1994).
Townsend, Richard F., *The Aztecs* (London, UK: Thames & Hudson, 2000).
Toynbee, Arnold J., *A Study of History,* 12 Volumes (Oxford, UK: Oxford University Press, 1934–1961).
United Nations (UN), *Growth of the World's Urban and Rural Population, 1920–2000* (New York, NY: Department of Economic and Social Affairs, Population Studies, 1969).
——, *World Urbanization Prospects: The 1996 Revision* (New York, NY: Department of Economic and Social Affairs, Population Division, 1998).
Waldman, Carl, *Atlas of the North American Indian* (New York, NY: Facts on File, Inc., 1985).
Wallerstein, Immanuel M., *The Modern World-System,* 3 Volumes (New York, NY: Academic Press, 1974–1980).
Wiegersma, Nancy, *Vietnam: Peasant Land, Peasant Revolution* (New York, NY: St. Martin's Press, 1988).
White, Lynn, Jr., *Medieval Technology and Social Change* (London, UK: Oxford University Press, 1962).
Wilks, Ivor, *Asante in the Nineteenth Century: The Structure and Evolution of a Political Order* (New York, NY: Cambridge University Press, 1975).
Williams, Eric E., *Capitalism and Slavery* (Chapel Hill, NC: University of North Carolina Press, 1944).
World Bank, *World Development Indicators* (CD rom and book) (Washington, DC: World Bank, 1999).
World Coal Institute, *Coal—Power for Progress,* on the web at http://www.wci-coal.com/pfp.htm, March (2000).
World Resources Institute, *World Resources: A Guide to the Global Environment,* with The United Nations Environment Programme, The United Nations Development Programme, and the World Bank (Oxford: Oxford University Press, 1998).
Wrigley, Edward A., *Continuity, Chance and Change* (Cambridge, UK: Cambridge University Press, 1988).

The Colonial Origins of Comparative Development: An Empirical Investigation

By DARON ACEMOGLU, SIMON JOHNSON, AND JAMES A. ROBINSON*

We exploit differences in European mortality rates to estimate the effect of institutions on economic performance. Europeans adopted very different colonization policies in different colonies, with different associated institutions. In places where Europeans faced high mortality rates, they could not settle and were more likely to set up extractive institutions. These institutions persisted to the present. Exploiting differences in European mortality rates as an instrument for current institutions, we estimate large effects of institutions on income per capita. Once the effect of institutions is controlled for, countries in Africa or those closer to the equator do not have lower incomes. (*JEL* O11, P16, P51)

What are the fundamental causes of the large differences in income per capita across countries? Although there is still little consensus on the answer to this question, differences in institutions and property rights have received considerable attention in recent years. Countries with better "*institutions,*" more secure property rights, and less distortionary policies will invest more in physical and human capital, and will use these factors more efficiently to achieve a greater level of income (e.g., Douglass C. North and Robert P. Thomas, 1973; Eric L. Jones, 1981; North, 1981). This view receives some support from cross-country correlations between measures of property rights and economic development (e.g., Stephen Knack and Philip Keefer, 1995; Paulo Mauro, 1995; Robert E. Hall and Charles I. Jones, 1999; Dani Rodrik, 1999), and from a few micro studies that investigate the relationship between property rights and investment or output (e.g., Timothy Besley, 1995; Christopher Mazingo, 1999; Johnson et al., 1999).

At some level it is obvious that institutions matter. Witness, for example, the divergent paths of North and South Korea, or East and West Germany, where one part of the country stagnated under central planning and collective ownership, while the other prospered with private property and a market economy. Nevertheless, we lack reliable estimates of the effect of institutions on economic performance. It is quite likely that rich economies choose or can afford better institutions. Perhaps more important, economies that are different for a variety of reasons will differ both

* Acemoglu: Department of Economics, E52-380b, Massachusetts Institute of Technology, Cambridge, MA 02319, and Canadian Institute for Advanced Research (e-mail: daron@mit.edu); Johnson: Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02319 (e-mail: sjohnson@mit.edu); Robinson: Department of Political Science and Department of Economics, 210 Barrows Hall, University of California, Berkeley, CA 94720 (e-mail: jamesar@socrates.berkeley.edu). We thank Joshua Angrist, Abhijit Banerjee, Esther Duflo, Stan Engerman, John Gallup, Claudia Goldin, Robert Hall, Chad Jones, Larry Katz, Richard Locke, Andrei Shleifer, Ken Sokoloff, Judith Tendler, three anonymous referees, and seminar participants at the University of California-Berkeley, Brown University, Canadian Institute for Advanced Research, Columbia University, Harvard University, Massachusetts Institute of Technology, National Bureau of Economic Research, Northwestern University, New York University, Princeton University, University of Rochester, Stanford University, Toulouse University, University of California-Los Angeles, and the World Bank for useful comments. We also thank Robert McCaa for guiding us to the data on bishops' mortality.

in their institutions and in their income per capita.

To estimate the impact of institutions on economic performance, we need a source of exogenous variation in institutions. In this paper, we propose a theory of institutional differences among countries colonized by Europeans,[1] and exploit this theory to derive a possible source of exogenous variation. Our theory rests on three premises:

1. There were different types of colonization policies which created different sets of institutions. At one extreme, European powers set up "extractive states," exemplified by the Belgian colonization of the Congo. These institutions did not introduce much protection for private property, nor did they provide checks and balances against government expropriation. In fact, the main purpose of the extractive state was to transfer as much of the resources of the colony to the colonizer.
   At the other extreme, many Europeans migrated and settled in a number of colonies, creating what the historian Alfred Crosby (1986) calls "Neo-Europes." The settlers tried to replicate European institutions, with strong emphasis on private property and checks against government power. Primary examples of this include Australia, New Zealand, Canada, and the United States.
2. The colonization strategy was influenced by the feasibility of settlements. In places where the disease environment was not favorable to European settlement, the cards were stacked against the creation of Neo-Europes, and the formation of the extractive state was more likely.
3. The colonial state and institutions persisted even after independence.

Based on these three premises, we use the mortality rates expected by the first European settlers in the colonies as an instrument for current institutions in these countries.[2] More specifically, our theory can be schematically summarized as

$$\text{(potential) settler mortality} \Rightarrow \text{settlements}$$

$$\Rightarrow \text{early institutions} \Rightarrow \text{current institutions}$$

$$\Rightarrow \text{current performance.}$$

We use data on the mortality rates of soldiers, bishops, and sailors stationed in the colonies between the seventeenth and nineteenth centuries, largely based on the work of the historian Philip D. Curtin. These give a good indication of the mortality rates faced by settlers. Europeans were well informed about these mortality rates at the time, even though they did not know how to control the diseases that caused these high mortality rates.

Figure 1 plots the logarithm of GDP per capita today against the logarithm of the settler mortality rates per thousand for a sample of 75 countries (see below for data details). It shows a strong negative relationship. Colonies where Europeans faced higher mortality rates are today substantially poorer than colonies that were healthy for Europeans. Our theory is that this relationship reflects the effect of settler mortality working through the institutions brought by Europeans. To substantiate this, we regress current performance on current institutions, and instrument the latter by settler mortality rates. Since our focus is on property rights and checks against government power, we use the protection against "risk of expropriation" index from Political Risk Services as a proxy for institutions. This variable measures differences in institutions originating from different types of states and state policies.[3] There is a strong

[1] By "colonial experience" we do not only mean the direct control of the colonies by European powers, but more generally, European influence on the rest of the world. So according to this definition, Sub-Saharan Africa was strongly affected by "colonialism" between the sixteenth and nineteenth centuries because of the Atlantic slave trade.

[2] Note that although only some countries were colonized, there is no selection bias here. This is because the question we are interested in is the effect of colonization policy *conditional* on being colonized.

[3] Government expropriation is not the only institutional feature that matters. Our view is that there is a "cluster of

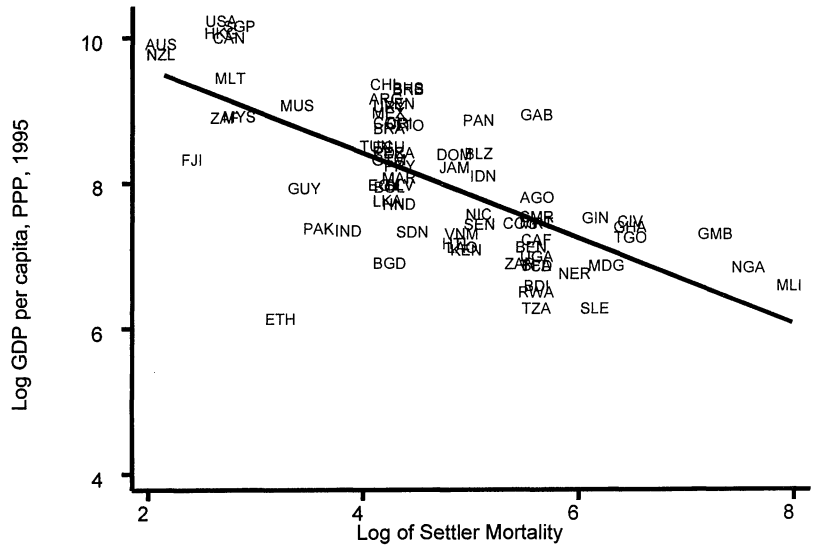


FIGURE 1. REDUCED-FORM RELATIONSHIP BETWEEN INCOME AND SETTLER MORTALITY

(first-stage) relationship between settler mortality rates and current institutions, which is interesting in its own right. The regression shows that mortality rates faced by the settlers more than 100 years ago explains over 25 percent of the variation in current institutions.[4] We also document that this relationship works through the channels we hypothesize: (potential) settler mortality rates were a major determinant of settlements; settlements were a major determinant of early institutions (in practice, institutions in 1900); and there is a strong correlation between early institutions and institutions today. Our two-stage least-squares estimate of the effect of institutions on performance is relatively precisely estimated and large. For example, it implies that improving Nigeria's institutions to the level of Chile could, in the long run, lead to as much as a 7-fold increase in Nigeria's income (in practice Chile is over 11 times as rich as Nigeria).

The exclusion restriction implied by our instrumental variable regression is that, conditional on the controls included in the regression, the mortality rates of European settlers more than 100 years ago have no effect on GDP per capita today, other than their effect through institutional development. The major concern with this exclusion restriction is that the mortality rates of settlers could be correlated with the current disease environment, which may have a direct effect on economic performance. In this case, our instrumental-variables estimates may be assigning the effect of diseases on income to institutions. We believe that this is unlikely to be the case and that our exclusion restriction is plausible. The great majority of European deaths in the colonies were caused by malaria and yellow fever. Although these diseases were fatal to Europeans who had no immunity, they had limited effect on indigenous adults who had developed various types of immunities. These diseases are therefore unlikely to be the reason why many countries in Africa and Asia are very poor today (see the discussion in Section III, subsection A). This notion is

institutions," including constraints on government expropriation, independent judiciary, property rights enforcement, and institutions providing equal access to education and ensuring civil liberties, that are important to encourage investment and growth. Expropriation risk is related to all these institutional features. In Acemoglu et al. (2000), we reported similar results with other institutions variables.

[4] Differences in mortality rates are *not* the only, or even the main, cause of variation in institutions. For our empirical approach to work, all we need is that they are *a source of* exogenous variation.

supported by the mortality rates of local people in these areas. For example, Curtin (1968 Table 2) reports that the annual mortality rates of local troops serving with the British army in Bengal and Madras were respectively 11 and 13 in 1,000. These numbers are quite comparable to, in fact lower than, the annual mortality rates of British troops serving in Britain, which were approximately 15 in 1,000. In contrast, the mortality rates of British troops serving in these colonies were much higher because of their lack of immunity. For example, mortality rates in Bengal and Madras for British troops were between 70 and 170 in 1,000. The view that the disease burden for indigenous adults was not unusual in places like Africa or India is also supported by the relatively high population densities in these places before Europeans arrived (Colin McEvedy and Richard Jones, 1975).

We document that our estimates of the effect of institutions on performance are not driven by outliers. For example, excluding Australia, New Zealand, Canada, and the United States does not change the results, nor does excluding Africa. Interestingly, we show that once the effect of institutions on economic performance is controlled for, neither distance from the equator nor the dummy for Africa is significant. These results suggest that Africa is poorer than the rest of the world not because of pure geographic or cultural factors, but because of worse institutions.

The validity of our approach—i.e., our exclusion restriction—is threatened if other factors correlated with the estimates of settler mortality affect income per capita. We adopt two strategies to substantiate that our results are not driven by omitted factors. First, we investigate whether institutions have a comparable effect on income once we control for a number of variables potentially correlated with settler mortality and economic outcomes. We find that none of these overturn our results; the estimates change remarkably little when we include controls for the identity of the main colonizer, legal origin, climate, religion, geography, natural resources, soil quality, and measures of ethnolinguistic fragmentation. Furthermore, the results are also robust to the inclusion of controls for the current disease environment (e.g., the prevalence of malaria, life expectancy, and infant mortality) and the current fraction of the population of European descent.

Naturally, it is impossible to control for all possible variables that might be correlated with settler mortality and economic outcomes. Furthermore, our empirical approach might capture the effect of settler mortality on economic performance, but working through other channels. We deal with these problems by using a simple overidentification test using measures of European migration to the colonies and early institutions as additional instruments. We then use overidentification tests to detect whether settler mortality has a direct effect on current performance. The results are encouraging for our approach; they generate no evidence for a direct effect of settler mortality on economic outcomes.

We are not aware of others who have pointed out the link between settler mortality and institutions, though scholars such as William H. McNeill (1976), Crosby (1986), and Jared M. Diamond (1997) have discussed the influence of diseases on human history. Diamond (1997), in particular, emphasizes comparative development, but his theory is based on the geographical determinants of the incidence of the neolithic revolution. He ignores both the importance of institutions and the potential causes of divergence in more recent development, which are the main focus of our paper. Work by Ronald E. Robinson and John Gallagher (1961), Lewis H. Gann and Peter Duignan (1962), Donald Denoon (1983), and Philip J. Cain and Anthony G. Hopkins (1993) emphasizes that settler colonies such as the United States and New Zealand are different from other colonies, and point out that these differences were important for their economic success. Nevertheless, this literature does not develop the link between mortality, settlements, and institutions.

Our argument is most closely related to work on the influence of colonial experience on institutions. Frederich A. von Hayek (1960) argued that the British common law tradition was superior to the French civil law, which was developed during the Napoleonic era to restrain judges' interference with state policies (see also Seymour M. Lipset, 1994). More recently, Rafael La Porta et al. (1998, 1999) emphasize the importance of colonial origin (the identity of

the colonizer) and legal origin on current institutions, and show that the common-law countries and former British colonies have better property rights and more developed financial markets. Similarly, David Landes (1998 Chapters 19 and 20) and North et al. (1998) argue that former British colonies prospered relative to former French, Spanish, and Portuguese colonies because of the good economic and political institutions and culture they inherited from Britain. In contrast to this approach which focuses on the identity of the colonizer, we emphasize *the conditions in the colonies.* Specifically, in our theory—and in the data—it is not the identity of the colonizer or legal origin that matters, but whether European colonialists could safely settle in a particular location: where they could not settle, they created worse institutions. In this respect, our argument is closely related to that of Stanley L. Engerman and Kenneth L. Sokoloff (1997) who also emphasize institutions, but link them to factor endowments and inequality.

Empirically, our work is related to a number of other attempts to uncover the link between institutions and development, as well as to Graziella Bertocchi and Fabio Canova (1996) and Robin M. Grier (1999), who investigate the effect of being a colony on postwar growth. Two papers deal with the endogeneity of institutions by using an instrumental variables approach as we do here. Mauro (1995) instruments for corruption using ethnolinguistic fragmentation. Hall and Jones (1999), in turn, use distance from the equator as an instrument for social infrastructure because, they argue, latitude is correlated with "Western influence," which leads to good institutions. The theoretical reasoning for these instruments is not entirely convincing. It is not easy to argue that the Belgian influence in the Congo, or Western influence in the Gold Coast during the era of slavery promoted good institutions. Ethnolinguistic fragmentation, on the other hand, seems endogenous, especially since such fragmentation almost completely disappeared in Europe during the era of growth when a centralized state and market emerged (see, e.g., Eugen J. Weber, 1976; Benedict Anderson, 1983). Econometrically, the problem with both studies is that their instruments can plausibly have a direct effect on performance. For example, William Easterly and Ross Levine (1997) argue that ethnolinguistic fragmentation can affect performance by creating political instability, while Charles de Montesquieu [1748] (1989) and more recently David E. Bloom and Jeffrey D. Sachs (1998) and John Gallup et al. (1998) argue for a direct effect of climate on performance. If, indeed, these variables have a direct effect, they are invalid instruments and do not establish that it is institutions that matter. The advantage of our approach is that conditional on the variables we already control for, settler mortality more than 100 years ago should have no effect on output today, other than through its effect on institutions. Interestingly, our results show that distance from the equator does not have an independent effect on economic performance, validating the use of this variable as an instrument in the work by Hall and Jones (1999).

The next section outlines our hypothesis and provides supporting historical evidence. Section II presents OLS regressions of GDP per capita on our index of institutions. Section III describes our key instrument for institutions, the mortality rates faced by potential settlers at the time of colonization. Section IV presents our main results. Section V investigates the robustness of our results, and Section VI concludes.

I. The Hypothesis and Historical Background

We hypothesize that settler mortality affected settlements; settlements affected early institutions; and early institutions persisted and formed the basis of current institutions. In this section, we discuss and substantiate this hypothesis. The next subsection discusses the link between mortality rates of settlers and settlement decisions, then we discuss differences in colonization policies, and finally, we turn to the causes of institutional persistence.

A. *Mortality and Settlements*

There is little doubt that mortality rates were a key determinant of European settlements. Curtin (1964, 1998) documents how both the British and French press informed the public of mortality rates in the colonies. Curtin (1964)

also documents how early British expectations for settlement in West Africa were dashed by very high mortality among early settlers, about half of whom could be expected to die in the first year. In the "Province of Freedom" (Sierra Leone), European mortality in the first year was 46 percent, in Bulama (April 1792–April 1793) there was 61-percent mortality among Europeans. In the first year of the Sierra Leone Company (1792–1793), 72 percent of the European settlers died. On Mungo Park's Second Expedition (May–November 1805), 87 percent of Europeans died during the overland trip from Gambia to the Niger, and all the Europeans died before completing the expedition.

An interesting example of the awareness of the disease environment comes from the Pilgrim fathers. They decided to migrate to the United States rather than Guyana because of the high mortality rates in Guyana (see Crosby, 1986 pp. 143–44). Another example comes from the Beauchamp Committee in 1795, set up to decide where to send British convicts who had previously been sent to the United States. One of the leading proposals was the island of Lemane, up the Gambia River. The committee rejected this possibility because they decided mortality rates would be too high even for the convicts. Southwest Africa was also rejected for health reasons. The final decision was to send convicts to Australia.

The eventual expansion of many of the colonies was also related to the living conditions there. In places where the early settlers faced high mortality rates, there would be less incentive for new settlers to come.[5]

B. *Types of Colonization and Settlements*

The historical evidence supports both the notion that there was a wide range of different types of colonization and that the presence or absence of European settlers was a key determinant of the form colonialism took. Historians, including Robinson and Gallagher (1961), Gann and Duignan (1962), Denoon (1983), and Cain and Hopkins (1993), have documented the development of "settler colonies," where Europeans settled in large numbers, and life was modeled after the home country. Denoon (1983) emphasizes that settler colonies had representative institutions which promoted what the settlers wanted and that what they wanted was freedom and the ability to get rich by engaging in trade. He argues that "there was undeniably something capitalist in the structure of these colonies. Private ownership of land and livestock was well established very early ..." (p. 35).

When the establishment of European-like institutions did not arise naturally, the settlers were ready to fight for them against the wishes of the home country. Australia is an interesting example here. Most of the early settlers in Australia were ex-convicts, but the land was owned largely by ex-jailors, and there was no legal protection against the arbitrary power of landowners. The settlers wanted institutions and political rights like those prevailing in England at the time. They demanded jury trials, freedom from arbitrary arrest, and electoral representation. Although the British government resisted at first, the settlers argued that they were British and deserved the same rights as in the home country (see Robert Hughes, 1987). Cain and Hopkins write (1993 p. 237) "from the late 1840s the British bowed to local pressures and, in line with observed constitutional changes taking place in Britain herself, accepted the idea that, in mature colonies, governors should in future form ministries from the majority elements in elected legislatures." They also suggest that "the enormous boom in public investment after 1870 [in New Zealand] ... was an attempt to build up an infrastructure ... to maintain high living standards in a country where voters expected politicians actively to promote their economic welfare." (p. 225).[6]

[5] Naturally, other factors also influenced settlements. For example, despite the relatively high mortality rates, many Europeans migrated to the Caribbean because of the very high incomes there at the time (see, e.g., Richard S. Dunn, 1972; David W. Galenson, 1996; Engerman and Sokoloff, 1997; David Eltis, 2000).

[6] Robert H. Bates (1983 Chapter 3) gives a nice example of the influence of settlers on policy in Africa. The British colonial government pursued many policies that depressed the price of cocoa, the main produce of the farmers in Ghana. In contrast, the British government supported the prices faced by the commercial cereal farmers in Kenya.

This is in sharp contrast to the colonial experience in Latin America during the seventeenth and eighteenth centuries, and in Asia and Africa during the nineteenth and early twentieth centuries. The main objective of the Spanish and the Portuguese colonization was to obtain gold and other valuables from America. Soon after the conquest, the Spanish crown granted rights to land and labor (the *encomienda*) and set up a complex mercantilist system of monopolies and trade regulations to extract resources from the colonies.[7]

Europeans developed the slave trade in Africa for similar reasons. Before the mid-nineteenth century, colonial powers were mostly restricted to the African coast and concentrated on monopolizing trade in slaves, gold, and other valuable commodities—witness the names used to describe West African countries: the Gold Coast, the Ivory Coast. Thereafter, colonial policy was driven in part by an element of superpower rivalry, but mostly by economic motives. Michael Crowder (1968 p. 50), for example, notes "it is significant that Britain's largest colony on the West Coast [Nigeria] should have been the one where her traders were most active and bears out the contention that, for Britain ... flag followed trade."[8] Lance E. Davis and Robert A. Huttenback (1987 p. 307) conclude that "the colonial Empire provides strong evidence for the belief that government was attuned to the interests of business and willing to divert resources to ends that the business community would have found profitable." They find that before 1885 investment in the British empire had a return 25 percent higher than that on domestic investment, though afterwards the two converged. Andrew Roberts (1976 p. 193) writes: "[from] ... 1930 to 1940 Britain had kept for itself 2,400,000 pounds in taxes from the Copperbelt, while Northern Rhodesia received from Britain only 136,000 pounds in grants for development." Similarly, Patrick Manning (1982) estimates that between 1905 and 1914, 50 percent of GDP in Dahomey was extracted by the French, and Crawford Young (1994 p. 125) notes that tax rates in Tunisia were four times as high as in France.

Probably the most extreme case of extraction was that of King Leopold of Belgium in the Congo. Gann and Duignan (1979 p. 30) argue that following the example of the Dutch in Indonesia, Leopold's philosophy was that "the colonies should be exploited, not by the operation of a market economy, but by state intervention and compulsory cultivation of cash crops to be sold to and distributed by the state at controlled prices." Jean-Philippe Peemans (1975) calculates that tax rates on Africans in the Congo approached 60 percent of their income during the 1920's and 1930's. Bogumil Jewsiewicki (1983) writes that during the period when Leopold was directly in charge, policy was "based on the violent exploitation of natural and human resources," with a consequent "destruction of economic and social life ... [and] ... dismemberment of political structures."

Overall, there were few constraints on state power in the nonsettler colonies. The colonial powers set up authoritarian and absolutist states with the purpose of solidifying their control and facilitating the extraction of resources. Young (1994 p. 101) quotes a French official in Africa: "the European commandant is not posted to observe nature, ... He has a mission ... to impose regulations, to limit individual liberties ... , to collect taxes." Manning (1988 p. 84) summarizes this as: "In Europe the theories of representative democracy won out over the theorists of absolutism ... . But in Africa, the European conquerors set up absolutist governments, based on reasoning similar to that of Louis XIV."

Bates shows that this was mainly because in Kenya, but not in Ghana, there were a significant number of European settler farmers, who exerted considerable pressure on policy.

[7] See James Lang (1975) and James Lockhart and Stuart B. Schwartz (1983). Migration to Spanish America was limited by the Spanish Crown, in part because of a desire to keep control of the colonists and limit their independence (see, for example, John H. Coatsworth, 1982). This also gives further support to our notion that settlers were able to influence the type of institutions set up in the colonies, even against the wishes of the home country government.

[8] Although in almost all cases the main objective of colonial policies was to protect economic interests and obtain profits, the recipients of these profits varied. In the Portuguese case, it was the state; in the Belgian case, it was King Leopold; and in the British case, it was often private enterprises who obtained concessions or monopoly trading rights in Africa (Crowder, 1968 Part III).

C. *Institutional Persistence*

There is a variety of historical evidence, as well as our regressions in Table 3 below, suggesting that the control structures set up in the nonsettler colonies during the colonial era persisted, while there is little doubt that the institutions of law and order and private property established during the early phases of colonialism in Australia, Canada, New Zealand, the United States, Hong Kong, and Singapore have formed the basis of the current-day institutions of these countries.[9]

Young emphasizes that the extractive institutions set up by the colonialists persisted long after the colonial regime ended. He writes "although we commonly described the independent polities as 'new states,' in reality they were successors to the colonial regime, inheriting its structures, its quotidian routines and practices, and its more hidden normative theories of governance" (1994 p. 283). An example of the persistence of extractive state institutions into the independence era is provided by the persistence of the most prominent extractive policies. In Latin America, the full panoply of monopolies and regulations, which had been created by Spain, remained intact after independence, for most of the nineteenth century. Forced labor policies persisted and were even intensified or reintroduced with the expansion of export agriculture in the latter part of the nineteenth century. Slavery persisted in Brazil until 1886, and during the sisal boom in Mexico, forced labor was reintroduced and persisted up to the start of the revolution in 1910. Forced labor was also reintroduced in Guatemala and El Salvador to provide labor for coffee growing. In the Guatemalan case, forced labor lasted until the creation of democracy in 1945. Similarly, forced labor was reinstated in many independent African countries, for example, by Mobutu in Zaire.

There are a number of economic mechanisms that will lead to institutional persistence of this type. Here, we discuss three possibilities.

(1) Setting up institutions that place restrictions on government power and enforce property rights is costly (see, e.g., Acemoglu and Thierry Verdier, 1998). If the costs of creating these institutions have been sunk by the colonial powers, then it may not pay the elites at independence to switch to extractive institutions. In contrast, when the new elites inherit extractive institutions, they may not want to incur the costs of introducing better institutions, and may instead prefer to exploit the existing extractive institutions for their own benefits.
(2) The gains to an extractive strategy may depend on the size of the ruling elite. When this elite is small, each member would have a larger share of the revenues, so the elite may have a greater incentive to be extractive. In many cases where European powers set up authoritarian institutions, they delegated the day-to-day running of the state to a small domestic elite. This narrow group often was the one to control the state after independence and favored extractive institutions.[10]
(3) If agents make irreversible investments that are complementary to a particular set of institutions, they will be more willing to support them, making these institutions persist (see, e.g., Acemoglu, 1995). For example, agents who have invested in human and physical capital will be in favor of spending

[9] The thesis that institutions persist for a long time goes back at least to Karl A. Wittfogel (1957), who argued that the control structures set up by the large "hydraulic" empires such as China, Russia, and the Ottoman Empire persisted for more than 500 years to the twentieth century. Engerman and Sokoloff (1997), La Porta et al. (1998, 1999), North et al. (1998), and Coatsworth (1999) also argue that colonial institutions persisted. Engerman et al. (1998) provide further evidence supporting this view.

[10] William Reno (1995), for example, argues that the governments of postindependence Sierra Leone adopted the tactics and institutions of the British colonizers to cement their political power and extract resources from the rest of society. Catherine Boone (1992) provides a similar analysis of the evolution of the modern state in Senegal. Most scholars also view the roots of authoritarianism under Mobutu in the colonial state practices in the Belgian Congo (e.g., Thomas M. Callaghy, 1984, or Thomas Turner and Young, 1985, especially p. 43). The situation in Latin America is similar. Independence of most Latin American countries came in the early nineteenth century as domestic elites took advantage of the invasion of Spain by Napoleon to capture the control of the state. But, the only thing that changed was the identity of the recipients of the rents (see, for example, Coatsworth, 1978, or John Lynch, 1986).

TABLE 1—DESCRIPTIVE STATISTICS

			By quartiles of mortality			
	Whole world	Base sample	(1)	(2)	(3)	(4)
Log GDP per capita (PPP) in 1995	8.3	8.05	8.9	8.4	7.73	7.2
	(1.1)	(1.1)				
Log output per worker in 1988 (with level of United States normalized to 1)	−1.70	−1.93	−1.03	−1.46	−2.20	−3.03
	(1.1)	(1.0)				
Average protection against expropriation risk, 1985–1995	7	6.5	7.9	6.5	6	5.9
	(1.8)	(1.5)				
Constraint on executive in 1990	3.6	4	5.3	5.1	3.3	2.3
	(2.3)	(2.3)				
Constraint on executive in 1900	1.9	2.3	3.7	3.4	1.1	1
	(1.8)	(2.1)				
Constraint on executive in first year of independence	3.6	3.3	4.8	2.4	3.1	3.4
	(2.4)	(2.4)				
Democracy in 1900	1.1	1.6	3.9	2.8	0.19	0
	(2.6)	(3.0)				
European settlements in 1900	0.31	0.16	0.32	0.26	0.08	0.005
	(0.4)	(0.3)				
Log European settler mortality	n.a.	4.7	3.0	4.3	4.9	6.3
		(1.1)				
Number of observations	163	64	14	18	17	15

Notes: Standard deviations are in parentheses. Mortality is potential settler mortality, measured in terms of deaths per annum per 1,000 "mean strength" (raw mortality numbers are adjusted to what they would be if a force of 1,000 living people were kept in place for a whole year, e.g., it is possible for this number to exceed 1,000 in episodes of extreme mortality as those who die are replaced with new arrivals). Sources and methods for mortality are described in Section III, subsection B, and in the unpublished Appendix (available from the authors; or see Acemoglu et al., 2000). Quartiles of mortality are for our base sample of 64 observations. These are: (1) less than 65.4; (2) greater than or equal to 65.4 and less than 78.1; (3) greater than or equal to 78.1 and less than 280; (4) greater than or equal to 280. The number of observations differs by variable; see Appendix Table A1 for details.

money to enforce property rights, while those who have less to lose may not be.

II. Institutions and Performance: OLS Estimates

A. *Data and Descriptive Statistics*

Table 1 provides descriptive statistics for the key variables of interest. The first column is for the whole world, and column (2) is for our base sample, limited to the 64 countries that were ex-colonies and for which we have settler mortality, protection against expropriation risk, and GDP data (this is smaller than the sample in Figure 1). The GDP per capita in 1995 is PPP adjusted (a more detailed discussion of all data sources is provided in Appendix Table A1). Income (GDP) per capita will be our measure of economic outcome. There are large differences in income per capita in both the world sample and our basic sample, and the standard deviation of log income per capita in both cases is 1.1. In row 3, we also give output per worker in 1988 from Hall and Jones (1999) as an alternative measure of income today. Hall and Jones (1999) prefer this measure since it explicitly refers to worker productivity. On the other hand, given the difficulty of measuring the formal labor force, it may be a more noisy measure of economic performance than income per capita.

We use a variety of variables to capture institutional differences. Our main variable, reported in the second row, is an index of protection against expropriation. These data are from Political Risk Services (see, e.g., William D. Coplin et al., 1991), and were first used in the economics and political science literatures by Knack and Keefer (1995). Political Risk Services reports a value between 0 and 10 for each country and year, with 0 corresponding to the

lowest protection against expropriation. We use the average value for each country between 1985 and 1995 (values are missing for many countries before 1985). This measure is appropriate for our purposes since the focus here is on differences in institutions originating from different types of states and state policies. We expect our notion of extractive state to correspond to a low value of this index, while the tradition of rule of law and well-enforced property rights should correspond to high values.[11] The next row gives an alternative measure, constraints on the executive in 1990, coded from the Polity III data set of Ted Robert Gurr and associates (an update of Gurr, 1997). Results using the constraints on the executive and other measures are reported in Acemoglu et al. (2000) and are not repeated here.

The next three rows give measures of early institutions from the same Gurr data set. The first is a measure of constraints on the executive in 1900 and the second is an index of democracy in 1900. This information is not available for countries that were still colonies in 1900, so we assign these countries the lowest possible score. In the following row, we report the mean and standard deviation of constraints on the executive in the first year of independence (i.e., the first year a country enters the Gurr data set) as an alternative measure of institutions. The second-to-last row gives the fraction of the population of European descent in 1900, which is our measure of European settlement in the colonies, constructed from McEvedy and Jones (1975) and Curtin et al. (1995). The final row gives the logarithm of the baseline settler mortality estimates; the raw data are in Appendix Table A2.

The remaining columns give descriptive statistics for groups of countries at different quartiles of the settler mortality distribution. This is useful since settler mortality is our instrument for institutions (this variable is described in more detail in the next section).

[11] The protection against expropriation variable is specifically for foreign investment, since Political and Risk Services construct these data for foreign investors. However, as noted by Knack and Keefer (1995), risk of expropriation of foreign and domestic investments are very highly correlated, and risk of expropriation of foreign investment may be more comparable across countries. In any case, all our results hold also with a variety of other measures of institutions (see Tables 4a, b, c, d, and e in Acemoglu et al., 2000, available from the authors).

B. *Ordinary Least-Squares Regressions*

Table 2 reports ordinary least-squares (OLS) regressions of log per capita income on the protection against expropriation variable in a variety of samples. The linear regressions are for the equation

$$(1) \qquad \log y_i = \mu + \alpha R_i + \mathbf{X}_i'\gamma + \varepsilon_i,$$

where y_i is income per capita in country i, R_i is the protection against expropriation measure, $\mathbf{X}_i$ is a vector of other covariates, and ε_i is a random error term. The coefficient of interest throughout the paper is α, the effect of institutions on income per capita.

Column (1) shows that in the whole world sample there is a strong correlation between our measure of institutions and income per capita. Column (2) shows that the impact of the institutions variable on income per capita in our base sample is quite similar to that in the whole world, and Figure 2 shows this relationship diagrammatically for our base sample consisting of 64 countries. The R^2 of the regression in column (1) indicates that over 50 percent of the variation in income per capita is associated with variation in this index of institutions. To get a sense of the magnitude of the effect of institutions on performance, let us compare two countries, Nigeria, which has approximately the 25th percentile of the institutional measure in this sample, 5.6, and Chile, which has approximately the 75th percentile of the institutions index, 7.8. The estimate in column (1), 0.52, indicates that there should be on average a 1.14-log-point difference between the log GDPs of the corresponding countries (or approximately a 2-fold difference—$e^{1.14} - 1 \approx 2.1$). In practice, this GDP gap is 253 log points (approximately 11-fold). Therefore, if the effect estimated in Table 2 were causal, it would imply a fairly large effect of institutions on performance, but still much less than the actual income gap between Nigeria and Chile.

Many social scientists, including Montesquieu [1784] (1989), Diamond (1997), and

TABLE 2—OLS REGRESSIONS

	Whole world (1)	Base sample (2)	Whole world (3)	Whole world (4)	Base sample (5)	Base sample (6)	Whole world (7)	Base sample (8)
	Dependent variable is log GDP per capita in 1995						Dependent variable is log output per worker in 1988	
Average protection against expropriation risk, 1985–1995	0.54 (0.04)	0.52 (0.06)	0.47 (0.06)	0.43 (0.05)	0.47 (0.06)	0.41 (0.06)	0.45 (0.04)	0.46 (0.06)
Latitude			0.89 (0.49)	0.37 (0.51)	1.60 (0.70)	0.92 (0.63)		
Asia dummy				−0.62 (0.19)		−0.60 (0.23)		
Africa dummy				−1.00 (0.15)		−0.90 (0.17)		
"Other" continent dummy				−0.25 (0.20)		−0.04 (0.32)		
R^2	0.62	0.54	0.63	0.73	0.56	0.69	0.55	0.49
Number of observations	110	64	110	110	64	64	108	61

Notes: Dependent variable: columns (1)–(6), log GDP per capita (PPP basis) in 1995, current prices (from the World Bank's World Development Indicators 1999); columns (7)–(8), log output per worker in 1988 from Hall and Jones (1999). Average protection against expropriation risk is measured on a scale from 0 to 10, where a higher score means more protection against expropriation, averaged over 1985 to 1995, from Political Risk Services. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable definitions and sources. Of the countries in our base sample, Hall and Jones do not report output per worker in the Bahamas, Ethiopia, and Vietnam.

Sachs and coauthors, have argued for a direct effect of climate on performance, and Gallup et al. (1998) and Hall and Jones (1999) document the correlation between distance from the equator and economic performance. To control for this, in columns (3)–(6), we add latitude as a regressor (we follow the literature in using the absolute value measure of latitude, i.e., distance from the equator, scaled between 0 and 1). This changes the coefficient of the index of institutions little. Latitude itself is also significant and has the sign found by the previous studies. In columns (4) and (6), we also add dummies for Africa, Asia, and other continents, with America as the omitted group. Although protection against expropriation risk remains significant, the continent dummies are also statistically and quantitatively significant. The Africa dummy in column (6) indicates that in our sample African countries are 90 log points (approximately 145 percent) poorer even after taking the effect of institutions into account. Finally, in columns (7) and (8), we repeat our basic regressions using the log of output per worker from Hall and Jones (1999), with very similar results.

Overall, the results in Table 2 show a strong correlation between institutions and economic performance. Nevertheless, there are a number of important reasons for not interpreting this relationship as causal. First, rich economies may be able to afford, or perhaps prefer, better institutions. Arguably more important than this reverse causality problem, there are many omitted determinants of income differences that will naturally be correlated with institutions. Finally, the measures of institutions are constructed *ex post,* and the analysts may have had a natural bias in seeing better institutions in richer places. As well as these problems introducing positive bias in the OLS estimates, the fact that the institutions variable is measured with considerable error and corresponds poorly to the "cluster of institutions" that matter in practice creates attenuation and may bias the OLS estimates

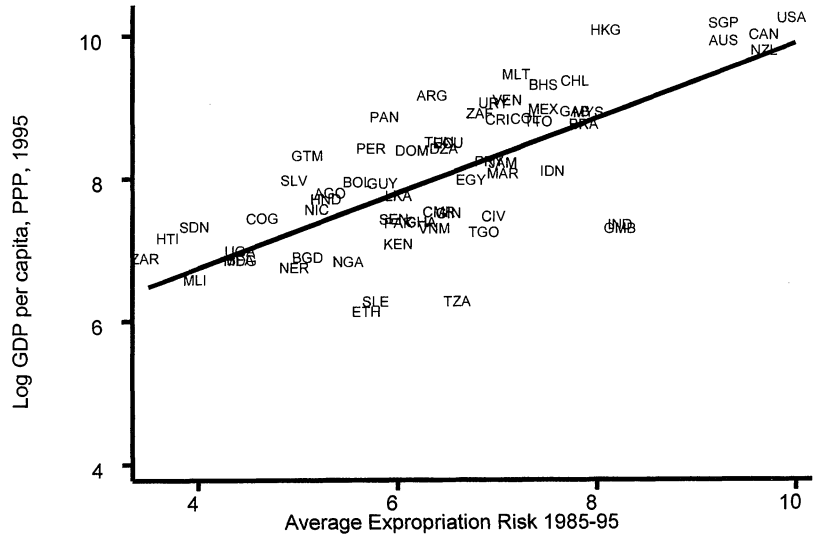


FIGURE 2. OLS RELATIONSHIP BETWEEN EXPROPRIATION RISK AND INCOME

downwards. All of these problems could be solved if we had an instrument for institutions. Such an instrument must be an important factor in accounting for the institutional variation that we observe, but have no direct effect on performance. Our discussion in Section I suggests that settler mortality during the time of colonization is a plausible instrument.

III. Mortality of Early Settlers

A. *Sources of European Mortality in the Colonies*

In this subsection, we give a brief overview of the sources of mortality facing potential settlers. Malaria (particularly Plasmodium falciporum) and yellow fever were the major sources of European mortality in the colonies. In the tropics, these two diseases accounted for 80 percent of European deaths, while gastrointestinal diseases accounted for another 15 percent (Curtin, 1989 p. 30). Throughout the nineteenth century, areas without malaria and yellow fever, such as New Zealand, were more healthy than Europe because the major causes of death in Europe—tuberculosis, pneumonia, and smallpox—were rare in these places (Curtin, 1989 p. 13).

Both malaria and yellow fever are transmitted by mosquito vectors. In the case of malaria, the main transmitter is the *Anopheles gambiae* complex and the mosquito *Anopheles funestus,* while the main carrier of yellow fever is *Aedes aegypti.* Both malaria and yellow fever vectors tend to live close to human habitation.

In places where the malaria vector is present, such as the West African savanna or forest, an individual can get as many as several hundred infectious mosquito bites a year. For a person without immunity, malaria (particularly Plasmodium falciporum) is often fatal, so Europeans in Africa, India, or the Caribbean faced very high death rates. In contrast, death rates for the adult local population were much lower (see Curtin [1964] and the discussion in our introduction above). Curtin (1998 pp. 7–8) describes this as follows:

> Children in West Africa ... would be infected with malaria parasites shortly after birth and were frequently reinfected afterwards; if they lived beyond the age of about five, they acquired an apparent immunity. The parasite remained with them, normally in the liver, but clinical symptoms were rare so long as they continued to be infected with the same species of P. falciporum.

The more recent books on malariology confirm this conclusion. For example, "In stable endemic areas a heavy toll of morbidity and mortality falls on young children but malaria is a relatively mild condition in adults" (Herbert M. Gilles and David A. Warrell, 1993 p. 64; see also the classic reference on this topic, Leonard J. Bruce-Chwatt, 1980 Chapter 4; Roy Porter, 1996).[12] Similarly, the World Health Organization (WHO) points out that in endemic malaria areas of Africa and the Western Pacific today "... the risk of malaria severity and death is almost exclusively limited to non-immunes, being most serious for young children over six months of age ... surviving children develop their own immunity between the age of 3–5 years" (Jose A. Najera and Joahim Hempel, 1996).

People in areas where malaria is endemic are also more likely to have genetic immunity against malaria. For example, they tend to have the sickle-cell trait, which discourages the multiplication of parasites in the blood, or deficiencies in *glucose-6-phosphate dehydrogenase* and *thalassaemia* traits, which also protect against malaria. Porter (1996 p. 34) writes: "In such a process, ... , close to 100 percent of Africans acquired a genetic trait that protects them against vivax malaria and probably against falciporum malaria as well." Overall, the WHO estimates that malaria kills about 1 million people per year, most of them children. It does not, however, generally kill adults who grew up in malaria-endemic areas (see Najera and Hempel, 1996).

Although yellow fever's epidemiology is quite different from malaria, it was also much more fatal to Europeans than to non-Europeans who grew up in areas where yellow fever commonly occurred.[13] Yellow fever leaves its surviving victims with a lifelong immunity, which also explains its epidemic pattern, relying on a concentrated nonimmune population. Curtin (1998 p. 10) writes: "Because most Africans had passed through a light case early in life, yellow fever in West Africa was a strangers' disease, attacking those who grew up elsewhere." Similarly, Michael B. A. Oldstone (1998 p. 49) writes:

> Most Black Africans and their descendants respond to yellow fever infection with mild to moderate symptoms such as headache, fever, nausea, and vomiting, and then recover in a few days. This outcome reflects the long relationship between the virus and its indigenous hosts, who through generations of exposure to the virus have evolved resistance.

In contrast, fatality rates among nonimmune adults, such as Europeans, could be as high as 90 percent.

Advances in medical science have reduced the danger posed by malaria and yellow fever. Yellow fever is mostly eradicated (Oldstone, 1998 Chapter 5), and malaria has been eradicated in many areas. Europeans developed methods of dealing with these diseases that gradually became more effective in the second half of the nineteenth century. For example, they came to understand that high doses of quinine, derived from the cinchona bark, acted as a prophylactic and prevented infection or reduced the severity of malaria. They also started to undertake serious mosquito eradication efforts and protect themselves against mosquito bites. Further, Europeans also learned that an often effective method of reducing mortality from yellow fever is flight from the area, since the transmitter mosquito, *Aedes aegypti,* has only a short range. Nevertheless, during much of the nineteenth century, there was almost a complete misunderstanding of the nature of both diseases. For example, the leading theory for malaria was that it was caused by "miasma" from swamps, and quinine was not used widely. The role of small collections of water to breed mosquitoes and transmit these diseases was not understood. It was only in the late nineteenth century that Europeans started to control these diseases.[14]

[12] Because malaria species are quite local, a person may have immunity to the local version of malaria, but be highly vulnerable to malaria a short distance away. This is probably the explanation for why Africans had such high mortality when they were forced to move by colonial powers. (Curtin et al., 1995 p. 463).

[13] Because yellow fever struck Europeans as an epidemic, many of the very high death rates we report below for European troops are from yellow fever.

[14] Even during the early twentieth century, there was much confusion about the causes of malaria and yellow

These considerations, together with the data we have on the mortality of local people and population densities before the arrival of Europeans, make us believe that settler mortality is a plausible instrument for institutional development: these diseases affected European settlement patterns and the type of institutions they set up, but had little effect on the health and economy of indigenous people.[15]

A final noteworthy feature, helpful in interpreting our results below, is that malaria prevalence depends as much on the microclimate of an area as on its temperature and humidity, or on whether it is in the tropics; high altitudes reduce the risk of infection, so in areas of high altitude, where "hill stations" could be set up, such as Bogota in Colombia, mortality rates were typically lower than in wet coastal areas. However, malaria could sometimes be more serious in high-altitude areas. For example, Curtin (1989 p. 47) points out that in Ceylon mortality was lower in the coast than the highlands because rains in the coast washed away the larvae of the transmitter mosquitoes. Similarly, in Madras many coastal regions were free of malaria, while northern India had high rates of infection. Curtin (1998 Chapter 7) also illustrates how there were marked differences in the prevalence of malaria within small regions of Madagascar. This suggests that mortality rates faced by Europeans are unlikely to be a proxy for some simple geographic or climactic feature of the country.

B. *Data on Potential Settler Mortality*

Our data on the mortality of European settlers come largely from the work of Philip Curtin. Systematic military medical record keeping began only after 1815, as an attempt to understand why so many soldiers were dying in some places. The first detailed studies were retrospective and dealt with British forces between 1817 and 1836. The United States and French governments quickly adopted similar methods (Curtin, 1989 pp. 3, 5). Some early data are also available for the Dutch East Indies. By the 1870's, most European countries published regular reports on the health of their soldiers.

The standard measure is annualized deaths per thousand mean strength. This measure reports the death rate among 1,000 soldiers where each death is replaced with a new soldier. Curtin (1989, 1998) reviews in detail the construction of these estimates for particular places and campaigns, and assesses which data should be considered reliable.

Curtin (1989), *Death by Migration,* deals primarily with the mortality of European troops from 1817 to 1848. At this time modern medicine was still in its infancy, and the European militaries did not yet understand how to control malaria and yellow fever. These mortality rates can therefore be interpreted as reasonable estimates of settler mortality. They are consistent with substantial evidence from other sources (see, for example, Curtin [1964, 1968]). Curtin (1998), *Disease and Empire,* adds similar data on the mortality of soldiers in the second half of the nineteenth century.[16] In all cases, we use the

fever. *The Washington Post* on Nov. 2, 1900 wrote: "Of all the silly and nonsensical rigmarole of yellow fever that has yet found its way into print ... the silliest beyond compare is to be found in the arguments and theories generated by a mosquito hypothesis" (quoted in Oldstone, 1998 pp. 64–65).

Many campaigns during the nineteenth century had very high mortality rates. For example, the French campaign in Madagascar during the 1890's and French attempts to build the Panama Canal during the 1880's were mortality disasters, the first due to malaria, the second due to yellow fever (see Curtin, 1998, and David McCullogh, 1977). In Panama, to stop ants the French used water pots under the legs of beds in barracks and hospitals. These pots provided an ideal milieu for the breeding of *Aedes aegypti,* causing very high rates of mortality (Oldstone, 1998 p. 66).

[15] In Acemoglu et al. (2001), we document that many of these areas in the tropical zone were richer and more densely settled in 1500 than the temperate areas later settled by the Europeans. This also supports the notion that the disease environment did not create an absolute disadvantage for these countries.

[16] These numbers have to be used with more care because there was a growing awareness of how to avoid epidemics of the worst tropical diseases, at least during short military campaigns. For example, the campaign in Ethiopia at the end of the nineteenth century had very low mortality rates because it was short and well managed (see Figure 1). Although the mortality rates from this successful campaign certainly underestimate the mortality rates faced

earliest available number for each country, reasoning that this is the best estimate of the mortality rates that settlers would have faced, at least until the twentieth century.

The main gap in the Curtin data is for South America since the Spanish and Portuguese militaries did not keep good records of mortality. Hector Gutierrez (1986) used Vatican records to construct estimates for the mortality rates of bishops in Latin America from 1604 to 1876. Because these data overlap with the Curtin estimates for several countries, we are able to construct a data series for South America.[17] Curtin (1964) also provides estimates of mortality in naval squadrons for different regions which we can use to generate alternative estimates of mortality in South America. Appendix B in Acemoglu et al. (2000), which is available from the authors, gives a detailed discussion of how these data are constructed, and Appendix Table A5 (available from the authors), shows that these alternative methods produce remarkably similar results. Appendix Table A2 lists our main estimates, and Table A1 gives information about sources.

IV. Institutions and Performance: IV Results

A. *Determinants of Current Institutions*

Equation (1) describes the relationship between current institutions and log GDP. In addition we have

$$R_i = \lambda_R + \beta_R C_i + \mathbf{X}_i'\gamma_R + \nu_{Ri}, \tag{2}$$

$$C_i = \lambda_C + \beta_C S_i + \mathbf{X}_i'\gamma_C + \nu_{Ci}, \tag{3}$$

$$S_i = \lambda_S + \beta_S \log M_i + \mathbf{X}_i'\gamma_S + \nu_{Si}, \tag{4}$$

where R is the measure of current institutions (protection against expropriation between 1985 and 1995), C is our measure of early (circa 1900) institutions, S is the measure of European settlements in the colony (fraction of the population with European descent in 1900), and M is mortality rates faced by settlers. $\mathbf{X}$ is a vector of covariates that affect all variables.

The simplest identification strategy might be to use S_i (or C_i) as an instrument for R_i in equation (1), and we report some of these regressions in Table 8. However, to the extent that settlers are more likely to migrate to richer areas and early institutions reflect other characteristics that are important for income today, this identification strategy would be invalid (i.e., C_i and S_i could be correlated with ε_i). Instead, we use the mortality rates faced by the settlers, log M_i, as an instrument for R_i. This identification strategy will be valid as long as log M_i is uncorrelated with ε_i—that is, if mortality rates of settlers between the seventeenth and nineteenth centuries have no effect on income today other than through their influence on institutional development. We argued above that this exclusion restriction is plausible.

Figure 3 illustrates the relationship between the (potential) settler mortality rates and the index of institutions. We use the logarithm of the settler mortality rates, since there are no theoretical reasons to prefer the level as a determinant of institutions rather than the log, and using the log ensures that the extreme African mortality rates do not play a disproportionate role. As it happens, there is an almost linear relationship between the log settler mortality and our measure of institutions. This relationship shows that ex-colonies where Europeans faced higher mortality rates have substantially worse institutions today.

In Table 3, we document that this relationship works through the channels hypothesized in Section I. In particular, we present OLS regressions of equations (2), (3), and (4). In the top panel, we regress the protection against expropriation variable on the other variables. Column (1) uses constraints faced by the executive in 1900 as the regressor, and shows a close association between early institutions and institutions today. For example, past institutions alone explain 20 percent of the variation in the index of current institutions. The second column adds the latitude variable,

by potential settlers in Ethiopia, we did not exclude this country because excluding it would have helped our hypothesis.

[17] Combining data from a variety of sources will introduce measurement error in our estimates of settler mortality. Nevertheless, since we are using settler mortality as an instrument, this measurement error does not lead to inconsistent estimates of the effect of institutions on performance.

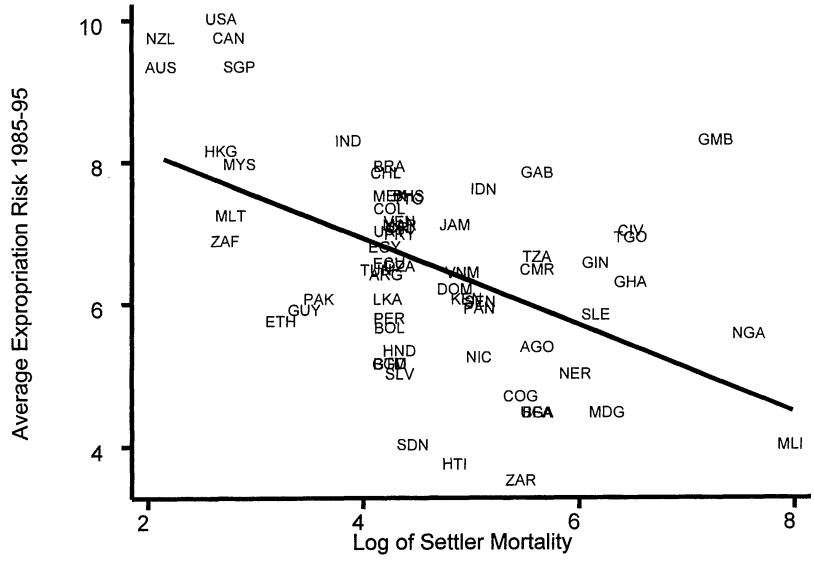


FIGURE 3. FIRST-STAGE RELATIONSHIP BETWEEN SETTLER MORTALITY AND EXPROPRIATION RISK

with little effect on the estimate. Columns (3) and (4) use the democracy index, and confirm the results in columns (1) and (2).

Both constraints on the executive and democracy indices assign low scores to countries that were colonies in 1900, and do not use the earliest postindependence information for Latin American countries and the Neo-Europes. In columns (5) and (6), we adopt an alternative approach and use the constraints on the executive in the first year of independence and also control separately for time since independence. The results are similar, and indicate that early institutions tend to persist.

Columns (7) and (8) show the association between protection against expropriation and European settlements. The fraction of Europeans in 1900 alone explains approximately 30 percent of the variation in our institutions variable today. Columns (9) and (10) show the relationship between the protection against expropriation variable and the mortality rates faced by settlers. This specification will be the first stage for our main two-stage least-squares estimates (2SLS). It shows that settler mortality alone explains 27 percent of the differences in institutions we observe today.

Panel B of Table 3 provides evidence in support of the hypothesis that early institutions were shaped, at least in part, by settlements, and that settlements were affected by mortality. Columns (1)–(2) and (5)–(6) relate our measure of constraint on the executive and democracy in 1900 to the measure of European settlements in 1900 (fraction of the population of European decent). Columns (3)–(4) and (7)–(8) relate the same variables to settler mortality. These regressions show that settlement patterns explain around 50 percent of the variation in early institutions. Finally, columns (9) and (10) show the relationship between settlements and mortality rates.

B. *Institutions and Economic Performance*

Two-stage least-squares estimates of equation (1) are presented in Table 4. Protection against expropriation variable, R_i, is treated as endogenous, and modeled as

$$(5) \qquad R_i = \zeta + \beta \log M_i + \mathbf{X}_i'\delta + v_i,$$

where M_i is the settler mortality rate in 1,000 mean strength. The exclusion restriction is that this variable does not appear in (1).

TABLE 3—DETERMINANTS OF INSTITUTIONS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A	Dependent Variable Is Average Protection Against Expropriation Risk in 1985–1995									
Constraint on executive in 1900	0.32	0.26								
	(0.08)	(0.09)								
Democracy in 1900			0.24	0.21						
			(0.06)	(0.07)						
Constraint on executive in first year of independence					0.25	0.22				
					(0.08)	(0.08)				
European settlements in 1900							3.20	3.00		
							(0.61)	(0.78)		
Log European settler mortality									−0.61	−0.51
									(0.13)	(0.14)
Latitude		2.20		1.60		2.70		0.58		2.00
		(1.40)		(1.50)		(1.40)		(1.51)		(1.34)
R^2	0.2	0.23	0.24	0.25	0.19	0.24	0.3	0.3	0.27	0.3
Number of observations	63	63	62	62	63	63	66	66	64	64

Panel B	Dependent Variable Is Constraint on Executive in 1900				Dependent Variable Is Democracy in 1900				Dependent Variable Is European Settlements in 1900	
European settlements in 1900	5.50	5.40			8.60	8.10				
	(0.73)	(0.93)			(0.90)	(1.20)				
Log European settler mortality			−0.82	−0.65			−1.22	−0.88	−0.11	−0.07
			(0.17)	(0.18)			(0.24)	(0.25)	(0.02)	(0.02)
Latitude		0.33		3.60		1.60		7.60		0.87
		(1.80)		(1.70)		(2.30)		(2.40)		(0.19)
R^2	0.46	0.46	0.25	0.29	0.57	0.57	0.28	0.37	0.31	0.47
Number of observations	70	70	75	75	67	67	68	68	73	73

Notes: All regressions are OLS. Standard errors are in parentheses. Regressions with constraint on executive in first year of independence also include years since independence as a regressor. Average protection against expropriation risk is on a scale from 0 to 10, where a higher score means more protection against expropriation of private investment by government, averaged over 1985 to 1995. Constraint on executive in 1900 is on a scale from 1 to 7, with a higher score indicating more constraints. Democracy in 1900 is on a scale from 0 to 10, with a higher score indicating more democracy. European settlements is percent of population that was European or of European descent in 1900. See Appendix Table A1 for more detailed variable definitions and sources.

Panel A of Table 4 reports 2SLS estimates of the coefficient of interest, α from equation (1) and Panel B gives the corresponding first stages.[18] Column (1) displays the strong first-stage relationship between (log) settler mortality and current institutions in our base sample, also shown in Table 3. The corresponding 2SLS estimate of the impact of institutions on income per capita is 0.94. This estimate is highly significant with a standard error of 0.16, and in fact larger than the OLS estimates reported in Table 2. This suggests that measurement error in the institutions variables that creates attenuation bias is likely to be more important than reverse causality and omitted variables biases. Here we are referring to "measurement error" broadly construed. In reality the set of institutions that matter for economic performance is very complex, and any single measure is bound to capture only part of the "true institutions,"

[18] We have also run these regressions with standard errors corrected for possible clustering of the mortality rates assigned to countries in the same disease environment. This clustering has little effect on the standard errors, and does not change our results.

TABLE 4—IV REGRESSIONS OF LOG GDP PER CAPITA

	Base sample (1)	Base sample (2)	Base sample without Neo-Europes (3)	Base sample without Neo-Europes (4)	Base sample without Africa (5)	Base sample without Africa (6)	Base sample with continent dummies (7)	Base sample with continent dummies (8)	Base sample, dependent variable is log output per worker (9)
Panel A: Two-Stage Least Squares									
Average protection against expropriation risk 1985–1995	0.94	1.00	1.28	1.21	0.58	0.58	0.98	1.10	0.98
	(0.16)	(0.22)	(0.36)	(0.35)	(0.10)	(0.12)	(0.30)	(0.46)	(0.17)
Latitude		−0.65		0.94		0.04		−1.20	
		(1.34)		(1.46)		(0.84)		(1.8)	
Asia dummy							−0.92	−1.10	
							(0.40)	(0.52)	
Africa dummy							−0.46	−0.44	
							(0.36)	(0.42)	
"Other" continent dummy							−0.94	−0.99	
							(0.85)	(1.0)	
Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995									
Log European settler mortality	−0.61	−0.51	−0.39	−0.39	−1.20	−1.10	−0.43	−0.34	−0.63
	(0.13)	(0.14)	(0.13)	(0.14)	(0.22)	(0.24)	(0.17)	(0.18)	(0.13)
Latitude		2.00		−0.11		0.99		2.00	
		(1.34)		(1.50)		(1.43)		(1.40)	
Asia dummy							0.33	0.47	
							(0.49)	(0.50)	
Africa dummy							−0.27	−0.26	
							(0.41)	(0.41)	
"Other" continent dummy							1.24	1.1	
							(0.84)	(0.84)	
R^2	0.27	0.30	0.13	0.13	0.47	0.47	0.30	0.33	0.28
Panel C: Ordinary Least Squares									
Average protection against expropriation risk 1985–1995	0.52	0.47	0.49	0.47	0.48	0.47	0.42	0.40	0.46
	(0.06)	(0.06)	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(0.06)
Number of observations	64	64	60	60	37	37	64	64	61

Notes: The dependent variable in columns (1)–(8) is log GDP per capita in 1995, PPP basis. The dependent variable in column (9) is log output per worker, from Hall and Jones (1999). "Average protection against expropriation risk 1985–1995" is measured on a scale from 0 to 10, where a higher score means more protection against risk of expropriation of investment by the government, from Political Risk Services. Panel A reports the two-stage least-squares estimates, instrumenting for protection against expropriation risk using log settler mortality; Panel B reports the corresponding first stage. Panel C reports the coefficient from an OLS regression of the dependent variable against average protection against expropriation risk. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable descriptions and sources.

creating a typical measurement error problem. Moreover, what matters for current income is presumably not only institutions today, but also institutions in the past. Our measure of institutions which refers to 1985–1995 will not be perfectly correlated with these.[19]

Does the 2SLS estimate make quantitative sense? Does it imply that institutional differences can explain a significant fraction of income dif-

[19] We can ascertain, to some degree, whether the difference between OLS and 2SLS estimates could be due to measurement error in the institutions variable by making use of an alternative measure of institutions, for example, the constraints on the executive measure. Using this measure as an instrument for the protection against expropriation index would solve the measurement error, but not the endogeneity problem. This exercise leads to an estimate of the effect of protection against expropriation equal to 0.87 (with standard error 0.16). This suggests that "measurement error" in the institutions variables (or the "signal-to-noise ratio" in the institutions variable) is of the right order of magnitude to explain the difference between the OLS and 2SLS estimates.

ferences across countries? Let us once again compare two "typical" countries with high and low expropriation risk, Nigeria and Chile (these countries are typical for the IV regression in the sense that they are practically on the regression line). Our 2SLS estimate, 0.94, implies that the 2.24 differences in expropriation risk between these two countries should translate into 206 log point (approximately 7-fold) difference. In practice, the presence of measurement error complicates this interpretation, because some of the difference between Nigeria and Chile's expropriation index may reflect measurement error. Therefore, the 7-fold difference is an upper bound. In any case, the estimates in Table 4 imply a substantial, but not implausibly large, effect of institutional differences on income per capita.

Column (2) shows that adding latitude does not change the relationship; the institutions coefficient is now 1.00 with a standard error of 0.22.[20] Remarkably, the latitude variable now has the "wrong" sign and is insignificant. This result suggests that many previous studies may have found latitude to be a significant determinant of economic performance because it is correlated with institutions (or with the exogenous component of institutions caused by early colonial experience).

Columns (3) and (4) document that our results are not driven by the Neo-Europes. When we exclude the United States, Canada, Australia, and New Zealand, the estimates remain highly significant, and in fact increase a little. For example, the coefficient for institutions is now 1.28 (s.e. = 0.36) without the latitude control, and 1.21 (s.e. = 0.35) when we control for latitude. Columns (5) and (6) show that our results are also robust to dropping all the African countries from our sample. The estimates without Africa are somewhat smaller, but also more precise. For example, the coefficient for institutions is 0.58 (s.e. = 0.1) without the latitude control, and still 0.58 (s.e. = 0.12) when we control for latitude.[21]

In columns (7) and (8), we add continent dummies to the regressions (for Africa, Asia, and other, with America as the omitted group). The addition of these dummies does not change the estimated effect of institutions, and the dummies are jointly insignificant at the 5-percent level, though the dummy for Asia is significantly different from that of America. The fact that the African dummy is insignificant suggests that the reason why African countries are poorer is not due to cultural or geographic factors, but mostly accounted for by the existence of worse institutions in Africa. Finally, in column (9) we repeat our basic regression using log of output per worker as calculated by Hall and Jones (1999). The result is very close to our baseline result. The 2SLS coefficient is 0.98 instead of 0.94 as in column (1).[22] This shows that whether we use income per capita or output per worker has little effect on our results. Overall, the results in Table 4 show a large effect of institutions on economic performance. In the rest of the paper, we investigate the robustness of these results.[23]

[20] In 2SLS estimation, all covariates that are included in the second stage, such as latitude, are also included in the first stage. When these first-stage effects are of no major significance for our argument, we do not report them in the tables to save space.

[21] We should note at this point that if we limit the sample to African countries only, the first-stage relationship using the protection against expropriation variable becomes considerably weaker, and the 2SLS effect of institutions is no longer significant. The 2SLS effect of institutions continue to be significant when we use some (but not all) measures of institutions. Therefore, we conclude that the relationship between settler mortality and institutions is weaker within Africa.

[22] The results with other covariates are also very similar. We repeated the same regressions using a variety of alternative measures of institutions, including constraints on the executive from the Polity III data set, an index of law and order tradition from Political Risk Services, a measure of property rights from the Heritage Foundation, a measure of rule of law from the Fraser Institute, and the efficiency of the judiciary from Business International. The results and the magnitudes are very similar to those reported in Table 4. We also obtained very similar results with the 1970 values for the constraints on the executive and income per capita in 1970, which show that the relationship between institutional measures and income per capita holds across time periods. These results are reported in the Appendix of the working paper version, and are also available from the authors.

[23] In the working paper version, we also investigated the robustness of our results in different subsamples with varying degrees of data quality and different methods of constructing the mortality estimates. The results change very little, for example, when we use data only from Curtin (1989), *Death by Migration,* when we do not assign mortality rates from neighboring disease environments, when

V. Robustness

A. *Additional Controls*

The validity of our 2SLS results in Table 4 depends on the assumption that settler mortality in the past has no direct effect on current economic performance. Although this presumption appears reasonable (at least to us), here we substantiate it further by directly controlling for many of the variables that could plausibly be correlated with both settler mortality and economic outcomes, and checking whether the addition of these variables affects our estimates.[24] Overall, we find that our results change remarkably little with the inclusion of these variables, and many variables emphasized in previous work become insignificant once the effect of institutions is controlled for.

La Porta et al. (1999) argue for the importance of colonial origin (identity of the main colonizing country) as a determinant of current institutions. The identity of the colonial power could also matter because it might have an effect through culture, as argued by David S. Landes (1998). In columns (1) and (2) of Table 5, we add dummies for British and French colonies (colonies of other nations are the omitted group). This has little affect on our results. Moreover, the French dummy in the first stage is estimated to be zero, while the British dummy is positive, and marginally significant. Therefore, as suggested by La Porta et al. (1998), British colonies appear to have better institutions, but this effect is much smaller and weaker than in a specification that does not control for the effect of settler mortality on institutional development.[25] Therefore, it appears that British colonies are found to perform substantially better in other studies in large part because Britain colonized places where settlements were possible, and this made British colonies inherit better institutions. To further investigate this issue, columns (3) and (4) estimate our basic regression for British colonies only. They show that both the relationship between settler mortality and institutions and that between institutions and income in this sample of 25 British colonies are very similar to those in our base sample. For example, the 2SLS estimate of the effect of institutions on income is now 1.07 (s.e. = 0.24) without controlling for latitude and 1.00 (s.e. = 0.22) with latitude. These results suggest that the identity of the colonizer is not an important determinant of colonization patterns and subsequent institutional development.

von Hayek (1960) and La Porta et al. (1999) also emphasize the importance of legal origin. In columns (5) and (6), we control for legal origin. In our sample, all countries have either French or British legal origins, so we simply add a dummy for French legal origin (many countries that are not French colonies nonetheless have French legal origin). Our estimate of the effect of institutions on income per capita is unaffected.[26]

An argument dating back to Max Weber views religion as a key determinant of economic performance. To control for this, in columns (7) and (8), we add the fraction of the populations that are Catholic, Muslim, and of other religions, with Protestants as the omitted group. In the table we report the joint significance level (p-value) of the corresponding F-statistic for these dummies as well as the 2SLS estimate of

the use data for Latin America from naval stations instead of bishops, and when we do not use data from small African samples. These results are available in Appendix Table A5 available from the authors, or in Acemoglu et al. (2000).

[24] Joseph N. Altonji et al. (2000) develop an econometric methodology to assess the importance of omitted variable bias. The basic idea is that if the estimate of the coefficient of interest does not change as additional covariates are included in the regression, it is less likely to change if we were able to add some of the missing omitted variables. Our methodology here is an informal version of this approach.

[25] Moreover, the British colonial dummy is negative and significant in the second stage. The net effect of being a British colony on income per capita is in fact *negative.* More specifically, British colonies have, on average, an index of institution that is 0.63 points lower. Given the 2SLS estimate of 1.10, this translates into 69 log points higher income per capita for British colonies ($1.10 \times 63 \approx 69$). The second-stage effect of being a British colony is -78 log points, implying -9 log point (approximately 10 percent) negative net effect of being a British colony. A possible explanation for this pattern is that (Anglo-Saxon?) researchers are overestimating how "bad" French institutions are, and the second-stage regression is correcting for this.

[26] The first stage shows that French legal origin is associated with worse institutions, but similarly, the net effect of having French legal origin is actually *positive:* $-67 \times 1.1 + 89 = 15$ log points (approximately 15 percent).

TABLE 5—IV REGRESSIONS OF LOG GDP PER CAPITA WITH ADDITIONAL CONTROLS

	Base sample (1)	Base sample (2)	British colonies only (3)	British colonies only (4)	Base sample (5)	Base sample (6)	Base sample (7)	Base sample (8)	Base sample (9)
Panel A: Two-Stage Least Squares									
Average protection against expropriation risk, 1985–1995	1.10 (0.22)	1.16 (0.34)	1.07 (0.24)	1.00 (0.22)	1.10 (0.19)	1.20 (0.29)	0.92 (0.15)	1.00 (0.25)	1.10 (0.29)
Latitude		−0.75 (1.70)				−1.10 (1.56)		−0.94 (1.50)	−1.70 (1.6)
British colonial dummy	−0.78 (0.35)	−0.80 (0.39)							
French colonial dummy	−0.12 (0.35)	−0.06 (0.42)							0.02 (0.69)
French legal origin dummy					0.89 (0.32)	0.96 (0.39)			0.51 (0.69)
p-value for religion variables							[0.001]	[0.004]	[0.42]
Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995									
Log European settler mortality	−0.53 (0.14)	−0.43 (0.16)	−0.59 (0.19)	−0.51 (0.14)	−0.54 (0.13)	−0.44 (0.14)	−0.58 (0.13)	−0.44 (0.15)	−0.48 (0.18)
Latitude		1.97 (1.40)				2.10 (1.30)		2.50 (1.50)	2.30 (1.60)
British colonial dummy	0.63 (0.37)	0.55 (0.37)							
French colonial dummy	0.05 (0.43)	−0.12 (0.44)							−0.25 (0.89)
French legal origin					−0.67 (0.33)	−0.7 (0.32)			−0.05 (0.91)
R^2	0.31	0.33	0.30	0.30	0.32	0.35	0.32	0.35	0.45
Panel C: Ordinary Least Squares									
Average protection against expropriation risk, 1985–1995	0.53 (0.19)	0.47 (0.07)	0.61 (0.09)	0.47 (0.06)	0.56 (0.06)	0.56 (0.06)	0.53 (0.06)	0.47 (0.06)	0.47 (0.06)
Number of observations	64	64	25	25	64	64	64	64	64

Notes: Panel A reports the two-stage least-squares estimates with log GDP per capita (PPP basis) in 1995 as dependent variable, and Panel B reports the corresponding first stage. The base case in columns (1) and (2) is all colonies that were neither French nor British. The religion variables are included in the first stage of columns (7) and (8) but not reported here (to save space). Panel C reports the OLS coefficient from regressing log GDP per capita on average protection against expropriation risk, with the other control variables indicated in that column (full results not reported to save space). Standard errors are in parentheses and *p*-values for joint significance tests are in brackets. The religion variables are percentage of population that are Catholics, Muslims, and "other" religions; Protestant is the base case. Our sample is all either French or British legal origin (as defined by La Porta et al., 1999).

the effect of institutions.[27] Finally, column (9) adds all the variables in this table simultaneously. Again, these controls have very little effect on our main estimate.

Another concern is that settler mortality is correlated with climate and other geographic characteristics. Our instrument may therefore be picking up the direct effect of these variables. We investigate this issue in Table 6. In columns (1) and (2), we add a set of temperature and humidity variables (all data from Philip M. Parker, 1997). In the table we report joint significance levels for these variables. Again, they have little effect on our estimates.

[27] The religion dummies are significant in the first stage, but once again they are estimated to have offsetting effects in the second stage, implying little net effect of religion on income.

TABLE 6—ROBUSTNESS CHECKS FOR IV REGRESSIONS OF LOG GDP PER CAPITA

	Base sample (1)	Base sample (2)	Base sample (3)	Base sample (4)	Base sample (5)	Base sample (6)	Base sample (7)	Base sample (8)	Base sample (9)
Panel A: Two-Stage Least Squares									
Average protection against expropriation risk, 1985–1995	0.84 (0.19)	0.83 (0.21)	0.96 (0.28)	0.99 (0.30)	1.10 (0.33)	1.30 (0.51)	0.74 (0.13)	0.79 (0.17)	0.71 (0.20)
Latitude		0.07 (1.60)		−0.67 (1.30)		−1.30 (2.30)		−0.89 (1.00)	−2.5 (1.60)
p-value for temperature variables	[0.96]	[0.97]							[0.77]
p-value for humidity variables	[0.54]	[0.54]							[0.62]
Percent of European descent in 1975			−0.08 (0.82)	0.03 (0.84)					0.3 (0.7)
p-value for soil quality					[0.79]	[0.85]			[0.46]
p-value for natural resources					[0.82]	[0.87]			[0.82]
Dummy for being landlocked					0.64 (0.63)	0.79 (0.83)			0.75 (0.47)
Ethnolinguistic fragmentation							−1.00 (0.32)	−1.10 (0.34)	−1.60 (0.47)
Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995									
Log European settler mortality	−0.64 (0.17)	−0.59 (0.17)	−0.41 (0.14)	−0.4 (0.15)	−0.44 (0.16)	−0.34 (0.17)	−0.64 (0.15)	−0.56 (0.15)	−0.59 (0.21)
Latitude		2.70 (2.00)		0.48 (1.50)		2.20 (1.50)		2.30 (1.40)	4.20 (2.60)
R^2	0.39	0.41	0.34	0.34	0.41	0.43	0.27	0.30	0.59
Panel C: Ordinary Least Squares									
Average protection against expropriation risk, 1985–1995	0.41 (0.06)	0.38 (0.06)	0.39 (0.06)	0.38 (0.06)	0.46 (0.07)	0.42 (0.07)	0.46 (0.05)	0.45 (0.06)	0.38 (0.06)

Notes: Panel A reports the two-stage least-squares estimates with log GDP per capita (PPP basis) in 1995, and Panel B reports the corresponding first stages. Panel C reports the OLS coefficient from regressing log GDP per capita on average protection against expropriation risk, with the other control variables indicated in that column (full results not reported to save space). Standard errors are in parentheses and *p*-values for joint significance tests are in brackets. All regressions have 64 observations, except those including natural resources, which have 63 observations. The temperature and humidity variables are: average, minimum, and maximum monthly high temperatures, and minimum and maximum monthly low temperatures, and morning minimum and maximum humidity, and afternoon minimum and maximum humidity (from Parker, 1997). Measures of natural resources are: percent of world gold reserves today, percent of world iron reserves today, percent of world zinc reserves today, number of minerals present in country, and oil resources (thousands of barrels per capita). Measures of soil quality/climate are steppe (low latitude), desert (low latitude), steppe (middle latitude), desert (middle latitude), dry steppe wasteland, desert dry winter, and highland. See Appendix Table A1 for more detailed variable definitions and sources.

A related concern is that in colonies where Europeans settled, the current population consists of a higher fraction of Europeans. One might be worried that we are capturing the direct effect of having more Europeans (who perhaps brought a "European culture" or special relations with Europe). To control for this, we add the fraction of the population of European descent in columns (3) and (4) of Table 6. This variable is insignificant, while the effect of institutions remains highly significant, with a coefficient of 0.96 (s.e. = 0.28). In columns (5) and (6), we control for measures of natural resources, soil quality (in practice soil types), and for whether the country is landlocked. All these controls are insignificant, and have little effect on our 2SLS estimate of the effect of institutions on income per capita.

In columns (7) and (8), we include ethnolinguistic fragmentation as another control and treat it as exogenous. Now the coefficient

of protection against expropriation is 0.74 (s.e. = 0.13), which is only slightly smaller than our baseline estimate. In Appendix A, we show that the inclusion of an endogenous variable positively correlated with income or institutions will bias the coefficient on institutions downwards. Since ethnolinguistic fragmentation is likely to be endogenous with respect to development (i.e., ethnolinguistic fragmentation tends to disappear after the formation of centralized markets; see Weber [1976] or Andersen [1983]) and is correlated with settler mortality, the estimate of 0.74 likely understates the effect of institutions on income. In column (9) of Table 6, we include all these variables together. Despite the large number of controls, protection against expropriation on income per capita is still highly significant, with a somewhat smaller coefficient of 0.71 (s.e. = 0.20), which is again likely to understate the effect of institutions on income because ethnolinguistic fragmentation is treated as exogenous.

Finally, in Table 7, we investigate whether our instrument could be capturing the general effect of disease on development. Sachs and a series of coauthors have argued for the importance of malaria and other diseases in explaining African poverty (see, for example, Bloom and Sachs, 1998; Gallup and Sachs, 1998; Gallup et al., 1998). Since malaria was one of the main causes of settler mortality, our estimate may be capturing the direct effect of malaria on economic performance. We are skeptical of this argument since malaria prevalence is highly endogenous; it is the poorer countries with worse institutions that have been unable to eradicate malaria.[28] While Sachs and coauthors argue that malaria reduces output through poor health, high mortality, and absenteeism, most people who live in high malaria areas have developed some immunity to the disease (see the discussion in Section III, subsection A). Malaria should therefore have little direct effect on economic performance (though, obviously, it will have very high social costs). In contrast, for Europeans, or anyone else who has not been exposed to malaria as a young child, malaria is usually fatal, making malaria prevalence a key determinant of European settlements and institutional development.

In any case, controlling for malaria does not change our results. We do this in columns (1) and (2) by controlling for the fraction of the population who live in an area where falciporum malaria is endemic in 1994 (as constructed and used by Gallup et al., 1998). Since malaria prevalence in 1994 is highly endogenous, the argument in Appendix A implies that controlling for it directly will underestimate the effect of institutions on performance. In fact, the coefficient on protection against expropriation is now estimated to be somewhat smaller, 0.69 instead of 0.94 as in Table 4. Nevertheless, the effect remains highly significant with a standard error of 0.25, while malaria itself is insignificant.

In a comment on the working paper version of our study, John W. McArthur and Sachs (2001) discuss the role of geography and institutions in determining economic performance. They accept our case for the importance of institutions, but argue that more general specifications show that the disease environment and health characteristics of countries (their "geography") matter for economic performance. In particular, they extend our work by controlling for life expectancy and infant mortality, and they also instrument for these health variables using geographic variables such as latitude and mean temperature. Table 7 also expands upon the specifications that McArthur and Sachs suggest. Columns (3)–(6) include life expectancy and infant mortality as exogenous controls. The estimates show a significant effect of institutions on income, similar to, but smaller than, our baseline estimates. Infant mortality is also marginally significant. Since health is highly endogenous, the coefficient on these variables will be biased up, while the coefficient of institutions will be biased down (see Appendix A). These estimates are therefore consistent with

[28] For example, the United States eliminated malaria from the Panama Canal Zone, and Australia eliminated it from Queensland (see Crosby, 1986 pp. 141–42). Even in Africa, there have been very successful campaigns against malaria, including those in Algeria and that conducted by the Rio-Tinto Zinc mining company in Zambia (then Northern Rhodesia). The WHO's Roll Back Malaria program contains a number of effective recommendations for controlling malaria that are relatively straightforward to implement if families have enough money (e.g., insecticide-treated bed nets).

TABLE 7—GEOGRAPHY AND HEALTH VARIABLES

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Instrumenting only for average protection against expropriation risk						Instrumenting for all right-hand-side variables			Yellow fever instrument for average protection against expropriation risk	
Panel A: Two-Stage Least Squares											
Average protection against expropriation risk, 1985–1995	0.69 (0.25)	0.72 (0.30)	0.63 (0.28)	0.68 (0.34)	0.55 (0.24)	0.56 (0.31)	0.69 (0.26)	0.74 (0.24)	0.68 (0.23)	0.91 (0.24)	0.90 (0.32)
Latitude		−0.57 (1.04)		−0.53 (0.97)		−0.1 (0.95)					
Malaria in 1994	−0.57 (0.47)	−0.60 (0.47)					−0.62 (0.68)				
Life expectancy			0.03 (0.02)	0.03 (0.02)				0.02 (0.02)			
Infant mortality					−0.01 (0.005)	−0.01 (0.006)			−0.01 (0.01)		
Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995											
Log European settler mortality	−0.42 (0.19)	−0.38 (0.19)	−0.34 (0.17)	−0.30 (0.18)	−0.36 (0.18)	−0.29 (0.19)	−0.41 (0.17)	−0.40 (0.17)	−0.40 (0.17)		
Latitude		1.70 (1.40)		1.10 (1.40)		1.60 (1.40)	−0.81 (1.80)	−0.84 (1.80)	−0.84 (1.80)		
Malaria in 1994	−0.79 (0.54)	−0.65 (0.55)									
Life expectancy			0.05 (0.02)	0.04 (0.02)							
Infant mortality					−0.01 (0.01)	−0.01 (0.01)					
Mean temperature							−0.12 (0.05)	−0.12 (0.05)	−0.12 (0.05)		
Distance from coast							0.57 (0.51)	0.55 (0.52)	0.55 (0.52)		
Yellow fever dummy										−1.10 (0.41)	−0.81 (0.38)
R^2	0.3	0.31	0.34	0.35	0.32	0.34	0.37	0.36	0.36	0.10	0.32
Panel C: Ordinary Least Squares											
Average protection against expropriation risk, 1985–1995	0.35 (0.06)	0.35 (0.06)	0.28 (0.05)	0.28 (0.05)	0.29 (0.05)	0.28 (0.05)	0.35 (0.06)	0.29 (0.05)	0.29 (0.05)	0.48 (0.06)	0.39 (0.06)
Number of observations	62	62	60	60	60	60	60	59	59	64	64

Notes: Panel A reports the two-stage least-squares estimates with log GDP per capita (PPP basis) in 1995, and Panel B reports the corresponding first stages. Panel C reports the coefficient from an OLS regression with log GDP per capita as the dependent variable and average protection against expropriation risk and the other control variables indicated in each column as independent variables (full results not reported to save space). Standard errors are in parentheses. Columns (1)–(6) instrument for average protection against expropriation risk using log mortality and assume that the other regressors are exogenous. Columns (7)–(9) include as instruments average temperature, amount of territory within 100 km of the coast, and latitude (from McArthur and Sachs, 2001). Columns (10) and (11) use a dummy variable for whether or not a country was subject to yellow fever epidemics before 1900 as an instrument for average protection against expropriation. See Appendix Table A1 for more detailed variable definitions and sources.

institutions being the major determinant of income per capita differences, with little effect from geography/health variables.

Columns (7)–(9) report estimates from models that treat both health and institutions as endogenous, and following McArthur and Sachs, instrument for them using latitude, mean temperature, and distance from the coast as instruments in addition to our instrument, settler mortality. McArthur and Sachs (2001) report that in these regressions the institution variable is still significant, but geography/health are also significant. In contrast to McArthur and Sachs' results, we find that only institutions are signif-

icant. This difference is due to the fact that McArthur and Sachs include Britain and France in their sample. Britain and France are not in our sample, which consists of only ex-colonies (there is no reason for variation in the mortality rates of British and French troops at home to be related to their institutional development). It turns out that once Britain and France are left out, the McArthur and Sachs' specification generates no evidence that geography/health variables have an important effect on economic performance.[29]

As a final strategy to see whether settler mortality could be proxying for the current disease environment, we estimated models using a yellow fever instrument. This is a dummy variable indicating whether the area was ever affected by yellow fever (from Oldstone, 1998; see Appendix Table A1). This is an attractive alternative strategy because yellow fever is mostly eradicated today, so this dummy should not be correlated with the current disease environment. The disadvantage of this approach is that there is less variation in this instrument than our settler mortality variable. Despite this, the yellow fever results, reported in columns (10) and (11) of Table 7, are encouraging. The estimate in our base sample is 0.91 (s.e. = 0.24) comparable to our baseline estimate of 0.95 reported in Table 4. Adding continent dummies in column (11) reduces this estimate slightly to 0.90 (s.e. = 0.32).[30]

[29] McArthur and Sachs (2001) also report specifications with more instruments. However, using six or seven instruments with only 64 observations leads to the "too-many-instruments" problem, typically biasing the IV estimate towards the OLS estimate (see John Bound et al., 1995). We therefore did not pursue these estimates further.

Finally, McArthur and Sachs also argue that our ex-colonies sample may not have enough geographic variation. In their view, this may be why we do not find a role for geographic variables. Nonetheless, there is substantial variation in the geography variables in our sample which includes countries such as Canada, the United States, New Zealand, and Australia. The standard deviation of distance from the equator in the world is 1.89, greater than 1.33 in our sample. This is mainly because there are a large number of European countries with high latitudes in the world sample, but not in our sample.

[30] If we drop the Neo-Europes (not reported here), the estimate is still similar and highly significant, 1.05 (s.e. = 0.35).

B. *Overidentification Tests*

We can also investigate the validity of our approach by using overidentification tests. According to our theory, settler mortality (M) affected settlements (S); settlements affected early institutions (C); and early institutions affected current institutions (R)—cf., equations (2), (3), and (4). We can test whether any of these variables, C, S, and M, has a direct effect on income per capita, log y, by using measures of C and S as additional instruments. The overidentification test presumes that one of these instruments, say S, is truly exogenous, and tests for the exogeneity of the others, such as settler mortality. This approach is useful since it is a direct test of our exclusion restriction. However, such tests may not lead to a rejection if all instruments are invalid, but still highly correlated with each other. Therefore, the results have to be interpreted with caution.

Overall, the overidentification test will reject the validity of our approach if either (i) the equation of interest, (1), does not have a constant coefficient, i.e., $\log y_i = \mu + \alpha_i R_i + \varepsilon_i$, where i denotes country, or (ii) C or S has a direct effect on income per capita, $\log y_i$ (i.e., either S or C is correlated with ε_i), or (iii) settler mortality, M, has an effect on $\log y_i$ that works through another variable, such as culture.

The data support the overidentifying restrictions implied by our approach.[31] This implies that, subject to the usual problems of power associated with overidentification tests, we can rule out all three of the above possibilities. This gives us additional confidence that settler mortality is a valid instrument and that we are estimating the effect of institutions on current performance with our instrumental-variable strategy (i.e., not capturing the effect of omitted variables).

[31] In some specifications, the overidentification tests using measures of early institutions reject at that 10-percent level (but not at the 5-percent level). There are in fact good reasons to expect institutions circa 1900 to have a direct effect on income today (and hence the overidentifying tests to reject our restrictions): these institutions should affect physical and human capital investments at the beginning of the century, and have some effect on current income levels through this channel.

TABLE 8—OVERIDENTIFICATION TESTS

	Base sample (1)	Base sample (2)	Base sample (3)	Base sample (4)	Base sample (5)	Base sample (6)	Base sample (7)	Base sample (8)	Base sample (9)	Base sample (10)
Panel A: Two-Stage Least Squares										
Average protection against expropriation risk, 1985–1995	0.87	0.92	0.71	0.68	0.72	0.69	0.60	0.61	0.55	0.56
	(0.14)	(0.20)	(0.15)	(0.20)	(0.14)	(0.19)	(0.14)	(0.17)	(0.12)	(0.14)
Latitude		−0.47		−0.34		0.31		−0.41		−0.16
		(1.20)		(1.10)		(1.05)		(0.92)		(0.81)
Panel B: First Stage for Average Protection Against Expropriation Risk										
European settlements in 1900	3.20	2.90								
	(0.62)	(0.83)								
Constraint on executive in 1900			0.32	0.26						
			(0.08)	(0.09)						
Democracy in 1900					0.24	0.20				
					(0.06)	(0.07)				
Constraint on executive in first year of independence							0.25	0.22		
							(0.08)	(0.08)		
Democracy in first year of independence									0.19	0.17
									(0.05)	(0.05)
R^2	0.30	0.30	0.20	0.24	0.24	0.26	0.19	0.25	0.26	0.30
Panel C: Results from Overidentification Test										
p-value (from chi-squared test)	[0.67]	[0.96]	[0.09]	[0.20]	[0.11]	[0.28]	[0.67]	[0.79]	[0.22]	[0.26]
Panel D: Second Stage with Log Mortality as Exogenous Variable										
Average protection against expropriation risk, 1985–1995	0.81	0.88	0.45	0.42	0.52	0.48	0.49	0.49	0.4	0.41
	(0.23)	(0.30)	(0.25)	(0.30)	(0.23)	(0.28)	(0.23)	(0.25)	(0.18)	(0.19)
Log European settler mortality	−0.07	−0.05	−0.25	−0.26	−0.21	−0.22	−0.14	−0.14	−0.19	−0.19
	(0.17)	(0.18)	(0.16)	(0.17)	(0.15)	(0.16)	(0.16)	(0.15)	(0.13)	(0.12)
Latitude		−0.52		0.38		0.28		−0.38		−0.17
		(1.15)		(0.89)		(0.86)		(0.84)		(0.73)

Notes: Panel A reports the two-stage least-squares estimates with log GDP per capita (PPP basis) in 1995 as the dependent variable, and Panel B reports the corresponding first stage (latitude is included in even-numbered columns but is never significant and not reported here to save space). Panel C reports the p-value for the null hypothesis that the coefficient on average protection against expropriation risk in the second-stage regression (i.e., Panel A) is the same as when instrumented using log mortality of settlers in addition to the indicated instruments. Panel D reports results from the regression in which log mortality is included as an exogenous variable and current institutions are instrumented using the alternative instrument indicated. Standard errors are in parentheses. All regressions with constraint on executive and democracy in first year of independence also include years since independence as a regressor. All regressions have 60 observations, except those with democracy in 1900 which have 59 observations and those with European settlements in 1900 which have 63 observations.

The results of the overidentification tests, and related results, are reported in Table 8. In the top panel, Panel A, we report the 2SLS estimates of the effect of protection against expropriation on GDP per capita using a variety of instruments other than mortality rates, while Panel B gives the corresponding first stages. These estimates are always quite close to those reported in Table 4. For example, in column (1), we use European settlements in 1900 as the *only* instrument for institutions. This results in an estimated effect of 0.87 (with standard error 0.14), as compared to our baseline estimate of 0.94. The other columns add latitude, and use other instruments such as constraint on the executive in 1900 and in the first year of independence, and democracy in 1900.

Panel D reports an easy-to-interpret version of the overidentification test. It adds the log of mortality as an exogenous regressor. If mortality rates faced by settlers had a direct effect on income per capita, we would expect this variable to come in negative and significant. In all cases, it is small and statistically insignificant. For example, in column (1), log mortality has a coefficient of −0.07 (with standard error 0.17). This confirms that the

impact of mortality rates faced by settlers likely works through their effect on institutions.

Finally, for completeness, in Panel C we report the p-value from the appropriate χ^2 overidentification test. This tests whether the 2SLS coefficients estimated with the instruments indicated in Panels A and B versus the coefficients estimated using (log) settler mortality in addition to the "true" instruments are significantly different (e.g., in the first column, the coefficient using European settlements alone is compared to the estimate using European settlements and log mortality as instruments). We never reject the hypothesis that they are equal at the 5-percent significance level. So these results also show no evidence that mortality rates faced by settlers have a direct effect—or an effect working through a variable other than institutions—on income per capita.

VI. Concluding Remarks

Many economists and social scientists believe that differences in institutions and state policies are at the root of large differences in income per capita across countries. There is little agreement, however, about what determines institutions and government attitudes towards economic progress, making it difficult to isolate exogenous sources of variation in institutions to estimate their effect on performance. In this paper we argued that differences in colonial experience could be a source of exogenous differences in institutions.

Our argument rests on the following premises: (1) Europeans adopted very different colonization strategies, with different associated institutions. In one extreme, as in the case of the United States, Australia, and New Zealand, they went and settled in the colonies and set up institutions that enforced the rule of law and encouraged investment. In the other extreme, as in the Congo or the Gold Coast, they set up extractive states with the intention of transferring resources rapidly to the metropole. These institutions were detrimental to investment and economic progress. (2) The colonization strategy was in part determined by the feasibility of European settlement. In places where Europeans faced very high mortality rates, they could not go and settle, and they were more likely to set up extractive states. (3) Finally, we argue that these early institutions persisted to the present. Determinants of whether Europeans could go and settle in the colonies, therefore, have an important effect on institutions today. We exploit these differences as a source of exogenous variation to estimate the impact of institutions on economic performance.

There is a high correlation between mortality rates faced by soldiers, bishops, and sailors in the colonies and European settlements; between European settlements and early measures of institutions; and between early institutions and institutions today. We estimate large effects of institutions on income per capita using this source of variation. We also document that this relationship is not driven by outliers, and is robust to controlling for latitude, climate, current disease environment, religion, natural resources, soil quality, ethnolinguistic fragmentation, and current racial composition.

It is useful to point out that our findings do not imply that institutions today are predetermined by colonial policies and cannot be changed. We emphasize colonial experience as one of the many factors affecting institutions. Since mortality rates faced by settlers are arguably exogenous, they are useful as an instrument to isolate the effect of institutions on performance. In fact, our reading is that these results suggest substantial economic gains from improving institutions, for example as in the case of Japan during the Meiji Restoration or South Korea during the 1960's.

There are many questions that our analysis does not address. Institutions are treated largely as a "black box": The results indicate that reducing expropriation risk (or improving other aspects of the "cluster of institutions") would result in significant gains in income per capita, but do not point out what concrete steps would lead to an improvement in these institutions. Institutional features, such as expropriation risk, property rights enforcement, or rule of law, should probably be interpreted as an equilibrium outcome, related to some more fundamental "institutions," e.g., presidential versus parliamentary system, which can be changed directly. A more detailed analysis of the effect of more fundamental institutions on property

rights and expropriation risk is an important area for future study.

Appendix A: Bias in the Effect of Institutions When Other Endogenous Variables Are Included

To simplify notation, suppose that R_i is exogenous, and another variable that is endogenous, z_i, such as prevalence of malaria or ethnolinguistic fragmentation, is added to the regression. Then, the simultaneous equations model becomes

$$Y_i = \mu_0 + \alpha R_i + \pi z_i + \varepsilon_i$$

$$z_i = \mu_1 + \phi Y_i + \eta_i,$$

where $Y_i = \log y_i$. We presume that $\alpha \geq 0$, $\phi < 0$, and $\pi < 0$, which implies that we interpret z_i as a negative influence on income. Moreover, this naturally implies that $\text{cov}(\eta_i, \varepsilon_i) < 0$ and $\text{cov}(z_i, R_i) < 0$, that is, the factor z_i is likely to be negatively correlated with positive influences on income.

Standard arguments imply that

$$\text{plim } \hat{\alpha} = \alpha + \frac{\text{cov}(\tilde{R}_i, \varepsilon_i)}{\text{var}(\tilde{R}_i)}$$

$$= \alpha - \kappa \cdot \frac{\text{cov}(z_i, \varepsilon_i)}{\text{var}(\tilde{R}_i)},$$

where κ and $\tilde{R}_i$ are the coefficient and the residual from the auxiliary equation, $R_i = \kappa_0 + \kappa z_i + \tilde{R}_i$, and so $\kappa = \text{cov}(z_i, R_i)/\text{var}(z_i) < 0$, which is negative due to the fact that $\text{cov}(R_i, z_i) < 0$. The reduced form for z_i is:

$$\text{(A1)} \qquad z_i = \frac{1}{1 - \phi\pi}\left((\mu + \phi\pi) + \phi\alpha R_i + \phi\varepsilon_i + \eta_i\right).$$

We impose the regularity condition $\phi \cdot \pi < 1$, so that an increase in the disturbance to the z-equation, η_i, actually increases z_i. Now using this reduced form, we can write

$$\text{(A2)} \qquad \text{plim } \hat{\alpha} = \alpha - \kappa \cdot \frac{\text{cov}(z_i, \varepsilon_i)}{\text{var}(\tilde{R}_i)}$$

$$= \alpha - \kappa \cdot \frac{(\sigma_{\varepsilon\eta} + \phi\sigma_\varepsilon^2)}{(1 - \phi\pi) \cdot \text{var}(\tilde{R}_i)}$$

where σ_ε^2 is the variance of ε, and $\sigma_{\varepsilon\eta}$ is the covariance of ε and η.

Substituting for κ in (A2), we obtain:

$$\text{plim } \hat{\alpha} = \alpha - \frac{(\sigma_{\varepsilon\eta} + \phi\sigma_\varepsilon^2)}{(1 - \phi\pi) \cdot \text{var}(\tilde{R}_i)} \cdot \frac{\text{cov}(z_i, R_i)}{\text{var}(z_i)}.$$

Recall that $\phi < 0$, $\sigma_{\varepsilon\eta} < 0$, and $\text{cov}(z_i, R_i) < 0$. Therefore, plim $\hat{\alpha} < \alpha$, and when we control for the endogenous variable z_i, the coefficient on our institution variable will be biased downwards.

APPENDIX TABLE A1: DATA DESCRIPTIONS AND SOURCES

Log GDP per capita, 1975 and 1995: Purchasing Power Parity Basis, from World Bank, World Development Indicators, CD-Rom, 1999.
Log output per worker, 1988: As used in Hall and Jones (1999), from www.stanford.edu/~chadj.
Average protection against expropriation risk, 1985–1995: Risk of expropriation of private foreign investment by government, from 0 to 10, where a higher score means less risk. Mean value for all years from 1985 to 1995. This data was previously used by Knack and Keefer (1995) and was organized in electronic form by the IRIS Center (University of Maryland); originally Political Risk Services.
Constraint on executive in 1900, 1970, 1990 and in first year of independence: Seven-category scale, from 1 to 7, with a higher score indicating more constraints. Score of 1 indicates unlimited authority; score of 3 indicates slight to moderate limitations; score of 5 indicates substantial limitations; score of 7 indicates executive parity or subordination. Equal to 1 if country was not independent at that date. Date of independence is the first year that the country appears in the Polity III data set. From the Polity III data set, downloaded from Inter-University Consortium for Political and Social Research. See Gurr (1997).
Democracy in 1900 and first year of independence: An 11-category scale, from 0 to 10, with a higher score indicating more democracy. Points from three dimensions: Competitiveness of Political Participation (from 1 to 3); Competitiveness of Executive Recruitment (from 1 to 2, with a bonus of 1 point if there is an election); and Constraints on Chief Executive (from 1 to 4). Equal to 1 if country not independent at that date. From the Polity III data set. See Gurr (1997).
European settlements in 1900 and percent of European descent 1975: Percent of population European or of European descent in 1900 and 1975. From McEvedy and Jones (1975) and other sources listed in Appendix Table A6 (available from the authors).
Ethnolinguistic fragmentation: Average of five different indices of ethnolinguistic fragmentation. Easterly and Levine (1997), as used in La Porta et al. (1999).
Religion variables: Percent of population that belonged to the three most widely spread religions of the world in 1980 (or for 1990–1995 for countries formed more recently). The four classifications are: Roman Catholic, Protestant, Muslim, and "other." From La Porta et al. (1999).
French legal origin dummy: Legal origin of the company law or commercial code of each country. Our base sample is all French Commercial Code or English Common Law Origin. From La Porta et al. (1999).
Colonial dummies: Dummy indicating whether country was a British, French, German, Spanish, Italian, Belgian, Dutch, or Portuguese colony. From La Porta et al. (1999).
Temperature variables: Average temperature, minimum monthly high, maximum monthly high, minimum monthly low, and maximum monthly low, all in centigrade. From Parker (1997).
Mean temperature: 1987 mean annual temperature in degrees Celsius. From McArthur and Sachs (2001).
Humidity variables: Morning minimum, morning maximum, afternoon minimum, and afternoon maximum, all in percent. From Parker (1997).
Soil quality: Dummies for steppe (low latitude), desert (low latitude), steppe (middle latitude), desert (middle latitude), dry steppe wasteland, desert dry winter, and highland. From Parker (1997).
Natural resources: Percent of world gold reserves today, percent of world iron reserves today, percent of world zinc reserves today, number of minerals present in country, and oil resources (thousands of barrels per capita.) From Parker (1997).
Dummy for landlocked: Equal to 1 if country does not adjoin the sea. From Parker (1997).
Malaria in 1994: Population living where falciporum malaria is endemic (percent). Gallup and Sachs (1998).
Latitude: Absolute value of the latitude of the country (i.e., a measure of distance from the equator), scaled to take values between 0 and 1, where 0 is the equator. From La Porta et al. (1999).
Log European settler mortality: See Appendix Table A2, reproduced below, and Appendix B (available from the authors).
Yellow fever: Dummy equal to 1 if yellow fever epidemics before 1900 and 0 otherwise. Oldstone (1998 p. 69) shows current habitat of the mosquito vector; these countries are coded equal to 1. In addition, countries in which there were epidemics in the nineteenth century, according to Curtin (1989, 1998) are also coded equal to 1.
Infant mortality: Infant mortality rate (deaths per 1,000 live births). From McArthur and Sachs (2001).
Life expectancy: Life expectancy at birth in 1995. From McArthur and Sachs (2001).
Distance from the coast: Proportion of land area within 100 km of the seacoast. From McArthur and Sachs (2001).

APPENDIX TABLE A2—DATA ON MORTALITY

Former colonies	Abbreviated name used in graphs	Log GDP per capita (PPP) in 1995	Average protection against expropriation risk 1985–1995	Main mortality estimate	Former colonies	Abbreviated name used in graphs	Log GDP per capita (PPP) in 1995	Average protection against expropriation risk 1985–1995	Main mortality estimate
Algeria	DZA	8.39	6.50	78.2	Jamaica	JAM	8.19	7.09	130
Angola	AGO	7.77	5.36	280	Kenya	KEN	7.06	6.05	145
Argentina	ARG	9.13	6.39	68.9	Madagascar	MDG	6.84	4.45	536.04
Australia	AUS	9.90	9.32	8.55	Malaysia	MYS	8.89	7.95	17.7
Bahamas	BHS	9.29	7.50	85	Mali	MLI	6.57	4.00	2940
Bangladesh	BGD	6.88	5.14	71.41	Malta	MLT	9.43	7.23	16.3
Bolivia	BOL	7.93	5.64	71	Mexico	MEX	8.94	7.50	71
Brazil	BRA	8.73	7.91	71	Morocco	MAR	8.04	7.09	78.2
Burkina Faso	BFA	6.85	4.45	280	New Zealand	NZL	9.76	9.73	8.55
Cameroon	CMR	7.50	6.45	280	Nicaragua	NIC	7.54	5.23	163.3
Canada	CAN	9.99	9.73	16.1	Niger	NER	6.73	5.00	400
Chile	CHL	9.34	7.82	68.9	Nigeria	NGA	6.81	5.55	2004
Colombia	COL	8.81	7.32	71	Pakistan	PAK	7.35	6.05	36.99
Congo (Brazzaville)	COG	7.42	4.68	240	Panama	PAN	8.84	5.91	163.3
Costa Rica	CRI	8.79	7.05	78.1	Paraguay	PRY	8.21	6.95	78.1
Côte d'Ivoire	CIV	7.44	7.00	668	Peru	PER	8.40	5.77	71
Dominican Republic	DOM	8.36	6.18	130	Senegal	SEN	7.40	6.00	164.66
Ecuador	ECU	8.47	6.55	71	Sierra Leone	SLE	6.25	5.82	483
Egypt	EGY	7.95	6.77	67.8	Singapore	SGP	10.15	9.32	17.7
El Salvador	SLV	7.95	5.00	78.1	South Africa	ZAF	8.89	6.86	15.5
Ethiopia	ETH	6.11	5.73	26	Sri Lanka	LKA	7.73	6.05	69.8
Gabon	GAB	8.90	7.82	280	Sudan	SDN	7.31	4.00	88.2
Gambia	GMB	7.27	8.27	1470	Tanzania	TZA	6.25	6.64	145
Ghana	GHA	7.37	6.27	668	Togo	TGO	7.22	6.91	668
Guatemala	GTM	8.29	5.14	71	Trinidad and Tobago	TTO	8.77	7.45	85
Guinea	GIN	7.49	6.55	483	Tunisia	TUN	8.48	6.45	63
Guyana	GUY	7.90	5.89	32.18	Uganda	UGA	6.97	4.45	280
Haiti	HTI	7.15	3.73	130	Uruguay	URY	9.03	7.00	71
Honduras	HND	7.69	5.32	78.1	USA	USA	10.22	10.00	15
Hong Kong	HKG	10.05	8.14	14.9	Venezuela	VEN	9.07	7.14	78.1
India	IND	7.33	8.27	48.63	Vietnam	VNM	7.28	6.41	140
Indonesia	IDN	8.07	7.59	170	Zaire	ZAR	6.87	3.50	240

REFERENCES

Acemoglu, Daron. "Reward Structures and the Allocation of Talent." *European Economic Review*, January 1995, *39*(1), pp. 17–33.

Acemoglu, Daron; Johnson, Simon and Robinson, James A. "The Colonial Origins of Comparative Development: An Empirical Investigation." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 7771, June 2000.

_____. "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 8460, April 2001.

Acemoglu, Daron and Verdier, Thierry. "Property Rights, Corruption and the Allocation of Talent: A General Equilibrium Approach." *Economic Journal*, September 1998, *108*(450), pp. 1381–403.

Altonji, Joseph; Elder, Todd and Taber, Christopher. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." Mimeo, Northwestern University, 2000.

Anderson, Benedict. *Imagined communities*. London: Verso, 1983.

Bates, Robert H. *Essays on the political economy of rural Africa*. Cambridge: Cambridge University Press, 1983.

Bertocchi, Graziella and Canova, Fabio. "Did Colonization Matter for Growth? An Empirical Exploration into the Historical Causes of Africa's Underdevelopment." Centre for Economic Policy Research

Discussion Paper No. 1444, September 1996.

Besley, Timothy. "Property Rights and Investment Incentives: Theory and Evidence from Ghana." *Journal of Political Economy*, October 1995, *103*(5), pp. 903–37.

Bloom, David E. and Sachs, Jeffrey D. "Geography, Demography, and Economic Growth in Africa." *Brookings Papers on Economic Activity*, 1998, (2), pp. 207–73.

Boone, Catherine. *Merchant capital and the roots of state power in Senegal, 1930–1985.* New York: Cambridge University Press, 1992.

Bound, John; Jaeger, David A. and Baker, Regina M. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association*, June 1995, *90*(430), pp. 443–50.

Bruce-Chwatt, Leonard J. *Essential malariology.* London: Wiley Medical Publications, 1980.

Cain, Philip J. and Hopkins, Anthony G. *British imperialism: Innovation and expansion 1688–1914.* New York: Longman, 1993.

Callaghy, Thomas M. *The state-society struggle: Zaire in comparative perspective.* New York: Columbia University Press, 1984.

Coatsworth, John H. "Obstacles to Economic Growth in 19th-Century Mexico." *American Historical Review*, February 1978, *83*(1), pp. 80–100.

_____. "The Limits of Colonial Absolutism: Mexico in the Eighteenth Century," in Karen Spalding, ed., *Essays in the political, economic and social history of Latin America.* Newark, DE: University of Delaware Press, 1982, pp. 25–51.

_____. "Economic and Institutional Trajectories in 19th Century Latin America," in John H. Coatsworth and Alan M. Taylor, eds., *Latin America and the world economy since 1800.* Cambridge, MA: Harvard University Press, 1999, pp. 23–54.

Coplin, William D.; O'Leary, Michael K. and Sealy, Tom. *A business guide to political risk for international decisions*, 2nd Ed. New York: Political Risk Services, 1991.

Crosby, Alfred. *Ecological imperialism: The biological expansion of Europe 900–1900.* New York: Cambridge University Press, 1986.

Crowder, Michael. *West Africa under colonial rule.* Chicago: Northwestern University Press, 1968.

Curtin, Philip D. *The image of Africa.* Madison, WI: University of Wisconsin Press, 1964.

_____. "Epidemiology and the Slave Trade." *Political Science Quarterly*, June 1968, *83*(2), pp. 181–216.

_____. *Death by migration: Europe's encounter with the tropical world in the 19th Century.* New York: Cambridge University Press, 1989.

_____. *Disease and empire: The health of European troops in the conquest of Africa.* New York: Cambridge University Press, 1998.

Curtin, Philip D.; Feierman, Steven; Thompson, Leonard and Vansina, Jan. *African history: From earliest times to independence*, 2nd Ed. London: Longman, 1995.

Davis, Lance E. and Huttenback, Robert A. *Mammon and the pursuit of empire: The political economy of British imperialism, 1860–1912.* Cambridge: Cambridge University Press, 1987.

Denoon, Donald. *Settler capitalism: The dynamics of dependent development in the southern hemisphere.* Oxford: Clarendon Press, 1983.

Diamond, Jared M. *Guns, germs and steel: The fate of human societies.* New York: W.W. Norton & Co., 1997.

Dunn, Richard S. *Sugar and slaves: The rise of the planter class in the English West Indies 1624–1713.* Chapel Hill, NC: University of North Carolina Press, 1972.

Easterly, William and Levine, Ross. "Africa's Growth Tragedy: Policies and Ethnic Divisions." *Quarterly Journal of Economics*, November 1997, *112*(4), pp. 1203–50.

Eltis, David. *The rise of African slavery in the Americas.* Cambridge: Cambridge University Press, 2000.

Engerman, Stanley L.; Mariscal, Elisa and Sokoloff, Kenneth L. "Schooling, Suffrage, and the Persistence of Inequality in the Americas, 1800–1945." Unpublished manuscript, UCLA, 1998.

Engerman, Stanley L. and Sokoloff, Kenneth L. "Factor Endowments, Institutions, and Differential Paths of Growth among New World

Economies," in Stephen Haber, ed., *How Latin America fell behind.* Stanford, CA: Stanford University Press, 1997, pp. 260–304.

Galenson, David W. "The Settlement and Growth of the Colonies: Population, Labor and Economic Development," in Stanley L. Engerman and Robert E. Gallman, eds., *The Cambridge economic history of the United States, volume I, the colonial era.* New York: Cambridge University Press, 1996, pp. 135–208.

Gallup, John L.; Mellinger, Andrew D. and Sachs, Jeffrey D. "Geography and Economic Development." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6849, 1998.

Gallup, John L. and Sachs, Jeffrey D. "The Economic Burden of Malaria." Unpublished manuscript, Harvard Center for International Development, October 1998.

Gann, Lewis H. and Duignan, Peter. *White settlers in tropical Africa.* Baltimore, MD: Penguin, 1962.

_____. *The rulers of Belgian Africa.* Princeton, NJ: Princeton University Press, 1979.

Gilles, Herbert M. and Warrell, David A. *Bruce-Chwatt's essential malariology*, 3rd Ed. London: Arnold, 1993.

Grier, Robin M. "Colonial Legacies and Economic Growth." *Public Choice*, March 1999, *98*(3–4), pp. 317–35.

Gurr, Ted Robert. "Polity II: Political Structures and Regime Change, 1800–1986." Unpublished manuscript, University of Colorado, Boulder, 1997.

Gutierrez, Hector. "La Mortalite des Eveques Latino-Americains aux XVIIe et XVIII Siecles." *Annales de Demographie Historique*, 1986, pp. 29–39.

Hall, Robert E. and Jones, Charles I. "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" *Quarterly Journal of Economics*, February 1999, *114*(1), pp. 83–116.

Hughes, Robert. *The fatal shore.* London: Collins Harvill, 1987.

Jewsiewicki, Bogumil. "Rural Society and the Belgian Colonial Economy," in D. Birmingham and P. M. Martin, eds., *The history of Central Africa, volume II.* New York: Longman, 1983, pp. 95–125.

Johnson, Simon; McMillan, John and Woodruff, Christopher. "Property Rights and Finance." Unpublished working paper, Massachusetts Institute of Technology and University of California, San Diego, 1999.

Jones, Eric L. *The European miracle: Environments, economies and geopolitics in the history of Europe and Asia.* Cambridge: Cambridge University Press, 1981.

Knack, Stephen and Keefer, Philip. "Institutions and Economic Performance: Cross-Country Tests Using Alternative Measures." *Economics and Politics*, November 1995, *7*(3), pp. 207–27.

Landes, David S. *The wealth and poverty of nations: Why some are so rich and some so poor.* New York: W.W. Norton & Co., 1998.

Lang, James. *Conquest and commerce: Spain and England in the Americas.* New York: Academic Press, 1975.

La Porta, Rafael; Lopez-de-Silanes, Florencio; Shleifer, Andrei and Vishny, Robert W. "Law and Finance." *Journal of Political Economy*, December 1998, *106*(6), pp. 1113–55.

_____. "The Quality of Government." *Journal of Law, Economics, and Organization*, April 1999, *15*(1), pp. 222–79.

Lipset, Seymour M. "The Social Requisites of Democracy Revisited." *American Sociological Review*, February 1994, *59*(1), pp. 1–22.

Lockhart, James and Schwartz, Stuart B. *Early Latin America.* New York: Cambridge University Press, 1983.

Lynch, John. *The Spanish American revolutions, 1808–1826.* New York: W.W. Norton & Co., 1986.

Manning, Patrick. *Slavery, colonialism, and economic growth in Dahomey, 1640–1980.* New York: Cambridge University Press, 1982.

_____. *Francophone Sub-Saharan Africa, 1880–1995.* New York: Cambridge University Press, 1988.

Mauro, Paulo. "Corruption and Growth." *Quarterly Journal of Economics*, August 1995, *110*(3), pp. 681–712.

Mazingo, Christopher. "Effects of Property Rights on Economic Activity: Lessons from the Stolypin Land Reform." Unpublished manuscript, Massachusetts Institute of Technology, 1999.

McArthur, John W. and Sachs, Jeffrey D. "Institu-

tions and Geography: Comment on Acemoglu, Johnson and Robinson (2000)." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 8114, February 2001.

McCullough, David. *The path between the seas: The creation of the Panama Canal 1870–1914.* New York: Simon and Schuster, 1977.

McEvedy, Colin and Jones, Richard. *Atlas of world population history.* New York: Facts on File, 1975.

McNeill, William H. *Plagues and peoples.* New York: Anchor Books, 1976.

Montesquieu, Charles de Secondat. *The spirit of the laws.* New York: Cambridge University Press, [1748] 1989.

Najera, Jose A. and Hempel, Joahim. "The Burden of Malaria." 1996, downloaded from the World Health Organization's Roll Back Malaria website, http://mosquito.who.int/docs/b.

North, Douglass C. *Structure and change in economic history.* New York: W.W. Norton & Co., 1981.

North, Douglass C.; Summerhill, William and Weingast, Barry. "Order, Disorder and Economic Change: Latin America vs. North America." Unpublished manuscript, Hoover Institution, Stanford University, 1998.

North, Douglass C. and Thomas, Robert P. *The rise of the western world: A new economic history.* Cambridge: Cambridge University Press, 1973.

Oldstone, Michael B. A. *Viruses, plagues, and history.* New York: Oxford University Press, 1998.

Parker, Philip M. *National cultures of the world: A statistical reference*, Cross-Cultural Statistical Encyclopedia of the World, Vol. 4. Westport, CT: Greenwood Press, 1997.

Peemans, Jean-Philippe. "Capital Accumulation in the Congo under Colonialism: The Role of the State," in Lewis H. Gann and Peter Duignan, eds., *Colonialism in Africa 1870–1960, volume 4, the economics of colonialism.* Stanford,CA: Hoover Institution Press, 1975, 165–212.

Porter, Roy, ed. *The Cambridge illustrated history of medicine.* Cambridge: Cambridge University Press, 1996.

Reno, William. *Corruption and state politics in Sierra Leone.* New York: Cambridge University Press, 1995.

Roberts, Andrew. *A history of Zambia.* London: Heinemann, 1976.

Robinson, Ronald E. and Gallagher, John. *Africa and the Victorians: The official mind of imperialism.* London: MacMillan, 1961.

Rodrik, Dani. "Where Did All the Growth Go?" *Journal of Economic Growth*, December 1999, *4*(4), pp. 385–412.

Turner, Thomas and Young, Crawford. *The rise and decline of the Zairian state.* Madison, WI: University of Wisconsin Press, 1985.

von Hayek, Frederich A. *The constitution of liberty.* Chicago: University of Chicago Press, 1960.

Weber, Eugen J. *Peasants into Frenchmen.* Stanford, CA: Stanford University Press, 1976.

Wittfogel, Karl A. *Oriental despotism: A comparative study of total power.* New Haven, CT: Yale University Press, 1957.

Young, Crawford. *The African colonial state in comparative perspective.* New Haven, CT: Yale University Press, 1994.

ELSEVIER

Journal of Development Economics 69 (2002) 205–226

JOURNAL OF Development ECONOMICS

www.elsevier.com/locate/econbase

A simple model of inequality, occupational choice, and development

Maitreesh Ghatak[a,*], Neville Nien-Huei Jiang[b]

[a] *Department of Economics, University of Chicago, Chicago, IL 60637, USA*
[b] *Department of Economics, Vanderbilt University, Nashville, TN 37235, USA*

Received 1 February 2000; accepted 1 September 2001

Abstract

We analyze a simple and tractable model of occupational choice in the presence of credit market imperfections. We examine the effect of parameters governing technology and transaction costs, and history, in terms of the initial wealth distribution, in determining the long-term wealth distribution and the level of per capita income of an economy.

JEL classification: D31; D82; O10
Keywords: Wealth inequality; Occupational choice; Poverty traps

1. Introduction

A well-known implication of neoclassical growth theory is that economies that have similar preferences and technologies converge to the same steady state per capita income.[1] In contrast, in development economics, we frequently encounter the idea of poverty traps: poor individuals and economies tend to remain poor because they start poor. One specific mechanism leading to the persistence of poverty that has recently received a lot of attention operates through borrowing constraints.[2] Because threats of punishment work less well against the poor, they face greater borrowing constraints. This in turn prevents

* Corresponding author.
E-mail address: m-ghatak@uchicago.edu (M. Ghatak).

[1] See Barro and Sala-i-Martin (1995).

[2] See Galor and Zeira (1993), Banerjee and Newman (1993, 1994), Aghion and Bolton (1997), Piketty (1997), and Mookherjee and Ray (2000).

0304-3878/02/$ - see front matter
PII: S0304-3878(02)00059-7

them from adopting efficient technologies or choosing profitable occupations, and hence they remain poor. At the aggregate level, this implies that unlike in neoclassical growth models, two economies that are identical in terms of all parameters may end up with different levels of per capita incomes in the steady state if initially they have different distributions of wealth and hence different sizes of the class of credit rationed. This argument is often invoked to explain the evidence from cross-country analysis suggesting that various measures of initial inequality are negatively correlated with growth.[3]

However, it turns out that the dynamic behavior of an economy in the presence of credit market imperfections is fairly complicated, and even under strong simplifying assumptions regarding technology, preferences and market structure, it is difficult to give clear-cut answers to questions such as when do initial conditions matter, and if they do, what is the relationship between initial inequality and the steady-state level of per capita income of an economy. In this paper we try to answer these questions by analyzing a simple and tractable dynamic model of occupational choice in the presence of credit market imperfections.

Our paper is closely related to the important contributions of Galor and Zeira (1993) and Banerjee and Newman (1993). They provide the following insight: in the presence of credit market imperfections, the current distribution of wealth will determine the proportion of credit-constrained individuals in the economy, which in turn may affect equilibrium returns to various occupations in a way that affects the future wealth distribution through intergenerational transfers. As a result, the transition of the wealth distribution for the economy as a whole is nonlinear and hence the wealth distribution dynamics is quite complex. In particular, it is difficult to say much except for multiple stationary wealth distributions may exist, and that the initial distribution of wealth may determine which steady-state equilibrium the economy converges to. Banerjee and Newman (1993) offer some simple examples to show instances of hysteresis. However, even in these examples, it is not always the case that the greater is the size of the poor relative to that of the rich in the initial distribution, the lower will be the steady-state level of income.

We consider a simplified version of the model of Banerjee and Newman (1993). In particular, we have a simpler occupational structure. It turns out, as a result of this one needs no more information about the wealth distribution than the proportion of people whose wealth is below the level needed to start an enterprise. Even though general results in this class of nonlinear dynamic models of wealth distribution are hard to obtain as demonstrated by the Banerjee–Newman model, this simplification allows us to characterize precisely all the steady-state equilibria corresponding to various configurations of parameters governing technology, preferences and transactions costs. It also allows us to calculate the effect of changes in parameters of interest and the initial distribution of wealth on steady state per capita income. However, as a result of this simplification, we lose some of the richness of the Banerjee–Newman model, which allows for alternative institutional forms associated with the modern technology that differ in terms of agency costs.

[3] See Benabou (1996) for a discussion of the empirical literature as well as other theoretical arguments consistent with the observed negative relationship between inequality and growth such as those based on political economy considerations.

Some of our findings are as follows: first, whether hysteresis occurs depends on the size of the threshold level of wealth needed to start an enterprise relative to the productivity of the modern and the subsistence technologies. In particular, the larger is the productivity difference between the modern and subsistence technologies, the greater is the likelihood of multiple steady states. Second, for parameter values under which initial conditions matter, the greater is the fraction of the population who are initially poor, the lower is the steady-state income. Third, while some forms of technological progress can eliminate poverty traps, all kinds of technological improvements do not necessarily increase steady-state income. For example, an increase in the productivity of the small scale or subsistence sector that pushes up wages can act as a drag on the growth of the modern sector.

The plan of the paper is as follows. In Section 2 we analyze the basic model. In Section 3 we extend the basic model, which is nonstochastic, by allowing the saving rate to be subject to random shocks. In Section 4 we make some concluding remarks and Appendix A contains some technical proofs.

2. The model

2.1. Demographics and preferences

Consider an economy inhabited by infinitely lived dynasties represented by successive generations of agents who live for one period. The population is large and its size is normalized to 1. There is no population growth. There are two goods in the economy, labor, and some final output which can serve both as a consumption good and a capital good. In period t a dynasty i is endowed with 1 unit of labor and an initial wealth $a_{i,t}$. It earns income by supplying labor and capital and the resulting income $y_{i,t}$ is divided at the end of the period between consumption $c_{i,t}$ and savings, or bequest to the next generation, $b_{i,t}$. Therefore,

$$a_{i,t+1} = b_{i,t}.$$

Following the literature, we assume that individuals have identical Cobb–Douglas utility functions over consumption and bequests, with $U^i(c_{i,t}, b_{i,t})=c_{i,t}^{1-s}b_{i,t}^{s}$, where $s\in(0, 1)$ and the budget constraint is $y_{i,t}=c_{i,t}+b_{i,t}$. This means that the current generation saves a constant fraction s of its income and leaves it as bequest:

$$a_{i,t+1} = sy_{i,t}.$$

We also assume that all agents are risk-neutral.

In period t, wealth is distributed according to the probability measure $\lambda_t(\cdot)$, and for convenience, we define

$$G_t(a)\equiv\lambda_t((-\infty, a)).$$

The function G_t is very similar to the distribution function except that it does not include the measure at point a.

2.2. Production technologies

There are two production technologies both of which are deterministic. One uses no capital and one unit of labor to produce $\underline{w}$ units of output. This will be described as a subsistence (or agricultural) technology. The other uses $I>0$ units of capital and two units of labor (one unit of supervisory labor and one unit of ordinary labor) to produce q units of output. One supervisor (or entrepreneur) can perfectly monitor one worker spending her entire labor endowment. This will be described as an entrepreneurial (or industrial) technology.[4]

Assumption 1. We assume that this technology is superior in the sense that the net output of using this technology is greater than were two units of labor using the subsistence technology. That is,

$$q - rI > 2\underline{w}$$

where r (≥ 1) is the exogenously given gross interest rate.[5]

2.3. Occupations

There are three possible occupations open to an individual who has inherited wealth $a_{i,t}$:

(a) Subsistence: The agent earns some income by using her labor endowment to produce $\underline{w}$ with the subsistence technology. She puts her inherited wealth in the bank, which yields $ra_{i,t}$. Therefore, her income is

$$y_{i,t}^{S} = \underline{w} + ra_{i,t}.$$

(b) Worker: The agent works for an entrepreneur for wage income w_t (which is determined endogenously). She puts her inherited wealth in the bank, which yields $ra_{i,t}$. Therefore, her income is

$$y_{i,t}^{W} = w_t + ra_{i,t}.$$

(c) Entrepreneur: The agent invests an amount I to start a firm and hires one worker to produce an output q with certainty. Her job is to monitor the worker. The agent's income as an entrepreneur is the output of the project less wage and capital costs:

$$y_{i,t}^{E} = q - w_t + r(a_{i,t} - I).$$

[4] In contrast in the Banerjee and Newman (1993) model, apart from these two types of technologies, there is a third one which involves some capital and one unit of labor ("self-employment").

[5] We can think of the credit market as an international market where the given economy is 'small'.

2.4. Credit and labor markets

The credit market is subject to transactions costs on the lending side due to imperfect enforcement of loan contracts.[6] This results in credit rationing of the following form: if an individual's wealth is below a certain minimum level, she would not get a loan no matter how high the interest rate she offers. Following Banerjee and Newman (1993), a simple way to generate this form of credit rationing is as follows: a borrower may default on her loan (namely, $r(I-a)$), but the cost of this action is that she gets caught with some probability π and then has to pay a fixed *nonmonetary* cost of F due to imprisonment or social sanctions. Thus, only those individuals get loans whose wealth satisfies the incentive compatibility constraint (ICC)[7]:

$$(q - w_t) - r(I - a_{i,t}) \geq q - w_t - \pi F$$

$$\text{or, } a_{i,t} \geq I - \frac{\pi F}{r}. \tag{1}$$

The lower is an individual's wealth, the greater is her incentive to default because she has to borrow a greater amount to start an enterprise, and the level of sanctions against default is the same for all borrowers. Hence, only those who have a certain minimum amount of wealth (namely, $I-\pi F/r$) can borrow.[8] Without loss of generality, we set $\pi=0$ so that only those who have enough wealth to fully finance their own enterprises are able to become entrepreneurs.

The wage rate at which entrepreneurs are indifferent between working as wage laborers and hiring workers is given by:

$$q - \bar{w} + r(a_{i,t} - I) = \bar{w} + r a_{i,t}$$

$$\text{or, } \bar{w} = \frac{q - rI}{2}.$$

By Assumption 1, $\underline{w}<\bar{w}$. Below we show that to ensure labor market equilibrium, the wage rate w must lie in the interval $[\underline{w},\bar{w}]$. Hence, the occupation of entrepreneurship earns no less than any other occupation for all wages (and strictly so for all $w<\bar{w}$). Given the features of the credit market, only those individuals who own enough capital ($a \geq I$) can become entrepreneurs even though everybody else would like to do so. We are going to

[6] We are assuming there are no imperfections on the deposit side of the credit market: there is a constant rate of return of r irrespective of the amount deposited.

[7] It is being assumed that even if a borrower gets caught trying to avoid repaying her debt, she gets to consume her profits.

[8] An implication of this form of credit rationing is that the threshold wealth level does not depend on the wage rate. Otherwise, the threshold wealth level will change with the wage rate. This tends to complicate the dynamics somewhat, but the basic results are not affected.

refer to those individuals whose wealth is less than I as capital-constrained, or simply, poor, and the rest as unconstrained, or rich.

The ICC tells us what fraction of the population is capital-constrained, namely, $G_t(I)$. Notice that this follows from our assumption that all entrepreneurs are self-financed and the credit market does not operate as π=0. Otherwise, the relevant fraction of the population that is capital-constrained would be $G_t(I-\pi F/r)$.

For $w_t<\underline{w}$, labor supply is zero, but for $w_t=\underline{w}$ labor supply jumps to $G_t(I)$ and as w_t goes above $\underline{w}$, the supply of labor grows until the wage rate is high enough, namely, $\bar{w}$, such that entrepreneurs are indifferent between working as wage laborers and hiring workers. Now we are ready to write down the supply curve of labor:

$$0 \text{ if } w_t < \underline{w}$$

$$[0, G_t(I)] \text{ if } w_t = \underline{w}$$

$$G_t(I) \text{ if } w_t \in (\underline{w}, \bar{w})$$

$$[G_t(I), 1] \text{ if } w_t = \bar{w}$$

$$1 \text{ if } w_t > \bar{w}.$$

Conversely, to derive the demand curve for labor, we notice that for $w_t>\bar{w}$ there is no demand for labor; as w_t falls to $\bar{w}$, the demand for labor jumps to any value between 0 and $1-G_t(I)$. When $w_t<\bar{w}$, the demand for labor is at a maximum, $1-G_t(I)$ and continues to remain so. Therefore, the demand for labor is:

$$0 \text{ if } w_t > \bar{w}$$

$$[0, 1 - G_t(I)] \text{ if } w_t = \bar{w}$$

$$1 - G_t(I) \text{ if } w_t < \bar{w}.$$

From the labor demand and supply schedules we can easily find the equilibrium wage rate in period t:

$$w_t^* = \begin{cases} \bar{w} \text{ if } G_t(I) < \dfrac{1}{2} \\ [\underline{w}, \bar{w}] \text{ if } G_t(I) = \dfrac{1}{2} \\ \underline{w} \text{ if } G_t(I) > \dfrac{1}{2}. \end{cases}$$

Since each entrepreneur hires exactly one worker, if there are more people who are capital-constrained (unconstrained), then the competition for entrepreneurs (workers)

among them will drive the equilibrium wage rate down (up) to its lower (upper) bound. When $G_t(I)=1/2$, the equilibrium wage rate is indeterminate, and throughout this paper, we are going to assume that the wage rate is equal to $\bar{w}$ in this case.

Notice that on one hand, the equilibrium wage rate depends on the current wealth distribution but on the other hand, it also influences next period's wealth distribution through the savings behavior of currently active agents.

2.5. Dynamics of individual wealth

Consider the factors governing dynasty i's bequest. First of all, the initial wealth level of an agent determines her capital income and her occupational choice. Secondly, the current wage rate is determined by the economy-wide wealth distribution. With the knowledge of an individual's occupational choice and that the wage rate can take only two values ($\underline{w}$ and $\bar{w}$), we can write down the difference equations describing the evolution of a dynasty i's wealth as:

$$\begin{aligned} a_{i,t+1}(a_{i,t} \mid w_t = \underline{w}) &= s[ra_{i,t} + \underline{w}] && \text{if } a_{i,t} < I \\ &= s[r(a_{i,t} - I) + q - \underline{w}] && \text{if } a_{i,t} \geq I \\ a_{i,t+1}(a_{i,t} \mid w_t = \bar{w}) &= s[ra_{i,t} + \bar{w}] && \forall a_{i,t}. \end{aligned}$$

Fig. 1 shows what these difference equations look like. Notice that there are two regimes of wealth transitions corresponding to the two wage levels. When the wage rate is low, an agent who is capital-constrained can only choose between being a worker and engaging in subsistence and in either case, her labor income is $\underline{w}$. A fraction s of the sum of her labor income and her capital income $ra_{i,t}$ is left for her next generation. An agent who is not credit-constrained will strictly prefer to be an entrepreneur and her total income will be $r(a_{i,t}-I)+q-\underline{w}$. When the wage rate is high, nobody will engage in subsistence and all agents will be indifferent between being entrepreneurs and workers.

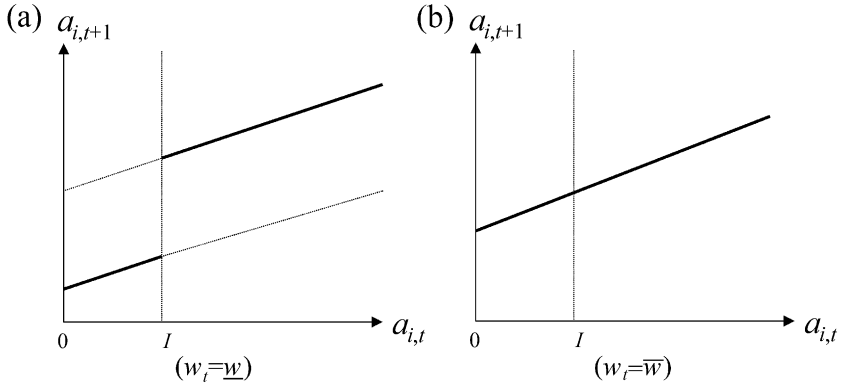


Fig. 1. Dynasty i's wealth transitions under different wage regimes.

Assumption 2. We assume that it is not possible for a dynasty to get arbitrarily rich over time merely by saving a constant fraction of its income every period and earning interest on it:

$$sr < 1$$

Assumptions 1 and 2 will be retained throughout this section.

2.6. Stationary wealth distributions and wages

In this section, we examine the long-run behavior of this economy. If the difference equations governing the wealth transitions are stable, it would be easy to prove the existence of a stationary wealth distribution. However, the fact that these difference equations depend on the wage levels raises the possibility that the process may not be stable. In particular, the concern here is that the wage rate may change infinitely often. The following lemma rules out this possibility.

Lemma 1. *The wage rate can change at most once.*

Proof: Notice that the difference equations are order-preserving. That is, $a_{i,t+1}>a_{j,t+1}$ if and only if $a_{i,t}>a_{j,t}$. Therefore, in order to study the wage dynamics, we can only look at the wealth dynamics of the dynasty which has the median wealth. Define $a_t^m \equiv \max\{a: G_t(a)\leq 1/2\}$. Note that a_t^m is well defined because $G(\cdot)$ is continuous from below according to our definition. Then $a_t^m \geq I \Leftrightarrow G_t(I)\leq \frac{1}{2}$ which implies $w_t=\bar{w}$. Similarly, $a_t^m < I \Leftrightarrow G_t(I) > \frac{1}{2}$ which implies $w_t=\underline{w}$. Now if $w_t=\underline{w}$ and $w_{t+1}=\bar{w}$, then we must have $a_t^m<I$ and $a_{t+1}^m \geq I$. This implies $s(ra_t^m + \underline{w})\geq I \Rightarrow s\underline{w}\geq(1-sr)I \Rightarrow s(ra+w)\geq I$ for all $a\geq I$ and $w\in\{\underline{w},\bar{w}\}$. That is, once the high-wage rate is reached, there will not be any downward mobility and hence the high wage will prevail forever. If $w_t=\bar{w}$ and $w_{t+1}=\underline{w}$, then we must have $a_t^m\geq I$ and $a_{t+1}^m< I$. This implies $s(ra_t^m + \bar{w}) < I \Rightarrow s\bar{w} < (1-sr)I \Rightarrow s(ra+w) < I$ for all $a<I$ and $w\in\{\underline{w},\bar{w}\}$. That is, there will not be any upward mobility and once the low-wage rate is reached, it will prevail forever. Therefore, we can conclude that starting with any initial distribution of wealth, the wage rate can change at most once. □

Lemma 1 shows that the wage rate is constant in the long run and rules out the possibility of cycles or chaotic wage dynamics. Once the wage rate switches from low to high, there will be no downward mobility and so the high wage prevails forever and similarly, once the wage rate switches from high to low, there will be no upward mobility and the low-wage prevails forever. As a result, although we have two regimes of the wealth transition process, there will not be infinite switches from one to the other. Only one of them will prevail in the long run. However, for the same parameter values, both wealth transition processes could be candidates for the long-run equilibrium and which one is arrived at could depend on initial conditions. Together with Assumption 2, which implies there exists a stationary point for each difference equation, we immediately have:

Proposition 1: *Given any initial wealth distribution, there exists a unique stationary wealth distribution to which it converges.*

By Lemma 1 in the long run the wage rate is constant and corresponding to this wage rate, one of the two possible wealth transition processes will prevail. The difference equations associated with these processes have unique stationary points and so the wealth

distribution of the economy will converge to a stationary distribution. This stationary wealth distribution will have all mass concentrated on one point (for the high-wage equilibrium) or two points (for the low-wage equilibrium) which is a consequence of the model being nonstochastic. Notice that the Lemma 1 and Proposition 1 do not suggest that given the parameters of the model there is a unique long-run wage rate, and a corresponding long-run stationary wealth distribution. Indeed, one of our main goals is to characterize parameter conditions under which multiple long-run equilibria could exist and to show which equilibrium the economy converges to depends on initial conditions. What these results do is to rule out cycles or chaotic behavior. Now we proceed to characterize how the long-run equilibrium of the economy depends on various parameters and the initial wealth distribution.

Let $a^J(w)$ be the stationary point of the difference equation describing the wealth transition of a dynasty engaged in occupation J (where J=S,W,E denotes the three occupations: subsistence, worker, and entrepreneur) when the wage rate is w. Then we have

$$a^{\mathrm{S}}(w) = \frac{s\underline{w}}{1 - sr} \text{ for all } w.$$

$$a^{\mathrm{W}}(\underline{w}) = \frac{s\underline{w}}{1 - sr}$$

$$a^{\mathrm{E}}(\underline{w}) = \frac{s(q - rI - \underline{w})}{1 - sr}$$

$$a^{\mathrm{W}}(\bar{w}) = a^{\mathrm{E}}(\bar{w}) = \frac{s(q - rI)}{2(1 - sr)}.$$

By Assumption 1, $a^{\mathrm{E}}(\underline{w})>a^{\mathrm{E}}(\bar{w})=a^{\mathrm{W}}(\bar{w})>a^{\mathrm{W}}(\underline{w})$.

Comparing the values of these threshold levels of wealth with I, we can completely characterize the long-run outcome (in terms of the stationary distribution of wealth, the equilibrium wage rate and the level of net output) of the economy.

Proposition 2: *The initial distribution of wealth matters in determining the stationary distribution of wealth and the long run equilibrium wage rate if and only if*

$$s(q - \underline{w}) \geq I > \frac{s\underline{w}}{1 - sr}.$$

Otherwise the economy converges to a high-wage equilibrium (if $I \leq s\underline{w}/(1-sr)$) or a subsistence equilibrium (if $I > s(q-\underline{w})$) irrespective of initial conditions.

Proof: The proof consists of the following two steps

Step 1. The following four cases characterize the steady-state equilibrium of the economy corresponding to various parameter values:

Case 1. $I > s(q - \underline{w}) \Leftrightarrow I > a^{\mathrm{E}}(\underline{w})$. This is a situation where the steady-state wealth of the entrepreneurial class cannot finance the operation of the industrial technology even

when wages are as low as possible. The only equilibrium in this economy is therefore one where everyone is engaged in subsistence production irrespective of the initial wealth distribution G_0. As a result the stationary wealth distribution displays no inequality.

Case 2. $s(q-\underline{w})\geq I>sq/(2-sr) \Leftrightarrow a^{E}(\underline{w})\geq I>a^{E}(\bar{w})=a^{W}(\bar{w})>a^{W}(\underline{w})$. The condition that $a^{E}(\underline{w})\geq I$ implies $s[r(a-I)+q-\underline{w}]\geq I\ \forall a\geq I$. It says when the wage rate is low, offspring of individuals who are able to start an enterprise in the current period will also be able to do so in the next period, i.e., there is no downward mobility. Similarly, $I>a^{W}(\underline{w})$ implies $s(ra+\underline{w})<I\ \forall a<I$, which means there is no upward mobility when the wage is low. If the economy starts out with the low-wage rate ($G_0(I)>1/2$), there will not be any mobility in either direction. This implies that the wage rate will always be equal to $\underline{w}$; the wealth of those dynasties that are initially capital-constrained will converge to $a^{W}(\underline{w})$; the wealth of those that are not will converge to $a^{E}(\underline{w})$; and there will be $1-G_0(I)$ firms operating in each period. Now suppose the economy starts out with the high-wage rate ($G_0(I)\leq 1/2$). The condition, $I>a^{E}(\bar{w})=a^{W}(\bar{w})$, implies $\bar{w}$ is not sustainable. There exists a finite τ such that $w_\tau=\bar{w}$ and $w_{\tau+1}=\underline{w}$. Thereafter the story is the same as above if we take $G_{\tau+1}(\cdot)$ as the initial wealth distribution in the new low-wage regime. And of course, $G_{\tau+1}$ depends on G_0.

Case 3. $sq/(2-sr)\geq I>s\underline{w}/(1-sr) \Leftrightarrow a^{E}(\underline{w})>a^{E}(\bar{w})=a^{W}(\bar{w})\geq I>a^{W}(\underline{w})$. Again, since $a^{E}(\underline{w})>I>a^{W}(\underline{w})$, there is no upward or downward mobility when wage rate is low. Therefore, if the economy starts out at low-wage rate ($G_0(I)>1/2$), the story is the same as in Case 2. However, the condition, $a^{E}(\bar{w})=a^{W}(\bar{w})\geq I$, implies $s(ra+((q-rI)/2))\geq I\ \forall a\geq I$. Hence, when the wage rate is high, people who are not capital-constrained will remain unconstrained, i.e., there is no downward mobility. Therefore, if the economy starts out with $G_0(I)\leq 1/2$, the high wage $\bar{w}$ will last forever. As a result, every dynasty's wealth will converge to $a^{E}(\bar{w})$.

Case 4. $s\underline{w}/(1-sr)\geq I \Leftrightarrow a^{W}(\underline{w})\geq I$. The high-wage equilibrium will result irrespective of G_0 because even when wages are low, the steady-state wealth level of the working class permits them to start a firm. As a result, the unique stationary wealth distribution displays no inequality.

Step 2. Next we show that the sets of parameter values that correspond to the four cases analyzed above are mutually exclusive and exhaustive with respect to the set of all admissible parameter values (i.e., those satisfying Assumptions 1 and 2).

Suppose $s\underline{w}/(1-sr)\leq I$. This inequality implies $[(2-sr)/(1-sr)]\underline{w}\leq 2\underline{w}+Ir$. As a result, Assumption 1, which guarantees $q>2\underline{w}+Ir$, also implies $q>[(2-sr)/(1-sr)]\underline{w}$, i.e., $q/(2-sr)>\underline{w}/(1-sr)$. The last inequality in turn implies, upon rearranging, $s(q-\underline{w})>sq/(2-sr)$ and $sq/(2-sr)>s\underline{w}/(1-sr)$. Thus, we have the following inequality which is derived from Assumptions 1 and 2:

$$s(q-\underline{w})>\frac{sq}{2-sr}>\frac{s\underline{w}}{1-sr}$$

which holds so long as $s\underline{w}/(1-sr)\leq I$. If instead, $I<s\underline{w}/(1-sr)$, then Case 4 always applies. That is, the only possible equilibrium is the high-wage equilibrium. □

Fig. 2 summarizes the four cases. Proposition 2 has several interesting economic implications which we discuss below.

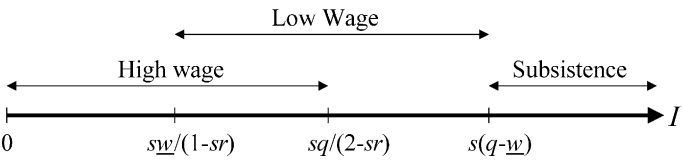


Fig. 2. Long-run wage rates under different parameter configurations (Non-Stochastic Model).

If there were no frictions in the credit market, so long as the modern technology is more productive than the subsistence technology (which is ensured by Assumption 1), it will be used by the entire economy. The initial distribution of wealth, the productivity of the subsistence technology or the propensity to save would not be relevant in determining total output. If credit markets are imperfect, Proposition 2 shows that the long-run equilibrium of the economy cannot be predicted by a simple comparison of the productivity of the two technologies. If the size of the wealth threshold needed to start an enterprise (I) is very high, then the economy will collapse to subsistence sector since the steady-state wealth level of even a rich dynasty in a low-wage equilibrium will fall short of it. Conversely, if I is very low, then the steady-state wealth level of even a poor dynasty in a low-wage equilibrium will exceed it. In this case, in the long run the economy will converge to a high-wage equilibrium where the whole population is engaged in the modern sector. For intermediate values of I, the long-run equilibrium of the economy cannot be predicted from the parameters governing technology and preferences only. The initial wealth distribution also matters. If the parameters are such that the low-wage equilibrium is the unique long-run equilibrium (i.e., this is the case where the steady-state wealth level of a dynasty under the high-wage equilibrium is less than I), the number of firms using the modern technology in the long-run equilibrium is the same as those at t=0 and this is how the initial wealth distribution matters. More interestingly, if the parameters are such that both the low and high-wage equilibrium are possible, then the initial distribution of wealth also determines which equilibrium will be chosen. If initially there are many dynasties who have wealth higher than I, then the high-wage equilibrium will result, and this will enable others to accumulate enough wealth so that in the long run everyone can become an entrepreneur. If on the other hand, if initially the credit-constrained dynasties are in a majority, they will push the wage down in the labor market which will continue to keep them poor in successive generations.

Proposition 2 also suggests that the effect of changes in parameter values regarding technology and preferences may depend on the initial wealth distribution, and in particular, can push the economy from one type of steady-state equilibrium to another. Let us consider the effects of changes in various parameters of the model.

An increase in the productivity of the modern technology q (as a result of technological change or economic policies, such as liberalizing the economy) will increase the income generated by existing enterprises using the modern technology. The effect of this on per capita income will depend on the initial wealth distribution under the low-wage equilibrium as that determines the number of firms using the modern technology, but not in the high-wage equilibrium. Moreover, if as a result of an increase in q the steady-state wealth level of some individuals are pushed above I, the number of enterprises using the modern technology in a steady-state equilibrium may increase. This will be the case if

initially $[s/(2-sr)]q<I$, and after the change $[s/(2-sr)]q\geq I$ (i.e., starting with a low-wage or a subsistence equilibrium, the high-wage equilibrium becomes feasible) or if initially $s(q-\underline{w})<I$, and after the change $s(q-\underline{w})\geq I$ (i.e., starting with a subsistence equilibrium, the low-wage equilibrium becomes feasible). This is an instance where technological change can eliminate a poverty trap without any redistributive measures.

An increase in the productivity of the subsistence technology $\underline{w}$ (as a result of technological change, or government policies such as minimum wage laws or subsidy to small-scale industry) will increase per capita income by raising the incomes of those engaged in the subsistence sector.[9] However, while an increase in $\underline{w}$ increases the steady-state wealth level of workers and those engaged in subsistence, it reduces the steady-state wealth level of entrepreneurs in a low-wage equilibrium. As a result, the effect of it on steady state income is ambiguous. For example, starting with a situation where $s(q-\underline{w})\geq I$ (so that the low-wage equilibrium exists), an increase in $\underline{w}$ can lead to $s(q-\underline{w})<I$ and, as a result, the economy can converge to a subsistence equilibrium. On the other hand, suppose initially $s\underline{w}/(1-sr)<I$ (i.e., the steady-state wealth level of workers or those engaged in subsistence is less than I) and the economy is in a low-wage equilibrium. If after the change $s\underline{w}/(1-sr)\geq I$, the economy will converge to a high-wage equilibrium instead of a low-wage equilibrium. This exercise suggests that an increase in the productivity of a technology does not necessarily raise steady state per capita income. Indeed, an increase in the productivity of the small-scale or subsistence sector that pushes up wages can act as a drag on the growth of the modern sector by reducing the steady-state income of entrepreneurs.

The effect of an increase in s is straightforward. It does not raise steady-state income directly in this model, but raises the steady-state wealth level of every dynasty. If an increase in s pushes the steady-state wealth level of some individuals above I, the number of enterprises using the modern technology in a steady-state equilibrium will increase.

Previously, we have assumed that the chance of being caught from default is zero (π=0). Therefore, there is actually no credit market in this economy—one needs to own the whole amount of capital required (I) to start up a modern firm. Now suppose $\pi>0$, so that only $(I-(\pi F/r))$ is needed to become an entrepreneur. Other things being equal, since it is easier to reach this threshold, the economy is more likely to end up with a high-wage equilibrium. For the same reason, an increase in the punishment (F) would have the same effect. Changes in the interest rate (r), however, have two opposite effects. A decrease in r would reduce the wealth threshold for borrowing on one hand, but on the other hand, it becomes harder to accumulate one's wealth. This suggests that improving the enforcement technology (i.e., increases in π and F) has an unambiguously positive role in eliminating poverty traps, whereas the effect of lower capital scarcity in the international credit market (i.e., a decrease in r) has an ambiguous effect.

Let us define the total income of the economy, the sum of wage and profit income, as:

$$Y = G(I)w + \{1 - G(I)\}(q - w - Ir)$$

The following result compares the equilibria in terms of total income.

[9] It will also increase the wages of workers engaged in the modern sector. But this will be matched by a decrease in the profits of entrepreneurs and there will be no effect on per capita income.

Proposition 3: *For parameter values for which initial conditions matter, the greater is the fraction of the population who are initially poor, the lower is steady-state income.*

Proof: Under a subsistence equilibrium, total income is $Y=\underline{w}$. In a low-wage equilibrium, total income is $Y=(q-Ir)\{1-G(I)\}-\{1-2G(I)\}\underline{w}$. Finally, in a high-wage equilibrium, total income is $Y=(q-Ir)/2$. Since $q-Ir>\underline{w}$ by Assumption 2, and under a low-wage equilibrium $G(I)\geq 1/2$

$$\frac{q-Ir}{2}\geq(q-Ir)\{1-G(I)\}-\{1-2G(I)\}\underline{w}>\underline{w}.$$

Hence, the total income of the economy under a high-wage equilibrium exceeds that under a low-wage equilibrium, which in turn exceeds that under a subsistence equilibrium. Proposition 2 shows that for the parameter values $s(q-\underline{w})\geq I\geq s\underline{w}/(1-sr)$ (corresponding to Cases 2 and 3), if $G_0(I)>1/2$, then the economy converges to a low-wage equilibrium where only $1-G_0(I)$ firms operate. Hence, the proposition follows. □

What this result shows is that even if the low-wage equilibrium is the unique equilibrium, the gain from having one less credit-constrained person is one more firm that uses the modern technology and generates greater income. When multiple equilibria exist, the long run gains from having a smaller number of people who are credit-constrained are much greater than in the previous case, since this might unleash market forces that push the economy to a high-wage equilibrium where the whole population is engaged in the modern sector.

The above result also shows that to the extent greater equality of the distribution of wealth reduces the fraction of the population who are capital-constrained, both greater equity and greater efficiency (in terms of total income) are achieved. As a result, *one-shot* redistributive policies can raise the total income of the economy permanently for parameter values for which the initial wealth distribution matters for the long-term performance of the economy, as Banerjee and Newman (1993) point out. To see this assume that the policy is implemented after the economy has settled down in a steady-state equilibrium. Suppose the government taxes bequests of rich dynasties and redistributes the revenue (so that the government budget is balanced) to poorer dynasties whose wealth is less than I with the goal of making as many individuals to be able to start their own enterprises as possible. Naturally, this policy will have no effect when the economy is in a high-wage or subsistence equilibrium because everyone has equal wealth to start with. For the case of low-wage equilibrium, it can have an effect. Consider Case 3. The policy moves everyone's wealth closer to the *mean*, whereas whether the wealth of the *median* person is greater than or less than I determines whether there is a high- or a low-wage equilibrium. Starting from a low-wage equilibrium, if the mean is greater than I, then such a redistributive policy will push the economy towards a high-wage equilibrium. Even if the mean is less than I in which case the high-wage equilibrium cannot be achieved, the policy will increase the number of enterprises that are operated and hence raise total income. Similarly, in Case 2 such a policy will increase the number of enterprises that are operated and hence, raise total income.

However, the implication of this exercise is not to support *any* egalitarian redistributive policy to increase total income, rather only those that increase the number of enterprises

operating in the economy. For example, in Case 3, if the mean wealth level is less than I, then a complete redistribution will push the economy to subsistence.

3. Extension: stochastic model with mobility

An important feature of the model in Section 2 is that the incomes of all agents, and the bequests of their progeny are all deterministic. This is unsatisfactory as the long-run wealth distribution has all probability mass concentrated on two points (for a low-wage equilibrium) or one point (the high-wage equilibrium or the subsistence equilibrium). As a result, there is no mobility across classes. In this section, we examine the implications of allowing upward and downward mobility through random shocks.

In particular, we assume that every individual's saving rate is subject to an idiosyncratic i.i.d. shock. In every period, each individual's saving rate could be high ($\bar{s}$) with probability p or low ($\underline{s}$) with probability $1-p$.[10] If $\bar{s}$ ($\underline{s}$) is high (low) enough, we will have upward (downward) mobility which is absent in the stationary distributions discussed in Section 2.[11]

We make the following assumptions about the parameters $\underline{s}$ and $\bar{s}$:

Assumption 3

$$\bar{s} > \frac{I}{\underline{w} + rI} \text{ and } \underline{s} = 0$$

The first part of the equation of Assumption 3 ensures there is upward mobility in this economy. Notice that $I/(\underline{w}+rI)<\bar{s}$ implies there exists an integer m such that

$$m = \min\left\{ n \in N : \ \bar{s}\left[\sum_{i=0}^{n-1} (r\bar{s})^i \underline{w}\right] \geq I \right\}.$$

That is, it takes at most m consecutive periods of good luck for a dynasty—even if it started with no initial wealth and even if wage rates remained low—to become rich.

The second part of the assumption, of course, ensures there is downward mobility in this economy, but more importantly, it greatly simplifies the analysis. By setting $\underline{s}=0$, an individual dynasty's wealth dynamics depends on the history only up to the last time it received a bad shock. Together with the first part, Assumption 3 implies the fraction of the poor, and therefore the current wage rate depends on the wage dynamics only up to the previous m periods. Together with the fact that the wage rate can only be high or low, this implies that the resulting wage dynamics must be either one of the following three types: always high wage, always low wage, or a cycle.

[10] These shocks could be taste shocks or shocks related to the technology of saving (e.g., a negative shock could be interpreted as an individual's savings being stolen or expropriated).

[11] An alternative way to introduce random shocks in the model would be to let production be stochastic (as in Banerjee and Newman, 1993). However, given our assumptions about the production technology and preferences, the contractual form of payment to workers will be indeterminate (for example, wage contracts or profit/output sharing contracts will be equivalent). This is unsatisfactory since the specific contractual form will be crucial in driving the extent of upward and downward mobility in the model.

To be more specific, given date t, we can define function $a_t(\cdot)$: $\{0, 1,\ldots,t\}\rightarrow R_+$ as

$$a_t(0) = 0,$$

$$a_t(n) = \bar{s}\left[\sum_{i=0}^{n-1}(r\bar{s})^i w_{t-i-1}\right] \qquad \text{if } n\in\{1,\ldots,t\}.$$

Therefore, $a_t(n)$ represents the initial wealth level at date t of a dynasty which received exactly n consecutive periods of good luck and was a wage-earner during these n periods. The distribution that $a_t(n)$ describes differs from the real wealth distribution at date t since wealthy people could be earning entrepreneurial profits instead of wages. However, for dynasties that were wage earners, $a_t(n)$ correctly represents their initial wealth levels at date t. Therefore, at date $t(\geq m)$, there will be a probability mass $p^n(1-p)$ at $a_t(n)$, $\forall n$=0, 1,. . .,$l(t)$, where $l(t)$=min$\{n: a_t(n)\geq I\}$. Since there is no poor dynasty whose wealth level is different from $a_t(n)$, $\forall n$=0, 1,. . ., $l(t)-1$,

$$G_t(I) = \sum_{n=0}^{l(t)-1} p^n(1-p).$$

Because $l(t)\leq m$, $\forall t$,$G_t(I)$ depends on at most $\{w_{t-i-1}\}_{i=0}^{m-1}$. Without loss of generality, we can use a function f: $W^m\rightarrow W$, where $W=\{\underline{w}, \bar{w}\}$, to describe the relationship between current wage rate and the wage rates in the previous m periods. Two results follow immediately:

Lemma 2. *The wage dynamics can be either stationary (with high or low wages), or display cycles*.

Proof: See Appendix A.

Lemma 3. *f is weakly increasing in each of its elements*.

Proof: See Appendix A. We divide our discussion for the rest of this section into two cases, constant-wage dynamics and cycles.

3.1. Constant-wage dynamics

In the following proposition, we show that if $\bar{s}$ is not too large, the stationary wage dynamics can only be of two types: always high wage or always low wage.

Proposition 4: *In addition to Assumption 3, if $\bar{s}\leq 1/r$, then the stationary wage dynamics can only be either always high wage or always low wage*.

Proof: See Appendix A.

Which case will emerge depends on how fast the poor become rich when wages are high or low, and in some cases, the initial wealth distribution. This characterization is provided by Proposition 5. If $r\bar{s}\leq 1$, in the expression for the wealth level of a currently poor dynasty, wage rates in the recent past receive greater weight than wage rates in the distant past. From the proof of Proposition 4, if $w_t=\bar{w}$, the wealth distribution of the poor

is going to remain the same or shifts to the right (first-order stochastic dominance), and in either case, $w_{t+1}=\bar{w}$. If $w_t=\underline{w}$, the opposite happens. The wealth distribution at date $t+1$ is the same as the wealth distribution at date t, or is first-order stochastically dominated by it and hence, $w_{t+1}=\underline{w}$.

Next, we ask under what conditions the economy will converge to the high-wage equilibrium. Intuitively, if the chance of receiving a high-saving shock is high (p large) and if it does not take long for a very poor dynasty to become rich (m small), the economy should end up with a high-wage equilibrium. On the other hand, the low-wage equilibrium would emerge if the chance of an individual to be born with no wealth is high (p small) and if it takes many periods to become rich. Between these two extremes, there should be cases where the initial distribution matters. We formally prove this in Lemma 4 and Proposition 5. Let us first define m', similar to m, as the number of periods needed for a zero-wealth dynasty to become rich under *high* wages. In other words,

$$m' = \min\left\{ n\in N : \bar{s}\left[\sum_{i=0}^{n-1}(r\bar{s})^i\bar{w}\right]\geq I\right\}.$$

Naturally, $m'\leq m$.

Lemma 4. *If $p^m\geq 1/2$, then the economy converges to the high-wage equilibrium. If $p^{m'}<1/2$, then the economy converge to the low-wage equilibrium.*

Proof: See Appendix A.

Given Lemma 4, we can prove the following proposition:

Proposition 5: *The initial distribution of wealth matters in determining the stationary distribution of wealth and the long-run equilibrium wage rate in the stochastic model if and only if*

$$p^{m'}\geq\frac{1}{2}>p^m.$$

Otherwise the economy converges to a high-wage equilibrium (if $p^{m'}\geq p^m>1/2$) or a low-wage equilibrium (if $1/2>p^{m'}\geq p^m$) irrespective of initial conditions.

Proof: See Appendix A.

In the case where the initial wealth distribution can matter, it is difficult to give conditions on initial distributions under which the economy would end up with a high- or low-wage equilibrium. Intuitively, if the economy starts out with many rich people, the high-wage rate would be likely to last for many periods. As a result, by the time most of those who were originally rich would be hit by a low savings shock, some of those who were originally poor would accumulate enough wealth. Therefore, the stationary distribution is more likely to be the one associated with the high wage. Conversely, if the economy starts out with many poor individuals, the low-wage rate would last for a long time. Then only a few lucky individuals will be able to accumulate enough wealth

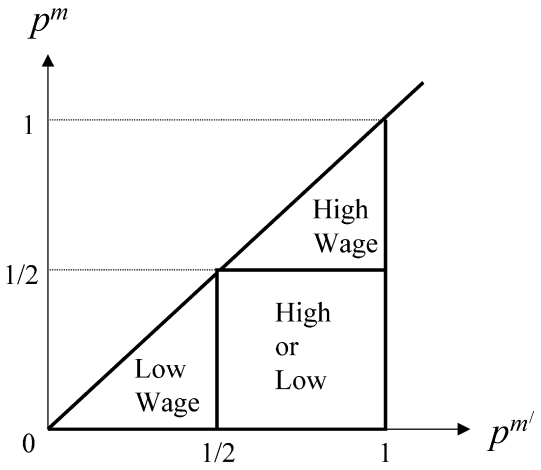


Fig. 3. Long-run wage rates under different parameter configurations (Stochastic Model).

before being hit by a low savings shock. Therefore, the low-wage equilibrium would result.

Proposition 5 provides conditions on parameters under which multiple stationary distributions may exist (also see Fig. 3). Other things being the same, the greater is the difference between the productivity of the modern and the subsistence technology (namely, $\bar{w}$ and $\underline{w}$), the greater will be the difference between m and m', and the more likely this case will occur. Also, this case is more likely with intermediate values of I. The higher (lower) is I the higher (lower) will be both m and m' and for given p the more likely the economy will end up in a low (high) wage equilibrium.

Since the number of firms operating in a low-wage equilibrium (p^m) is less than that under a high-wage equilibrium for parameter values for which initial conditions matter, the greater is the fraction of the population who are initially poor, the more likely the economy will end up in a low-wage equilibrium with a lower level of long run per capita income. This is similar in spirit to Proposition 3 in the nonstochastic model. However, in the low-wage equilibrium of the nonstochastic model, the long-run number of firms depends on the parameters of the model as well as the initial fraction of poor individuals, whereas in the stochastic model it depends only on the parameters.

3.2. Cycles

From Proposition 4 and Lemma 3, if $\bar{s}$ is large enough ($\bar{s}>1/r$) and if $p^{m'} \geq 1/2 > p^m$, the economy does not necessarily converge to a constant-wage equilibrium; the wage dynamics might display cycles.[12] It is difficult to provide general results for this case

[12] It turns out that we do not need $\bar{s} \leq 1/r$ to prove Lemma 4 and Proposition 5. In other words, they are true even when $\bar{s}>1/r$.

and we restrict our discussion around a simple example of a cycle, the simplest one we can find, where the high wage and the low wage alternate with each other.

Example: Suppose $\bar{s}>1/r$, $m=3$, $m'=2$, and $p^2\geq 1/2>p^3$. If $\bar{s}[(r\bar{s})\underline{w}+\bar{w}]<I\leq\bar{s}[(r\bar{s})\bar{w}+\underline{w}]$, then the wage dynamics might display a two-period cycle where high wage and low wage alternate with each other.[13]

Starting with a high-wage period $w_t=\bar{w}$, the wealth distribution must display a probability mass

(i) $(1-p)$ at $a=0$ consisting of those who received a bad saving shock last period;
(ii) $p(1-p)$ at $a=\bar{s}\underline{w}$ consisting of those who received a bad saving shock in the period before last period, but a good shock in the last period (notice that the wage rate in the previous period was low);
(iii) $p^2(1-p)$ at $a=\bar{s}[(r\bar{s})\bar{w}+\underline{w}]$, consisting of those who received a good saving shock in the last two periods (notice that the wage rate in the period before the previous period was high);
(iv) p^3 consisting of those with $a>\bar{s}[(r\bar{s})\bar{w}+\underline{w}]$.

Since $\bar{s}[(r\bar{s})\bar{w}+\underline{w}]\geq I$, and, by assumption $w_t=\bar{w}$ in the current period, the fraction of the poor cannot be more than half. This is indeed the case as $1-p^2\leq 1/2<1-p^3$. In the next period, the wealth distribution has a probability mass $(1-p)$ at 0, $p(1-p)$ at $\bar{s}\bar{w}$, and $p^2(1-p)$ at $\bar{s}[(r\bar{s})\underline{w}+\bar{w}]$, and p^3 consisting of those with $a>\bar{s}[(r\bar{s})\underline{w}+\bar{w}]$. Again, since $1-p^2\leq 1/2<1-p^3$ and $\bar{s}[(r\bar{s})\underline{w}+\bar{w}]<I$ the wage rate drops back to $\underline{w}$. In period $t+2$, the wealth distribution changes back to that of period t which results in a high-wage rate and the same process goes on forever.

Cycles in our model can occur in the special case when the positive savings shock is very high—so high that a dynasty that receives only positive saving shocks will become infinitely rich just by saving, however small the initial amount it started off with. Then current wages will have a large impact on future income. When wages are high the richest among the poor stay poor because past wages (which are low) play a dominant role. Since there is only downward mobility but no upward mobility, the wage rate switches from high to low. But then, even though the wage is low, the richest dynasties among the poor become rich since the high wage they experienced previously is weighted by $(\bar{s}r)^2$. The parameter configuration we assume ensures that there is more upward mobility than downward mobility so that the wage rate becomes high again. This process will go on forever and the economy will display cycles. Aghion et al. (1999) also show the possibility of endogenous cycles in a model with imperfect credit markets, but the mechanism there is very different. In their model, high investment generates high future profits and investment, but it also pushes up the interest rate, which reduces future profits and investment. If the second effect is strong enough relative to the first, output will display negative serial correlation.

[13] This is not the only possible outcome. Depending on the initial distribution we could also get a stationary equilibrium with low or high wages.

4. Conclusion

In this paper, we analyzed a simple dynamic model of occupational choice in the presence of credit market imperfections where wealth inequality and returns to various occupations are endogenous. We examined conditions under which multiple steady-state equilibria exist and characterized how initial conditions affect which equilibrium the economy converges to. We conclude with two observations both of which suggest directions for future research. First, there are many interesting questions regarding the relationship between credit market imperfections and economic development that the current model or models similar to it (such as Banerjee and Newman, 1993, on which it is based, and also Galor and Zeira, 1993; Piketty, 1997) cannot address. As examples, one can mention recent research studying consequences of dynastic utility maximization in a similar framework and a richer set of possible occupations (see Mookherjee and Ray, 2000), allowing entrepreneurs to have heterogeneous talent (Bernhardt and Lloyd-Ellis, 2000), and the interaction between credit market imperfections and incentives and contracting in the labor market (see Ghatak et al., 2001). Second, while there is some cross-country evidence on the negative effect on inequality and measures of credit market imperfections on growth (Benabou, 1996) that is consistent with the prediction of this model, more micro-level evidence on the effect of borrowing constraints on economic mobility is clearly needed.[14]

Acknowledgements

We thank the anonymous referees for helpful comments, and Christian Ahlin, Abhijit V. Banerjee, Charles Hunter, Joseph Kabowski, Alexander Karaivanov, and Andreas Lehnert for useful discussions. We are responsible for all remaining errors.

Appendix A

Proof of Lemma 2: First, for any $\mathbf{w}=(w_1, w_2, \cdots, w_m)\in W^m$, we can define a function M: $W^m \rightarrow W^m$ as

$$M(\mathbf{w}) = (w_2, w_3, \cdots, w_m, f(\mathbf{w})).$$

Second, we compare $\mathbf{w}$ with $M(\mathbf{w})$: if $\mathbf{w}=M(\mathbf{w})$, we stop. If $\mathbf{w}\neq M(\mathbf{w})$, then we calculate $M^2(\mathbf{w})\equiv M(M(\mathbf{w}))$ and check if it is equal to either $\mathbf{w}$ or $M(\mathbf{w})$. If yes, we stop; if no, then we compare $M^3(\mathbf{w})$ with $\mathbf{w}$, $M(\mathbf{w})$, and $M^2(\mathbf{w})$, and so on. Since wage rate w only takes two

[14] There has been some work on this area using panel data sets from the US and the UK (see, for example Evans and Leighton, 1989; Blanchflower and Oswald, 1998). But very little is known about developing countries where borrowing constraints are presumably much more severe.

values ($\underline{w}$ and $\bar{w}$) and m is finite, this process cannot go on forever. There must exist some k and k' with $0 \le k < k' \le 2^m - 1$ such that $M^k(\mathbf{w}) = M^{k'}(\mathbf{w})$. Third, since $M^k(\mathbf{w}) = M^{k'}(\mathbf{w})$ implies $M^{k+t}(\mathbf{w}) = M^{k'+t}(\mathbf{w})$ which in turn implies $f(M^{k+t}(\mathbf{w})) = f(M^{(k+t)+(k'-k)}(\mathbf{w}))$ for all $t = 0,1,2,\ldots$, the wage dynamics displays a $(k'-k)$-period cycle. A special case is when $k'-k=1$ where we have constant-wage dynamics. □

Proof of Lemma 3: Let $\mathbf{w}, \mathbf{w}' \in W^m$ and $\mathbf{w} \le \mathbf{w}'$. Then

$$\bar{s}\left[\sum_{i=0}^{n-1}(r\bar{s})^i w_{m-i}\right] \le \bar{s}\left[\sum_{i=0}^{n-1}(r\bar{s})^i w'_{m-i}\right] \qquad \text{for } n = 1,2,\ldots,m$$

$$\Rightarrow a(n) \le a'(n) \qquad \text{for } n = 1,2,\ldots m$$

$$\Rightarrow l \ge l'$$

$$\Rightarrow \sum_{n=0}^{l-1} p^n(1-p) \ge \sum_{n=0}^{l'-1} p^n(1-p)$$

$$\Rightarrow G(I) \ge G'(I)$$

$$\Rightarrow f(\mathbf{w}) \le f(\mathbf{w}')$$

□

Proof of Proposition 4: For any $t \ge m$, there is a probability mass $p^n(1-p)$ at $a_t(n)$, $\forall n = 0,1,\ldots,l(t)$. Compare $a_{t+1}(n)$ with $a_t(n)$,

$$a_{t+1}(n) - a_t(n) = \bar{s}\left[\sum_{i=0}^{n-1}(r\bar{s})^i w_{t-i}\right] - \bar{s}\left[\sum_{i=0}^{n-1}(r\bar{s})^i w_{t-i-1}\right]$$

$$= \bar{s}\left\{w_t - (1-r\bar{s})\left[\sum_{i=0}^{n-2}(r\bar{s})^i w_{t-i-1}\right] - (r\bar{s})^{n-1} w_{t-n}\right\}$$

$$\in [\bar{s}(w_t - \bar{w}), \bar{s}(w_t - \underline{w})].$$

If $w_t = \bar{w}$, $a_{t+1}(n) - a_t(n) \ge 0$. This implies $l(t+1) \le l(t)$ and since $G_t(I) = \sum_{i=0}^{l(t)-1} p^i(1-p) \le 1/2$, $G_{t+1}(I) = \sum_{i=0}^{l(t+1)-1} p^i(1-p) \le 1/2$. That is, $w_{t+1} = w_t = \bar{w}$. If

$w_t=\underline{w}$, $a_{t+1}(n)-a_t(n)\leq 0$. This implies $l(t+1)\geq l(t)$ and since $G_t(I)=\sum_{i=0}^{l(t)-1}p^i(1-p)>1/2$, $G_{t+1}(I)=\sum_{i=0}^{l(t+1)-1}p^i(1-p)>1/2$. That is, $w_{t+1}=w_t=\underline{w}$. □

Proof of Lemma 4: If $p^m\geq 1/2$, then $\sum_{i=0}^{m-1}p^i(1-p)\leq 1/2$, which in turn implies that $f(\underline{w}, \underline{w}, \ldots,\underline{w})=\bar{w}$. Therefore, from Lemma 3, $f(\cdot)=\bar{w}$ for any element in W^m. If $p^{m'}<1/2$ then $\sum_{i=0}^{m'-1}p^i(1-p)>1/2$ which in turn implies that $f(\bar{w}, \bar{w}, \ldots,\bar{w})=\underline{w}$. Therefore, from Lemma 3, $f(\cdot)=\underline{w}$ for any element in W^m. □

Proof of Proposition 5: The "only if" part is implied by the previous Lemma. To prove the "if" part, it suffices to find two wealth distributions such that one is consistent with the high-wage equilibrium, whereas the other is consistent with the low-wage equilibrium. The obvious choice for such a distribution is one in the steady state. First, for the high-wage equilibrium, consider a wealth distribution at date t that has a probability mass

$$(1-p) \quad \text{at} \quad 0$$

$$p^n(1-p) \quad \text{at} \quad \bar{s}\left[\sum_{i=0}^{n-1}(r\bar{s})\bar{w}\right] \qquad \forall n=1,2,\ldots$$

From the definition of m', we know

$$\bar{s}\left[\sum_{i=0}^{m'-1}(r\bar{s})\bar{w}\right]<I\leq\bar{s}\left[\sum_{i=0}^{m'}(r\bar{s})\bar{w}\right]$$

and therefore

$$G_t(I)=\sum_{n=0}^{m'-1}p^n(1-p)=1-p^{m'}\leq\frac{1}{2}.$$

This implies $w_t=\bar{w}$ which in turn implies the next-period wealth distribution remains exactly the same. By changing m', $\bar{w}$ to m, $\underline{w}$, one can show that the low-wage equilibrium may also emerge in a similar way. The only difference is that the wealth distribution similarly constructed is not the steady-state distribution since the rich can earn entrepreneurial profits instead of wages. But since it is the wealth transition of the poor that matters in determining the wage rate, the argument is still valid. □

References

Aghion, P., Bolton, P., 1997. A trickle-down theory of growth and development with debt overhang. Review of Economic Studies 64 (2), 151–172.

Aghion, P., Banerjee, A., Piketty, T., 1999. Dualism and macroeconomic volatility. Quarterly Journal of Economics 114 (4), 1359–1398.

Banerjee, A.V., Newman, A., 1993. Occupational choice and the process of development. Journal of Political Economy 101 (2), 274–298.

Banerjee, A.V., Newman, A., 1994. Poverty, incentives, and development. American Economic Review Papers and Proceedings 84 (2), 211–215.

Barro, R., Sala-i-Martin, X., 1995. Economic Growth. McGraw-Hill, New York.

Benabou, R., 1996. Inequality and growth. In: Bernanke, B.S., Rotemberg, J.J. (Eds.), NBER Macroeconomic Annual, 1996. MIT Press, Cambridge, pp. 11–74.

Bernhardt, D., Lloyd-Ellis, H., 2000. Enterprise, inequality and economic development. Review of Economic Studies 67 (1), 147–168.

Blanchflower, D.G., Oswald, A.J., 1998. What makes an entrepreneur? Journal of Labor Economics 16 (1), 26–60.

Evans, D., Leighton, L., 1989. Some empirical aspects about entrepreneurship. American Economic Review 79 (3), 519–535.

Galor, O., Zeira, J., 1993. Income distribution and macroeconomics. Review of Economic Studies 60 (1), 35–52.

Ghatak, M., Morelli, M., Sjöström, T., 2001. Occupational choice and dynamic incentives. Review of Economic Studies 68 (4), 781–810.

Mookherjee, D., Ray, D., 2000. Persistent inequality. Institute for Economic Development Discussion Paper #108, Boston University.

Piketty, T., 1997. The dynamics of wealth distribution and the interest rate with credit rationing. Review of Economic Studies 64 (2), 173–189.

History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India

By ABHIJIT BANERJEE AND LAKSHMI IYER*

We analyze the colonial land revenue institutions set up by the British in India, and show that differences in historical property rights institutions lead to sustained differences in economic outcomes. Areas in which proprietary rights in land were historically given to landlords have significantly lower agricultural investments and productivity in the post-independence period than areas in which these rights were given to the cultivators. These areas also have significantly lower investments in health and education. These differences are not driven by omitted variables or endogeneity problems; they probably arise because differences in historical institutions lead to very different policy choices. (*JEL* O11, P16, P51)

There is renewed interest among economists in the question of whether history, through its effect on the pattern of institutional development, has a persistent effect on economic performance. In a recent series of papers, Rafael La Porta et al. (1998, 1999, 2000) have argued that the historical fact of being colonized by the British, rather than any of the other colonial powers, has a strong effect on the legal system of the country and, through that, on economic performance. The role of history in determining the shape of present-day institutions is also at the heart of two recent sets of papers, one by Daron Acemoglu et al. (2001, 2002) and the other by Stanley Engerman and Kenneth Sokoloff (1997, 2000, 2002). Acemoglu et al. show that mortality rates among early European settlers is a strong predictor of whether these countries end up with what economists today call "good" institutions (which protect private property rights) and whether their economies are doing well today. Engerman and Sokoloff argue that the reason why Brazil is where it is today, and the United States is where it is, has a lot to do with the fact that in the early years after European conquest Brazil was deemed to be suitable for growing sugar and the United States was not. Since sugar cultivation demanded the use of slave labor, Brazil ended up with a much larger slave population, and this, they argue, meant that Brazilian society was much more hierarchical than American society, causing a divergence in the types of institutions that evolved in these two countries and eventually a divergence in the rates of growth.

This paper is a part of the same broad research agenda. Where it differs is in focusing on one very specific historical institution—the system for collecting land revenue—in one specific country—India. We compare the present-day economic performance of different districts of India, which were placed under different land revenue systems by British colonial rulers as a result of certain historical accidents. We show that districts in India where the collection of land revenue from the cultivators was assigned to a class of landlords systematically underperform the districts where this type of intermediation was avoided, after controlling for a wide range of geographical differences. The differences show up in agricultural investment and yields, in various measures of public investment in education and health, as well as in health and educational outcomes. For example, the average yield of wheat is 23 percent higher and infant mortality is 40 percent lower in non-landlord districts. The non-landlord effect remains sig-

* Banerjee: Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02139 (e-mail: banerjee@mit.edu); Iyer: Harvard Business School, Soldiers Field, Boston, MA 02163 (e-mail: liyer@hbs.edu). We thank Daron Acemoglu, Sam Bowles, Esther Duflo, Maitreesh Ghatak, Karla Hoff, Kaivan Munshi, Raghuram Rajan, Andrei Shleifer, two anonymous referees, and numerous seminar participants for helpful comments. We also thank Nabeela Alam and Theresa Cheng for research assistance and Michael Kremer for help in accessing historical land tenure data.

nificant even when we restrict our data analysis to a set of 35 districts, chosen so that a landlord district always borders a non-landlord district. Finally, in all the data we have from the earlier period, i.e., from the nineteenth and early twentieth centuries, there is no evidence of landlord districts being at a disadvantage.

An obvious advantage of focusing on one specific institution in one particular country is that it makes it easy to locate the source of the difference, relative to the case where there is a complex of institutions that are all different. Another advantage is that we have access to a very detailed history of how the institutional variation came about, which makes it easier to argue for exogeneity of specific pieces of the variation. In particular, we will argue, based on historical facts, that areas where the land revenue collection was taken over by the British between 1820 and 1856 (but not before or after) are much more likely to have a non-landlord system, for reasons that have nothing to do with factors that directly influence agricultural investment and yields. We will therefore use the fact of being conquered in this period as an instrument for having a non-landlord system. We allow for the possibility that areas that were conquered in this period may have had a different experience simply because, for example, they were conquered later than most other areas, by including controls for the length of British rule. The instrumental variable estimates confirm the OLS results.

A third advantage of this particular experiment is that the land revenue systems introduced by the British departed with the British: there are no direct taxes on agricultural incomes in independent India. Our results therefore tell us that the system for land revenue collection established by the British 150 years ago or more continues to have an effect, long after it was abolished. We therefore have a pure example of institutional overhang, underscoring how hard it is to reform the institutional environment.[1]

The one disadvantage of a very specific experiment like ours is the suspicion that it reflects the peculiarity of the Indian experience. In other words, our results would be more interesting if we could identify the reasons for this extreme persistence. While our data do not allow us to identify exactly the channel through which the historical land revenue system continues to have an effect, there are a number of clues. When the British left, areas where landlords collected the revenue had an elite class that had enjoyed a great deal of economic and political power for over a century; there was no counterpart to this class in the non-landlord areas. This meant that these areas inherited a more unequal land distribution at the time of independence, and a very specific set of social cleavages, absent elsewhere.

Our data suggest, however, that in the post-independence period there is substantial convergence in inequality between the landlord and non-landlord areas, probably because states with landlord-dominated areas tend to enact a greater number of land reforms. This makes it unlikely that the persistence of the landlord effect is mainly through its effect on the contemporaneous land distribution.

On the other hand, it seems that, despite the abolition of the formal structure of landlordism, the class-based antagonism that it created within the communities in these areas persisted well into the post-independence period. The conflictual environment this created is likely to have limited the possibility of collective action in these areas. This collective action–based view is consistent with the fact that the gap between the non-landlord and landlord districts grows particularly fast in the period 1965–1980 when there is extensive public investment in rural areas. We find that states with a higher proportion of landlord districts have much lower levels of public development expenditures and that a substantial part of the gap between landlord and non-landlord districts in health, education, and agricultural technology investments can be explained by this difference in public spending. This suggests that the key to what happened may lie in the relative inability of the landlord districts to claim their fair share of public investment.

The paper is structured as follows: Section I describes the historical background and the land tenure system under British rule. We discuss the reasons why the tenure system varies from district to district, and argue that the choice of tenure system can be reasonably regarded as a source of exogenous variation. Section II outlines different mechanisms through which

[1] This distinguishes this work from the recent empirical literature on the effects of current land reform on current economic outcomes (see Banerjee et al., 2002; Timothy Besley and Robin Burgess, 2000; Justin Y. Lin, 1992, among others).

historical land tenure might affect long-term outcomes. Sections III and IV describe our data and empirical strategy. Our main empirical results are described in Section V. Section VI concludes by discussing potential mechanisms that might explain the persistence of the effect of British land tenure systems.

I. Historical Background

A. *British Political Control*

The British Empire in India lasted for nearly two hundred years. The British first arrived as traders: the English East India Company received a permit in 1613 from the Mughal emperor, Jahangir, to build a factory at Surat. Their empire building began with their victories in the battle of Plassey in 1757 and the battle of Buxar in 1764, as a result of which they obtained political control of the modern states of Bengal and Bihar (formerly Bengal Presidency). The British were formally granted revenue-collection rights in these areas in 1765. After 1818, the British were the major political power in India and by 1860 a large part of the territories of modern India, Pakistan, and Bangladesh were part of the British Empire. There were also a large number of princely states in different parts of the country, all of which were under British political control but had autonomy in administrative matters.

Different parts of the country came under British rule in different periods. While the Bengal Presidency came into British hands in 1765, the rest of eastern India was conquered much later. Some parts of the modern state of Orissa were conquered in 1803 and Assam was conquered between 1824 and 1826. Meanwhile, in south India, the British obtained four districts (the "Northern Circars") as a grant from the Mughal emperor in 1765. These and other areas conquered between 1792 and 1801 came to form the Madras Presidency. Parts of the western state of Gujarat were conquered in 1803 and the rest, along with large parts of Bombay Presidency, were obtained after conquering the Marathas in 1817–1818. Some of these areas formed part of the Central Provinces, to which other parts were added over a long period until 1860. In the north, large parts of the North-West Provinces were obtained from the Nawab of Oudh in 1801–1803, but Oudh itself was not annexed by the British until 1856. The northwestern state of Punjab was annexed after the Sikh wars in 1846 and 1849. Table 1 in the Web Appendix (http://www.e-aer.org/data/sept05_app_banerjee.pdf) provides district-wide details on the date and mode of acquisition by the British.

The rule of the East India Company came to an end after the Mutiny of 1857, when Indian troops revolted against their British officers. The revolt was soon suppressed, but it forced the British government to bring India under its direct control. The British left India in 1947, when the Indian Empire was partitioned into India and Pakistan.[2] Large parts of former Bengal Presidency and Panjab Province are now in Bangladesh and Pakistan, respectively.

B. *Pre-British and British Systems of Land Revenue*

Land revenue, or land tax, was the major source of revenue for all governments of India, including the British. During the period of Mughal rule in the sixteenth and seventeenth centuries, land revenue was collected by nonhereditary, transferable state officials (the *mansabdari* system introduced by Emperor Akbar). After the decline of Mughal power in the early eighteenth century, these officials and others grabbed power where they could and became de facto hereditary landlords and petty chiefs in their local areas. As a result, by the time British rule was firmly established in India (toward the end of the eighteenth century), it was very hard to tell what the "original land revenue systems" of India had been, and different British administrators could come to very different conclusions about it.

Land revenue, or land tax, continued to be the major source of government revenue during British times as well. In 1841, it constituted 60 percent of total British government revenue, although this proportion decreased over time as the British developed additional tax resources. Not surprisingly, land revenue and its collection were the most important issues in policy debates during this period. We use the terms "land revenue systems" or "land tenure systems" to refer

[2] Bangladesh, formerly East Pakistan, became an independent nation in 1975.

TABLE 1—STATE-WISE DISTRIBUTION OF LANDLORD AND NON-LANDLORD DISTRICTS

State	Mean non-landlord proportion	Classification of revenue systems				Total districts
		Landlord based	Individual based	Village bodies		
				Landlord	Non-landlord	
Andhra Pradesh	0.66	2	8	0	0	10
Bihar	0.00	12	0	0	0	12
Gujarat	1.00	0	7	0	0	7
Haryana	0.85	0	0	0	5	5
Karnataka	1.00	0	15	0	0	15
Madhya Pradesh	0.10	14	1	0	0	15
Maharashtra	0.78	4	14	0	0	18
Orissa	0.32	6	2	0	0	8
Punjab	0.87	0	0	0	6	6
Rajasthan	0.00	1	0	0	0	1
Tamil Nadu	0.75	2	9	0	0	11
Uttar Pradesh	0.42	0	0	12	35	47
West Bengal	0.00	11	0	0	0	11
Total	0.51	52	56	12	46	166

Notes: This table lists only districts that used to be part of British India. Areas where the British did not set up the land revenue system are excluded. Districts of British India currently in Pakistan, Bangladesh, or Burma are excluded. The table also excludes the states of Assam and Kerala, for which agricultural data are not available in the World Bank dataset. The table lists 1960 districts, some of which were split into two or more districts over time. We use unsplit districts in all our analyses.

to the arrangements made by the British administration to collect the land revenue from the cultivators of the land. These systems defined who had the liability to pay the land tax to the British. Up to a first approximation, all cultivable land in British India fell under one of three alternative systems: (a) a landlord-based system (also known as *zamindari* or *malguzari*), (b) an individual cultivator-based system (*raiyatwari*), and (c) a village-based system (*mahalwari*). Table 1 gives the number of districts in each category for the states in our data. The map in Figure 1 illustrates the geographic distribution of these areas.

In the landlord areas, the revenue liability for a village or a group of villages lay with a single landlord. The landlord was free to set the revenue terms for the peasants under his jurisdiction and to dispossess any peasants who did not pay the landlord what they owed him.[3] Whatever remained after paying the British revenue demand was for the landlord to keep. These revenue-collecting rights could be bequeathed, as well as bought and sold (Kumar, 1982). In this sense, the landlord effectively had property rights on the land. Landlord systems were established mainly in Bengal, Bihar, Orissa, the Central Provinces (modern Madhya Pradesh state), and some parts of Madras Presidency (modern Tamil Nadu and Andhra Pradesh states). In some of these areas, the British declared the landlords' revenue commitments to the government to be fixed in perpetuity (the "Permanent Settlement" of 1793). In other areas, a "temporary" settlement was implemented whereby the revenue was fixed for a certain number of years, after which it was subject to revision.

In most areas of Madras and Bombay Presidencies, and in Assam, the *raiyatwari* system was adopted under which the revenue settlement was made directly with the individual *raiyat* or cultivator. In these areas, an extensive cadastral survey of the land was done and a detailed record-of-rights was prepared, which served as the legal title to the land for the cultivator. Unlike the Permanent Settlement areas, the revenue commitment was not fixed; it was usually calculated as the money value of a share of the estimated average annual output. This share typically varied from place to place, was different for different soil types, and was adjusted periodically in response to changes in the productivity of the land.

In the North-West Provinces and Panjab, the

[3] Some measures for protecting the rights of tenants and subproprietors were introduced in later years.

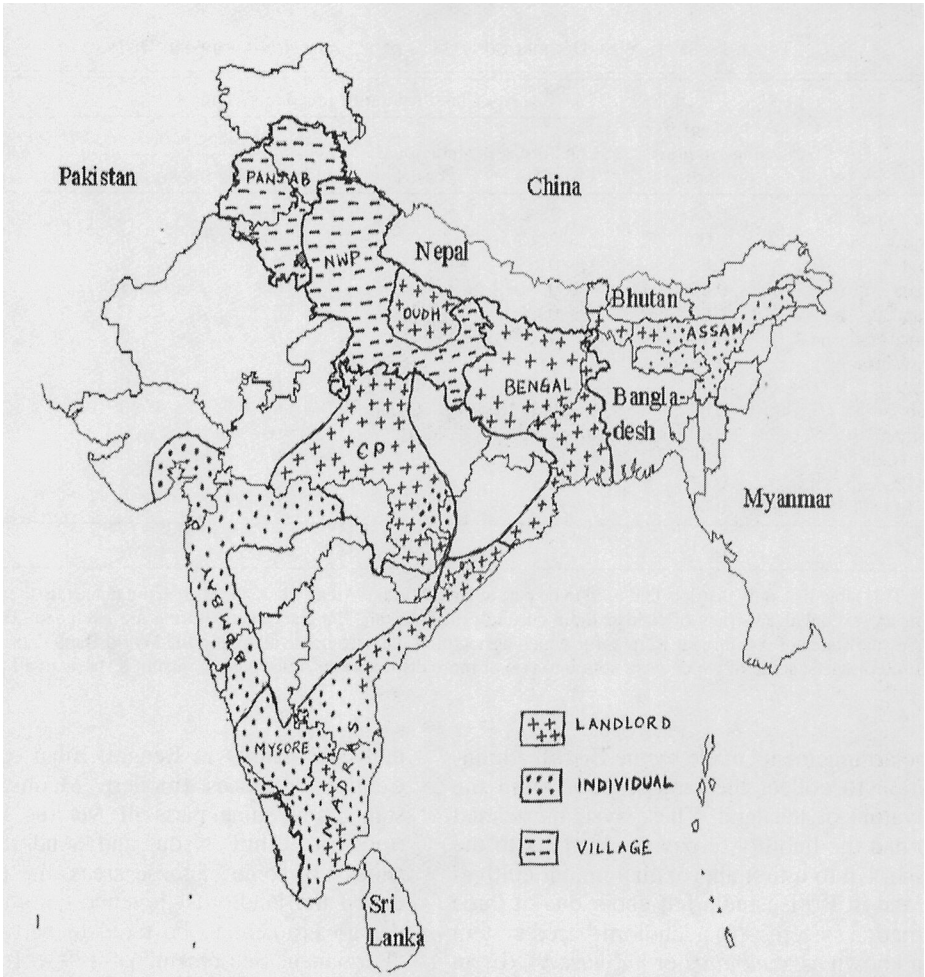


FIGURE 1. MAP OF INDIA

village-based (*mahalwari*) system was adopted in which village bodies which jointly owned the village were responsible for the land revenue. Village bodies could be in charge of varying areas, from part of a village to several villages. The composition of the village body also varied from place to place. In some areas it was a single person or family that made up the village body and hence was very much like the Bengal landlord system (*zamindari*), while in other areas the village body had a large number of members with each person being responsible for a fixed share of the revenue. This share was either determined by ancestry (the *pattidari* system), or based on actual possession of the land (the *bhaiachara* system), the latter being very much like the individual-based *raiyatwari* system. The revenue rates in these areas were determined on fairly ad hoc grounds, based on a diverse set of factors, including: "an examination of rents recorded in the *jamabandis,* the rates which were actually paid by the various classes of tenants and the rates which were considered fair on each class of soil. ... These estimates are based primarily on soils, and secondly on consideration of the caste of the tenant, capabilities of irrigation, command of manure &c, all of which points received attention" (F. W. Porter, 1878, p. 108).[4]

[4] Except in the areas under the Permanent Settlement, the amount of revenue actually paid was often less than the stated revenue liability, due to remissions being granted in

C. *Choice of Land Revenue System*

Why did the British choose different systems in different areas? It is broadly agreed that their major motivation was to ensure a large and steady source of revenue for the government, while maintaining a certain political equilibrium. It is also clear, however, that they often faced a lack of hard information and based their decision on a priori arguments. For instance, Sir Thomas Munro argued for the establishment of an individual cultivator system in Madras on the grounds that it would raise agricultural productivity by improving incentives; that the cultivators would be less subject to arbitrary expropriation than under a landlord; that they would have a measure of insurance (via government revenue remissions in bad times); that the government would be assured of its revenue (since small peasants are less able to resist paying their dues); and that this was the mode of land tenure prevailing in South India from ancient times. The Madras Board of Revenue, in its turn, used more or less the same arguments (in reverse, of course) for favoring landlords. Large landlords would have the capacity to invest more and therefore productivity would be higher; the peasants' long-term relationship with the landlord would result in less expropriation than the short-term one with a government official; a big landlord would provide insurance for small farmers; a steady revenue would be assured because the landlords would be wealthy and could make up an occasional shortfall from their own resources; and this was the mode of tenure prevailing from ancient times (Nilmani Mukherjee, 1962)! While the British often invoked history to justify the choices they made, they frequently misread history. For example, one reason they favored landlords in Bengal is because they found landlords in Bengal when they arrived. As has been pointed out by a number of scholars,[5] however, these landlords were really local chieftains and not the large farmers that the British had thought them to be.

Decisions were therefore often taken on the basis of some general principle, and the ideology of the individual decision maker and contemporary economic doctrines played an important role in combination with the exigencies of the moment. Table 2 of the Web Appendix provides details of how different land revenue systems came to be established in different provinces of British India. Here, we summarize the main channels of influence.

Influence of Individual Administrators.—The ideas and political influence of particular administrators sometimes determined land revenue systems in whole provinces. For instance, in the Madras land tenure debate cited above, the Board of Revenue initially prevailed over Sir Thomas Munro, and all the villages were put under village-level landlords with renewable leases. Munro, traveled to London, however, and managed to convince the Court of Directors of the East India Company of the merits of the individual-based *raiyatwari* system; they then ordered the Madras Board of Revenue to implement this policy all over the province after 1820, on the expiration of the landlord leases. Similarly, the individual system was tried out in Bombay Presidency quite early, mainly because the governor, Lord Elphinstone, was in favor of it and had been a supporter of Munro during the debate in Madras.

Another instance of individual influence occurred in the North-West Provinces. Landlord systems with short-term leases were implemented there initially, and there was considerable debate as to whether or not there should be a Permanent Settlement along the lines of that prevailing in Bengal. In 1819, however, Holt Mackenzie, the Secretary of the Board of Revenue, wrote a famous Minute, which claimed that historically every village had had a proprietary village body and felt that no settlement that did not give proper recognition to such customary rights should be declared in perpetuity. This became the basis for Regulation VII of 1822, which laid the basis for village-level settlements (B. R. Misra, 1942). The previous actions, however, could not always be undone and in several places the previously appointed large landlords (*talukdars*) retained their positions.[6]

times of bad harvests and other hardships. Our focus here is not on the actual revenue paid or the revenue rates, which prevailed at various points of time, but on the allocation of revenue and control rights in land.

[5] See Tirthankar Roy (2000) and Ratnalekha Ray (1979).

[6] For instance, the Aligarh settlement officer writes, "So far indeed had the action of our first officials sanctioned the

Political Events.—The most notable example of this occurred in Oudh province. This region was annexed by the British in 1856 and merged with the North-West Provinces to form the United Provinces (state of Uttar Pradesh today). Since the North-West Provinces had a village-based revenue system, it was proposed to extend the same to Oudh, and a cadastral survey that would form the basis of this settlement was under way when the mutiny broke out in 1857. After it was successfully subdued, the British felt that having the large landlords (*talukdars*) on their side would be politically advantageous. Thus, there was a reversal of policy and several landlords whose land had been taken away under the village-based settlement had the land given back to them, and in 1859 they were declared to have a permanent, hereditary, and transferable proprietary right. Districts that used to be a part of Oudh thus came to have a larger area under landlord control than the other districts of Uttar Pradesh.

Date of Conquest.—There are at least three reasons why areas that came under British revenue administration at later dates were in general more likely to have non-landlord systems. First, areas conquered later had some non-landlord precedents to follow and these made it easier to make the case for the non-landlord system. For instance, Berar was put under an individual-based system because neighboring Bombay had been; and similarly Panjab adopted the village-based system already in place in the North-West Provinces. In fact, once Munro's victory over the Board of Revenue in Madras was sealed by a widespread conversion of landlord areas into *raiyatwari* areas, and Holt Mackenzie had succeeded in making the case for village bodies, there were to be no new landlord areas until the reversal in Oudh. Second, landlord-based systems required much less administrative machinery to be set up by the British, and so areas conquered in the early periods of British rule were likely to have landlord-based systems. Once a landlord-based system was established, however, it was costly to change the system (this was most obviously true where there was a Permanent Settlement) and hence the landlord system survived. Finally, the increasing popularity of dealing directly with the peasant mirrored shifts in the views of economists and others in Britain. In the 1790s, under the shadow of the French Revolution across the Channel, the British elites were inclined to side with the landlords. In the 1820s, with peasant power long defeated and half forgotten, they were more inclined to be sympathetic to the utilitarians and others who were arguing for dealing directly with peasants.[7,8]

Presence of a Landlord Class before the British Took Over.—This was probably one of the factors leading to the landlord system being favored, at least in Bengal. As the historian Tapan Ray Chaudhuri says, "... in terms of rights and obligations, there was a clear line of continuity in the *zamindari* system of Bengal between the pre- and the post-Permanent Settlement era" (Dharma Kumar, 1982). This was not, however, always the case. For instance, it was decided to have a landlord-based system in the Central Provinces, even though there was no existing landlord class.[9]

D. *Post-Independence Developments in Land Policy*

Under the constitution of independent India, states were granted the power to enact land reforms. Several states passed legislation in the early 1950s, formally abolishing landlords and other intermediaries between the government and the cultivator. Other laws have also been passed by different states at different times regarding tenancy reform, ceiling on land holdings, and land consolidation measures. Besley and Burgess (2000) provide a good review of

usurpations of the Talukdars, that among other cases they granted to Raja Bhagwant Singh a lease for life of the whole of the pargana Mursan for Rs.80,000 leaving the old communities entirely at his mercy ..." (W. H. Smith, 1882).

[7] James Mill actually worked for the East India Company, and George Wingate, who helped set up the individual-cultivator system in Bombay, was heavily influenced by him.

[8] For a discussion of the role of ideology and economic doctrines in the formation of the land revenue systems, see Ranajit Guha (1963) and Eric Stokes (1959, 1978a).

[9] B. H. Baden-Powell (1892) states: "In the Central Provinces we find an almost wholly artificial tenure, created by our revenue-system and by the policy of the Government of the day."

these laws and their impact on state-level poverty rates.

II. Why Should the Historical Land System Matter?

Why would we expect productivity and investment (including public investment in infrastructure) to differ between areas having a greater or lesser extent of landlord control? Why would these differences persist and not be wiped out as soon as the landlord class is formally abolished? In this section, we list some potential answers to these questions, postponing to Section VI any discussion of the empirical plausibility of these answers.

A. *Differences in the Distribution of Wealth*

Under landlord-based systems, the landlords were given a more or less free hand to set the terms for the tenants[10] and, as a result, they were in a position to appropriate most of the gains in productivity. Moreover, landlord areas were also the only areas subject to the Permanent Settlement of 1793 (which fixed the landlord's dues permanently in nominal terms), and even where the settlement was not permanent, the political power of the landlord class made it less likely that their rates would be raised when their surplus grew. As the nineteenth century was a period of significant productivity growth and inflation, the landlord class grew rich over this period and inequality went up. By contrast, in the individual cultivator areas, rents were raised frequently by the British in an attempt to extract as much as possible from the tenant. There was, as a result, comparatively little differentiation within the rural population of these areas until, in the latter years of the nineteenth century, the focus of the British moved away from extracting as much as they could from the peasants. At this point, there was indeed increasing differentiation within the peasant class, but overall one would expect less inequality in the non-landlord areas.

In fact, this is what the limited historical data we have suggest. The provinces with a higher non-landlord proportion have lower Gini measures of land inequality in 1885 (Figure 2A). Further, the differences in inequality persist until the end of the colonial period. In 1948, the districts of Uttar Pradesh that had a higher landlord proportion had a much higher proportion of land revenue being paid by very large landlords and a correspondingly higher measure of inequality (Figure 2B).

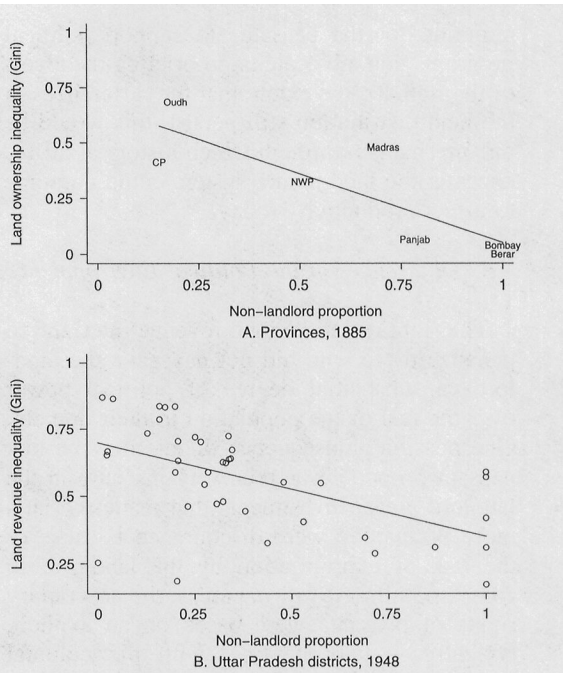


FIGURE 2. LAND TENURE AND LAND INEQUALITY

The distribution of wealth is important for three reasons: first, because it determines the size of the group within the peasantry that has enough land and other wealth to be able to make the many somewhat lumpy and/or risky investments necessary to raise productivity;[11] second, because it affects the balance between those who cultivate mainly their own land and those who cultivate other people's land (as is well-known, cultivating other people's land generates incentive problems, which reduces investment and productivity); finally, because it made it likely that the political interests of the rural masses would diverge substantially from that of the elite. In particular, it made it very

[10] Under the *Haftam* regulation of 1799 and the *Panjam* regulation of 1812.

[11] See Banerjee and Andrew F. Newman (1993) and Oded Galor and Joseph Zeira (1993) for theoretical models of the link between income distribution and long-run development.

tempting for the peasants to support political programs that advocate expropriating the assets of the rich. To the extent that the differences in the land distribution still persist, this would be one mechanism through which historical differences in the land tenure system could continue to affect productivity today.

B. *Differences in the Political Environment*

The right to set the land revenue rates and to penalize those who did not pay gave the landlords a substantial degree of political power over the rest of the population in their domain. One possible consequence of this may be that peasant property was relatively insecure in the landlord areas. Investments that made the land more productive were discouraged because of the risk of expropriation by the landlord. In contrast, in the *raiyatwari* areas, the proprietary rights of peasants were based on an explicit, typically written, contract with the colonial state, which the colonial state was broadly committed to honor. This may have resulted in better incentives for the peasants in the non-landlord areas in the colonial period.

The exercise of this type of more or less arbitrary power by the landlord over the property and not, infrequently, the body of the peasant, created a political ethos of class-based resentment in these areas, which persisted well into the post-independence period.[12] Those familiar with post-independence India will recognize, for example, that the areas most associated with Maoist peasant uprisings (known as "Naxalite" movements)—clearly the most extreme form of the politics of class conflict in India—are West Bengal, Bihar, and the Srikakulam district of Andhra Pradesh, all landlord areas. Paul R. Brass (1994, pp. 326–27) argues explicitly that these peasant movements had their roots in the history of exploitation and oppression of peasants by landlords. Moreover, these class-based conflicts go back to the colonial period. Kathleen Gough (1974) studies 77 peasant struggles from the end of the Mughal era until today and suggests that at least a third of these originated in Bengal, the oldest and best established of the landlord areas. Along the same lines, Partha Chatterjee (1984) has argued, based on the pattern of voting on the Tenancy Act Amendment in the Bengal Legislative Council in 1928, that the representatives of the peasants voted largely in a block against the landlords, and vice versa.

Given this history, it is no surprise that the elites and the masses in these areas rarely shared the trust that is essential for being able to act together in the collective interest.[13] It is quite plausible that, in the post-independence period, the political energies of the masses were directed more toward expropriating from the rich (via land reforms, for example) than toward trying to get more public goods (schools, tap water, electricity) from the state, while the political energies of the rich were aimed at trying to ensure that the poor did not get their way.[14] Moreover, it was not uncommon for the rural elites in the landlord areas to be quite disassociated from the actual business of agriculture, since they typically were more likely to be rent collectors than farmers, and even the rent collection rights were often leased out. This would tend to weaken the political pressure on the state to deliver public goods that were important to farmers. Moreover, they were often physically absent, preferring to live in the city and simply collect their rents, and as a result had only rather limited stakes in improving the living conditions in rural areas.

C. *Differences in the Relationship with the Colonial State*

Since it was easier for the colonial government to raise rents in non-landlord areas, it meant that the state could capture some of the productivity gains from these areas, and hence had more reason to invest in irrigation, railways, schools, and other infrastructure in these areas during the colonial period.[15] In this context, we

[12] See Sugato Bose (1993) for an account of the rise of class-based agrarian politics in colonial Bengal (a landlord area) and its subsequent influence on the politics of independent West Bengal.

[13] See Alberto Alesina and Dani Rodrik (1994) and Torsten Persson and Guido Tabellini (1994) for models where collective action fails in the presence of groups with misaligned interests.

[14] For instance, the rich could undercut democratic processes and resist public policies that would empower the poor, very much along the lines taken by the Latin American elites (see Engerman and Sokoloff, 2002).

[15] Amiya K. Bagchi (1976) also makes this point.

should note that almost all canals constructed by the British were in non-landlord areas. If, indeed, these areas had better public goods when the British left, it is plausible that they could continue to have some advantage even now.

III. Data

We use a combination of historical and recent data for our analysis. All data are at the district level, a district in India being an administrative unit within a state. In 1991, India had 415 districts in 17 major states, a district on average having an area of 7,500 square kilometers and a population of 1.5 million.

We chose to use district-level rather than state-level data for three major reasons. First, modern Indian state boundaries are completely different from older British province boundaries due to the linguistic reorganization of states in 1956. Although district boundaries have also changed a little over time, it is still possible unambiguously to match current districts to older districts—the main source of change is that some of the older districts have been split into two or more districts over time. Second, because of the integration of several princely states in 1947, nearly all the states in our data are composed of both British-ruled districts and districts that were ruled by Indian kings in the colonial period. Since we have historical data on land tenure only for British districts, it is hard to compute a good state-level measure of historical institutions. Third, using district-level data gives us a larger sample size. The drawback is that we are limited in the kind of data that we can get. For instance, we do not have measures of GDP or average income per capita at the district level. We will thus be using other correlates or proxies of economic prosperity for which we have data at the district level: agricultural investment outcomes (the proportion of irrigated gross cropped area, quantity of fertilizer used per hectare of gross cropped area, and the proportion of area sown with high-yielding varieties (HYV) of rice, wheat, and other cereals); agricultural productivity (crop yields); and the stock of health and education infrastructure (schools and health centers).

The district-level data on agricultural investments and productivity come from the India Agriculture and Climate Data Set assembled by the World Bank and cover the period 1956–1987. This dataset has information on 271 districts in 13 major states.[16] All data are at the 1961 district level, aggregating over subsequent splits in districts. We also have data for health and education infrastructure from the 1981 Census. We matched each modern district to an older British district using old and new maps, and retained only the districts where the land tenure system was established by the British, because we do not have detailed information on the land systems in districts that were under native princes or tribal chiefs.[17] For each district of British India,[18] we then proceed to compute a measure of non-landlord control in the colonial period as follows: for many areas (the states of Andhra Pradesh, Madhya Pradesh, Panjab, Tamil Nadu, and Uttar Pradesh), we have district-level information on the proportion of villages, estates, or land area, not under the revenue liability of landlords; for other areas where we do not have the exact proportion (Bihar, Karnataka, Maharashtra, Orissa, West Bengal), we assign the non-landlord measure as being either zero or one, depending on the dominant land revenue system. In all cases, the measure of non-landlord control is computed based on data from the 1870s or 1880s. The details of the data sources and the construction of this variable are in Table 3 of the Web Appendix.

IV. Empirical Approach

We will compare agricultural investments and productivity between landlord and non-landlord areas by running regressions of the form

[16] The states included in the dataset are Andhra Pradesh, Bihar, Gujarat, Haryana, Karnataka, Madhya Pradesh, Maharashtra, Orissa, Punjab, Rajasthan, Tamil Nadu, Uttar Pradesh, and West Bengal. Assam, Himachal Pradesh, Jammu and Kashmir, and Kerala are the large states not covered.

[17] This usually corresponds to the areas under direct British administrative control, with one exception. In the princely state of Mysore (part of modern Karnataka state), the British took over the administration in 1831 and ruled for 50 years, before reinstating the royal family in 1881. During this time, the British instituted an individual-based land revenue system, which the ruler was obliged to continue after his reinstatement.

[18] We dropped districts currently in Pakistan and Bangladesh.

$$(1)\quad y_{it} = constant + \alpha_t + \beta \mathrm{NL}_i + X_{it}\gamma + \varepsilon_{it}$$

where y_{it} is our outcome variable of interest (investment, productivity, etc.) in district i and year t, α_t is a year-fixed effect, NL_i is the historical measure of the non-landlord control in district i, and X_{it} are other control variables. Our coefficient of interest is β, which captures the average difference between a non-landlord district and a landlord district in the post-independence period.

In all our regressions, we control for such geographic variables as latitude, altitude, soil type, mean annual rainfall, and a dummy for whether the district is on the coast or not. In addition, we also control for the length of time under British rule (or, equivalently, the date of British conquest), which may have independent effects, because early British rule was particularly rapacious or because the best (or the worst) districts fell to the British first. Note that we do not include district fixed effects in this regression, since NL_i is fixed for district i over time (it is the historical land arrangement). We do adjust our standard errors for within-district correlation, however, since our data consist of repeated observations for each district over time. We also do not use state fixed effects in our base specification because the within-state variation in non-landlord proportion is much less than the cross-state variation.[19] More importantly, the modern states were formed at a later date than our non-landlord proportion and we would like to see how far historical factors can account for the widely varied performance of Indian states in the post-independence period.

As mentioned in the introduction, we will try to deal with concerns about exogeneity, first by looking only at the difference between neighboring districts, and second by adopting an instrumental variables approach. After establishing the robustness of the differences in investment and productivity between landlord and non-landlord areas, we estimate some additional specifications. First we reestimate the yield equations after controlling for various measures of investment in agriculture (fertilizer use, irrigation, etc.) to check whether there is a non-landlord effect over and above the effect on investment. Then we allow the non-landlord coefficient to vary over time to see whether we can demonstrate how the gap between landlord and non-landlord areas has evolved over time.

[19] In our later regressions with state fixed effects, we are in effect dropping the states of Bihar, Gujarat, Karnataka, Rajasthan, and West Bengal.

V. The Impact on Agricultural Outcomes

A. *Differences in Geography and Other Differences*

There are significant geographical differences between landlord areas and non-landlord areas (Table 2). Landlord areas have somewhat lower altitudes, higher rainfall, and fewer areas with black soil as compared to non-landlord areas. In particular, we note that landlord areas have a greater depth of topsoil, which together with the greater rainfall and lower altitudes seems to indicate that these areas might be inherently more fertile and productive. Landlord areas have a slightly higher total population and a significantly higher population density than non-landlord areas. This is consistent with the fact they seem to be more fertile areas. They have a greater proportion of minorities, such as castes that were discriminated against historically and are formally listed as "Scheduled Castes" in the Indian Constitution, and more people living in rural areas. Further, landlord areas have a greater proportion of the workforce engaged in farming, and devote more area to food crops like rice and wheat and less to cash crops like cotton, oilseeds, tobacco, and sugarcane. This could be due simply to different climatic conditions or could reflect an endogenous shift toward commercial agriculture in non-landlord areas.

We have very limited historical data on yields. Looking at data for rice yields in ten districts of Madras Presidency, and rice and wheat yields for 17 districts of Uttar Pradesh during the colonial period, we see in Figure 3 that yields were in fact *lower* in non-landlord areas during this period.[20] Given the size of the sample, we cannot hope to control for geographical differences between the districts. These yield differences may therefore reflect differences in geography. The only point we are

[20] The yield data for Uttar Pradesh come from the same settlement reports of the 1870s and 1880s that we use to calculate our non-landlord proportion. Very few of the reports contain data on yields, resulting in a very small sample. We also have data for ten Tamil Nadu districts from Haruka Yanagisawa (1996).

TABLE 2—DIFFERENCES IN GEOGRAPHY AND DEMOGRAPHICS

		Mean	Standard deviation	Difference[a]	Standard error of difference
Geography					
Latitude		22.19	5.60	−4.35***	(0.961)
Altitude		366.41	148.14	93.64***	(25.98)
				—	
Mean annual rainfall (mm)		1263.09	471.64	373.99***	(80.83)
Coastal dummy		0.1497	0.3579	0.084	(0.065)
Top 2 soil types	Black soil	0.2096	0.4082	0.244***	(0.072)
	Alluvial soil	0.1677	0.3747	−0.135**	(0.067)
	Red soil	0.5689	0.4967	0.075	(0.090)
Top-soil depth	<25 cm	0.0181	0.1336	0.016	(0.024)
	25–50 cm	0.1145	0.3193	−0.076	(0.058)
	50–100 cm	0.2289	0.4214	0.193	(0.075)
	100–300 cm	0.0904	0.2876	0.135***	(0.051)
	>300 cm	0.5482	0.4991	−0.268***	(0.088)
Area share of various crops: 1956–1987					
Area share of rice		0.366	0.298	−0.194***	(0.054)
Area share of wheat		0.149	0.157	−0.058**	(0.026)
Area share of other cereals		0.205	0.172	0.128***	(0.031)
Area share of oilseeds		0.067	0.088	0.065***	(0.013)
Area share of cotton		0.041	0.096	0.066***	(0.018)
Area share of tobacco		0.003	0.015	0.005**	(0.002)
Area share of sugarcane		0.031	0.053	0.005	(0.008)
Cash crops-to-cereals ratio		0.149	0.257	0.152***	(0.048)
Demographics: 1961, 1971, 1981, 1991					
Log (Population)		14.26	0.634	−0.088	(0.109)
Population density		36.44	85.92	−11.22**	(4.02)
Proportion of scheduled castes		0.1598	0.0733	−0.034**	(0.014)
Proportion of scheduled tribes		0.0980	0.1630	−0.010	(0.031)
Proportion rural		0.8102	0.1237	−0.066***	(0.023)
Proportion of working population in farming		0.7119	0.1352	−0.050*	(0.027)

Notes: Standard errors in parentheses, corrected for district-level clustering. * Significant at 10-percent level; ** significant at 5-percent level; *** significant at 1-percent level. For the area under different crops and demographics, the difference is calculated after controlling for year fixed effects.

[a] Difference represents the average difference between non-landlord and landlord districts, computed as the regression coefficient on the non-landlord proportion.

making here is that the landlord districts did not start with a disadvantage.

B. *Differences in Agricultural Investments and Productivity*

We mainly investigate investment and productivity differences in the 1956–1985 period. Table 3 documents large and significant differences in measures of agricultural investments and productivity between landlord and non-landlord areas in the post-independence period. Each entry in this table represents the regression coefficient from a regression of the dependent variable on the non-landlord proportion, controlling for year fixed effects, geographical variables (latitude, altitude, mean annual rainfall, and soil types), length of British rule, and within-district clustering of errors. We show the detailed regression specification for one of the dependent variables (log agricultural yield) in Table 4 in the Web Appendix, listing the coefficients on all our control variables. Our base specification in column (1) shows that non-landlord districts have a 24-percent-higher proportion of irrigated area and 43-percent-higher levels of fertilizer use. They have a 27-percent-higher proportion of rice area and 18 percent more wheat area under high-yielding varieties.

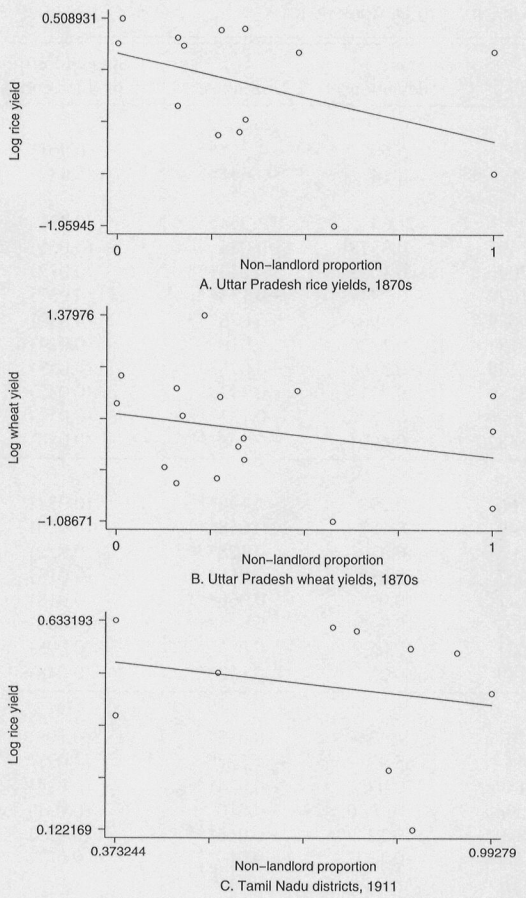


FIGURE 3. AGRICULTURAL YIELDS IN COLONIAL PERIOD

Overall agricultural yields are 16 percent higher, rice yields are 17 percent higher, and wheat yields are 23 percent higher. Further, column 2 shows that these differences are slightly bigger if we exclude the states of West Bengal and Bihar, the two states that have the highest proportion of landlord districts and the first to be conquered by the British. (We wanted to be sure that something idiosyncratic about these states was not driving our results.)

It is worth noting that these differences are driven neither by substitution away from agriculture in landlord districts nor by a greater shift toward crops other than rice or wheat. As we see in Table 2, landlord areas have a higher proportion of their working population engaged in farming, and they also devote a lower proportion of area to growing cash crops.

C. *Results Using Binary Measures of Non-Landlord Control*

Our results are robust to using a binary landlord/non-landlord classification rather than the continuous measure. We construct this classification as follows: a district is classified as "landlord" if it was under a landlord-based system, if it was under a landlord-based system and only partly converted to a different system (several districts of Madras), or if it was in Oudh, which we have argued had a higher proportion of landlords due to the reversal of policy after 1856. All other individual-based or village-based systems are classified as "non-landlord." Column 3 shows that our results are relatively robust to using a binary classification. A few coefficients are no longer significant here, probably because we are deliberately mismeasuring our regressor—some of the "non-landlord" districts in our binary classification nevertheless have large areas under landlords.[21] We also compute results using a more restricted sample: since we might not be fully sure of the classification of village-based districts, we exclude them and do a comparison of only landlord districts with individual-based districts. Some of the coefficients in this specification are larger than our base specification (column 4). This is probably because when we leave out the village-based districts, we are comparing almost wholly landlord areas with the other extreme, the individual-cultivator areas.

D. *Results Using Neighboring Districts*

Obviously, our interpretation of these results has to be tempered by the possibility that the non-landlord gap might reflect omitted variables. One strategy to control for possible omitted variables is to consider an extremely restricted sample: we consider only those districts that happen to be geographical neighbors (i.e., share a common border), but that

[21] In this classification, the "landlord" districts have at most 40 percent of land under non-landlord control, while some of our so-called "non-landlord" districts in fact have less than 20 percent of their land under non-landlord control. We have also tried an alternative specification where the binary variable takes the value one if the non-landlord proportion is greater than 0.5, and zero otherwise. Our results are robust to this specification as well (results not shown).

TABLE 3—DIFFERENCES IN AGRICULTURAL INVESTMENTS AND YIELDS
(*Mean non-landlord proportion = 0.5051 (s.d. = 0.4274)*)

		Coefficient on non-landlord proportion		Coefficient on non-landlord dummy	
Dependent variable	Mean of dependent variable	OLS Full sample (1)	OLS Excluding Bengal and Bihar (2)	OLS Full sample (3)	OLS Excluding village-based districts (4)
Agricultural investments					
Proportion of gross cropped area irrigated	0.276	0.065*	0.066*	0.077***	0.005
		(0.034)	(0.035)	(0.027)	(0.032)
Fertilizer use (kg/ha)	24.64	10.708***	10.992***	9.988***	10.695***
		(3.345)	(3.406)	(2.301)	(3.040)
Proportion of rice area under HYV	0.298	0.079*	0.094**	0.016	0.074*
		(0.044)	(0.043)	(0.032)	(0.038)
Proportion of wheat area under HYV	0.518	0.092**	0.119***	0.031	0.107**
		(0.046)	(0.045)	(0.036)	(0.052)
Proportion of other cereals area under HYV	0.196	0.057*	0.084***	−0.035	0.109***
		(0.031)	(0.024)	(0.025)	(0.041)
Agricultural productivity					
log (yield of 15 major crops)		0.157**	0.152**	0.173***	0.089
		(0.071)	(0.074)	(0.053)	(0.085)
log (rice yield)		0.171**	0.195**	0.099	0.173**
		(0.081)	(0.081)	(0.062)	(0.079)
log (wheat yield)		0.229***	0.228***	0.188***	0.143
		(0.067)	(0.070)	(0.054)	(0.098)
No. of districts		166	143	166	109
Year fixed effects		YES	YES	YES	YES
Geographic controls		YES	YES	YES	YES
Date of British land revenue control		YES	YES	YES	YES

Notes: Standard errors in parentheses, corrected for district-level clustering. * Significant at 10-percent level; ** significant at 5-percent level; *** significant at 1-percent level. Each cell represents the coefficient from a regression of the dependent variable on the measure of non-landlord control. Data are from 1956 to 1987. Data for area under high-yielding varieties (HYV) is after 1965. Geographic controls are altitude, latitude, mean annual rainfall, and dummies for soil type and coastal regions. The non-landlord dummy is assigned as follows: the dummy equals one for all individual-based districts and all village-based districts except those in Oudh. For landlord-based districts and the village-based districts of Oudh, the dummy is zero.

happened to have different historical land systems. (These districts and the historical reasons for their land system differences are listed in Table 5 of the Web Appendix.) We expect that there would be fewer differences in omitted variables, if any, in this sample of geographic neighbors than in our overall sample, and we verify that there are no significant differences in our observed geographic and demographic variables between these districts (results available upon request).

Even when we restrict our sample to this small set of 35 geographically neighboring districts, we still see large and significant differences between landlord and non-landlord districts in agricultural investments and outcomes (Table 4, panel A, column 1). In particular, total yields are 15 percent higher and wheat yields 25 percent higher in non-landlord areas than in landlord areas. These estimates are very close to the estimates in our base specification. The differences in fertilizer use and HYV adoption for wheat are also fairly close to the magnitudes obtained in our base specification. These results serve to confirm that our original results were not caused primarily by some unobserved district characteristics.

E. *Results Using Instrumental Variables*

As discussed above, our results might also be biased if the British decision regarding which land tenure system to adopt depended on other characteristics of the area in systematic ways. We would like to highlight a few facts in this

TABLE 4—ROBUSTNESS OF OLS RESULTS

Panel A: Robustness checks

Dependent variable	Coefficient on non-landlord proportion: OLS Neighbors only (1)	IV Full sample (2)
Agricultural investments		
Proportion of gross cropped area irrigated	0.101**	0.216
	(0.041)	(0.137)
Fertilizer use (kg/ha)	10.589**	26.198**
	(4.979)	(13.244)
Proportion of rice area under HYV	−0.015	0.411**
	(0.083)	(0.163)
Proportion of wheat area under HYV	0.078**	0.584***
	(0.034)	(0.163)
Proportion of other cereals area under HYV	−0.025	0.526***
	(0.024)	(0.129)
Agricultural productivity		
log (yield of 15 major crops)	0.145**	0.409
	(0.061)	(0.261)
log (rice yield)	0.126	0.554*
	(0.098)	(0.285)
log (wheat yield)	0.253***	0.706***
	(0.084)	(0.214)
No. of districts	35	166
Year fixed effects	YES	YES
Geographic controls	YES	YES
Date of British land revenue control	YES	YES

Panel B: First-stage regressions for IV
Dependent variable: Non-landlord proportion

Coefficient on	(1)	(2)	(3)
Instrument (=1 if date of British revenue control is between 1820 and 1856)	0.331***	0.430***	0.419***
	(0.086)	(0.092)	(0.087)
R-squared	0.40	0.43	0.63
No. of observations	166	166	166
Geographic controls	YES	YES	YES
Date of British land revenue control	YES	YES	YES
Date of British land revenue control squared	NO	YES	NO
State fixed effects	NO	NO	YES

Notes: Standard errors in parentheses, corrected for district-level clustering. * Significant at 10-percent level; ** significant at 5-percent level; *** significant at 1-percent level. Each cell in Panel A represents the coefficient from a regression of the dependent variable on the non-landlord proportion. Data are from 1956–1987. Data for area under high-yielding varieties (HYV) is after 1965. Geographic controls are altitude, latitude, mean annual rainfall, and dummies for soil type and coastal regions. Instrument is a dummy that equals one if the date of British revenue control is after 1820 and before 1856.

regard. First, we do not expect the choice of land tenure system to be very highly correlated with local district characteristics, since the choice of land tenure system was made for large contiguous areas at the same time and was often based on very little information regarding local conditions. Second, as explained in Section I C, places that were conquered earlier tended to have landlord-based systems. If British annexation policy was selectively directed toward the more productive places,[22] then landlord-controlled areas are likely to be inherently more productive. Third, *zamindari* areas were usually highly fertile areas which created enough rent to support a landlord-tenant-laborer hierarchy (Roy, 2000). In some areas, where landlord

[22] See Iyer (2005) for some evidence in support of this hypothesis.

defaults were excessive, these were later changed to different forms of settlement. Therefore, areas that ended up with non-landlord systems are more likely to be inherently less productive, or at least were less productive in colonial times. Another way to deal with this potential problem of omitted variables is to use an instrumental variables strategy. This has the additional advantage of helping us deal with the problem of measurement error in our non-landlord proportion variable, caused by district boundary changes and the fact that the historical record tends to be impressionistic (in any case, reflects the impression of one observer at one point of time).

Our instrumental variables strategy is based on the observation, mentioned in Section I, that areas that came under British revenue administration after 1820 have predominantly non-landlord systems, except for the policy reversal which occurred in Oudh (taken over in 1856) after the revolt of 1857. We believe that the source of this variation is in part due to the success of Munro and Mackenzie in establishing non-landlord systems in Madras and the North-West Provinces (starting around 1820), which created the all-important precedents that were followed in the districts conquered after 1820, as well as a broader shift in ideology in England. Therefore, the fact that areas conquered between 1820 and 1856 got non-landlord systems does not depend on the characteristics of the district, and a dummy for the date of conquest being between 1820 and 1856 is a valid instrument for the non-landlord proportion, especially after we control for the date of British conquest to take into account any direct effects of a longer period of British rule.[23]

Figure 4 demonstrates the basis for our instrumental variable strategy. In this figure, we plot the kernel regression of the non-landlord proportion and the mean log agricultural yield against the date of conquest. It is clear that there is a good fit in the shape of the two graphs and that both curves are highly nonlinear. Therefore, the co-movement in the two graphs is not driven by the fact that both are trending up or down, making it less likely that the relation between the two reflects the direct effect of the date of conquest. The figure also demonstrates that the non-landlord proportion is significantly higher for areas conquered between 1820 and 1856 compared to areas conquered earlier or later. This is exactly what we would have expected given the discussion above.[24] Panel B in Table 4 shows the first-stage coefficients of our IV strategy; we should note that the first-stage relationship remains significant even when we include a quadratic control for the length of British rule, as well as when we include state fixed effects.

Our IV results confirm that non-landlord systems indeed have a large and significant impact on current outcomes (Table 4, panel A, column 2). In fact, all the IV coefficients are larger than their OLS counterparts, although the difference between the two estimates is not statistically significant.[25] The standard errors for the IV estimates are also larger than the OLS standard errors, but the non-landlord effect remains statistically significant in the case of HYV adoption, as well as in fertilizer usage and wheat yields. Rice yields are significantly greater at the 10-percent level. Specifications involving a quadratic control for the length of British rule typically give coefficients that are smaller in magnitude, but generally of the same level of significance (results not shown).

The fact that the IV results are larger than the OLS results suggests that the OLS results are biased downward. This is the direction of bias we would have expected, given our discussion above, especially the fact that landlord areas, which were not productive enough to sustain a landlord class, tended to become non-landlord. It is also the direction of bias suggested by the presence of classical measurement error. Since our non-landlord variable is limited to being between 0 and

[23] By "date of conquest," we mean the date when the district came under British land revenue administration. The two dates are usually the same, with two exceptions. The first is the kingdom of Mysore, which was under British administration for the period when the land revenue systems were put in place, but was never part of the British Empire. The second is the kingdom of Nagpur, which was formally annexed in 1854, but had been under British revenue control in 1818.

[24] The other "hump" (or mode) on the left is mainly due to the districts of Madras Presidency, which were conquered fairly early, but which switched over to a non-landlord system after 1820.

[25] A Hausman test does not reject the null hypothesis that the OLS and IV coefficients are equal.

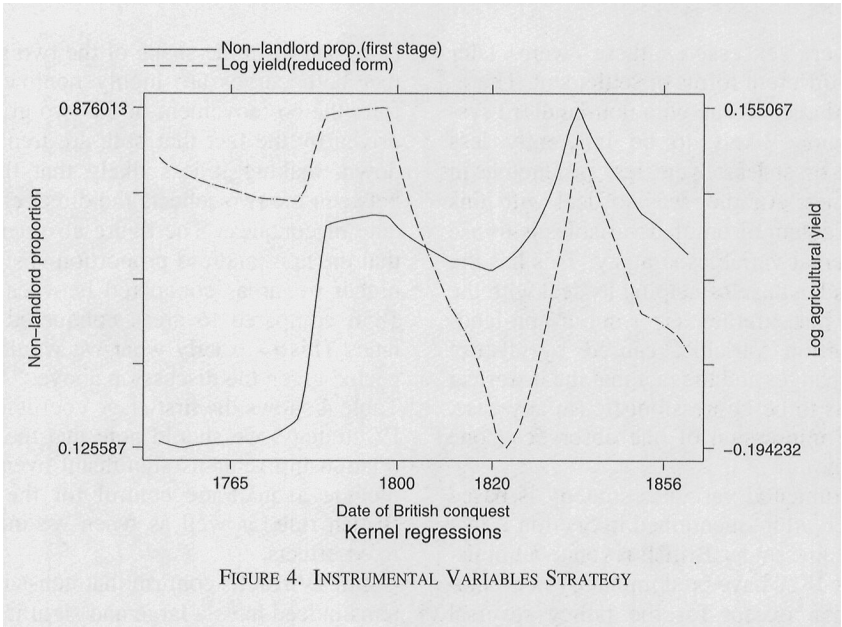


FIGURE 4. INSTRUMENTAL VARIABLES STRATEGY

1, however, we have nonclassical measurement error. Even then, for the special case of a binary regressor and no covariates, (nonclassical) measurement error will still bias the OLS coefficient downward, but will also bias the IV coefficient upward (Kane et al., 1999). For this special case, we verify that measurement error is not the only source of the difference between the OLS and IV estimates.[26]

Our IV results, together with the results on neighboring districts and the historical data, lead us to conclude that our OLS results are not biased upward due to omitted district characteristics. Because of the possibility of upward bias in the IV estimates, however, we will continue to treat the OLS results as benchmark estimates of the difference between landlord and non-landlord districts.

[26] We run the regressions with the binary regressor (defined in Section V C) and no covariates. If there were only measurement error, the OLS would be biased downward, the IV would be biased upward but have the same sign as the OLS coefficient, and the ratio of the two would be the same for all the outcome variables. We find that, of the eight outcome variables, the IV coefficient is larger than the OLS for five, the IV is smaller in magnitude than the OLS for one, and for the remaining two outcomes, the OLS coefficient is negative while the IV is positive. This suggests that measurement error is not the only problem.

F. *Does Land Tenure Have an Independent Effect on Productivity?*

We have established large and robust differences between landlord and non-landlord districts in terms of agricultural investments and productivity, with the non-landlord districts showing better performance in all of these measures. In Table 5, we argue that the differences in productivity are due largely to differences in investments. We do this by regressing productivity measures on the proportion of non-landlord control, as well as on the measures of investment. All the measures of investment (irrigation, fertilizer use, and adoption of HYV) are positive and strongly significant, as we would expect. The addition of these measures reduces the coefficient on the non-landlord proportion by 78 percent for total yields, 59 percent for rice yields, and 52 percent for wheat yields. The non-landlord variable is also no longer statistically significant.

G. *When Do the Differences Arise?*

As shown before, non-landlord districts were not more productive than landlord-based districts in the colonial period. Figure 5 indicates that the differences in investments (irrigation, fertilizer) and yields widen in the mid-1960s. Table 6 (panel A) formally estab-

TABLE 5—ARE YIELDS EXPLAINED BY INVESTMENTS?

	Dependent variables		
	Log total yield OLS (1)	Log rice yield OLS (2)	Log wheat yield OLS (3)
Proportion non-landlord	0.035	0.070	0.109
	(0.053)	(0.063)	(0.063)
Proportion of gross cropped area irrigated	0.693**	0.439**	0.435**
	(0.112)	(0.096)	(0.117)
Fertilizer use (kg/ha)	0.007**	0.004**	0.001
	(0.001)	(0.001)	(0.001)
Percent area under HYV	4.274**	0.580**	0.618**
	(1.122)	(0.063)	(0.070)
Adjusted *R*-squared	0.60	0.52	0.56
No. of districts	166	166	166
Year fixed effects	YES	YES	YES
Geographic controls	YES	YES	YES
Date of British land revenue control	YES	YES	YES

Notes: Standard errors in parentheses, corrected for district-level clustering. * Significant at 10-percent level; ** significant at 5-percent level; *** significant at 1-percent level. Data are from 1956–1987. Data for area under high-yielding varieties (HYV) is after 1965. Geographic controls are altitude, latitude, mean annual rainfall, and dummies for soil type and coastal regions.

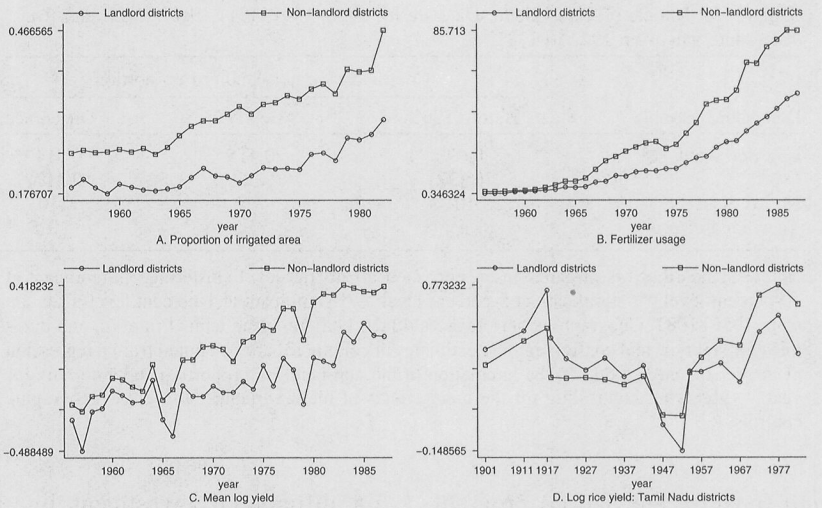


FIGURE 5. INVESTMENT AND PRODUCTIVITY TIME SERIES

lishes that the gap between landlord and non-landlord areas is larger after 1965 than in the 1956–1965 period. We also have data on rice yields for a limited sample of ten districts of Tamil Nadu from the colonial period onward. Figure 5D indicates that the non-landlord areas overtake the landlord areas during the mid-1960s. Table 6 (panel B) also checks this formally by computing the gap in the pre-1965 and post-1965 period.

VI. Why do the Landlord Districts Fall Behind?

The period after 1965 saw the state in India becoming much more active in rural areas, through the Intensive Rural Development Programs, the efforts to disseminate new high-yielding varieties of crops (resulting in the "Green Revolution"), and the building of public infrastructure (including fertilizer delivery systems) in rural areas under the 1971

TABLE 6—WHEN DO THE DIFFERENCES APPEAR?

Panel A: Full sample

Dependent variable	Coefficient on non-landlord proportion: 1956–1965 (1)	After 1965 (2)	Difference (3)
Agricultural investments			
Proportion of gross cropped area irrigated	0.046	0.079**	0.033**
	(0.033)	(0.036)	(0.016)
Fertilizer use (kg/ha)	1.026**	15.581***	14.55***
	(0.425)	(4.763)	(4.44)
Agricultural productivity			
log (yield of 15 major crops)	0.066	0.201***	0.135***
	(0.065)	(0.076)	(0.033)
log (rice yield)	0.108	0.196**	0.088**
	(0.069)	(0.089)	(0.044)
log (wheat yield)	0.146**	0.268***	0.122*
	(0.058)	(0.079)	(0.063)
No. of districts	166	166	166
Year fixed effects	YES	YES	YES
Geographic controls	YES	YES	YES
Date of British land revenue control	YES	YES	YES

Panel B: Rice yields for Tamil Nadu districts

Sample: 10 districts of Tamil Nadu. Data are for 1870, 1901, 1911, 1917, 1919, and five-yearly intervals from 1922 to 1982.

Dependent variable	Coefficient on non-landlord proportion: Before 1965	After 1965	Difference
Log rice yield	−0.099	0.415	0.514**
	(0.172)	(0.366)	(0.217)
No. of districts	10	10	10
Year fixed effects	YES	YES	YES

Notes: Standard errors in parentheses, corrected for district-level clustering. * Significant at 10-percent level; ** significant at 5-percent level; *** significant at 1-percent level. Data are from 1956–1987. Geographic controls are altitude, latitude, mean annual rainfall, and dummies for soil type and coastal regions. Estimates in column (3) are computed from a regression of the dependent variable on the interaction of the non-landlord proportion and a dummy for year >1965, after controlling for the main effects of these variables, as well as geographic controls.

Garibi Hatao (poverty alleviation) program. As we have seen, the landlord areas were slower in the adoption of high-yielding varieties. They also seem to have benefited less from the growth in public investment in irrigation, though our numbers do not distinguish between public and private irrigation facilities. Why were landlord areas unable to take advantage of the new opportunities that presented themselves after the mid-1960s? We discussed some potential answers in Section II, and we assess their empirical relevance here.

Of the three alternative classes of explanations discussed earlier, the explanation based on differential investment by the colonial state is probably the least compelling, given that the major differences between the landlord and non-landlord areas arose after 1965 (Table 6). In principle, one could still argue that the advantage they got from these early public investments continues to help them in the post-independence period.[27] The fact that

[27] Tirthankar Roy (2002) makes the argument that the areas that gained from the Green Revolution were those that showed improvements during the colonial period as well.

the main source of the non-landlord advantage does not come from the *mahalwari* districts of northern India,[28] which were the main beneficiaries of the canal construction during the colonial period, makes it harder to believe that this is the source of the entire difference.

We noted in Section II that the landlord-controlled areas had higher levels of land inequality in the colonial period. It therefore comes as no surprise that the major landlord-dominated states enacted an average of 6.5 land reform measures in the period between 1957 and 1992, while non-landlord states had an average of 3.5.[29] Besley and Burgess (2000) report that states that enacted a larger number of land reforms had a somewhat greater decline in the Gini coefficient of land inequality. This does not mean that there has been complete convergence in the land distribution in the two areas. As late as 1990, 64 percent of all land holdings in landlord areas were classified as "marginal" (less than one hectare), which is about eight percentage points higher than the corresponding figure in non-landlord areas.[30] Further, 48 percent of all holdings are small to medium sized (one to ten hectares) in individual-based areas, but only 35 percent in landlord areas. There is no significant difference in the proportion of extremely large holdings, which is probably due to the impact of land ceiling laws passed after independence.

These differences in the land distribution, however, cannot explain our results. For instance, if we were to ascribe the entire difference of 16 percent in agricultural yields to the fact that there are more marginal holdings in landlord areas, on the grounds that these holdings are less productive because they underinvest, we would have to accept that the small holdings are only about 12 percent as productive as larger holdings, which seems implausibly low.[31] This also contradicts the evidence from developing countries, which suggests that small farms are, if anything, more productive than large farms (Binswanger and Rosenzweig, 1986). Further, our results do not change when we control directly for the Gini coefficient of land holdings in 1971 or the number of land reforms passed by the state. If we use consumption inequality as a better measure of wealth inequality, we find that landlord areas show significantly larger declines in consumption inequality between 1972 and 1987 than non-landlord areas (Table 6 in the Web Appendix). In fact, by 1987 the landlord districts show significantly lower consumption inequality.[32]

We therefore feel that the biggest piece of the story is probably the differences in the political environment. If the effect of the political environment operated mainly through the insecurity of peasant property in the landlord areas, however, we would have observed convergence rather than divergence after independence, since peasant property clearly became less insecure once the landlords lost their formal authority. This suggests that the important difference in the political environment probably has to do with the nature of collective action in the two areas. We find that in addition to placing a greater emphasis on land reform measures, states with a higher proportion of landlord areas spent less on development expenditure. Between 1960 and 1965, the landlord states spent 13 rupees per capita on development expenditure, compared to 19 rupees in the non-landlord states. This spending gap is higher in the post-

[28] Table 3, column 4, shows that leaving them out makes the non-landlord coefficient larger for some of the outcomes.

[29] Data on state-level land reforms comes from Besley and Burgess (2000). We classify Bihar, Madhya Pradesh, Orissa, Rajasthan, Uttar Pradesh, and West Bengal as "landlord" states, and Andhra Pradesh, Assam, Gujarat, Karnataka, Kerala, Maharashtra, Punjab, and Tamil Nadu as "non-landlord" states.

[30] The difference of eight percentage points is obtained by regressing the proportion of marginal (less than one hectare) holdings on the non-landlord proportion, after controlling for geographic variables.

[31] Suppose small farms are δ times as productive as large farms, z is the share of small farms and total productivity is simply the sum of large farm and small farm productivity. Then the percentage productivity difference between non-landlord and landlord areas equals $\{[(1-\delta)\Delta z]/[1-(1-\delta)z_{landlord}]\}$. Using productivity difference $= 0.16$, $\Delta z = 0.08$ and $z_{landlord} = 0.64$, we obtain $\delta \approx 0.12$.

[32] These measures are computed using household survey data from the National Sample Surveys (NSS). We should keep in mind that these data are not at the district level but at the NSS region level, usually consisting of three to ten districts. Our standard errors for these regressions are clustered at the NSS region level to take care of this aspect of our data.

1965 period, just when new technologies were appearing in the agricultural sector: landlord states spent 29 rupees per capita, while the non-landlord states spent a much higher 49 rupees per capita (Table 7 in the Appendix). This is not simply because of lack of resources: development expenditure as a proportion of state domestic product is also lower in the landlord states, and the difference in per capita spending persists even after controlling for state domestic product per capita (Appendix Table 7, column 3). Given that the difference in the number of land reforms is also mainly from the post-1965 period, one way to characterize the difference in the nature of public action is to say that landlord-dominated states were busy carrying out land reform exactly when the non-landlord states started focusing on development.

This difference in public spending turns out to be important in explaining our results. When we add development expenditure per capita as an explanatory variable in our base regressions, we find that it sharply reduces the magnitude of the non-landlord coefficient for the measures of HYV adoption (Table 7, column 2). The idea of state policy priorities as the major channel of influence is consistent with what we find when we estimate the investment and yield equations after including a fixed effect for each state. This reduces the estimated coefficient on the non-landlord share substantially (by 50 percent or so), though the signs are unaltered and several remain significant (Table 7, column 3).[33] The differences in state policies are also reflected in the substantial difference between landlord and non-landlord areas in the provision of educational and health facilities: landlord areas had 21 percent fewer villages (15 percentage points) equipped with primary schools, while the gap in middle school and high school availability are 61 percent and 63 percent, respectively. Given these differences in investments, it is not surprising that literacy rates are 5 percentage points higher in non-landlord areas, while infant mortality rates are 40 percent lower; both these differences are statistically significant (Table 7, panel D).[34] A large part of these differences can be attributed to the difference in state development expenditure (column 2).

Why are the political priorities so different in these two areas? As already suggested in Section II, the masses in the landlord areas, with their memories of an oppressive and often absentee landlord class, may perceive their interests as being opposed to that of the local elite, while those in the non-landlord areas may be more interested in working with that elite. The existence of a highly conflictual environment is consistent with our results on crime rates (Appendix, Table 8). Landlord districts have significantly higher levels of violent crime (such as murder, rape, kidnap, armed robbery, and riots), but not of economic crimes like cheating or counterfeiting.

The perception of a large divergence of interests between the masses and the elite in landlord areas may not, however, be necessarily correct. The final empirical exercise in this paper compares poverty reduction in the landlord and non-landlord areas. While the head count ratio falls in both areas between 1972 and 1987 (the mean reduction is about 11 percentage points), the decline in poverty according to our OLS estimates is about seven percentage points higher in non-landlord areas (Appendix, Table 6). The difference in poverty reduction is five percentage points for the sample of neighboring districts and is robust to the inclusion of a state fixed effect. The IV estimate, however, is completely insignificant and has the opposite sign. In sum, there is no evidence that the masses fare better in the landlord areas, and there is some evidence that they do worse. If we were prepared to attribute the change in poverty to the differences in political priorities and the resulting differences in policies, these results would suggest that the masses could perhaps have done a little better, or at least no worse, by focusing on what they had in common with the elites.

[33] We need to be a little cautious when interpreting these results. Adding state fixed effects effectively drops the states that have no within-state variation in non-landlord proportion. These states (Bihar, Gujarat, Karnataka, Rajasthan, and West Bengal) account for about one-fourth of our sample, so putting in state fixed effects results in a lack of power in our estimation.

[34] IV estimates of these differences are larger in magnitude than the OLS estimates for literacy, infant mortality, and primary school provision (results not shown).

TABLE 7—IMPACT OF STATE POLICY

Dependent variables	Mean of dependent variable	Coefficient on non-landlord proportion: OLS Base specification (1)	OLS Control for state dev exp per capita (2)	OLS State FE (3)
Panel A: Agricultural investments				
Proportion of gross cropped area irrigated	0.276	0.065*	0.074**	0.028
		(0.034)	(0.035)	(0.036)
Fertilizer use (kg/ha)	24.64	10.708***	10.805***	4.297
		(3.345)	(3.717)	(3.308)
Proportion of rice area under HYV	0.298	0.079*	0.007	0.000
		(0.044)	(0.040)	(0.042)
Proportion of wheat area under HYV	0.518	0.092**	0.061	0.028
		(0.046)	(0.047)	(0.039)
Proportion of other cereals area under HYV	0.196	0.057*	0.025	0.043*
		(0.031)	(0.030)	(0.026)
Panel B: Agricultural productivity				
log (yield of 15 major crops)		0.157**	0.174**	0.059
		(0.071)	(0.076)	(0.072)
log (rice yield)		0.171**	0.083	0.016
		(0.081)	(0.082)	(0.078)
log (wheat yield)		0.229***	0.243***	0.150***
		(0.067)	(0.072)	(0.045)
Panel C: Education and health investments, 1981				
Proportion of villages having:				
Primary school	0.745	0.154***	0.062*	0.102***
		(0.036)	(0.037)	(0.039)
Middle school	0.204	0.125***	0.093***	0.064***
		(0.023)	(0.021)	(0.018)
High school	0.082	0.052***	0.019	0.030**
		(0.018)	(0.014)	(0.013)
Primary health center	0.023	0.011***	0.002	0.012***
		(0.004)	(0.004)	(0.004)
Primary health subcenter	0.031	0.033***	0.011	0.006
		(0.011)	(0.009)	(0.006)
Panel D: Education and health outcomes				
Literacy rate (1961, 1971, 1981, 1991)	0.2945	0.0524**	0.0290*	0.0241
		(0.0190)	(0.0171)	(0.0176)
Infant mortality rate (1991)	82.17	−32.71***	−25.43***	−15.81***
		(5.38)	(5.28)	(5.40)
State fixed effects		NO	NO	YES
Year fixed effects		YES	YES	YES
Geographic controls		YES	YES	YES
Date of British land revenue control		YES	YES	YES

Notes: Standard errors in parentheses, corrected for district-level clustering. * Significant at 10-percent level; ** significant at 5-percent level; *** significant at 1-percent level. Geographic controls are altitude, latitude, mean annual rainfall, and dummies for soil type and coastal regions.

REFERENCES

Acemoglu, Daron; Johnson, Simon and Robinson, James A. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, 2001, *91*(5), pp. 1369–1401.

Acemoglu, Daron; Johnson, Simon and Robinson, James A. "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution." *Quarterly Journal of Economics*, 2002, *117*(4), pp. 1231–94.

Alesina, Alberto and Rodrik, Dani. "Distributive Politics and Economic Growth." *Quarterly Journal of Economics*, 1994, *109*(2), pp. 465–90.

Baden-Powell, Baden H. *The land-systems of*

British India, 3 Volumes. Oxford: Clarendon Press, 1892.

Bagchi, Amiya K. "Reflections on Patterns of Regional Growth in India under British Rule." *Bengal Past and Present*, 1976, *95*(1), pp. 247–89.

Banerjee, Abhijit V.; Gertler, Paul J. and Ghatak, Maitreesh. "Empowerment and Efficiency: Tenancy Reform in West Bengal." *Journal of Political Economy*, 2002, *110*(2), pp. 239–80.

Banerjee, Abhijit V. and Newman, Andrew F. "Occupational Choice and the Process of Development." *Journal of Political Economy*, 1993, *101*(2), pp. 274–98.

Besley, Timothy and Burgess, Robin. "Land Reform, Poverty Reduction, and Growth: Evidence from India." *Quarterly Journal of Economics*, 2000, *115*(2), pp. 389–430.

Binswanger, Hans P. and Rosenzweig, Mark R. "Behavioural and Material Determinants of Production Relations in Agriculture." *Journal of Development Studies*, 1986, *22*(3), pp. 503–39.

Bose, Sugato. *Peasant labour and colonial capital: Rural Bengal since 1770*. Cambridge: Cambridge University Press, 1993.

Brass, Paul R. *The Politics of India since independence*. Cambridge: Cambridge University Press, 1994.

Chatterjee, Partha. *Bengal 1920–1947, Vol. 1: The land question*. Calcutta: K. P. Bagchi and South Asia Books, 1984.

Engerman, Stanley L. and Sokoloff, Kenneth L. "Factor Endowments, Institutions, and Differential Paths of Growth among New World Economies: A View from Economic Historians of the United States," in Steven Haber, ed., *How Latin America fell behind: Essays on the economic histories of Brazil and Mexico, 1800–1914*. Stanford: Stanford University Press, 1997, pp. 260–304.

Engerman, Stanley L. and Sokoloff, Kenneth L. "Factor Endowments, Inequality, and Paths of Development among New World Economies." *Economia: Journal of the Latin American and Caribbean Economic Association*, 2002, *3*(1), pp. 41–88.

Galor, Oded and Zeira, Joseph. "Income Distribution and Macroeconomics." *Review of Economic Studies*, 1993, *60*(1), pp. 35–52.

Gough, Kathleen. "Indian Peasant Uprisings." *Economic and Political Weekly*, 1974, *9*(13), pp. 1391–1412.

Guha, Ranajit. *A rule of property for Bengal: An essay on the idea of permanent settlement*. Paris: Mouton & Co., 1963.

Gupta, Rai M. N. *Land system of Bengal*. Calcutta: University of Calcutta, 1940.

Iyer, Lakshmi. "The Long-Term Impact of Colonial Rule: Evidence from India." Harvard University, Harvard Business School Working Papers: No 05-041, 2005.

Kane, Thomas J.; Rouse, Cecelia Elena and Staiger, Douglas. "Estimating Returns to Schooling When Schooling Is Misreported." National Bureau of Economic Research, Inc., NBER Working Papers: No. 7235, 1999.

Kumar, Dharma. *The Cambridge economic history of India, Vol. 2:c. 1757-c. 1970*. Cambridge: Cambridge University Press, 1982.

La Porta, Rafael; López de Silanes, Florencio; Shleifer, Andrei and Vishny, Robert. "Law and Finance." *Journal of Political Economy*, 1998, *106*(6), pp. 1113–55.

La Porta, Rafael; López de Silanes, Florencio; Shleifer, Andrei and Vishny, Robert. "The Quality of Government." *Journal of Law, Economics, and Organization*, 1999, *15*(1), pp. 222–79.

La Porta, Rafael; López de Silanes, Florencio; Shleifer, Andrei and Vishny, Robert. "Investor Protection and Corporate Governance." *Journal of Financial Economics*, 2000, *58*(1-2), pp. 3–27.

Lin, Justin Y. "Rural Reforms and Agricultural Growth in China." *American Economic Review*, 1992, *82*(1), pp. 34–51.

Misra, Babu Ram. *Land revenue policy in the united provinces under British rule*. Benares: Nand Kishore & Brothers, 1942.

Mukherjee, Nilmani. *The Ryotwari system in Madras 1792–1827*. Calcutta: Firma K. L. Mukhopadhyay, 1962.

Patel, Govindlal Dalsukhbhai *The land problem of re-organized Bombay State*. Bombay: N.M. Tripathi Pvt. Ltd., 1957.

Persson, Torsten and Tabellini, Guido. "Is Inequality Harmful for Growth?" *American Economic Review*, 1994, *84*(3), pp. 600–21.

Porter, F. W. *Final settlement report of the Allahabad district*. Allahabad: North-Western Provinces and Oudh Government Press, 1878.

Ray, Ratnalekha. *Change in Bengal agrarian society 1760–1850*. New Delhi: Manohar, 1979.

Roy, Tirthankar. *The economic history of India, 1857–1947.* Oxford: Oxford University Press, 2000.

Roy, Tirthankar. "Economic History and Modern India: Redefining the Link." *Journal of Economic Perspectives*, 2002, *16*(3), pp. 109–30.

Smith, W. H. *Final report on the revision of settlement in the district of Aligarh.* Allahabad: North-Western Provinces and Oudh Government Press, 1882.

Sokoloff, Kenneth L. and Engerman, Stanley L. "Institutions, Factor Endowments, and Paths of Development in the New World." *Journal of Economic Perspectives*, 2000, *14*(3), pp. 217–32.

Stokes, Eric. *The English utilitarians and India.* Oxford: Clarendon Press, 1959.

Stokes, Eric. "The Land Revenue Systems of the North-Western Provinces and Bombay Deccan 1830–1948: Ideology and the Official Mind." in Eric Stokes, ed., *The peasant and the Raj: Studies in agrarian society and peasant rebellion in colonial India.* Cambridge: Cambridge University Press, 1978a.

Stokes, Eric. "The Structure of Landholding in Uttar Pradesh 1860–1948." in Eric Stokes, ed., *The peasant and the Raj: Studies in agrarian society and peasant rebellion in colonial India.* Cambridge: Cambridge University Press, 1978b, ch. 9.

Yanagisawa, Haruka. *A century of change: Caste and irrigated lands in Tamil Nadu 1860's–1870's.* New Delhi: Manohar, 1996.

Econometrica, Vol. 78, No. 6 (November, 2010), 1863–1903

THE PERSISTENT EFFECTS OF PERU'S MINING *MITA*

BY MELISSA DELL[1]

This study utilizes regression discontinuity to examine the long-run impacts of the *mita*, an extensive forced mining labor system in effect in Peru and Bolivia between 1573 and 1812. Results indicate that a *mita* effect lowers household consumption by around 25% and increases the prevalence of stunted growth in children by around 6 percentage points in subjected districts today. Using data from the Spanish Empire and Peruvian Republic to trace channels of institutional persistence, I show that the *mita*'s influence has persisted through its impacts on land tenure and public goods provision. *Mita* districts historically had fewer large landowners and lower educational attainment. Today, they are less integrated into road networks and their residents are substantially more likely to be subsistence farmers.

KEYWORDS: Forced labor, land tenure, public goods.

1. INTRODUCTION

THE ROLE OF HISTORICAL INSTITUTIONS in explaining contemporary underdevelopment has generated significant debate in recent years.[2] Studies find quantitative support for an impact of history on current economic outcomes (Nunn (2008), Glaeser and Shleifer (2002), Acemoglu, Johnson, and Robinson (2001, 2002), Hall and Jones (1999)), but have not focused on channels of persistence. Existing empirical evidence offers little guidance in distinguishing a variety of potential mechanisms, such as property rights enforcement, inequality, ethnic fractionalization, barriers to entry, and public goods. This paper uses variation in the assignment of an historical institution in Peru to identify land tenure and public goods as channels through which its effects persist.

Specifically, I examine the long-run impacts of the mining *mita*, a forced labor system instituted by the Spanish government in Peru and Bolivia in 1573 and abolished in 1812. The *mita* required over 200 indigenous communities to send one-seventh of their adult male population to work in the Potosí silver and Huancavelica mercury mines (Figure 1). The contribution of *mita* conscripts changed discretely at the boundary of the subjected region: on one side, all communities sent the same percentage of their population, while on the other side, all communities were exempt.

[1]I am grateful to Daron Acemoglu, Bob Allen, Josh Angrist, Abhijit Banerjee, John Coatsworth, David Cook, Knick Harley, Austin Huang, Nils Jacobsen, Alan Manning, Ben Olken, James Robinson, Peter Temin, Gary Urton, Heidi Williams, Jeff Williamson, and seminar participants at City University of Hong Kong, Chinese University of Hong Kong, Harvard, MIT, Oxford, Stanford Institute of Theoretical Economics, and Warwick for helpful comments and suggestions. I also thank Javier Escobal and Jennifer Jaw for assistance in accessing data. Research funding was provided by the George Webb Medley Fund (Oxford University).

[2]See, for example, Coatsworth (2005), Glaeser et al. (2004), Easterly and Levine (2003), Acemoglu, Johnson, and Robinson (2001, 2002), Sachs (2001), and Engerman and Sokoloff (1997).

 DOI: 10.3982/ECTA8121

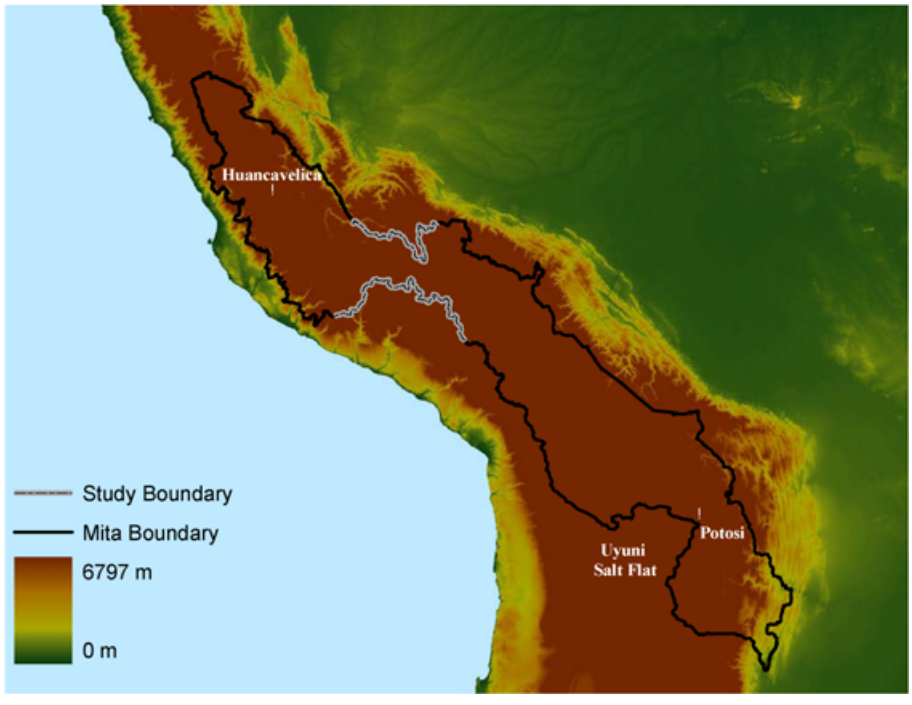


FIGURE 1.—The *mita* boundary is in black and the study boundary in light gray. Districts falling inside the contiguous area formed by the *mita* boundary contributed to the *mita*. Elevation is shown in the background.

This discrete change suggests a regression discontinuity (RD) approach for evaluating the long-term effects of the *mita*, with the *mita* boundary forming a multidimensional discontinuity in longitude–latitude space. Because validity of the RD design requires all relevant factors besides treatment to vary smoothly at the *mita* boundary, I focus exclusively on the portion that transects the Andean range in southern Peru. Much of the boundary tightly follows the steep Andean precipice, and hence has elevation and the ethnic distribution of the population changing discretely at the boundary. In contrast, elevation, the ethnic distribution, and other observables are statistically identical across the segment of the boundary on which this study focuses. Moreover, specification checks using detailed census data on local tribute (tax) rates, the allocation of tribute revenue, and demography—collected just prior to the *mita*'s institution in 1573—do not find differences across this segment. The multidimensional nature of the discontinuity raises interesting and important questions about how to specify the RD polynomial, which will be explored in detail.

Using the RD approach and household survey data, I estimate that a long-run *mita* effect lowers equivalent household consumption by around 25% in

subjected districts today. Although the household survey provides little power for estimating relatively flexible models, the magnitude of the estimated *mita* effect is robust to a number of alternative specifications. Moreover, data from a national height census of school children provide robust evidence that the *mita*'s persistent impact increases childhood stunting by around 6 percentage points in subjected districts today. These baseline results support the well known hypothesis that extractive historical institutions influence long-run economic prosperity (Acemoglu, Johnson, and Robinson (2002)). More generally, they provide microeconomic evidence consistent with studies establishing a relationship between historical institutions and contemporary economic outcomes using aggregate data (Nunn (2008), Banerjee and Iyer (2005), Glaeser and Shleifer (2002)).

After examining contemporary living standards, I use data from the Spanish Empire and Peruvian Republic, combined with the RD approach, to investigate channels of persistence. Although a number of channels may be relevant, to provide a parsimonious yet informative picture, I focus on three that the historical literature and fieldwork highlight as important. First, using district-level data collected in 1689, I document that *haciendas*—rural estates with an attached labor force—developed primarily outside the *mita* catchment. At the time of the *mita*'s enactment, a landed elite had not yet formed. To minimize the competition the state faced in accessing scarce *mita* labor, colonial policy restricted the formation of *haciendas* in *mita* districts, promoting communal land tenure instead (Garrett (2005), Larson (1988)). The *mita*'s effect on *hacienda* concentration remained negative and significant in 1940. Second, econometric evidence indicates that a *mita* effect lowered education historically, and today *mita* districts remain less integrated into road networks. Finally, data from the most recent agricultural census provide evidence that a long-run *mita* impact increases the prevalence of subsistence farming.

Based on the quantitative and historical evidence, I hypothesize that the long-term presence of large landowners in non-*mita* districts provided a stable land tenure system that encouraged public goods provision. The property rights of large landowners remained secure from the 17th century onward. In contrast, the Peruvian government abolished the communal land tenure that had predominated in *mita* districts soon after the *mita* ended, but did not replace it with a system of enforceable peasant titling (Jacobsen (1993), Dancuart and Rodriguez (1902, Vol. 2, p. 136)). As a result, extensive confiscation of peasant lands, numerous responding peasant rebellions as well as banditry and livestock rustling were concentrated in *mita* districts during the late 19th and 20th centuries (Jacobsen (1993), Bustamante Otero (1987, pp. 126–130), Flores Galindo (1987, p. 240), Ramos Zambrano (1984, pp. 29–34)). Because established landowners in non-*mita* districts enjoyed more secure title to their property, it is probable that they received higher returns from investing in public goods. Moreover, historical evidence indicates that well established landowners possessed the political connections required to secure public goods

(Stein (1980)). For example, the *hacienda* elite lobbied successfully for roads, obtaining government funds for engineering expertise and equipment, and organizing labor provided by local citizens and *hacienda* peons (Stein (1980, p. 59)). These roads remain and allow small-scale agricultural producers to access markets today, although *haciendas* were subdivided in the 1970s.

The positive association between historical *haciendas* and contemporary economic development contrasts with the well known hypothesis that historically high land inequality is the fundamental cause of Latin America's poor long-run growth performance (Engerman and Sokoloff (1997)). Engerman and Sokoloff argued that high historical inequality *lowered* subsequent investments in public goods, leading to worse outcomes in areas of the Americas that developed high land inequality during the colonial period. This theory's implicit counterfactual to large landowners is secure, enfranchised smallholders of the sort that predominated in some parts of North America. This is not an appropriate counterfactual for Peru or many other places in Latin America, because institutional structures largely in place before the formation of the landed elite did not provide secure property rights, protection from exploitation, or a host of other guarantees to potential smallholders.[3] The evidence in this study indicates that large landowners—while they did not aim to promote economic prosperity for the masses—did shield individuals from exploitation by a highly extractive state and ensure public goods. Thus, it is unclear whether the Peruvian masses would have been better off if initial land inequality had been lower, and it is doubtful that initial land inequality is the most useful foundation for a theory of long-run growth. Rather, the Peruvian example suggests that exploring constraints on how the state can be used to shape economic interactions, for example, the extent to which elites can employ state machinery to coerce labor or citizens can use state guarantees to protect their property, could provide a particularly useful starting point for modeling Latin America's long-run growth trajectory.

In the next section, I provide an overview of the *mita*. Section 3 discusses identification and tests whether the *mita* affects contemporary living standards. Section 4 examines channels empirically. Finally, Section 5 offers concluding remarks.

2. THE MINING *MITA*

2.1. *Historical Introduction*

The Potosí mines, discovered in 1545, contained the largest deposits of silver in the Spanish Empire, and the state-owned Huancavelica mines provided the

[3]This argument is consistent with evidence on long-run inequality from other Latin American countries, notably Acemoglu et al. (2008) on Cundinamarca and Colombia and Coatsworth (2005) on Mexico.

mercury required to refine silver ore. Beginning in 1573, indigenous villages located within a contiguous region were required to provide one-seventh of their adult male population as rotating *mita* laborers to Potosí or Huancavelica, and the region subjected remained constant from 1578 onward.[4] The *mita* assigned 14,181 conscripts from southern Peru and Bolivia to Potosí and 3280 conscripts from central and southern Peru to Huancavelica (Bakewell (1984, p. 83)).[5] Using population estimates from the early 17th century (Cook (1981)), I calculate that around 3% of adult males living within the current boundaries of Peru were conscripted to the *mita* at a given point in time. The percentage of males who at some point participated was considerably higher, as men in subjected districts were supposed to serve once every 7 years.[6]

Local native elites were responsible for collecting conscripts, delivering them to the mines, and ensuring that they reported for mine duties (Cole (1985, p. 15), Bakewell (1984)). If community leaders were unable to provide their allotment of conscripts, they were required to pay in silver the sum needed to hire wage laborers instead. Historical evidence suggests that this rule was strictly enforced (Garrett (2005, p. 126), Cole (1985, p. 44), Zavala (1980), Sanchez-Albornoz (1978)). Some communities did commonly meet *mita* obligations through payment in silver, particularly those in present-day Bolivia who had relatively easy access to coinage due to their proximity to Potosí (Cole (1985)). Detailed records of *mita* contributions from the 17th, 18th, and early 19th centuries indicate that communities in the region that this paper examines contributed primarily in people (Tandeter (1993, pp. 56, 66), Zavala (1980, Vol. II, pp. 67–70)). This is corroborated by population data collected in a 1689 parish census (Villanueva Urteaga (1982)), described in the Supplemental Material (Dell (2010)), which shows that the male–female ratio was 22% lower in *mita* districts (a difference significant at the 1% level).[7]

[4]The term *mita* was first used by the Incas to describe the system of labor obligations, primarily in local agriculture, that supported the Inca state (D'Altoy (2002, p. 266), Rowe (1946, pp. 267–269)). While the Spanish coopted this phrase, historical evidence strongly supports independent assignment. Centrally, the Inca *m'ita* required every married adult male in the Inca Empire (besides leaders of large communities), spanning an area far more extensive than the region I examine, to provide several months of labor services for the state each year (D'Altoy (2002, p. 266), Cieza de León (1551)).

[5]Individuals could attempt to escape *mita* service by fleeing their communities, and a number pursued this strategy (Wightman (1990)). Yet fleeing had costs: giving up access to land, community, and family; facing severe punishment if caught; and either paying additional taxes in the destination location as a "foreigner" (*forastero*) or attaching oneself to a *hacienda*.

[6]*Mita* districts contain 17% of the Peruvian population today (Instituto Nacional de Estadística e Información de Perú (INEI) (1993)).

[7]While colonial observers highlighted the deleterious effects of the *mita* on demography and well-being in subjected communities, there are some features that could have promoted relatively better outcomes. For example, *mita* conscripts sold locally produced goods in Potosí, generating trade linkages.

With silver deposits depleted, the *mita* was abolished in 1812, after nearly 240 years of operation. Sections 3 and 4 discuss historical and empirical evidence showing divergent histories of *mita* and non-*mita* districts.

2.2. *The Mita's Assignment*

Why did Spanish authorities require only a portion of districts in Peru to contribute to the *mita* and how did they determine which districts to subject? The aim of the Crown was to revive silver production to levels attained using free labor in the 1550s, before epidemic disease had substantially reduced labor supply and increased wages. Yet coercing labor imposed costs: administrative and enforcement costs, compensation to conscripts for traveling as much as 1000 kilometers (km) each way to and from the mines, and the risk of decimating Peru's indigenous population, as had occurred in earlier Spanish mining ventures in the Caribbean (Tandeter (1993, p. 61), Cole (1985, pp. 3, 31), Cañete (1794), Levillier (1921, Vol. 4, p. 108)). To establish the minimum number of conscripts needed to revive production to 1550s levels, Viceroy Francisco Toledo commissioned a detailed inventory of mines and production processes in Potosí and elsewhere in 1571 (Bakewell (1984, pp. 76–78), Levillier (1921, Vol. 4)). These numbers were used, together with census data collected in the early 1570s, to enumerate the *mita* assignments. The limit that the *mita* subject no more than one-seventh of a community's adult male population at a given time was already an established rule that regulated local labor drafts in Peru (Glave (1989)). Together with estimates of the required number of conscripts, this rule roughly determined what fraction of Andean Peru's districts would need to be subjected to the *mita*.

Historical documents and scholarship reveal two criteria used to assign the *mita*: distance to the mines at Potosí and Huancavelica and elevation. Important costs of administering the *mita*, such as travel wages and enforcement costs, were increasing in distance to the mines (Tandeter (1993, p. 60), Cole (1985, p. 31)). Moreover, Spanish officials believed that only highland peoples could survive intensive physical labor in the mines, located at over 4000 meters (13,000 feet) (Golte (1980)). The geographic extent of the *mita* is consistent with the application of these two criteria, as can be seen in Figure 1.[8] This study focuses on the portion of the *mita* boundary that transects the Andean range, which this figure highlights in white, and the districts along this portion are termed the study region (see Supplemental Material Figure A1 for a detailed view). Here, exempt districts were those located farthest from

[8]An elevation constraint was binding along the eastern and western *mita* boundaries, which tightly follow the steep Andean precipice. The southern Potosí *mita* boundary was also constrained, by the border between Peru and the Viceroyalty of Rio de la Plata (Argentina), and by the geographic divide between agricultural lands and an uninhabitable salt flat.

the mining centers given road networks at the time (Hyslop (1984)).[9] While historical documents do not mention additional criteria, concerns remain that other underlying characteristics may have influenced *mita* assignment. This will be examined further in Section 3.2.

3. THE *MITA* AND LONG-RUN DEVELOPMENT

3.1. *Data*

I examine the *mita*'s long-run impact on economic development by testing whether it affects living standards today. A list of districts subjected to the *mita* is obtained from Saignes (1984) and Amat y Junient (1947) and matched to modern districts as detailed in the Supplemental Material, Table A.I. Peruvian districts are in most cases small political units that consist of a population center (the district capital) and its surrounding countryside. *Mita* assignment varies at the district level.

I measure living standards using two independent data sets, both georeferenced to the district. Household consumption data are taken from the 2001 Peruvian National Household Survey (Encuesta Nacional de Hogares (ENAHO)) collected by the National Institute of Statistics (INEI). To construct a measure of household consumption that reflects productive capacity, I subtract the transfers received by the household from total household consumption and normalize to Lima metropolitan prices using the deflation factor provided in ENAHO. I also utilize a microcensus data set, obtained from the Ministry of Education, that records the heights of all 6- to 9-year-old school children in the region. Following international standards, children whose heights are more than 2 standard deviations below their age-specific median are classified as stunted, with the medians and standard deviations calculated by the World Health Organization from an international reference population. Because stunting is related to malnutrition, to the extent that living standards are lower in *mita* districts, we would also expect stunting to be more common there. The height census has the advantage of providing substantially

[9]This discussion suggests that exempt districts were those located relatively far from both Potosí and Huancavelica. The correlation between distance to Potosí and distance to Huancavelica is −0.996, making it impossible to separately identify the effect of distance to each mine on the probability of receiving treatment. Thus, I divide the sample into two groups—municipalities to the east and those to the west of the dividing line between the Potosí and Huancavelica *mita* catchment areas. When considering districts to the west (Potosí side) of the dividing line, a flexible specification of *mita* treatment on a cubic in distance to Potosí, a cubic in elevation, and their linear interaction shows that being 100 additional kilometers from Potosí lowers the probability of treatment by 0.873, with a standard error of 0.244. Being 100 meters higher increases the probability of treatment by 0.061, with a standard error of 0.027. When looking at districts to the east (Huancavelica side) of the dividing line and using an analogous specification with a polynomial in distance to Huancavelica, the marginal effect of distance to Huancavelica is negative but not statistically significant.

more observations from about four times more districts than the household consumption sample. While the height census includes only children enrolled in school, 2005 data on primary school enrollment and completion rates do not show statistically significant differences across the *mita* boundary, with primary school enrollment rates exceeding 95% throughout the region examined (Ministro de Educación del Perú (MINEDU) (2005b)). Finally, to obtain controls for exogenous geographic characteristics, I calculate the mean area weighted elevation of each district by overlaying a map of Peruvian districts on 30 arc second (1 km) resolution elevation data produced by NASA's Shuttle Radar Topography Mission (SRTM (National Aeronautics and Space Administration and the National Geospatial-Intelligence Agency) (2000)), and I employ a similar procedure to obtain each district's mean area weighted slope. The Supplemental Material contains more detailed information about these data and the living standards data, as well as the data examined in Section 4.

3.2. *Estimation Framework*

Mita treatment is a deterministic and discontinuous function of known covariates, longitude and latitude, which suggests estimating the *mita*'s impacts using a regression discontinuity approach. The *mita* boundary forms a multidimensional discontinuity in longitude–latitude space, which differs from the single-dimensional thresholds typically examined in RD applications. While the identifying assumptions are identical to those in a single-dimensional RD, the multidimensional discontinuity raises interesting and important methodological issues about how to specify the RD polynomial, as discussed below. Before considering this and other identification issues in detail, let us introduce the basic regression form:

$$c_{idb} = \alpha + \gamma mita_d + X'_{id}\beta + f(\text{geographic location}_d) + \boldsymbol{\phi}_b + \varepsilon_{idb}, \tag{1}$$

where c_{idb} is the outcome variable of interest for observation i in district d along segment b of the *mita* boundary, and $mita_d$ is an indicator equal to 1 if district d contributed to the *mita* and equal to 0 otherwise; X_{id} is a vector of covariates that includes the mean area weighted elevation and slope for district d and (in regressions with equivalent household consumption on the left-hand side) demographic variables giving the number of infants, children, and adults in the household; $f(\text{geographic location}_d)$ is the RD polynomial, which controls for smooth functions of geographic location. Various forms will be explored. Finally, $\boldsymbol{\phi}_b$ is a set of boundary segment fixed effects that denote which of four equal length segments of the boundary is the closest to the observation's district capital.[10] To be conservative, all analysis excludes metropolitan Cusco. Metropolitan Cusco is composed of seven non-*mita* and two *mita*

[10]Results (available upon request) are robust to allowing the running variable to have heterogeneous effects by including a full set of interactions between the boundary segment fixed effects

districts located along the *mita* boundary and was the capital of the Inca Empire (Cook (1981, pp. 212–214), Cieza de León (1959, pp. 144–148)). I exclude Cusco because part of its relative prosperity today likely relates to its pre-*mita* heritage as the Inca capital. When Cusco is included, the impacts of the *mita* are estimated to be even larger.

The RD approach used in this paper requires two identifying assumptions. First, all relevant factors besides treatment must vary smoothly at the *mita* boundary. That is, letting c_1 and c_0 denote potential outcomes under treatment and control, x denote longitude, and y denote latitude, identification requires that $E[c_1|x, y]$ and $E[c_0|x, y]$ are continuous at the discontinuity threshold. This assumption is needed for individuals located just outside the *mita* catchment to be an appropriate counterfactual for those located just inside it. To assess the plausibility of this assumption, I examine the following potentially important characteristics: elevation, terrain ruggedness, soil fertility, rainfall, ethnicity, preexisting settlement patterns, local 1572 tribute (tax) rates, and allocation of 1572 tribute revenues.

To examine elevation—the principal determinant of climate and crop choice in Peru—as well as terrain ruggedness, I divide the study region into 20×20 km grid cells, approximately equal to the mean size of the districts in my sample, and calculate the mean elevation and slope within each grid cell using the SRTM data.[11] These geographic data are spatially correlated, and hence I report standard errors corrected for spatial correlation in square brackets. Following Conley (1999), I allow for spatial dependence of an unknown form. For comparison, I report robust standard errors in parentheses. The first set of columns of Table I restricts the sample to fall within 100 km of the *mita* boundary; the second, third, and fourth sets of columns restrict it to fall within 75, 50, and 25 km, respectively. The first row shows that elevation is statistically identical across the *mita* boundary.[12] I next look at terrain ruggedness, using the SRTM data to calculate the mean uphill slope in each grid cell. In contrast to elevation, there are some statistically significant, but relatively small, differences in slope, with *mita* districts being *less* rugged.[13]

---

and $f(\text{geographic location}_d)$. They are also robust to including soil type indicators, which I do not include in the main specification because they are highly collinear with the longitude–latitude polynomial used for one specification of $f(\text{geographic location}_d)$.

[11] All results are similar if the district is used as the unit of observation instead of using grid cells.

[12] Elevation remains identical across the *mita* boundary if I restrict the sample to inhabitable areas (<4800 m) or weight by population, rural population, or urban population data (Center for International Earth Science Information (2004, SEDAC)).

[13] I also examined data on district soil quality and rainfall (results available upon request; see the data appendix in the Supplemental Materials for more details). Data from the Peruvian Instituto Nacional de Recursos Naturales (INRENA (1997)) reveal *higher* soil quality in *mita* districts. I do not emphasize soil quality because it is endogenous to land usage. While climate is exogenous, high resolution data are not available and interpolated climate estimates are notoriously

TABLE I

SUMMARY STATISTICS[a]

	Sample Falls Within											
	<100 km of *Mita* Boundary			<75 km of *Mita* Boundary			<50 km of *Mita* Boundary			<25 km of *Mita* Boundary		
	Inside	Outside	s.e.	Inside	Outside	s.e.	Inside	Outside	s.e.	Inside	Outside	s.e.
GIS Measures												
Elevation	4042	4018	[188.77] (85.54)	4085	4103	[166.92] (82.75)	4117	4096	[169.45] (89.61)	4135	4060	[146.16] (115.15)
Slope	5.54	7.21	[0.88]* (0.49)***	5.75	7.02	[0.86] (0.52)**	5.87	6.95	[0.95] (0.58)*	5.77	7.21	[0.90] (0.79)*
Observations	177	95		144	86		104	73		48	52	
% Indigenous	63.59	58.84	[11.19] (9.76)	71.00	64.55	[8.04] (8.14)	71.01	64.54	[8.42] (8.43)	74.47	63.35	[10.87] (10.52)
Observations	1112	366		831	330		683	330		329	251	
Log 1572 tribute rate	1.57	1.60	[0.04] (0.03)	1.57	1.60	[0.04] (0.03)	1.58	1.61	[0.05] (0.04)	1.65	1.61	[0.02]* (0.03)

(*Continues*)

TABLE I—*Continued*

	Sample Falls Within											
	<100 km of *Mita* Boundary			<75 km of *Mita* Boundary			<50 km of *Mita* Boundary			<25 km of *Mita* Boundary		
	Inside	Outside	s.e.	Inside	Outside	s.e.	Inside	Outside	s.e.	Inside	Outside	s.e.
% 1572 tribute to Spanish Nobility	59.80	63.82	[1.39]*** (1.36)***	59.98	63.69	[1.56]** (1.53)**	62.01	63.07	[1.12] (1.34)	61.01	63.17	[1.58] (2.21)
Spanish Priests	21.05	19.10	[0.90]** (0.94)**	21.90	19.45	[1.02]** (1.02)**	20.59	19.93	[0.76] (0.92)	21.45	19.98	[1.01] (1.33)
Spanish Justices	13.36	12.58	[0.53] (0.48)*	13.31	12.46	[0.65] (0.60)	12.81	12.48	[0.43] (0.55)	13.06	12.37	[0.56] (0.79)
Indigenous Mayors	5.67	4.40	[0.78] (0.85)	4.55	4.29	[0.26] (0.29)	4.42	4.47	[0.34] (0.33)	4.48	4.42	[0.29] (0.39)
Observations	63	41		47	37		35	30		18	24	

[a]The unit of observation is 20 × 20 km grid cells for the geospatial measures, the household for % indigenous, and the district for the 1572 tribute data. Conley standard errors for the difference in means between *mita* and non-*mita* observations are in brackets. Robust standard errors for the difference in means are in parentheses. For % indigenous, the robust standard errors are corrected for clustering at the district level. The geospatial measures are calculated using elevation data at 30 arc second (1 km) resolution (SRTM (2000)). The unit of measure for elevation is 1000 meters and for slope is degrees. A household is indigenous if its members primarily speak an indigenous language in the home (ENAHO (2001)). The tribute data are taken from Miranda (1583). In the first three columns, the sample includes only observations located less than 100 km from the *mita* boundary, and this threshold is reduced to 75, 50, and finally 25 km in the succeeding columns. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

The third row examines ethnicity using data from the 2001 Peruvian National Household Survey (ENAHO). A household is defined as indigenous if the primary language spoken in the household is an indigenous language (usually Quechua). Results show no statistically significant differences in ethnic identification across the *mita* boundary.

Spanish authorities could have based *mita* assignment on settlement patterns, instituting the *mita* in densely populated areas and claiming land for themselves in sparsely inhabited regions where it was easier to usurp. A detailed review by Bauer and Covey (2002) of all archaeological surveys in the region surrounding the Cusco basin, covering much of the study region, indicates no large differences in settlement density at the date of Spanish Conquest. Moreover, there is no evidence suggesting differential rates of population decline in the 40 years between conquest and enactment of the *mita* (Cook (1981, pp. 108–114)).

Spanish officials blamed demographic collapse on excessive, unregulated rates of tribute extraction by local Hispanic elites (*encomenderos*), who received the right to collect tribute from the indigenous population in return for their role in Peru's military conquests. Thus Viceroy Francisco Toledo coordinated an in-depth inspection of Peru, Bolivia, and Ecuador in the early 1570s to evaluate the maximum tribute that could be demanded from local groups without threatening subsistence. Based on their assessment of ability to pay, authorities assigned varying tribute obligations at the level of the district socioeconomic group, with each district containing one or two socioeconomic groups. (See the Supplemental Material for more details on the tribute assessment.) These per capita contributions, preserved for all districts in the study region, provide a measure of Spanish authorities' best estimates of local prosperity. The forth row of Table I shows average tribute contributions per adult male (women, children, and those over age 50 were not taxed). Simple means comparisons across the *mita* boundary do not find statistically significant differences. The fifth through eight examine district level data on how Spanish authorities allocated these tribute revenues, divided between rents for Spanish nobility (*encomenderos*, fith row), salaries for Spanish priests (sixth row), salaries for local Spanish administrators (*justicias*, seventh row), and salaries for indigenous mayors (*caciques*, eigth row). The data on tribute revenue allocation are informative about the financing of local government, about the

inaccurate for the mountainous region examined in this study (Hijmans et al. (2005)). Temperature is primarily determined by altitude (Golte (1980), Pulgar-Vidal (1950)), and thus is unlikely to differ substantially across the *mita* boundary. To examine precipitation, I use station data from the Global Historical Climatology Network, Version 2 (Peterson and Vose (1997)). Using all available data (from stations in 50 districts located within 100 km of the *mita* boundary), *mita* districts appear to receive somewhat *higher* average annual precipitation, and these differences disappear when comparing districts closer to the *mita* boundary. When using only stations with at least 20 years of data (to ensure a long-run average), which provides observations from 20 different stations (11 outside the *mita* catchment and 9 inside), the difference declines somewhat in magnitude and is not statistically significant.

extent to which Spain extracted local revenues, and about the relative power of competing local administrators to obtain tribute revenues. Table I reveals some modest differences: when the sample is limited to fall within 100 km or 75 km from the *mita* boundary, we see that Spanish nobility received a slightly lower share of tribute revenue inside the *mita* catchment than outside (60% versus 64%), whereas Spanish priests received a slightly higher share (21% versus 19%). All differences disappear as the sample is limited to fall closer to the *mita* boundary.

In the ideal RD setup, the treatment effect is identified using only the variation at the discontinuity. Nonparametric RD techniques can be applied to approximate this setup in contexts with a large number of observations very near the treatment threshold (Imbens and Lemieux (2008)). While nonparametric techniques have the advantage of not relying on functional form assumptions, the data requirements that they pose are particularly high in the geographic RD context, as a convincing nonparametric RD would probably require precise georeferencing: for example, each observation's longitude–latitude coordinates or address.[14] This information is rarely made available due to confidentiality restrictions, and none of the available Peruvian micro data sets contains it. Moreover, many of the data sets required to investigate the *mita*'s potential long-run effects do not provide sufficiently large sample sizes to employ nonparametric techniques. Thus, I use a semiparametric RD approach that limits the sample to districts within 50 km of the *mita* boundary. This approach identifies causal effects by using a regression model to distinguish the treatment indicator, which is a nonlinear and discontinuous function of longitude (x) and latitude (y), from the smooth effects of geographic location. It is important for the regression model to approximate these effects well, so that a nonlinearity in the counterfactual conditional mean function $E[c_0|x, y]$ is not mistaken for a discontinuity, or vice versa (Angrist and Pischke (2009)). To the best of my knowledge, this is the first study to utilize a multidimensional, semiparametric RD approach.

Because approaches to specifying a multidimensional RD polynomial have not been widely explored, I report estimates from three baseline specifications of $f(\text{geographic location}_d)$. The first approach uses a cubic polynomial in latitude and longitude.[15] This parametrization is relatively flexible; it is analogous to the standard single-dimensional RD approach; and the RD plots, drawn in "x–y outcome" space, allow a transparent visual assessment of the data.

[14] A notable example of a multidimensional nonparametric RD is Black's (1999) study of the value that parents place on school quality. Black compared housing prices on either side of school attendance district boundaries in Massachusetts. Because she employs a large and precisely georeferenced data set, Black was able to include many boundary segment fixed effects and limit the sample to observations located within 0.15 miles of the boundary, ensuring comparison of observations in extremely close proximity.

[15] Letting x denote longitude and y denote latitude, this polynomial is $x + y + x^2 + y^2 + xy + x^3 + y^3 + x^2y + xy^2$.

For these reasons, this approach appears preferable to projecting the running variable into a lower-dimensional space—as I do in the other two baseline specifications—when power permits its precise estimation. One drawback is that some of the necessary datasets do not provide enough power to precisely estimate this flexible specification. The multidimensional RD polynomial also increases concerns about overfitting at the discontinuity, as a given order of a multidimensional polynomial has more degrees of freedom than the same order one-dimensional polynomial. This point is discussed using a concrete example in Section 4.3. Finally, there is no a priori reason why a polynomial form will do a good job of modeling the interactions between longitude and latitude. I partially address this concern by examining robustness to different orders of RD polynomials.

Given these concerns, I also report two baseline specifications that project geographic location into a single dimension. These single-dimensional specifications can be precisely estimated across the paper's data sets and provide useful checks on the multidimensional RD. One controls for a cubic polynomial in Euclidean distance to Potosí, a dimension which historical evidence identifies as particularly important. During much of the colonial period, Potosí was the largest city in the Western Hemisphere and one of the largest in the world, with a population exceeding 200,000. Historical studies document distance to Potosí as an important determinant of local production and trading activities, and access to coinage (Tandeter (1993, p. 56), Glave (1989), Cole (1985)).[16] Thus, a polynomial in distance to Potosí is likely to capture variation in relevant unobservables. However, this approach does not map well into the traditional RD setup, although it is similar in controlling for smooth variation and requiring all factors to change smoothly at the boundary. Thus I also examine a specification that controls for a cubic polynomial in distance to the *mita* boundary. I report this specification because it is similar to traditional one-dimensional RD designs, but to the best of my knowledge neither historical nor qualitative evidence suggests that distance to the *mita* boundary is economically important. Thus, this specification is most informative when examined in conjunction with the other two.

In addition to the two identifying assumptions already discussed, an additional assumption often employed in RD is no selective sorting across the treatment threshold. This would be violated if a direct *mita* effect provoked substantial out-migration of relatively productive individuals, leading to a larger indirect effect. Because this assumption may not be fully reasonable, I do not emphasize it. Rather I explore the possibility of migration as an interesting channel of persistence, to the extent that the data permit. During the past 130 years, migration appears to have been low. Data from the 1876, 1940,

[16]Potosí traded extensively with the surrounding region, given that it was located in a desert 14,000 feet above sea level and that it supported one of the world's largest urban populations during the colonial period.

and 1993 population censuses show a district level population correlation of 0.87 between 1940 and 1993 for both *mita* and non-*mita* districts.[17] Similarly, the population correlation between 1876 and 1940 is 0.80 in *mita* districts and 0.85 in non-*mita* districts. While a constant aggregate population distribution does not preclude extensive sorting, this is unlikely given the relatively closed nature of indigenous communities and the stable linkages between *haciendas* and their attached peasantry (Morner (1978)). Moreover, the 1993 Population Census (INEI (1993)) does not show statistically significant differences in rates of out-migration between *mita* and non-*mita* districts, although the rate of in-migration is 4.8% higher outside the *mita* catchment. In considering why individuals do not arbitrage income differences between *mita* and non-*mita* districts, it is useful to note that over half of the population in the region I examine lives in formally recognized indigenous communities. It tends to be difficult to gain membership and land in a different indigenous community, making large cities—which have various disamenities—the primary feasible destination for most migrants (INEI (1993)).

In contrast, out-migration from *mita* districts during the period that the *mita* was in force may have been substantial. Both Spanish authorities and indigenous leaders of *mita* communities had incentives to prevent migration, which made it harder for local leaders to meet *mita* quotas that were fixed in the medium run and threatened the *mita*'s feasibility in the longer run. Spanish authorities required individuals to reside in the communities to which the colonial state had assigned their ancestors soon after Peru's conquest to receive citizenship and access to agricultural land. Indigenous community leaders attempted to forcibly restrict migration. Despite these efforts, the state's capacity to restrict migration was limited, and 17th century population data—available for 15 *mita* and 14 non-*mita* districts—provide evidence consistent with the hypothesis that individuals migrated disproportionately from *mita* to non-*mita* districts.[18] To the extent that flight was selective and certain cognitive skills, physical strength, or other relevant characteristics are highly heritable, so that initial differences could persist over several hundred years, historical migration could contribute to the estimated *mita* effect. The paucity of data and complex patterns of heritability that would link historically selective migration to the present unfortunately place further investigation substantially beyond the scope of the current paper.

I begin by estimating the *mita*'s impact on living standards today; see Table II. First, I test for a *mita* effect on household consumption, using the log of equivalent household consumption, net transfers, in 2001 as the dependent variable. Following Deaton (1997), I assume that children aged 0 to 4 are equal

[17]The 2005 Population Census was methodologically flawed and thus I use 1993.

[18]According to data from the 1689 Cusco parish reports (see the Supplemental Material), in the 14 non-*mita* districts, 52.5% of individuals had ancestors who had not been assigned to their current district of residence, as compared to 35% in the 15 *mita* districts.

TABLE II

LIVING STANDARDS[a]

	Dependent Variable						
	Log Equiv. Household Consumption (2001)			Stunted Growth, Children 6–9 (2005)			
Sample Within:	<100 km of Bound. (1)	<75 km of Bound. (2)	<50 km of Bound. (3)	<100 km of Bound. (4)	<75 km of Bound. (5)	<50 km of Bound. (6)	Border District (7)
			Panel A. Cubic Polynomial in Latitude and Longitude				
Mita	−0.284	−0.216	−0.331	0.070	0.084*	0.087*	0.114**
	(0.198)	(0.207)	(0.219)	(0.043)	(0.046)	(0.048)	(0.049)
R^2	0.060	0.060	0.069	0.051	0.020	0.017	0.050
			Panel B. Cubic Polynomial in Distance to Potosí				
Mita	−0.337***	−0.307***	−0.329***	0.080***	0.078***	0.078***	0.063*
	(0.087)	(0.101)	(0.096)	(0.021)	(0.022)	(0.024)	(0.032)
R^2	0.046	0.036	0.047	0.049	0.017	0.013	0.047
			Panel C. Cubic Polynomial in Distance to *Mita* Boundary				
Mita	−0.277***	−0.230**	−0.224**	0.073***	0.061***	0.064***	0.055*
	(0.078)	(0.089)	(0.092)	(0.023)	(0.022)	(0.023)	(0.030)
R^2	0.044	0.042	0.040	0.040	0.015	0.013	0.043
Geo. controls	yes	yes	yes	yes	yes	yes	yes
Boundary F.E.s	yes	yes	yes	yes	yes	yes	yes
Clusters	71	60	52	289	239	185	63
Observations	1478	1161	1013	158,848	115,761	100,446	37,421

[a]The unit of observation is the household in columns 1–3 and the individual in columns 4–7. Robust standard errors, adjusted for clustering by district, are in parentheses. The dependent variable is log equivalent household consumption (ENAHO (2001)) in columns 1–3, and a dummy equal to 1 if the child has stunted growth and equal to 0 otherwise in columns 4–7 (Ministro de Educación (2005a)). *Mita* is an indicator equal to 1 if the household's district contributed to the *mita* and equal to 0 otherwise (Saignes (1984), Amat y Juniet (1947, pp. 249, 284)). Panel A includes a cubic polynomial in the latitude and longitude of the observation's district capital, panel B includes a cubic polynomial in Euclidean distance from the observation's district capital to Potosí, and panel C includes a cubic polynomial in Euclidean distance to the nearest point on the *mita* boundary. All regressions include controls for elevation and slope, as well as boundary segment fixed effects (F.E.s). Columns 1–3 include demographic controls for the number of infants, children, and adults in the household. In columns 1 and 4, the sample includes observations whose district capitals are located within 100 km of the *mita* boundary, and this threshold is reduced to 75 and 50 km in the succeeding columns. Column 7 includes only observations whose districts border the *mita* boundary. 78% of the observations are in *mita* districts in column 1, 71% in column 2, 68% in column 3, 78% in column 4, 71% in column 5, 68% in column 6, and 58% in column 7. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

to 0.4 adults and children aged 5 to 14 are equal to 0.5 adults. Panel A reports the specification that includes a cubic polynomial in latitude and longitude, panel B reports the specification that uses a cubic polynomial in distance to Potosí, and panel C reports the specification that includes a cubic polynomial in distance to the *mita* boundary. Column 1 of Table II limits the sample to districts within 100 km of the *mita* boundary, and columns 2 and 3 restrict it to fall within 75 and 50 km, respectively.[19] Columns 4–7 repeat this exercise, using as the dependent variable a dummy equal to 1 if the child's growth is stunted and equal to 0 otherwise. Column 4 limits the sample to districts within 100 km of the *mita* boundary, and columns 5 and 6 restrict it to fall within 75 and 50 km, respectively. Column 7 limits the sample to only those districts bordering the *mita* boundary. In combination with the inclusion of boundary segment fixed effects, this ensures that I am comparing observations in close geographic proximity.

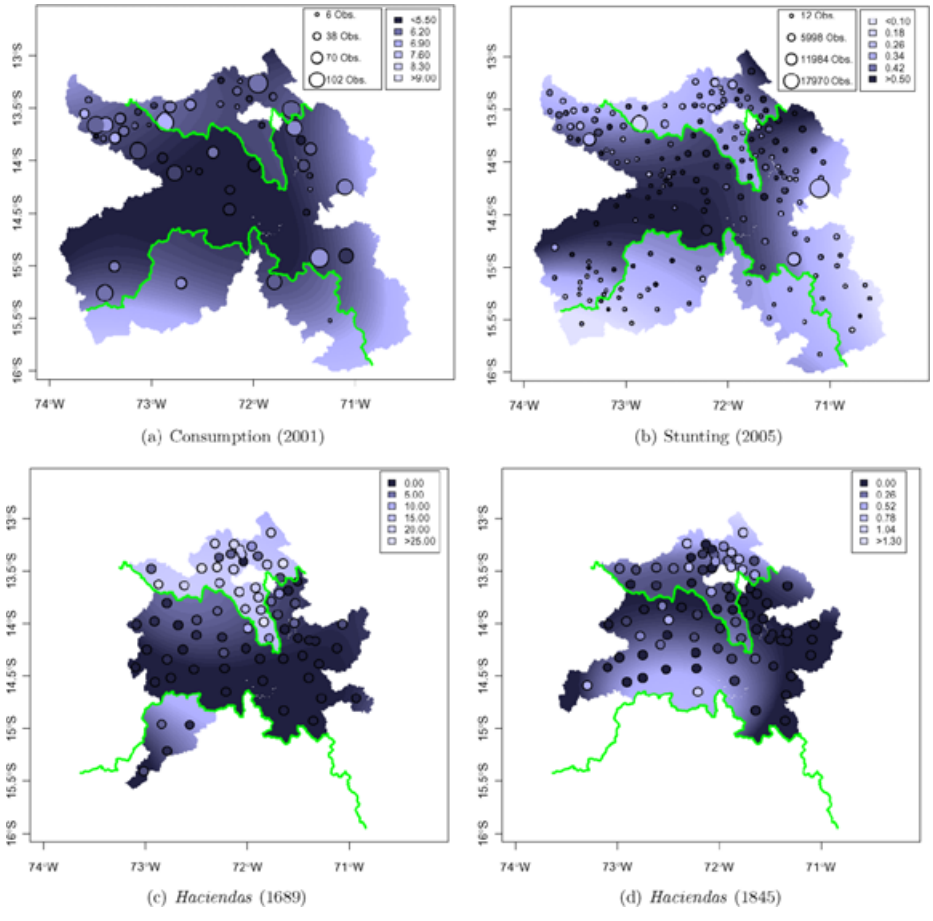
3.3. *Estimation Results*

Columns 1–3 of Table II estimate that a long-run *mita* effect lowers household consumption in 2001 by around 25% in subjected districts. The point estimates remain fairly stable as the sample is restricted to fall within narrower bands of the *mita* boundary. Moreover, the *mita* coefficients are economically similar across the three specifications of the RD polynomial, and I am unable to reject that they are statistically identical. All of the *mita* coefficients in panels B and C, which report the single-dimensional RD estimates, are statistically significant at the 1% or 5% level. In contrast, the point estimates using a cubic polynomial in latitude and longitude (panel A) are not statistically significant. This imprecision likely results from the relative flexibility of the specification, the small number of observations and clusters (the household survey samples only around one-quarter of districts), and measurement error in the dependent variable (Deaton (1997)).

Columns 4–7 of Table II examine census data on stunting in children, an alternative measure of living standards which offers a substantially larger sample. When using only observations in districts that border the *mita* boundary, point estimates of the *mita* effect on stunting range from 0.055 (s.e. = 0.030) to 0.114 (s.e. = 0.049) percentage points. This compares to a mean prevalence of stunting of 40% throughout the region examined.[20] Of the 12 point estimates reported in Table II, 11 are statistically significant, and I cannot reject at the 10% level that the estimates are the same across specifications.

[19]The single-dimensional specifications produce similar estimates when the sample is limited to fall within 25 km of the *mita* boundary. The multidimensional specification produces a very large and imprecisely estimated *mita* coefficient because of the small sample size.

[20]A similar picture emerges when I use height in centimeters as the dependent variable and include quarter × year of birth dummies, a gender dummy, and their interactions on the right-hand side.

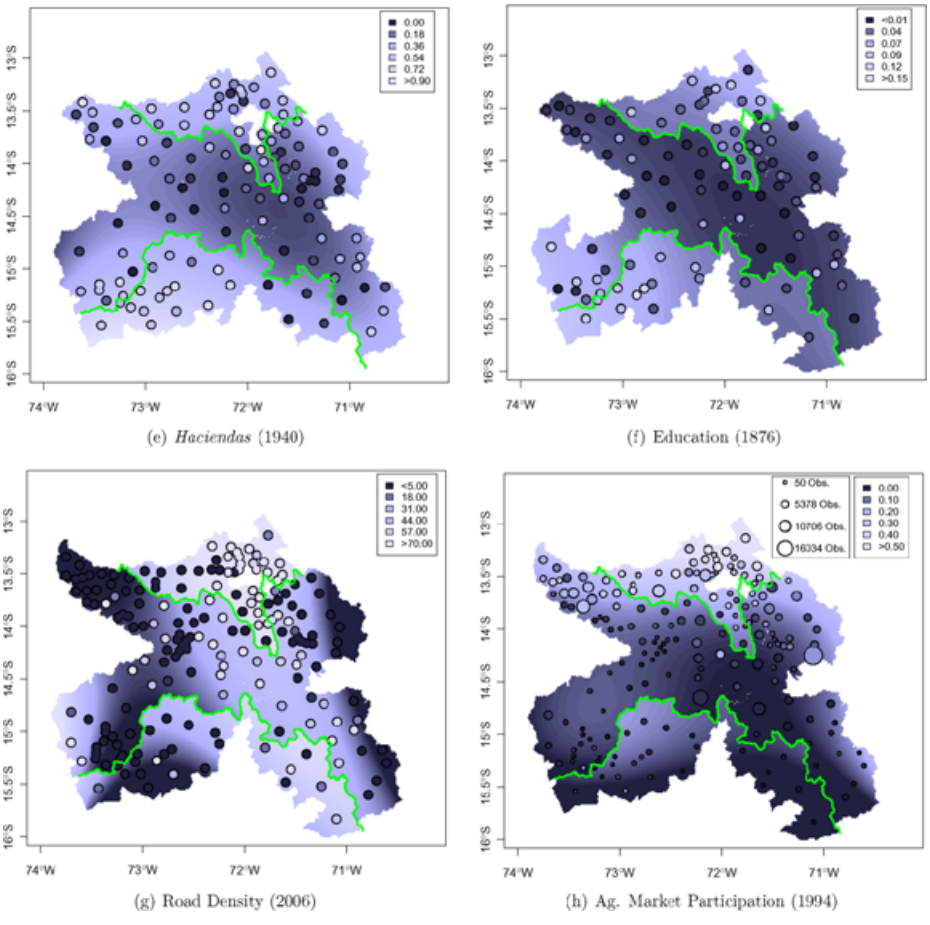


(a) Consumption (2001) (b) Stunting (2005)

(c) *Haciendas* (1689) (d) *Haciendas* (1845)

FIGURE 2.—Plots of various outcomes against longitude and latitude. See the text for a detailed description.

The results can be seen graphically in Figure 2. Each subfigure shows a district-level scatter plot for one of the paper's main outcome variables. These plots are the three-dimensional analogues to standard two-dimensional RD plots, with each district capital's longitude on the x axis, its latitude on the y axis, and the data value for that district shown using an evenly spaced monochromatic color scale, as described in the legends. When the underlying data are at the microlevel, I take district-level averages, and the size of the dot indicates the number of observations in each district. Importantly, the scaling on these dots, which is specified in the legend, is nonlinear, as otherwise some would be microscopic and others too large to display. The background in each plot shows predicted values, for a finely spaced grid of longitude–latitude co-



(e) *Haciendas* (1940) (f) Education (1876)

(g) Road Density (2006) (h) Ag. Market Participation (1994)

FIGURE 2.—*Continued*.

ordinates, from a regression of the outcome variable under consideration on a cubic polynomial in longitude–latitude and the *mita* dummy. In the typical RD context, the predicted value plot is a two-dimensional curve, whereas here it is a three-dimensional surface, with the third dimension indicated by the color gradient.[21] The shades of the data points can be compared to the shades of the predicted values behind them to judge whether the RD has done an adequate job of averaging the data across space. The majority of the population in the region is clustered along the upper segment of the *mita* boundary, giving these

[21]Three-dimensional surface plots of the predicted values are shown in Figure A2 in the Supplemental Material, and contour plots are available upon request.

TABLE III

SPECIFICATION TESTS[a]

	Dependent Variable						
	Log Equiv. Hausehold Consumption (2001)			Stunted Growth, Children 6–9 (2005)			
Sample Within:	<100 km of Bound. (1)	<75 km of Bound. (2)	<50 km of Bound. (3)	<100 km of Bound. (4)	<75 km of Bound. (5)	<50 km of Bound. (6)	Border District (7)
Alternative Functional Forms for RD Polynomial: Baseline I							
Linear polynomial in latitude and longitude							
Mita	−0.294***	−0.199	−0.143	0.064***	0.054**	0.062**	0.068**
	(0.092)	(0.126)	(0.128)	(0.021)	(0.022)	(0.026)	(0.031)
Quadratic polynomial in latitude and longitude							
Mita	−0.151	−0.247	−0.361	0.073*	0.091**	0.106**	0.087**
	(0.189)	(0.209)	(0.216)	(0.040)	(0.043)	(0.047)	(0.041)
Quartic polynomial in latitude and longitude							
Mita	−0.392*	−0.324	−0.342	0.073	0.072	0.057	0.104**
	(0.225)	(0.231)	(0.260)	(0.056)	(0.050)	(0.048)	(0.042)
Alternative Functional Forms for RD Polynomial: Baseline II							
Linear polynomial in distance to Potosí							
Mita	−0.297***	−0.273***	−0.220**	0.050**	0.048**	0.049**	0.071**
	(0.079)	(0.093)	(0.092)	(0.022)	(0.022)	(0.024)	(0.031)
Quadratic polynomial in distance to Potosí							
Mita	−0.345***	−0.262***	−0.309***	0.072***	0.064***	0.072***	0.060*
	(0.086)	(0.095)	(0.100)	(0.023)	(0.022)	(0.023)	(0.032)
Quartic polynomial in distance to Potosí							
Mita	−0.331***	−0.310***	−0.330***	0.078***	0.075***	0.071***	0.053*
	(0.086)	(0.100)	(0.097)	(0.021)	(0.020)	(0.021)	(0.031)
Interacted linear polynomial in distance to Potosí							
Mita	−0.307***	−0.280***	−0.227**	0.051**	0.048**	0.043*	0.076***
	(0.092)	(0.094)	(0.095)	(0.022)	(0.021)	(0.022)	(0.029)
Interacted quadratic polynomial in distance to Potosí							
Mita	−0.264***	−0.177*	−0.285**	0.033	0.027	0.039*	0.036
	(0.087)	(0.096)	(0.111)	(0.024)	(0.023)	(0.023)	(0.024)

(*Continues*)

districts substantially more weight in figures showing predicted values from microlevel regressions.

Table III examines robustness to 14 different specifications of the RD polynomial, documenting *mita* effects on household consumption and stunting that are generally similar across specifications. The first three rows report results from alternative specifications of the RD polynomial in longitude–latitude: linear, quadratic, and quartic. The next five rows report alternative specifications using distance to Potosí: linear, quadratic, quartic, and the *mita* dummy inter-

TABLE III—*Continued*

	Dependent Variable						
	Log Equiv. Household Consumption (2001)			Stunted Growth, Children 6–9 (2005)			
Sample Within:	<100 km of Bound. (1)	<75 km of Bound. (2)	<50 km of Bound. (3)	<100 km of Bound. (4)	<75 km of Bound. (5)	<50 km of Bound. (6)	Border District (7)
Alternative Functional Forms for RD Polynomial: Baseline III							
Linear polynomial in distance to *mita* boundary							
Mita	−0.299***	−0.227**	−0.223**	0.072***	0.060***	0.058**	0.056*
	(0.082)	(0.089)	(0.091)	(0.024)	(0.022)	(0.023)	(0.032)
Quadratic polynomial in distance to *mita* boundary							
Mita	−0.277***	−0.227**	−0.224**	0.072***	0.060***	0.061***	0.056*
	(0.078)	(0.089)	(0.092)	(0.023)	(0.022)	(0.023)	(0.030)
Quartic polynomial in distance to *mita* boundary							
Mita	−0.251***	−0.229**	−0.246***	0.073***	0.064***	0.063***	0.055*
	(0.078)	(0.089)	(0.088)	(0.023)	(0.022)	(0.023)	(0.030)
Interacted linear polynomial in distance to *mita* boundary							
Mita	−0.301*	−0.277	−0.385*	0.082	0.087	0.095	0.132**
	(0.174)	(0.190)	(0.210)	(0.054)	(0.055)	(0.065)	(0.053)
Interacted quadratic polynomial in distance to *mita* boundary							
Mita	−0.351	−0.505	−0.295	0.140*	0.132	0.136	0.121*
	(0.260)	(0.319)	(0.366)	(0.082)	(0.084)	(0.086)	(0.064)
Ordinary Least Squares							
Mita	−0.294***	−0.288***	−0.227**	0.057**	0.048*	0.049*	0.055*
	(0.083)	(0.089)	(0.090)	(0.025)	(0.024)	(0.026)	(0.031)
Geo. controls	yes	yes	yes	yes	yes	yes	yes
Boundary F.E.s	yes	yes	yes	yes	yes	yes	yes
Clusters	71	60	52	289	239	185	63
Observations	1478	1161	1013	158,848	115,761	100,446	37,421

[a]Robust standard errors, adjusted for clustering by district, are in parentheses. All regressions include geographic controls and boundary segment fixed effects (F.E.s). Columns 1–3 include demographic controls for the number of infants, children, and adults in the household. Coefficients significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

acted with a linear or quadratic polynomial in distance to Potosí.[22] Next, the ninth trough thirteenth rows examine robustness to the same set of specifications, using distance to the *mita* boundary as the running variable. Finally, the fourteenth row reports estimates from a specification using ordinary least squares. The *mita* effect on consumption is always statistically significant in

[22]The *mita* effect is evaluated at the mean distance to Potosí for observations very near (<10 km from) the *mita* boundary. Results are broadly robust to evaluating the *mita* effect at different average distances to Potosí, that is, for districts <25 km from the boundary, for bordering districts, or for all districts.

the relatively parsimonious specifications: those that use noninteracted, single-dimensional RD polynomials and ordinary least squares. In the more flexible specifications—the longitude–latitude regressions and those that interact the RD polynomial with the *mita* dummy—the *mita* coefficients in the consumption regression tend to be imprecisely estimated. As in Table II, the household survey does not provide enough power to precisely estimate relatively flexible specifications, but the coefficients are similar in magnitude to those estimated using a more parsimonious approach. Estimates of the *mita*'s impact on stunting are statistically significant across most specifications and samples.[23]

Given broad robustness to functional form assumptions, Table IV reports a number of additional robustness checks using the three baseline specifications of the RD polynomial. To conserve space, I report estimates only from the sample that contains districts within 50 km of the *mita* boundary. Columns 1–7 examine the household consumption data and columns 8–12 examine the stunting data. For comparison purposes, columns 1 and 8 present the baseline estimates from Table II. Column 2 adds a control for ethnicity, equal to 1 if an indigenous language is spoken in the household and 0 otherwise. Next, columns 3 and 9 include metropolitan Cusco. In response to the potential endogeneity of the *mita* to Inca landholding patterns, columns 4 and 10 exclude districts that contained Inca royal estates, which served sacred as opposed to productive purposes (Niles (1987, p. 13)). Similarly, columns 5 and 11 exclude districts falling along portions of the *mita* boundary formed by rivers to account for one way in which the boundary could be endogenous to geography. Column 6 estimates consumption equivalence flexibly, using log household consumption as the dependent variable, and controlling for the ratio of children to adults and the log of household size. In all cases, point estimates and significance levels tend to be similar to those in Table II. As expected, the point estimates are somewhat larger when metropolitan Cusco is included.

Table IV investigates whether differential rates of migration today may be responsible for living standards differences between *mita* and non-*mita* districts. Given that in-migration in non-*mita* districts is about 4.8% higher than in *mita* districts (whereas rates of out-migration are statistically and economically similar), I omit the 4.8% of the non-*mita* sample with the highest equivalent household consumption and least stunting, respectively. Estimates in columns 7 and 12 remain of similar magnitude and statistical significance, documenting that migration today is not the primary force responsible for the *mita* effect.

If the RD specification is estimating the *mita*'s long-run effect as opposed to some other underlying difference, being inside the *mita* catchment should not affect economic prosperity, institutions, or demographics prior to the *mita*'s enactment. In a series of specification checks, I first regress the log of the

[23]Results (not shown) are also robust to including higher order polynomials in elevation and slope.

TABLE IV

ADDITIONAL SPECIFICATION TESTS[a]

	Log Equivalent Household Consumption (2001)							Stunted Growth, Children 6–9 (2005)				
	Baseline (1)	Control for Ethnicity (2)	Includes Cusco (3)	Excludes Districts With Inca Estates (4)	Excludes Portions of Boundary Formed by Rivers (5)	Flexible Estimation of Consump. Equivalence (6)	Migration (7)	Baseline (8)	Includes Cusco (9)	Excludes Districts With Inca Estates (10)	Excludes Portions of Boundary Formed by Rivers (11)	Migration (12)
Panel A. Cubic Polynomial in Latitude and Longitude												
Mita	−0.331	−0.202	−0.465**	−0.281	−0.322	−0.326	−0.223	0.087*	0.147***	0.093*	0.090*	0.069
	(0.219)	(0.157)	(0.207)	(0.265)	(0.215)	(0.230)	(0.198)	(0.048)	(0.048)	(0.048)	(0.048)	(0.049)
R^2	0.069	0.154	0.104	0.065	0.070	0.292	0.067	0.017	0.046	0.019	0.018	0.016
Panel B. Cubic Polynomial in Distance to Potosí												
Mita	−0.329***	−0.282***	−0.450***	−0.354***	−0.376***	−0.328***	−0.263***	0.078***	0.146***	0.077***	0.081***	0.060**
	(0.096)	(0.073)	(0.096)	(0.101)	(0.114)	(0.099)	(0.095)	(0.024)	(0.030)	(0.026)	(0.024)	(0.025)
R^2	0.047	0.140	0.087	0.036	0.049	0.275	0.042	0.013	0.039	0.014	0.013	0.012
Panel C. Cubic Polynomial in Distance to *Mita* Boundary												
Mita	−0.224**	−0.195***	−0.333***	−0.255**	−0.217**	−0.224**	−0.161*	0.064***	0.132***	0.066***	0.065***	0.046*
	(0.092)	(0.070)	(0.087)	(0.110)	(0.098)	(0.095)	(0.088)	(0.023)	(0.027)	(0.025)	(0.023)	(0.024)
R^2	0.040	0.135	0.088	0.047	0.039	0.270	0.037	0.013	0.042	0.014	0.013	0.012
Geo. controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Bound. F.E.s	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Clusters	52	52	57	47	51	52	52	185	195	180	183	185
Observations	1013	1013	1173	930	992	1013	997	100,446	127,259	96,440	99,940	98,922

[a]Robust standard errors, adjusted for clustering by district, are in parentheses. All regressions include soil type indicators and boundary segment fixed effects (F.E.s). Columns 1–5 and 7 include demographic controls for the number of infants, children, and adults in the household. Column (6) includes controls for the log of household size and the ratio of children to household members, using the log of household consumption as the dependent variable. The samples include observations whose district capitals are less than 50 km from the *mita* boundary. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

mean district 1572 tribute contribution per adult male on the variables used in the stunting regressions in Table II. I then examine the shares of 1572 tribute revenues allocated to rents for Spanish nobility, salaries for Spanish priests, salaries for local Spanish administrators, and salaries for indigenous mayors. Finally, also using data from the 1572 census, I investigate demographics, with the population shares of tribute paying males (those aged 18–50), boys, and women as the dependent variables. These regressions, reported in Table V, do not show statistically significant differences across the *mita* boundary, and the estimated *mita* coefficients are small.

TABLE V

1572 TRIBUTE AND POPULATION[a]

	Dependent Variable							
		Share of Tribute Revenues				Percent		
	Log Mean Tribute	Spanish Nobility	Spanish Priests	Spanish Justices	Indig. Mayors	Men	Boys	Females
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A. Cubic Polynomial in Latitude and Longitude							
Mita	0.020	−0.010	0.004	0.004	0.003	−0.006	0.011	−0.009
	(0.031)	(0.030)	(0.019)	(0.010)	(0.005)	(0.009)	(0.012)	(0.016)
R^2	0.762	0.109	0.090	0.228	0.266	0.596	0.377	0.599
	Panel B. Cubic Polynomial in Distance to Potosí							
Mita	0.019	−0.013	0.008	0.006	−0.001	−0.012	0.005	−0.011
	(0.029)	(0.025)	(0.015)	(0.009)	(0.004)	(0.008)	(0.010)	(0.012)
R^2	0.597	0.058	0.073	0.151	0.132	0.315	0.139	0.401
	Panel C. Cubic Polynomial in Distance to *Mita* Boundary							
Mita	0.040	−0.009	0.005	0.003	−0.001	−0.011	0.001	−0.008
	(0.030)	(0.018)	(0.012)	(0.006)	(0.004)	(0.007)	(0.008)	(0.010)
R^2	0.406	0.062	0.096	0.118	0.162	0.267	0.190	0.361
Geo. controls	yes	yes	yes	yes	yes	yes	yes	yes
Boundary F.E.s	yes	yes	yes	yes	yes	yes	yes	yes
Mean dep. var.	1.591	0.625	0.203	0.127	0.044	0.193	0.204	0.544
Observations	65	65	65	65	65	65	65	65

[a]The dependent variable in column 1 is the log of the district's mean 1572 tribute rate (Miranda (1583)). In columns 2–5, it is the share of tribute revenue allocated to Spanish nobility (*encomenderos*), Spanish priests, Spanish justices, and indigenous mayors (*caciques*), respectively. In columns 6–8, it is the share of 1572 district population composed of males (aged 18–50), boys, and females (of all ages), respectively. Panel A includes a cubic polynomial in longitude and latitude, panel B includes a cubic polynomial in Euclidean distance from the observation's district capital to Potosí, and panel C includes a cubic polynomial in Euclidean distance to the nearest point on the *mita* boundary. All regressions include geographic controls and boundary segment fixed effects. The samples include districts whose capitals are less than 50 km from the *mita* boundary. Column 1 weights by the square root of the district's tributary population and columns 6–8 weight by the square root of the district's total population. 66% of the observations are from *mita* districts. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

To achieve credible identification, I exploit variation across observations located near the *mita* boundary. If the boundary is an unusual place, these estimates may have little external validity. To examine this issue further, I use ordinary least squares to estimate the correlation between the *mita* and the main outcome variables (including those that will be examined in Section 4), limiting the sample to districts located *between* 25 and 100 km from the *mita* boundary. The estimates are quite similar to those obtained from the RD specifications (results available upon request). Moreover, correlations between the *mita* and living standards (measured by both consumption and stunting) calculated along the entire *mita* boundary within Peru are consistent in magnitude with the effects documented above.[24] In summary, the RD evidence appears informative about the *mita*'s overall impacts.

Why would the *mita* affect economic prosperity nearly 200 years after its abolition? To open this black box, I turn to an investigation of channels of persistence.

4. CHANNELS OF PERSISTENCE

This section uses data from the Spanish Empire and Peruvian Republic to test channels of persistence. There exist many potential channels, but to provide a picture that is both parsimonious and informative, I focus on three that the historical literature and fieldwork suggest are important: land tenure, public goods, and market participation. The results document that the *mita* limited the establishment of large landowners inside the *mita* catchment and, combined with historical evidence, suggest that land tenure has in turn affected public goods provision and smallholder participation in agricultural markets.

The tables in the main text report three specifications, which use a cubic polynomial in latitude and longitude, a cubic polynomial in distance to Potosí, or a cubic polynomial in distance to the *mita* boundary. Table A.III in the Supplemental Material reports results from the 14 additional specifications examined in Table III. In most cases, the point estimates across these specifications are similar. When not, I note it explicitly.[25]

4.1. *Land Tenure and Labor Systems*

This section examines the impact of the *mita* on the formation of *haciendas*—rural estates with an attached labor force permanently settled on the estate (Keith (1971, p. 437)). Critically, when authorities instituted the *mita*

[24]When considering observations in Peru within 50 km of any point on the *mita* boundary, being inside the *mita* catchment is associated with 28.4 percent lower equivalent household consumption and an increase of 16.4 percentage points in the prevalence of stunting.

[25]As in Table III, the more flexible specifications in Table A.III are less likely than the parsimonious ones to estimate statistically significant effects.

in 1573 (40 years after the Spanish conquest of Peru), a landed elite had not yet formed. At the time, Peru was parceled into *encomiendas*, pieces of territory in which appointed Spaniards exercised the right to collect tribute and labor services from the indigenous population but did not hold title to land (Keith (1971, p. 433)). Rivalries between *encomenderos* provoked civil wars in the years following Peru's conquest, and thus the Crown began to dismantle the *encomienda* system during the 1570s. This opened the possibility for manipulating land tenure to promote other policy goals, in particular, the *mita*.[26]

Specifically, Spanish land tenure policy aimed to minimize the establishment of landed elites in *mita* districts, as large landowners—who unsurprisingly opposed yielding their attached labor for a year of *mita* service—formed the state's principal labor market competition (Larson (1988), Sanchez-Albornoz (1978)).[27] Centrally, as Bolivian historian Larson (1988, p. 171) concisely articulated, "*Haciendas* secluded peasants from the extractive institutions of colonial society." Moreover, by protecting native access to agricultural lands, the state promoted the ability of the indigenous community to subsidize *mita* conscripts, who were paid substantially below subsistence wages (Garrett (2005, p. 120), Tandeter (1993, pp. 58–60), Cole (1985, p. 31)). Similarly, authorities believed that protecting access to land could be an effective means of staving off demographic collapse (Larson (1982, p. 11), Cook (1981, pp. 108–114, 250), Morner (1978)). Finally, in return for ensuring the delivery of conscripts, local authorities were permitted to extract surplus that would have otherwise been claimed by large landowners (Garrett (2005, p. 115)).

I now examine the concentration of *haciendas* in 1689, 1845, and 1940. The 1689 data are contained in parish reports commissioned by Bishop Manuel de Mollinedo and submitted by all parishes in the bishopric of Cusco, which encompassed most of the study region. The reports list the number of *haciendas* and the population within each subdivision of the parish, and were compiled by Horacio Villanueva Urteaga (1982). For *haciendas* in 1845, I employ data collected by the Cusco regional government, which had jurisdiction over a substantial fraction of the study region, on the percentage of the rural tributary population residing in *haciendas* (Peralta Ruiz (1991)). Data from 1845, 1846, and 1850 are combined to form the circa 1845 data set.[28] Finally, data from the 1940 Peruvian Population Census are aggregated to the district level to calculate the percentage of the rural population residing in *haciendas*.

[26]Throughout the colonial period, royal policy aimed to minimize the power of the (potentially revolutionary) landed class: landowners did not acquired the same political clout as mine owners, the most powerful colonial interest group (Tandeter (1993), Cole (1985)).

[27]For example, land sales under Philip VI between 1634 and 1648 and by royal charter in 1654 played a central role in *hacienda* formation and were almost exclusively concentrated in non-*mita* districts (Brisseau (1981, p. 146), Glave and Remy (1978, p. 1)).

[28]When data are available for more than one year, figures change little, and I use the earliest observation.

TABLE VI

LAND TENURE AND LABOR SYSTEMS[a]

	Dependent Variable				
	Haciendas per District in 1689 (1)	*Haciendas* per 1000 District Residents in 1689 (2)	Percent of Rural Tributary Population in *Haciendas* in ca. 1845 (3)	Percent of Rural Population in *Haciendas* in 1940 (4)	Land Gini in 1994 (5)
	Panel A. Cubic Polynomial in Latitude and Longitude				
Mita	−12.683***	−6.453**	−0.127*	−0.066	0.078
	(3.221)	(2.490)	(0.067)	(0.086)	(0.053)
R^2	0.538	0.582	0.410	0.421	0.245
	Panel B. Cubic Polynomial in Distance to Potosí				
Mita	−10.316***	−7.570***	−0.204**	−0.143***	0.107***
	(2.057)	(1.478)	(0.082)	(0.051)	(0.036)
R^2	0.494	0.514	0.308	0.346	0.194
	Panel C. Cubic Polynomial in Distance to *Mita* Boundary				
Mita	−11.336***	−8.516***	−0.212***	−0.120***	0.124***
	(2.074)	(1.665)	(0.060)	(0.045)	(0.033)
R^2	0.494	0.497	0.316	0.336	0.226
Geo. controls	yes	yes	yes	yes	yes
Boundary F.E.s	yes	yes	yes	yes	yes
Mean dep. var.	6.500	5.336	0.135	0.263	0.783
Observations	74	74	81	119	181

[a]The unit of observation is the district. Robust standard errors are in parentheses. The dependent variable in column 1 is *haciendas* per district in 1689 and in column 2 is *haciendas* per 1000 district residents in 1689 (Villanueva Urteaga (1982)). In column 3 it is the percentage of the district's tributary population residing in *haciendas* ca. 1845 (Peralta Ruiz (1991)), in column 4 it is the percentage of the district's rural population residing in *haciendas* in 1940 (Dirección de Estadística del Perú (1944)), and in column 5 it is the district land gini (INEI (1994)). Panel A includes a cubic polynomial in the latitude and longitude of the observation's district capital, panel B includes a cubic polynomial in Euclidean distance from the observation's district capital to Potosí, and panel C includes a cubic polynomial in Euclidean distance to the nearest point on the *mita* boundary. All regressions include geographic controls and boundary segment fixed effects. The samples include districts whose capitals are less than 50 km from the *mita* boundary. Column 3 is weighted by the square root of the district's rural tributary population and column 4 is weighted by the square root of the district's rural population. 58% of the observations are in *mita* districts in columns 1 and 2, 59% in column 3, 62% in column 4, and 66% in column 5. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

In Table VI, column 1 (number of *haciendas* per district) and column 2 (number of *haciendas* per 1000 district residents) show a very large *mita* effect on the concentration of *haciendas* in the 17th century, of similar magnitude and highly significant across specifications.[29] The median coefficient from column 1, con-

[29]Given the *mita*'s role in provoking population collapse (Wightman (1990, p. 72)), the latter measure is likely endogenous, but nevertheless provides a useful robustness check.

tained in panel C, estimates that the *mita* lowered the number of *haciendas* in subjected districts by 11.3 (s.e. = 2.1), a sizeable effect given that on average *mita* districts contained only one *hacienda*. Figure 2, panel (c) clearly demonstrates the discontinuity. Moreover, Table VI provides reasonably robust support for a persistent impact. Column 3 estimates that the *mita* lowered the percentage of the rural tributary population in *haciendas* in 1845 by around 20 percentage points (with estimates ranging from 0.13 to 0.21), an effect that is statistically significant across specifications. Column 4 suggests that disparities persisted into the 20th century, with an estimated effect on the percentage of the rural labor force in *haciendas* that is somewhat smaller for 1940 than for 1845—as can be seen by comparing panels (d) and (e) of Figure 2—and not quite as robust. The median point estimate is −0.12 (s.e. = 0.045) in panel C; the point estimates are statistically significant at the 1% level in panels B and C, but the longitude–latitude specification estimates an effect that is smaller, at −0.07, and imprecise.

Table VI also documents that the percentage of the rural population in *haciendas* nearly doubled between 1845 and 1940, paralleling historical evidence for a rapid expansion of *haciendas* in the late 19th and early 20th centuries. This expansion was spurred by a large increase in land values due to globalization and seems to have been particularly coercive inside the *mita* catchment (Jacobsen (1993, pp. 226–237), Favre (1967, p. 243), Nuñez (1913, p. 11)). No longer needing to ensure *mita* conscripts, Peru abolished the communal land tenure predominant in *mita* districts in 1821, but did not replace it with enforceable peasant titling (Jacobsen (1993), Dancuart and Rodriguez (1902, Vol. 2, p. 136)). This opened the door to tactics such as the *interdicto de adquirir*, a judicial procedure which allowed aspiring landowners to legally claim "abandoned" lands that in reality belonged to peasants. *Hacienda* expansion also occurred through violence, with cattle nustling, grazing estate cattle on peasant lands, looting, and physical abuse used as strategies to intimidate peasants into signing bills of sale (Avila (1952, p. 22), Roca-Sanchez (1935, pp. 242–243)). Numerous peasant rebellions engulfed *mita* districts during the 1910s and 1920s, and indiscriminate banditry and livestock rustling remained prevalent in some *mita* districts for decades (Jacobsen (1993), Ramos Zambrano (1984), Tamayo Herrera (1982), Hazen (1974, pp. 170–178)). In contrast, large landowners had been established since the early 17th century in non-*mita* districts, which remained relatively stable (Flores Galindo (1987, p. 240)).

In 1969, the Peruvian government enacted an agrarian reform bill mandating the complete dissolution of *haciendas*. As a result, the *hacienda* elite were deposed and lands formerly belonging to *haciendas* were divided into Agricultural Societies of Social Interest (SAIS) during the early 1970s (Flores Galindo (1987)). In SAIS, neighboring indigenous communities and the producers acted as collective owners. By the late 1970s, attempts to impose collective ownership through SAIS had failed, and many SAIS were divided and allocated to individuals (Mar and Mejia (1980)). The 1994 Agricultural Census

(INEI (1994)) documents that when considering districts within 50 km of the *mita* boundary, 20% of household heads outside the *mita* catchment received their land in the 1970s through the agrarian reform, versus only 9% inside the *mita* catchment. Column 5, using data from the 1994 Agricultural Census, documents somewhat lower land inequality in non-*mita* districts. This finding is consistent with those in columns 1–4, given that non-*mita* districts had more large properties that could be distributed to smallholders during the agrarian reform.[30]

4.2. *Public Goods*

Table VII examines the *mita*'s impact on education in 1876, 1940, and 2001, providing two sets of interesting results.[31] First, there is some evidence that the *mita* lowered access to education historically, although point estimates are imprecisely estimated by the longitude–latitude RD polynomial. In column 1, the dependent variable is the district's mean literacy rate, obtained from the 1876 Population Census (Dirección de Estadística del Perú (1878)). Individuals are defined as literate if they could read, write, or both. Panels B and C show a highly significant *mita* effect of around 2 percentage points, as compared to an average literacy rate of 3.6% in the region I examine. The estimated effect is smaller, at around one percentage point, and not statistically significant, when estimated using the more flexible longitude–latitude specification.[32] In column 2, the dependent variable is mean years of schooling by district, from the 1940 Population Census (Dirección de Estadística del Perú 1944). The specifications reported in panels A–C suggest a long-run negative *mita* effect of around 0.2 years, as compared to a mean schooling attainment of 0.47 years throughout the study region, which again is statistically significant in panels B and C. While this provides support for a *mita* effect on education historically, the evidence for an effect today is weak. In column 3, the dependent variable is individual years of schooling, obtained from ENAHO (2001). The *mita* coefficient is negative in all panels, but is of substantial magnitude and marginally significant only in panel A.[33] It is also statistically insignificant in most specifications in Table A.III. This evidence is consistent with studies of the Peruvian educational

[30]The 1994 Agricultural Census also documents that a similar percentage of households across the *mita* boundary held formal titles to their land.

[31]Education, roads, and irrigation are the three public goods traditionally provided in Peru (Portocarrero, Beltran, and Zimmerman (1988)). Irrigation has been almost exclusively concentrated along the coast.

[32]In some of the specifications in Table A.III in the Supplemental Material that interact the RD polynomial with the *mita* dummy, the estimated *mita* effect is near 0. This discrepancy is explained by two *mita* districts with relatively high literacy located near the *mita* boundary, to which these specifications are sensitive. When these two observations are dropped, the magnitude of the effect is similar across specifications.

[33]Data from the 1981 Population Census (INEI (1981)) likewise do not show a *mita* effect on years of schooling. Moreover, data collected by the Ministro de Educación in 2005 reveal no sys-

TABLE VII
EDUCATION[a]

	Dependent Variable		
	Literacy 1876 (1)	Mean Years of Schooling 1940 (2)	Mean Years of Schooling 2001 (3)
	Panel A. Cubic Polynomial in Latitude and Longitude		
Mita	−0.015	−0.265	−1.479*
	(0.012)	(0.177)	(0.872)
R^2	0.401	0.280	0.020
	Panel B. Cubic Polynomial in Distance to Potosí		
Mita	−0.020***	−0.181**	−0.341
	(0.007)	(0.078)	(0.451)
R^2	0.345	0.187	0.007
	Panel C. Cubic Polynomial in Distance to *Mita* Boundary		
Mita	−0.022***	−0.209***	−0.111
	(0.006)	(0.076)	(0.429)
R^2	0.301	0.234	0.004
Geo. controls	yes	yes	yes
Boundary F.E.s	yes	yes	yes
Mean dep. var.	0.036	0.470	4.457
Clusters	95	118	52
Observations	95	118	4038

[a]The unit of observation is the district in columns 1 and 2 and the individual in column 3. Robust standard errors, adjusted for clustering by district, are in parentheses. The dependent variable is mean literacy in 1876 in column 1 (Dirección de Estadística del Perú (1878)), mean years of schooling in 1940 in column 2 (Dirección de Estadística del Perú (1944)), and individual years of schooling in 2001 in column 3 (ENAHO (2001)). Panel A includes a cubic polynomial in the latitude and longitude of the observation's district capital, panel B includes a cubic polynomial in Euclidean distance from the observation's district capital to Potosí, and panel C includes a cubic polynomial in Euclidean distance to the nearest point on the *mita* boundary. All regressions include geographic controls and boundary segment fixed effects. The samples include districts whose capitals are less than 50 km from the *mita* boundary. Columns 1 and 2 are weighted by the square root of the district's population. 64% of the observations are in *mita* districts in column 1, 63% in column 2, and 67% in column 3. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

sector, which emphasize near-universal access (Saavedra and Suárez (2002), Portocarrero and Oliart (1989)).

What about roads, the other principal public good in Peru? I estimate the *mita*'s impact using a GIS road map of Peru produced by the Ministro de Transporte (2006). The map classifies roads as paved, gravel, nongravel, and *trocha*

tematic differences in primary or secondary school enrollment or completion rates. Examination of data from a 2006 census of schools likewise showed little evidence for a causal impact of the *mita* on school infrastructure or the student-to-teacher ratio.

TABLE VIII

ROADS[a]

	Dependent Variable		
	Density of Local Road Networks (1)	Density of Regional Road Networks (2)	Density of Paved/Gravel Regional Roads (3)
	Panel A. Cubic Polynomial in Latitude and Longitude		
Mita	0.464	−29.276*	−22.426*
	(18.575)	(16.038)	(12.178)
R^2	0.232	0.293	0.271
	Panel B. Cubic Polynomial in Distance to Potosí		
Mita	−1.522	−32.644***	−30.698***
	(12.101)	(8.988)	(8.155)
R^2	0.217	0.271	0.256
	Panel C. Cubic Polynomial in Distance to *Mita* Boundary		
Mita	0.535	−35.831***	−32.458***
	(12.227)	(9.386)	(8.638)
R^2	0.213	0.226	0.208
Geo. controls	yes	yes	yes
Boundary F.E.s	yes	yes	yes
Mean dep. var.	85.34	33.55	22.51
Observations	185	185	185

[a]The unit of observation is the district. Robust standard errors are in parentheses. The road densities are defined as total length in meters of the respective road type in each district divided by the district's surface area, in kilometers squared. They are calculated using a GIS map of Peru's road networks (Ministro de Transporte (2006)). Panel A includes a cubic polynomial in the latitude and longitude of the observation's district capital, panel B includes a cubic polynomial in Euclidean distance from the observation's district capital to Potosí, and panel C includes a cubic polynomial in Euclidean distance to the nearest point on the *mita* boundary. All regressions include geographic controls and boundary segment fixed effects. The samples include districts whose capitals are less than 50 km from the *mita* boundary. 66% of the observations are in *mita* districts. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

carrozable, which translates as "narrow path, often through wild vegetation ... that a vehicle can be driven on with great difficulty" (Real Academia Española (2006)). The total length (in meters) of district roads is divided by the district surface area (in kilometers squared) to obtain a road network density.

Column 1 of Table VIII suggests that the *mita* does not impact local road networks, which consist primarily of nongravel and *trocha* roads. Care is required in interpreting this result, as the World Bank's Rural Roads program, operating since 1997, has worked to reduce disparities in local road networks in marginalized areas of Peru. In contrast, there are significant disparities in regional road networks, which connect population centers to each other. Column 2 in panel A estimates that a *mita* effect lowers the density of regional

roads by a statistically significant −29.3 meters of roadway for every square kilometer of district surface area (s.e. = 16.0). In panels B and C, the coefficients are similar, at −32.6 and −35.8, respectively, and are significant at the 1% level. This large effect compares to an average road density in *mita* districts of 20. Column 3 breaks down the result by looking only at the two highest quality road types—paved and gravel—and a similar picture emerges.[34]

If substantial population and economic activity endogenously clustered along roads, the relative poverty of *mita* districts would not be that surprising. While many of Peru's roads were built or paved in the interlude between 1940 and 1990, aggregate population responses appear minimal. The correlation between 1940 district population density and the density of paved and gravel roads, measured in 2006, is 0.58; when looking at this correlation using 1993 population density, it remains at 0.58.

In summary, while I find little evidence that a *mita* effect persists through access to schooling, there are pronounced disparities in road networks across the *mita* boundary. Consistent with this evidence, I hypothesize that the long-term presence of large landowners provided a stable land tenure system that encouraged public goods provision.[35] Because established landowners in non-*mita* districts controlled a large percentage of the productive factors and because their property rights were secure, it is probable that they received higher returns to investing in public goods than those inside the *mita* catchment. Moreover, historical evidence indicates that these landowners were better able to secure roads, through lobbying for government resources and organizing local labor, and these roads remain today (Stein (1980, p. 59)).[36]

4.3. *Proximate Determinants of Household Consumption*

This section examines the *mita*'s long-run effects on the proximate determinants of consumption. The limited available evidence does not suggest differences in investment, so I focus on the labor force and market participation.[37] Agriculture is an important economic activity, providing primary employment for around 70% of the population in the region examined. Thus, Table IX begins by looking at the percentage of the district labor force whose primary occupation is agriculture, taken from the 1993 Population Census. The median

[34] 18% of *mita* districts can be accessed by paved roads versus 40% of non-*mita* districts (INEI (2004)).

[35] The elasticity of equivalent consumption in 2001 with respect to *haciendas* per capita in 1689, in non-*mita* districts, is 0.036 (s.e. = 0.022).

[36] The first modern road building campaigns occurred in the 1920s and many of the region's roads were constructed in the 1950s (Stein (1980), Capuñay (1951, pp. 197–199)).

[37] Data from the 1994 Agricultural Census on utilization of 15 types of capital goods and 12 types of infrastructure for agricultural production do not show differences across the *mita* boundary, nor is the length of fallowing different. I am not aware of data on private investment outside of agriculture.

TABLE IX

CONSUMPTION CHANNELS[a]

	Dependent Variable		
	Percent of District Labor Force in Agriculture—1993 (1)	Agricultural Household Sells Part of Produce in Markets—1994 (2)	Household Member Employed Outside the Agricultural Unit—1994 (3)
	Panel A. Cubic Polynomial in Latitude and Longitude		
Mita	0.211	−0.074**	−0.013
	(0.140)	(0.036)	(0.032)
R^2	0.177	0.176	0.010
	Panel B. Cubic Polynomial in Distance to Potosí		
Mita	0.101	−0.208***	−0.033
	(0.061)	(0.030)	(0.020)
R^2	0.112	0.144	0.008
	Panel C. Cubic Polynomial in Distance to *Mita* Boundary		
Mita	0.092*	−0.225***	−0.038**
	(0.054)	(0.032)	(0.018)
R^2	0.213	0.136	0.006
Geo. controls	yes	yes	yes
Boundary F.E.s	yes	yes	yes
Mean dep. var.	0.697	0.173	0.245
Clusters	179	178	182
Observations	179	160,990	183,596

[a]Robust standard errors, adjusted for clustering by district in columns 2 and 3, are in parentheses. The dependent variable in column 1 is the percentage of the district's labor force engaged in agriculture as a primary occupation (INEI (1993)), in column 2 it is an indicator equal to 1 if the agricultural unit sells at least part of its produce in markets, and in column 3 it is an indicator equal to 1 if at least one member of the household pursues secondary employment outside the agricultural unit (INEI (1994)). Panel A includes a cubic polynomial in the latitude and longitude of the observation's district capital, panel B includes a cubic polynomial in Euclidean distance from the observation's district capital to Potosí, and panel C includes a cubic polynomial in Euclidean distance to the nearest point on the *mita* boundary. All regressions include geographic controls and boundary segment fixed effects. Column 1 is weighted by the square root of the district's population. 66% of the observations in column 1 are in *mita* districts, 68% in column 2, and 69% in column 3. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

point estimate on $mita_d$ is equal to 0.10 and marginally significant only in panel C, providing some weak evidence for a *mita* effect on employment in agriculture. Further results (not shown) do not find an effect on male and female labor force participation and hours worked.

The dependent variable in column 2, from the 1994 Agricultural Census, is a dummy equal to 1 if the agricultural household sells at least part of its produce in market. The corpus of evidence suggests we can be confident that the *mita*'s effects persist in part through an economically meaningful impact on

agricultural market participation, although the precise magnitude of this effect is difficult to convincingly establish given the properties of the data and the mechanics of RD. The cubic longitude–latitude regression estimates a long-run *mita* effect of −0.074 (s.e. = 0.036), which is significant at the 5% level and compares to a mean market participation rate in the study region of 0.17. The magnitude of this estimate differs substantially from estimates that use a cubic polynomial in distance to Potosí (panel B, −0.208, s.e. = 0.030) and a cubic polynomial in distance to the *mita* boundary (panel C, −0.225, s.e. = 0.032). It also contrasts to the estimate from ordinary least squares limiting the sample to districts bordering the boundary (−0.178, s.e. = 0.050).

The surface plots in Figure 3 shed some light on why the cubic longitude–latitude point estimate is smaller. They show predicted values in "longitude–latitude–market participation rate" space from regressing the market participation dummy on the *mita* dummy (upper left), the *mita* dummy and a linear polynomial in longitude–latitude (upper right), the *mita* dummy and a quadratic polynomial in longitude–latitude (lower left), or the *mita* dummy and a cubic polynomial in longitude–latitude (lower right).[38] The *mita* region is seen from the side, appearing as a "canyon" with lower market participation values. In the surface plot with the cubic polynomial, which is analogous to the regression in panel A, the function increases smoothly and steeply, by orders of magnitude, near the *mita* boundary. In contrast, the other plots model less of the steep variation near the boundary as smooth and thus estimate a larger discontinuity. The single-dimensional RDs likewise have fewer degrees of freedom to model the variation near the boundary as smooth. It is not obvious which specification produces the most accurate results, as a more flexible specification will not necessarily yield a more reliable estimate. For example, consider the stylized case of an equation that includes the *mita* dummy and a polynomial with as many terms as observations. This has a solution that perfectly fits the data with a discontinuity term of zero, regardless of how large the true *mita* effect is. On the other hand, flexibility is important if parsimonious specifications do not have enough degrees of freedom to accurately model smoothly changing unobservables. While there is not, for example, a large urban area at the peak of the cubic polynomial causing market participation to increase steeply in this region, it is difficult to conclusively argue that the variation is attributable to the discontinuity and not to unobservables, or vice versa.[39] The estimates in Tables IX and A.III are most useful for determining a range of

[38]I show three-dimensional surface plots, instead of shaded plots as in Figure 2, because the predicted values can be seen more clearly and it is not necessary to plot the data points.

[39]Note, however, that the relatively large (*mita*) urban area of Ayacucho, while outside the study region, is near the cluster of *mita* districts with high market participation in the upper left corner of the *mita* area.

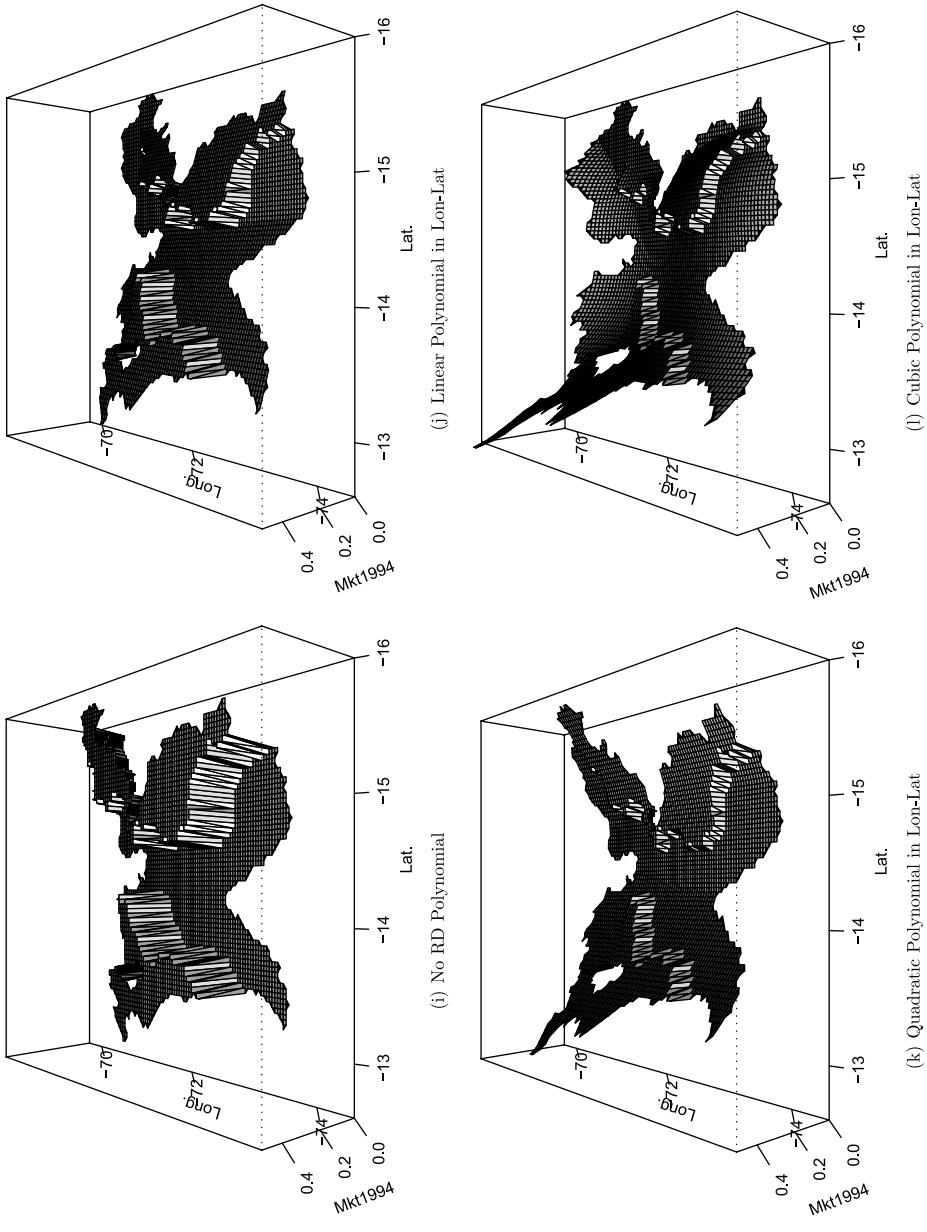


FIGURE 3.—Plots of predicted values from regressing a market participation dummy on the *mita* dummy and various degrees of polynomials in longitude and latitude. See the text for a detailed description.

possible *mita* effects consistent with the data, and this range supports an economically meaningful *mita* effect on market participation.[40]

A *mita* effect on market participation is consistent with the findings on road networks, particularly given that recent studies on Andean Peru empirically connect poor road infrastructure to higher transaction costs, lower market participation, and reduced household income (Escobal and Ponce (2002), Escobal (2001), Agreda and Escobal (1998)).[41] An alternative hypothesis is that agricultural producers in *mita* districts supplement their income by working as wage laborers rather than by producing for markets. In column 3, the dependent variable is an indicator equal to 1 if a member of the agricultural household participates in secondary employment outside the agricultural unit, also taken from the 1994 Agricultural Census. Estimates suggest that, if anything, the *mita* effect on participation in secondary employment is negative.

Could residents in *mita* districts have less desire to participate in the market economy, rather than being constrained by poor road infrastructure? While Shining Path, a Maoist guerilla movement, gained a strong foothold in the region during the 1980s, this hypothesis seems unlikely.[42] Shining Path's rise to power occurred against a backdrop of limited support for Maoist ideology, and the movement's attempts to reduce participation in markets were unpopular and unsuccessful where attempted (McClintock (1998), Palmer (1994)).

Recent qualitative evidence also underscores roads and market access. The citizens I spoke with while visiting eight primarily *mita* and six primarily non-*mita* provinces were acutely aware that some areas are more prosperous than

[40]The specifications interacting the *mita* dummy with a linear or quadratic polynomial in distance to the *mita* boundary, reported in Table A.III, do not estimate a significant *mita* effect. Graphical evidence suggests that these specifications are sensitive to outliers near the boundary.

[41]In my sample, 33% of agricultural households in districts with paved road density above the median participate in markets, as compared to 13% in districts with paved road density below the median. Of course, there may also exist other channels through which a *mita* effect lowers market participation. Data from the 1994 Agricultural Census reveal that the median size of household landholdings is somewhat lower inside the *mita* catchment (at 1.2 hectares) than outside (at 1.4 hectares). If marketing agricultural produce involves fixed costs, a broader group of small farmers in non-*mita* districts may find it profitable.

[42]Many of the factors linked to the *mita* (poor infrastructure, limited access to markets, poorly defined property rights, and poverty) are heavily emphasized as the leading factors promoting Shining Path (Comisión de la Verdad y Reconciliación (2003, Vol. 1, p. 94), McClintock (1998), Palmer (1994)). Thus, I tested whether there was a *mita* effect on Shining Path (results available upon request). To measure the intensity of Shining Path, I exploit a loophole in the Peruvian constitution that stipulates that when more than two-thirds of votes cast are blank or null, authorities cannot be renewed (Pareja and Gatti (1990)). In an attempt to sabotage the 1989 municipal elections, Shining Path operatives encouraged citizens to cast blank or null (secret) ballots (McClintock (1998, p. 79)). I find that a *mita* effect increased blank/null votes by 10.7 percentage points (s.e. = 0.031), suggesting greater support for and intimidation by Shining Path in *mita* districts. Moreover, estimates show that a *mita* effect increased the probability that authorities were not renewed by a highly significant 43.5 percentage points. I also look at blank/null votes in 2002, 10 years after Shining Path's defeat, and there is no longer an effect.

others. When discussing the factors leading to the observed income differences, a common theme was that it is difficult to transport crops to markets. Thus, most residents in *mita* districts are engaged in subsistence farming. Agrarian scientist Gonzales Castro (2006) argued, "Some provinces have been favored, with the government—particularly during the large road building campaign in the early 1950s—choosing to construct roads in some provinces and completely ignore others." At the forefront of the local government's mission in the (primarily *mita*) province of Espinar is "to advocate effectively for a system of modern roads to regional markets" Espinar Municipal Government (2008). Popular demands have also centered on roads and markets. In 2004, (the *mita* district) Ilave made international headlines when demonstrations involving over 10,000 protestors culminated with the lynching of Ilave's mayor, whom protestors accused of failing to deliver on promises to pave the town's access road and build a local market (Shifter (2004)).

5. CONCLUDING REMARKS

This paper documents and exploits plausible exogenous variation in the assignment of the *mita* to identify channels through which it influences contemporary economic development. I estimate that its long-run effects lower household consumption by around 25% and increase stunting in children by around 6 percentage points. I then document land tenure, public goods, and market participation as channels through which its impacts persist.

In existing theories about land inequality and long-run growth, the implicit counterfactual to large landowners in Latin America is secure, enfranchised smallholders (Engerman and Sokoloff (1997)). This is not an appropriate counterfactual for Peru, or many other places in Latin America, because institutional structures largely in place before the formation of the landed elite did not provide secure property rights, protection from exploitation, or a host of other guarantees to potential smallholders. Large landowners—while they did not aim to promote economic prosperity for the masses—did shield individuals from exploitation by a highly extractive state and did ensure public goods. This evidence suggests that exploring constraints on how the state can be used to shape economic interactions—for example, the extent to which elites can employ state machinery to coerce labor or citizens can use state guarantees to protect their property—is a more useful starting point than land inequality for modeling Latin America's long-run growth trajectory. The development of general models of institutional evolution and empirical investigation of how these constraints are influenced by forces promoting change are particularly central areas for future research.

REFERENCES

ACEMOGLU, D., M. A. BAUTISTA, P. QEURUBIN, AND J. A. ROBINSON (2008): "Economic and Political Inequality in Development: The Case of Cundinamarca, Colombia," in *Institutions*

and Economic Performance, ed. by E. Helpman. Cambridge, MA: Harvard University Press. [1866]

ACEMOGLU, D., S. JOHNSON, AND J. A. ROBINSON (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91, 1369–1401. [1863]

——— (2002): "Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution," *Quarterly Journal of Economics*, 117, 1231–1294. [1863,1865]

AGREDA, V., AND J. ESCOBAL (1998): "Analasís de la comercialización agrícola en el Perú," *Boletín de opinión*, 33. [1898]

AMAT Y JUNIENT, M. (1947): *Memoria de Gobierno*. Sevilla: Escuela de Estudios Hispano-Americanos. [1869,1878]

ANGRIST, J., AND J. PISCHKE (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press. [1875]

AVILA, J. (1952): "Exposición e causas que justifican la necesidad de la reforma agraria en el distrito de Azángaro de la provincia del mismo nombre, Puno," Ph.D. Thesis, Universidad Nacional del Cusco. [1890]

BAKEWELL, P. (1984): *Miners of the Red Mountain. Indian Labor in Potosi, 1545–1650*. Albuquerque: University of New Mexico Press. [1867,1868]

BANERJEE, A., AND L. IYER (2005): "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India," *American Economic Review*, 95, 1190–1213. [1865]

BAUER, B., AND A. COVEY (2002): "Processes of State Formation in the Inca Heartland (Cusco, Peru)," *American Anthropologist*, 104, 846–864. [1874]

BLACK, S. (1999): "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, 114, 577–599. [1875]

BRISSEAU, J. (1981): *Le Cuzco dans sa région: Étude de l'aire d'influence d'une ville andine*. Lima: Institut français d'études andines. [1888]

BUSTAMANTE OTERO, L. H. (1987): "Mita y realidad: Teodomerio Gutierrez Cuevas o Rumi Maqui en el marco de la sublevación campesina de Azángaro: 1915–1916," Ph.D. Thesis, Pontificia Universidad Católica del Perú. [1865]

CAÑETE, P. V. (1794): *El codigo carolino de ordenanzas reales de las minas de Potos y demas provincias del Río de la Plata*. Buenos Aires: E Martiré. Reprinted in 1973. [1868]

CAPUÑAY, M. (1951): *Leguia, vida y obra del constructor del gran Perú*. Lima: Enrique Bustamante y Bellivian. [1894]

CENTER FOR INTERNATIONAL EARTH SCIENCE INFORMATION (2004): *Global Rural-Urban Mapping Project, Alpha Version: Population Grids*. Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. Available at http://sedac.ciesin.columbia.edu/gpw (May 10th, 2007). [1871]

CIEZA DE LEÓN, P. (1551): *El Senorio de los Incas; 2a. Parte de la Crónica del Perú*. Lima: Instituto de Estudios Peruanos. Reprinted in 1967. [1867]

——— (1959): *The Incas*. Norman: University of Oklahoma Press. [1871]

COATSWORTH, J. (2005): "Structures, Endowments, and Institutions in the Economic History of Latin America," *Latin American Research Review*, 40, 126–144. [1863,1866]

COLE, J. (1985): *The Potosí Mita 1573–1700. Compulsory Indian Labor in the Andes*. Stanford: Stanford University Press. [1867,1868,1876,1888]

COMISIÓN DE LA VERDAD Y RECONCILIACIÓN (2003): *Informe Final*. Lima: Comision de la Verdad y Reconciliacion. [1898]

CONLEY, T. (1999): "GMM Estimation With Cross Sectional Dependence," *Journal of Econometrics*, 92, 1–45. [1871]

COOK, N. D. (1981): *Demographic Collapse: Colonial Peru 1520–1620*. Cambridge: Cambridge University Press. [1867,1871,1874,1888]

D'ALTOY, T. (2002): *The Incas*. Oxford: Blackwell. [1867]

DANCUART, E., AND J. RODRIGUEZ (1902): *Anales de la hacienda pública del Perú*, 24 Vols. Lima: Imprenta de La Revista y Imprenta de Guillermo Stolte. [1865,1890]

DEATON, A. (1997): *The Analysis of Household Surveys: A Microeconomic Approach to Development Policy*. Baltimore: Johns Hopkins University Press. [1877,1879]

DELL, M. (2010): "Supplement to 'The Persistent Effects of Peru's Mining *Mita*'," *Econometrica Supplementary Material*, 78, http://www.econometricsociety.org/ecta/Supmat/8121_data description.pdf; http://www.econometricsociety.org/ecta/Supmat/8121_data_and_programs.zip. [1867]

DIRECCIÓN DE ESTADÍSTICA DEL PERÚ (1878): *Censo general de la república del Perú, formatdo en 1876*. Lima: Imp. del Teatro. [1891,1892]

——— (1944): *Censo Nacional de Población y Ocupación, 1940*. Lima: Imprenta Torres Aguirree. [1889,1891,1892]

EASTERLY, W., AND R. LEVINE (2003): "Tropics, Germs, and Crops: How Endowments Influence Economic Development," *Journal of Monetary Economics*, 50, 3–39. [1863]

ENGERMAN, S., AND K. SOKOLOFF (1997): "Factors Endowments, Institutions, and Differential Paths of Growth Among New World Economies," in *How Latin American Fell Behind*, ed. by S. Haber. Stanford: Stanford University Press, 260–304. [1863,1866,1899]

ESCOBAL, J. (2001): "The Benefits of Roads in Rural Peru: A Transaction Costs Approach," Working Paper, Grupo de Analisis para el Desarrollo, Lima. [1898]

ESCOBAL, J., AND C. PONCE (2002): "The Benefits of Rural Roads: Enhancing Income Opportunities for the Rural Poor," Working Paper 40, Grupo de Analisis para el Desarrollo, Lima. [1898]

ESPINAR MUNICIPAL GOVERNMENT (2008): "Vision," available at www.muniespinar.gob.pe/vision.php. [1899]

FAVRE, H. (1967): "Evolución y situación de las haciendas en la region de Huancavelica, Perú," in *La Hacienda en el Perú*, ed. by H. Favre, C. Collin-Delavaud, and J. M. Mar. Lima: Instituto de Estudios Peruanos, 237–257. [1890]

FLORES GALINDO, A. (1987): *Buscando un inca: Identidad y utopia en los Andes*. Lima: Instituto de Apoyo Agrario. [1865,1890]

GARRETT, D. (2005): *Shadows of Empire: The Indian Nobility of Cusco, 1750–1825*. Cambridge: Cambridge University Press. [1865,1867,1888]

GLAESER, E., AND A. SHLEIFER (2002): "Legal Origins," *Quarterly Journal of Economics*, 117, 1193–1230. [1863,1865]

GLAESER, E., R. LAPORTA, F. L. DE SILANES, AND A. SHLEIFER (2004): "Do Institutions Cause Growth?" *Journal of Economic Growth*, 9, 271–303. [1863]

GLAVE, L. M. (1989): *Trajinantes: Caminos indígenas de la sociedad colonial*. Lima: Instituto de Apoyo Agraria. [1868,1876]

GLAVE, L. M., AND M. I. REMY (1978): *Estructura Agraria e historia rural cuzqueña: A proposito de las haciendas del Valle Sagrado: 1780–1930*. Cuzco: Archivo Historico del Cuzco. [1888]

GOLTE, J. (1980): *La racionalidad de la organización andina*. Lima: Insituto de Estudios Peruanos. [1868,1874]

GONZALES CASTRO, E. (2006): Personal Interview. 14 December. [1899]

HALL, R., AND C. JONES (1999): "Why do Some Countries Produce so Much More Output per Worker Than Others?" *Quarterly Journal of Economics*, 114, 83–116. [1863]

HAZEN, D. (1974): "The Awakening of Puno: Government Policy and the Indian Problem in Southern Peru, 1900–1955," Ph.D. Thesis, Yale University, New Haven. [1890]

HIJMANS, R. ET AL. (2005): "Very High Resolution Interpolated Climate Surfaces for Global Land Area," *International Journal of Climatology*, 25, 1965–1978. [1874]

HYSLOP, J. (1984): *The Inka Road System*. New York: Academic Press. [1869]

IMBENS, G., AND T. LEMIEUX (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615–635. [1875]

INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMACIÓN DE PERÚ (INEI) (1981): *VIII Censo de Población*. Lima, Peru: Instituto Nacional de Estadistica e Informacion, Direccion Nacional de Censos y Encuestas. [1891]

——— (1993): *IX Censo de Población*. Lima, Peru: Instituto Nacional de Estadistica e Informacion, Direccion Nacional de Censos y Encuestas. [1867,1877,1895]

——— (1994): *Tercer Censo Nacional Agropecuario*. Lima, Peru: Instituto Nacional de Estadistica e Informacion, Direccion Nacional de Censos y Encuestas. [1889,1891,1895]

——— (2001): *Encuesta Nacional de Hogares (ENAHO)*. Lima, Peru: Instituto Nacional de Estadistica e Informacion, Direccion Nacional de Censos y Encuestas. [1873,1878,1891,1892]

——— (2004): *Registro Nacional de Municipalidades (REMANU)*. Lima, Peru: Instituto Nacional de Estadistica e Informacion, Direccion Nacional de Censos y Encuestas. [1894]

INSTITUTO NACIONAL DE RECURSOS NATURALES (INRENA) (1997): *Potencial Bioclimatica*. Lima: INRENA. [1871]

JACOBSEN, N. (1993): *Mirages of Transition: The Peruvian Altiplano, 1780–1930*. Berkeley: University of California Press. [1865,1890]

KEITH, R. G. (1971): "Encomienda, Hacienda and Corregimiento in Spanish America: A Structural Analysis," *Hispanic American Historical Review*, 51, 431–446. [1887,1888]

LARSON, B. (1982): *Exploitación Agraria y Resistencia Campesina en Cochabamba*. La Paz: Centro de Estudios de la Realidad Economico y Social. [1888]

——— (1988): *Colonialism and Agrarian Transformation in Bolivia: Cochabamba, 1550–1900*. Princeton: Princeton University Press. [1865,1888]

LEVILLIER, R. (1921): *Gobernantes del Perú, cartas y papeles, siglo XVI: Documentos del Archivo de Indias*. Madrid: Sucesores de Rivadeneyra. [1868]

MAR, M., AND M. MEJIA (1980): *La reforma agraria en el Perú*. Lima: Instituto de Estudios Peruanos. [1890]

MCCLINTOCK, C. (1998): *Revolutionary Movements in Latin America: El Salvador's FMLN and Peru's Shining Path*. Washington, DC: United States Institute of Peace Press. [1898]

MINISTRO DE EDUCACIÓN DEL PERÚ (MINEDU) (2005a): *Censo de Talla*. Lima: Ministro de Educacion del Peru. [1878]

——— (2005b): *Indicadores de cobertura y culminación de la educación básica*. Lima: Ministro de Educacion del Peru. Available at www.minedu.gob.pe. [1870]

MINISTRO DE TRANSPORTE (2006): "Red Vial en GIS," unpublished data compiled by Ministro de Transporte, Peru. [1892,1893]

MIRANDA, C. (1583): *Tasa de la visita general de Francisco Toledo*. Lima, Peru: Universidad Nacional de San Marcos, Dirección Universitaria de Biblioteca y Publicaciones. [1873,1886]

MORNER, M. (1978): *Perfil de la sociedad rural del Cuzco a fines de la colonia*. Lima: Universidad del Pacifico. [1877,1888]

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION AND THE NATIONAL GEOSPATIAL-INTELLIGENCE AGENCY (2000): *Shuttle Radar Topography Mission 30 Arc Second Finished Data*. Available at http://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10008_1. [1870,1873]

NILES, S. (1987): *Callachaca: Style and Status in an Inca Community*. Iowa City: University of Iowa Press. [1884]

NUÑEZ, J. T. (1913): *Memoria leida de la ceremonía de apertura del año judicial de 1913 por el Presidente de la Ilustrisima Corte Superior de los departamentos de Puno y Madre de Dios, Dr. J. Teofilo Nuñez*. Puno: Imprenta del Seminario. [1890]

NUNN, N. (2008): "The Long-Term Effects of Africa's Slave Trades," *Quarterly Journal of Economics*, 123, 139–176. [1863,1865]

PALMER, D. S. (1994): *The Shining Path of Peru*. New York: St. Martin's Press. [1898]

PAREJA PFLUCKER, P., AND A. G. GATTI MURRIEL (1990): *Evaluación de las elecciones municipales de 1989: Impacto politico de la violencia terrorista*. Lima: Instituto Nacional de Planificación. [1898]

PERALTA RUÍZ, V. (1991): *En pos del tributo: Burocracia estatal, elite regional y comunidades indígenas en el Cusco rural (1826–1854)*. Cusco: Casa Bartolome de las Casas. [1888,1889]

PETERSON, T. C., AND R. S. VOSE (1997): "An Overview of the Global Historical Climatology Network Temperature Data Base," *Bulletin of the American Meteorological Society*, 78, 2837–2849. [1874]

PORTOCARRERO, F., A. BELTRAN, AND A. ZIMMERMAN (1988): *Inversiones públicas en el Perú (1900–1968). Una aproximación cuantitativa*. Lima: CIUP. [1891]
PORTOCARRERO, G., AND P. OLIART (1989): *El Perú desde la escuela*. Lima: Instituto de Apoyo Agrario. [1892]
PULGAR VIDAL, J. (1950): *Geografía del Perú: Las ocho regiones naturales del Perú*. Lima: Editorial Universo. [1874]
RAMOS ZAMBRANO, A. (1984): *La Rebelión de Huancane: 1923–1924*. Puno: Editorial Samuel Frisancho Pineda. [1865,1890]
REAL ACADEMIA ESPAÑOLA (2006): *Diccionario de la lengua Española*. Madrid: Editorial Espasa Calpe. [1893]
ROCA SANCHEZ, P. E. (1935): *Por la clase indígena*. Lima: Pedro Barrantes Castro. [1890]
ROWE, J. (1946): "Inca Culture at the Time of the Spanish Conquest," in *Handbook of South American Indians*, Vol. 2, ed. by J. Steward. Washington, DC: Bureau of American Ethnology, 183–330. [1867]
SAAVEDRA, J., AND P. SUÁREZ (2002): *El financiamiento de la educación pública en el Perú: El rol de las familias*. Lima, Peru: GRADE. [1892]
SACHS, J. (2001): "Tropical Underdevelopment," Working Paper 8119, NBER. [1863]
SAIGNES, T. (1984): "Las etnias de Charcas frente al sistema colonial (Siglo XVII): Ausentismo y fugas en el debate sobre la mano de obra indígena, 1595–1665," *Jahrbuchfr Geschichte (JAS)*, 21, 27–75. [1869,1878]
SANCHEZ-ALBORNOZ, N. (1978): *Indios y tributos en el Alto Perú*. Lima: Instituo de Estudios Peruanos. [1867,1888]
SHIFTER, M. (2004): "Breakdown in the Andes," *Foreign Affairs*, September/October, 126–138. [1899]
STEIN, S. (1980): *Populism in Peru: The Emergence of the Masses and the Politics of Social Control*. Madison: University of Wisconsin Press. [1866,1894]
TAMAYO HERRERA, J. (1982): *Historia social e indigenismo en el Altiplano*. Lima: Ediciones Treintaitres. [1890]
TANDETER, E. (1993): *Coercion and Market: Silver Mining in Colonial Potosí, 1692–1826*. Albuquerque: University of New Mexico Press. [1867,1868,1876,1888]
VILLANUEVA URTEAGA, H. (1982): *Cuzco 1689: Informes de los párrocos al obispo Mollinedo*. Cusco: Casa Bartolomé de las Casas. [1867,1888,1889]
WIGHTMAN, A. (1990): *Indigenous Migration and Social Change: The Forasteros of Cuzco, 1570–1720*. Durham: Duke University Press. [1867,1889]
ZAVALA, S. (1980): *El servicio personal de los indios en el Perú*. Mexico City: El Colegio de México. [1867]

Dept. of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, E52, Cambridge, MA 02142, U.S.A.; mdell@mit.edu.

Manuscript received September, 2008; final revision received January, 2010.

A THEORY OF MISGOVERNANCE*

ABHIJIT V. BANERJEE

This paper tries to explain why government bureaucracies are often associated with red tape, corruption, and lack of incentives. The paper identifies two specific ingredients that together can provide an explanation: the fact that governments often act precisely in situations where markets fail and the presence of agency problems within the government. We show that these problems are exacerbated at low levels of development and in bureaucracies dealing with poor people. We also argue that we need to posit the existence of a welfare-oriented constituency within the government in order to explain red tape and corruption.

I. INTRODUCTION

I.A. Goals of a Theory of Misgovernance

The stereotypical view of government bureaucrats, as articulated in the press for example, is that they are lacking in incentives, obsessed with red tape, and probably corrupt. The point of departure of this paper is that while such views may well be correct, it is worth understanding to what extent these phenomena can be explained without departing from the standard paradigm where the government is a benevolent social planner. In other words, we are looking for an explanation of government failures that makes no reference to the rapacity of governments, their monopoly of state power, or the unique sociological status of governments.[1]

To pose the problem in this way is not to deny that some

*This paper was inspired by many conversations with Andrei Shleifer. Two anonymous referees made extremely helpful comments. I have also profited from comments by Daron Acemoglu, Andres Almazan, Dipak Banerjee, Tuli Banerjee, Gary Becker, Douglas Bernheim, Christopher Clague, Peter Diamond, Avinash Dixit, Drew Fudenberg, Oliver Hart, Patrick Legros, Eric Maskin, Andrew Newman, Thomas Piketty, Rohini Somanathan, Jean Tirole, Jorgen Weibull, and seminar participants at Princeton University, the University of British Columbia, the Kennedy School of Government at Harvard University, the University of Chicago, and Harvard University. Some of these arguments were in the notes that were circulated some time ago as "The Costs and Benefits of Corruption." This work was carried out when the author was an IPR junior fellow. It was supported by financial assistance from IRIS and the National Science Foundation. However, the views expressed here are strictly the author's own. The author also acknowledges the hospitality of DELTA in Paris where this work was started.

1. Theories of the government failures based on the government's rapacity and its monopoly of state power abound. Perhaps the most articulate statement is to be found in the works of Mancur Olson and his followers (see, for example, Olson [1993]). A very different theory of government failures which emphasizes the unique sociological status of modern governments and the consequent limits on what the government can and cannot do, is in Wilson [1989]. See also the formalization of the Wilson's ideas in Dixit [1996].

The Quarterly Journal of Economics, November 1997.

governments are extremely rapacious. Nor is it to deny that the sociological status of governments is both important and interesting. But it is to emphasize that a significant part of what we see as government failures may exist even when a government has the best of intentions and is subject to no special sociological constraints.

To overlook this simple point runs the danger, in our view, of limiting our understanding of where and under what circumstances governments perform relatively well and therefore biasing our policy stances. To take a simple instance, if we observe a high degree of corruption in a particular government bureaucracy and assume that all other bureaucracies in the same government will be equally corrupt, we may recommend against specific forms of government activism that may in fact work well.

The basic claim of this paper is that it is possible to develop a theory of misgovernance by a benevolent government based on two eminently reasonable premises: one, that a substantial part of what governments do is to respond to market failures; and two, like all other organizations, the government has agents who are more interested in their own welfare than in any collective goals. And, perhaps more importantly, the theory set up for the sake of this explanation has sensible and useful implications about the performance of different government bureaucracies under different circumstances.

The model we set up is extremely simple. There are three types of agents: the government, bureaucrats, and the people outside. The government in our model has a set of publicly provided private goods that the people want. It is interested in allocating them in a way that maximizes social welfare. These goods may be educational opportunities; beds in hospitals; licenses to produce, import, or pollute; or even irrigation water.[2] To avoid being unnecessarily specific, we will just call them slots.

These slots are scarce in the sense that the number of people who want them exceeds the number of goods. Not all the people who want these slots value them equally; we assume that there are two types of which one has a higher willingness to pay for the slots. Clearly, in an efficient allocation people of this type should get the slots ahead of the others, and we would typically expect the market to deliver this outcome. However, here this is not nec-

2. Wade [1982] provides a fascinating description of the process of allocation of irrigation water (in Southern India) by a public bureaucracy.

essarily the case because we assume that at least for some people, the willingness to pay is higher than the ability to pay. The obvious reason for such a discrepancy would be a credit market imperfection, but it could also arise out of a labor market imperfection that limits the number of hours someone can work (most jobs actually do this to a greater or a lesser extent).

This assumption of a capital market imperfection is relatively uncontroversial in the context of education or health. It is less obvious that those who are bidding for trade or production licenses are generally credit-constrained, but it may not be unreasonable to assume that this constraint binds for at least some of them. Certainly in the early years of development planning, limited and unequal access to credit was often the stated justification for the licensing of industrial production, imports, exports, and access to foreign exchange.[3]

The fact that the market may fail to allocate the slots efficiently is going to be key to our model. It explains both why the government is involved in the allocation of these goods as well as why imitating the market will not be the best way to allocate them.

In our model the actual allocation of the slots is the responsibility of a bureaucrat. We assume that the bureaucrat cannot observe the value put on a slot by each person who demands it. We also assume that the bureaucrat cares only about his own welfare and that government cannot perfectly monitor the mechanism used by the bureaucrat to allocate the slots. Therefore, there are really two potential incentive problems: the applicants for the slots may lie to the bureaucrat about their willingness to pay, and the bureaucrat may lie to the government about the mechanism he is using.[4]

As we will show, the combination of these quite elementary assumptions yields a model that has a rich set of predictions:

3. While this form of government intervention eventually proved to be a constraint on development and was probably based on an excessive mistrust of the price system, there is little reason to believe that the arguments in their favor were disingenuous. In other words, the eventual abandonment of these systems does not imply that the initial decision to adopt them was not ex ante social welfare maximizing, given the information and the understanding that the government then had.

4. The mechanism design problem that the bureaucrat solves is of some independent interest. There is now a growing literature on general mechanism design problems with credit-constrained agents. See, for example, Aghion and Burgess [1993], Bolton and Roland [1992], Che and Gale [1994], and Lewis and Sappington [1996]. Our paper departs from these in emphasizing the role of red tape in designing such mechanisms.

first, it can explain why bureaucrats will want to use red tape, interpreted as completely pointless bureaucratic procedures that one has to endure in dealing with bureaucracies. Second, the model can explain corruption. Here it is worth emphasizing that in order to explain corruption one needs to explain more than moneymaking by government bureaucrats: one needs to explain *illegal* moneymaking. And to do so, one needs to explain why the government makes it illegal to make money.[5] Third, the model explains why, under certain circumstances, the government will give bureaucrats very low powered incentives or no incentives at all.

At a very different level the model also allows us to ask what would change if the government were interested in making money rather than in social welfare. It turns out that in this case there would be no red tape at all, unless there were unobservable differences in the ability to pay and even when there are such unobservable differences, there will be *less* red tape in this case than in the case where the government is welfare-minded. The same is true of corruption: there would be no corruption in the world of this model if the government did not care about social welfare. In other words, the assumption that the government is rapacious makes it harder to explain red tape and corruption. This is less paradoxical than it appears: as will be explained in the following pages, both corruption and less obviously, red tape, arise out of the government's efforts to control the bureaucrat in the social interest. If the government did not have society's interest at heart, there would be no need to have such controls.

It is also worth asking whether the assumption of agency problems within the government is necessary for our specific results. To check this, we also consider the case where both the government and the bureaucrat are welfare-minded. We show that in this case there will be no red tape and (obviously) no corruption. In other words, a conflict of interest within the government is key to our story.

Finally, the model gives a number of predictions about the determinants of red tape and corruption. In particular, we show that, on the whole, red tape and corruption are more likely to arise where ability to pay is low relative to the willingness to pay,

5. This insistence on explaining why the government makes corruptible rules is what distinguishes our framework from much of the existing literature on corruption (see, for example, Shleifer and Vishny [1993]).

where the goods being allocated are particularly scarce and where there is inequality in the ability to pay. We also find that it is precisely in these environments that bureaucrats may face weak incentives. We interpret these as saying that government failures are most likely in bureaucracies dealing with poorer section of society and in poor countries.[6]

We postpone providing intuition for these results until we have presented the key ingredients of the model. This is the subject of the next subsection. Once the model is presented, we will present some relatively loose analysis that will explain the basic properties of the model and provide intuition for the results claimed above. More formal analysis is provided in the later sections of the paper.

I.B. The Model

We assume that the set of slots being allocated is of Lebesgue measure 1 and the population of applicants to be of Lebesgue measure $N > 1$. The applicants can be of two types, L and H, or alternatively low and high. The low type generates a return L if awarded the slot, while the high type generates a return of H. We assume that these are both the social and private returns and that $L < H$. We assume that the fraction of type H applicants is $N_H < 1$ and that of type L is N_L. Finally, we assume that the applicants are risk-neutral and have quasi-linear preference over slots and money; i.e., if an applicant gets a slot worth H with probability π and pays an amount p_H for it his net utility will be $\pi H - p_H$.

The applicants for the slots are cash-constrained in the sense that their valuation of these slots may exceed their ability to pay for them. We model the cash constraint as an upper bound, y, on each applicant's ability to pay. We do not allow the government to relax this constraint by giving people money on the grounds that if the government started giving away money a lot of people may claim that they want a slot in order to get the money. In this

6. This is consistent with the evidence presented in Mauro [1995] about the correlation between government failures and level of development. We are aware, of course, that there are other reasons why bureaucrats in poorer countries are corrupt. For example, the salaries paid to responsible government servants in many LDCs do not seem to be commensurate with their responsibilities. In other words, it is possible that the bureaucrats in these countries are corrupt because they get paid less than their efficiency wages. However, this begs the question of why the government sets salaries that are so low. Our model has the advantage of giving reasons for why the government may choose to let the bureaucrat make money.

section and the next two we will assume that y is the same for all applicants. This assumption will be relaxed in Section IV.

The slots belong to the government, but the actual allocation of the slots is the responsibility of a bureaucrat. This distinction between the government and the bureaucrat is central to the argument we make here: in our model the bureaucrat chooses the mechanism that is used for allocating the slots, while the government is responsible for rewarding and punishing the bureaucrat.[7]

Regarding the preferences of our two main actors—the bureaucrat and the government—for most of the paper we make the assumption that the bureaucrat cares only about the total amount of money he makes, less the costs of implementing red tape and any other costs, while the government cares only about social welfare.[8] These preferences make the most sense if we assume that both the government and the bureaucrat are risk-neutral and face no liquidity constraint. In this case the government can always satisfy the bureaucrat's participation constraint by making him a lump sum transfer, and, on the other side, if the government feels that the bureaucrat is making too much money and wants to recoup some of the revenue from the sale of the slots, all it has to do is to set a fixed fee for each slot. We will, however, also consider what happens if both the government and the bureaucrat are only interested in making money, as well as the case where both are welfare-minded.

The mechanism chosen by the bureaucrat for allocating the slots will typically combine prices and what we call red tape. In other words, an applicant who wants a slot will have to pay a certain amount and also go through a certain amount of red tape before he gets the slot. We model red tape as a pure waste of time.[9] We assume that going through a unit of red tape costs the

7. The distinction we make here between the government and the bureaucrat parallels the distinction made by Laffont and Tirole [1993] between the constitution-maker and the regulatory agency.

8. The two preferences we have specified are clearly both quite extreme. In reality, a welfare-oriented government may also care about revenue because of budgetary concerns. However, allowing the government to put a small weight on revenue does not change our results. Also the way we have modeled the welfare-oriented preferences assumes that even a welfare-oriented government does not care about how the allocation of the slots affects the distribution of wealth. This is deliberate; allowing the government a more complex objective makes it easier to explain why it might generate inefficient outcomes: our present formulation therefore provides the sharpest test of our theory.

9. Nothing essential would change if we assumed, instead, that red tape actually produces information. Also, despite being a waste of time, screening is an important social function, and therefore we do not interpret the use of red tape per se to be a sign of inefficiency. It is rather the red tape that is in excess of the socially necessary amount that we view as a measure of governmental inefficiency.

applicant δ. These costs may be thought of as the losses in productivity from delays, time costs of waiting in lines, or simply the emotional costs of being harassed. We will assume that this is a nonmonetary cost in the sense that having to bear it does not reduce the applicant's ability to pay.[10] We also assume that the cost per unit of time to the bureaucrat of inflicting red tape on an applicant is v, where v/δ is small.

To complete the model, we need to specify the ways in which the government can provide incentives for the bureaucrat. For the time being, we will assume that the government does not observe the mechanism used by the bureaucrat to allocate the slots: it observes neither the amount of red tape nor the prices charged by the bureaucrat. This assumption is relaxed in Section III, where we allow the government to punish the bureaucrat for using the wrong mechanism but put a bound on such punishments.

However, we do allow the government the possibility of providing the bureaucrat with some incentives on the basis of how the bureaucrat has allocated the slots that were given to him to allocate. There are several alternative ways of introducing such incentives that give more or less equivalent results. Here we choose a formulation that is analytically convenient at the cost of being somewhat crude. We assume the following.

(i) The government samples a small fraction of those who are given slots by the bureaucrat and determines their types. Because of the assumption that the number of slots forms a continuum, the sample tells the government the exact number of slots that went to type L applicants.[11]

10. This is more than we really need to assume: our results only require that the wasted time does not reduce the applicant's ability to pay one for one. Interpreted in this way this assumption seems to be quite consistent with our suggested interpretations.

11. This of course requires that the government can tell who are type L applicants. It is legitimate to ask why, if we allow the government access to a technology for determining the type of the applicant, we also do not do so for the bureaucrats. However, the situation we have in mind is one where it is quite costly to directly establish the applicant's type, and therefore a bureaucrat will not want to do so (especially since, as will become evident, there are cheaper ways to screen). On the other hand, we imagine that each bureaucrat allocates many slots, and therefore, if the government can influence the allocation of all these slots by sampling a small fraction of those who get the slots and determining their types, it may very well be worthwhile.

It may also be the case that it is much more difficult to discover the applicant's true type at the time the slots are being allocated than it is in the long run: information has a way of leaking out on its own over time. Since the bureaucrat typically has a long-term relationship with the government, the government may be able to use this information against the bureaucrat much more easily than the bureaucrat can use it against the person who got the slot.

It is also clear that, ideally, all these arguments should be modeled formally, but we do not see any way of doing this without making the paper unreadable.

(ii) The government imposes a fine F on the bureaucrat for each slot in excess of $1 - N_H$ which goes to an L-type applicant, where $1 - N_H$ is both the fraction of slots that would go to type L applicants in the first-best allocation and the minimum fraction of slots that must go to type L applicants in any allocation. In other words, the bureaucrat who gives slots to N'_L type L applicants, pays a total fine of $(N'_L - 1 + N_H)F$.

(iii) We assume that the government gets to choose F, and until Section III we do not impose any bound on how large F can be.

This particular formulation is, admittedly, crude. However, note that while we could allow the government to use more sophisticated incentive schemes, this would not expand the set of implementable outcomes or reduce the cost of implementing them.[12] Intuitively, what matters from the point of view of the bureaucrat's incentives is the marginal cost of giving an additional slot to a type L applicant. In this formulation this marginal cost turns out to be just F, which, by assumption, the government can set at any level it wants.

We also assume that the government can always control the number of slots that the bureaucrat allocates in order to avoid the possibility of an additional monopoly inefficiency that arises because the bureaucrat rations the slots to raise their price. This is an additional complication that is unimportant to our basic line of argument and therefore, we feel, best avoided.

To end the description of the model, the sequencing of the actions is as follows. The government first chooses F. Then, given F, the bureaucrat chooses the mechanism for allocating the slots. The applicants make their choices taking the mechanism as given.

I.C. Some Rudimentary Analysis

In order to understand the logic of our model, we start with a special case where the analysis is extremely straightforward. The bureaucrat in this case is only allowed to charge a price to those who receive the slot. We will call such mechanisms *winner-pay* mechanisms and distinguish them from *all-pay* mechanisms, which are mechanisms where all participants have to pay, irrespective of whether or not they get slots.

12. Strictly, this is only true when all bureaucrats are identical in terms of their preferences, which is true in all sections of the paper except Section III.

Within this special model, first consider a situation where both the bureaucrat and the government are welfare-oriented. In this case, so long as y is not too low, the first-best outcome in which all the high types get a slot and nobody suffers any red tape, can be implemented by using a price mechanism. Essentially all we have to do is offer the low type a sufficient discount on what the high type is paying, and then the low type will be willing to accept the lower probability of getting the good. The only problem arises when y is very low; then it is impossible to give the low type a large enough discount (this is obvious when $y = 0$). We state the precise claim in the following.

CLAIM 1. Under the assumption that the government and the bureaucrat are both social welfare maximizers, the first-best allocation can be achieved if $y \geq L - L(1 - N_H)/N_L$, by using the following allocative mechanism.
If $y > L$, those who declare themselves to be a type H pay a price $p_H = \min(y, H - (H - L)(1 - N_H)/N_L)$ and always get the slot. Those who claim to be type L get the slot with probability $(1 - N_H)/N_L$ and pay a price $p_L = L$ when they get a slot.
If $y \leq L$, those who declare themselves to be a type H pay a price $p_H = y$ and always get the slot. Those who claim to be type L get the slot with probability $(1 - N_H)/N_L$ and pay a price $p_L = L - (L - y)N_L/(1 - N_H)$ when they get a slot.

We omit a formal proof of this proposition since it is a simple extension of the verbal argument given in the text.

One practical way to implement the mechanism proposed here is for the government to sell N_H slots at the market price to the type H's (they are after all paying what would be the market price) and to reserve the rest for allocation to the type L's at a lower than market price. In fact, the task of allocating slots to the high types may even be given over to the private sector in order to reduce the bureaucratic burden on the government. One observes many examples of such mechanisms in the real world (for example, certain medicines may be sold both on the market and through the public distribution system).

Consider next the other extreme case—where both the government and the bureaucrat are only interested in making money. In this case it is in the government's interest to allow the bureaucrat to freely maximize profits (i.e., to set $F = 0$) and then collect the revenue from the bureaucrat as a lump sum fee (or equivalently, by charging the bureaucrat a high enough price per

slot). Now as long as $y < L$, the maximum profit the bureaucrat can get is y per slot.[13] This can be achieved by setting a single price equal to y and then offering everybody an equal chance of buying the slot at that price. No red tape will be used. In other words, a purely rapacious government will also avoid red tape (at the cost of generating a poor final allocation).

Finally, let us consider the intermediate case in which there is a conflict of objectives. Given our assumptions, the government can always induce the bureaucrat to give a slot to each high type person—simply by setting F sufficiently high. However, the bureaucrat will not want to use a mechanism of the type described in Claim 1; he makes too little money on the low type. Rather he would want to set the price to both types equal to y (at least as long as $y < L$). However, if both types are paying the same and those who declare themselves to be the high type are getting the slot for sure, everyone will claim to be the high type. To restore incentive compatibility, the bureaucrat will have to threaten anybody who claims to be a high type with enough red tape; i.e., the amount of red tape, T_H will have to satisfy

$$L - y - \delta T_H = (L - y)(1 - N_H)/N_L. \tag{1}$$

This solution will be optimal for the bureaucrat so long as red tape does not cost him too much; i.e., ν is small relative to δ.

This argument assumes that $y < L$. No red tape would arise if $y \geq L$: the bureaucrat could simply charge the type H applicants $p_H > L$ and the type L's L, and incentive compatibility would be automatic (see Section II for a formal statement of this claim).

Finally, observe that, in the case where $\nu = 0$, for *any* positive value of F the bureaucrat will use the mechanism described in the previous paragraph and give a slot to every type H while charging both types a price y. Screening will be achieved entirely by the use of red tape. This follows from the fact that by using this mechanism the bureaucrat is getting as much money as he can ever get; every slot is earning the maximum amount y. Therefore, he loses nothing by using red tape to do all the screening (especially since $\nu = 0$, but a similar result holds when ν is close to 0).

13. Since we do not allow him to charge those who do not get the slot.

I.D. What Do These Results Tell Us?

The results in the previous section offer a number of useful insights. We present them below, numbered, to emphasize the various distinct points.

1. The first implication of these results is that even though red tape is always wasteful, it may be used by the bureaucrat. This is because red tape relaxes the low type's incentive constraint and thereby allows the bureaucrat to charge the low type a higher price.

Red tape in our model is deliberately created by the bureaucrat in order to make money. This contrasts with the view taken by Wilson [1989], among others, who sees red tape as resulting from a set of highly rigid rules set up by the principal in order to limit corruption in the bureaucracy. There is some reason, however, to believe that this cannot be the whole picture. First, in many situations it at least appears that the bureaucrat is going out of his way to generate extra red tape which seems inconsistent with the view that red tape is just a constraint on the bureaucrat. Second, if one takes this view, one still needs to explain why, given that agency problems are ubiquitous, we should not observe the same kind of excessive red tape in private firms as well.[14] By contrast, our view of red tape explains both why bureaucrats favor red tape and why government bureaucracies have more red tape.

While the two views of red tape are very different, it can be argued that they work to reinforce each other. Thus, a rule set up by the principal to limit corruption may be used by a corrupt bureaucrat as an excuse for wasting an applicant's time. To take a concrete and familiar example, most government offices have the rule that anyone who wants anything from the office has to fill out a number of forms. The aim of this rule is to reduce favoritism. Yet the same rule is often invoked by bureaucrats who want to harass certain applicants. They simply ask the applicant to fill out these forms (usually in a large number of copies) and then find small errors in the way the forms were filled out in order to reject the forms so that the applicant has to go through the same procedure again.

14. There is an explanation for this in Wilson [1989], but it relies on the premise that for sociological reasons the government faces certain unique constraints.

2. The second implication of the model is that there would be no red tape if people could pay enough for the slots; i.e., $y \geq L$. In this situation, profit maximization leads to the efficient outcome, and therefore there is no conflict of interest between the bureaucrat and the government. A market failure, then, is necessary for there to be red tape, and of course the same market failure is also the reason why the government is involved in the allocative process.

3. The third implication of the results in the previous section is that, in the world of this model, red tape does not arise because bureaucrats lack incentives. In fact, there is most red tape precisely where the incentives are the strongest; i.e., where F is the largest. This is less paradoxical than it sounds: it is an example of the important observation made in Holmstrom and Milgrom [1991] that increasing the incentives along a dimension of performance that is measurable (here, the share of slots going to the low type) will distort incentives along a nonmeasurable dimension (here, the amount of red tape). In other words, the problem is not that the bureaucrat lacks incentives but that there is a lack of balance between his incentives along different dimensions.

4. A related point is that the most red tape does not arise where the government is the most cynical. If the government were simply interested in making money, it would always set $F = 0$ and allow the bureaucrat to choose the mechanism that maximizes his own income. The government would then recoup the money by charging the bureaucrat a very high price for the slots. We already know that in this scenario there will be no red tape.

This also implies that if the same bureaucracy was a part of a profit-maximizing firm, there would be no red tape.

There will also be no red tape if the bureaucrat shared the government's objective of maximizing social welfare (this is what Claim 1 tells us). It is in the intermediate case, where a welfare-oriented government is trying to control a money-minded bureaucrat, that we would expect to see most red tape. In other words, while a lot of red tape is evidence for some moneymaking by government bureaucrats, it is also evidence that there is some constituency inside the government which is interested in social welfare.[15]

15. A referee has pointed out that this result relies on the assumption that a self-serving government has access to a nondistorting mechanism for extracting revenue from the bureaucrat. Absent such a mechanism, even a self-serving gov-

5. High-powered incentives for bureaucrats (high F) in our model lead to better allocations (more H-types get slots) at the cost of higher levels of red tape. In fact, as we remark at the end of the previous subsection, when the cost of red tape to the bureaucrat is small (which seems plausible), even very weak incentives for the bureaucrats can lead to a lot of red tape. This result illustrates a more general point: when goods are being allocated among people who cannot necessarily afford to pay their full value, people will often get goods that are worth more than they have paid for them. As a result, the bureaucrat who is in charge of allocating those goods may be able to make the people who want the goods do something purely wasteful (like enduring some red tape) without reducing what they are willing to pay him. In other words, the bureaucrat has the option of imposing a substantial social cost on his clients at little or no cost to himself. This makes it substantially harder to design proper incentives for the bureaucrat.

6. A consequence of the previous observation is that if the social cost of red tape is sufficiently large, it may be optimal for the government to opt for very low-powered incentives for the bureaucrat. This observation may shed some light on why we do not usually observe explicit high-powered incentives for bureaucrats,[16] and later in the paper (in subsection II.C) we argue that this may be especially true of government bureaucrats in LDCs.

7. Another result follows from equation (1). It is easily checked that T_H is decreasing in y. In other words, red tape will be high where the average person's ability to pay is low. This is because when the ability to pay is low, type H applicants earn very large rents, and therefore a type L applicant is more likely tempted to claim that he is a type H. Therefore, more red tape is needed to discourage him.

Equation (1) also tells us that an increase in N resulting from equiproportional increases in N_H and N_L leads to a rise in red tape. This tells us that red tape will be higher when the slots

ernment may want to set a high value of F just to extract some extra revenue from the bureaucrat. However, while the assumption of a perfectly nondistorting transfer is an idealization, it seems reasonable to assume that since the government and the bureaucrat typically have a long-term relation the transfers between them should be relatively nondistorting even if the bureaucrat is risk-averse and or cash-constrained. As a result, while our results may not hold exactly in a more realistic model, the results from that model should be more or less similar.

16. For other explanations see, for example, Tirole [1992].

are relatively scarcer. This is intuitive: as the slots get scarcer, it becomes more attractive to claim to be a type H (who, as long as $F > 0$, are guaranteed slots).

Both these results hold for any fixed nonzero value of F (when $F = 0$, there is no red tape). The problem is that the assumption of a fixed F is at odds with the structure of the model, since F is actually chosen by the government and typically it will choose different values of F for different levels of the scarcity of the slots and the ability to pay.

The full analysis of the case where F is endogenous is left until subsection II.B. The results we get there are somewhat weaker but along the same lines: the relation between red tape and the ability to pay is still broadly negative, and the relation between red tape and scarcity of the slot is broadly positive.

How do we interpret these relationships? One interpretation is that we are comparing bureaucracies within the same economy who allocate different kinds of goods. Under this interpretation our result for y says that bureaucracies that deal with a population in which the mismatch between the ability to pay and the willingness to pay is the largest[17] will have the most red tape. In particular, this may argue for a lot of red tape in bureaucracies that deal with very poor people.

An alternative interpretation would be to think of low levels of y as representing poorer countries or communities. However, this is not necessarily correct since what matters is the value of y relative to the values of H and L, and while y tends to be lower in poorer countries, H and L may also be lower.

However, as long as we interpret the slots to be beds in a hospital, H and L are naturally interpreted as the value put on life or good health and this, a priori, may be just as high in a poor country as it is in a rich country. If we think of the slots as opportunities for higher education, once again there may not be a tight connection between y and H and L since the latter two numbers are presumably determined, at least in part, in the world market.

There is another reason why y may be low in poorer countries relative to H and L: capital markets work less well in poor coun-

17. This statement is somewhat loose since we do not say how we measure the mismatch. The natural measure is probably the ratio of the two, but this would be strictly correct only if there were no level effects, i.e., if it were true that if we scale down y, L, and H in the same proportion the amount of red tape will be unchanged. However, this is not true for the obvious reason that if the good is not worth very much, no one will be willing to go through much red tape to get it. The interpretation given in the text is therefore less than completely precise.

tries and as a result the ability to pay will tend to be low relative to the willingness to pay.

If we grant the premise that low values of y go with low levels of development, our results suggest a possible explanation of the high correlation, mentioned above, between low levels of development and poor governmental performance.

The interpretation of the results about the effects of an increase in scarcity is more straightforward: bureaucracies that allocate goods that are particularly scarce will be associated with high levels of red tape. In addition, it seems reasonable to think that at least a certain class of publicly provided private good will be scarcer in poorer countries: richer countries will find it easier to expand the supply if there is a perceived scarcity. Thus, in every OECD country every child has access to schooling of a certain minimum quality, but this is palpably not true in LDCs.

8. Finally, the model allows us to give a partial explanation of why government bureaucracies are associated with corruption. As we say in the Introduction, corruption in the government is not inevitable even with self-serving bureaucrats. What causes corruption is the combination of the fact that the bureaucrats want to make money and the fact that governments make laws to prevent them from doing so. It is therefore natural to ask why governments make such laws. One simple answer to this question comes from the model we develop here: red tape in our model results from the fact that the bureaucrats are trying to make money while satisfying the government's imperative of giving every H-type a slot. Therefore, if the government can discourage the bureaucrats from making money by making it illegal to do so, it would also end up controlling the amount of red tape.

Our model thus provides us with a reason why the government would like to impose controls on the prices that the bureaucrat can charge those who want the slots.[18] The model so far does not permit the government to impose such controls, but in Section III we extend the model to allow for them. However, as is reasonable, we do not permit the controls to be perfect, and we put limits on how severely those who breach the controls can be punished. Consequently, unless the controls are essentially nonbinding, some fraction of bureaucrats will charge prices that are above the permitted prices: this is what we call corruption.

We can now investigate the determinants of corruption. In-

18. Holmstrom and Milgrom [1991] make a related argument about why firms may discourage moneymaking by their agents.

tuitively, it would seem that high levels of red tape reflect extreme divergence between the bureaucrat's objectives and what society wants him to do, and therefore it is precisely where red tape is high that we would expect the most corruption. This intuition turns out to be broadly correct, but because of the endogeneity of the government's choice of what kinds of controls to impose on bureaucrats, it is also sometimes possible for red tape and corruption to move in opposite directions.

To the extent that red tape and corruption do move together, our discussion of the determinants of red tape suggests that corruption is most likely in bureaucracies that deal with poor people, in bureaucracies in poor countries, and in bureaucracies that allocate goods that are scarce.

I.E. Plan of the Paper

The exposition of the workings of the model presented in the preceding subsections is misleading in one important respect. We have assumed that the bureaucrat uses winner-pay mechanisms, but because the winner typically values the slot more than he can afford to pay, all-pay rather than winner-pay mechanisms will maximize the bureaucrat's income.

The next section shows that all the results in this section generalize to the case where we allow the bureaucrat to use this broader class of mechanisms. With that assurance at hand, we then return to the case where the bureaucrat only uses winner-pay mechanisms, but we extend the model in other directions. A reader who is impatient about getting to the results may therefore opt to skip Section II on the first reading.

In Section III we look at the case where the government can (imperfectly) observe the payments made to the bureaucrats. This allows us to analyze the determinants of corruption. In Section IV we look at an extension of the basic model where we allow for inequality in the abilities to pay. We conclude in Section V with some discussion of some deficiencies of our model.

II. Analysis of the General Model

II.A. Solving the Bureaucrat's Problem

In this section we will solve the bureaucrat's problem assuming that he cares only about his own net income and does not care about social welfare. The other extreme case where the bureau-

crat cares only about social welfare is already addressed in Claim 1.

In solving the bureaucrat's problem, we will take as given the value of the punishment for misallocation, F. By doing so, we can accommodate a range of preferences for the government. For example, in the case where the government itself is money-minded and colludes with the bureaucrat to make money, it would set $F = 0$ so as to not place any additional constraints on the ability of the bureaucrat to make money. On the other hand, by setting F to be very large, the government can essentially force the bureaucrat to allocate a slot to every H-type (though it cannot still control red tape).

The mechanism design problem faced by the bureaucrat is potentially quite complex; however, in a previous version of the paper we show that the optimal mechanism always has a specific form.[19] It can be described by six numbers $(p_H, p_L, \pi_H, \pi_L, T_H, T_L)$ of which the first two represent the price charged to everyone who claims to be a high type or a low type, the second two are the probabilities that a person would get the slot conditional on the person's declared type, and the last pair are the amounts of red tape suffered—once again conditional on the person's declared type.

We can use the fact that each and every slot has to be allocated to eliminate π_L, and as result we can replace π_H by π. With this notation the bureaucrat's maximization problem [MB] can be written as

$$\begin{aligned}
&\text{Choose } p_H, p_L, \pi, T_H, T_L \text{ to maximize}\\
&N_H p_H + N_L p_L - N_H \nu T_H - N_L \nu T_L - (1-\pi) N_H F\\
&\text{subject to the constraints}\\
\text{(ICH)}\quad & H\cdot\pi - p_H - \delta T_H \geq H\cdot(1-\pi N_H)/N_L - p_L - \delta T_L,\\
\text{(ICL)}\quad & L\cdot(1-\pi N_H)/N_L - p_L - \delta T_L \geq L\cdot\pi - p_H - \delta T_H,\\
\text{(IRH)}\quad & H\cdot\pi - p_H - \delta T_H \geq 0,\\
\text{(IRL)}\quad & L\cdot(1-\pi N_H)/N_L - p_L - \delta T_L \geq 0,\\
& 0 \leq p_L \leq y,\ 0 \leq p_H \leq y,\ 0 \leq \pi \leq 1,\ T_H, T_L \geq 0.
\end{aligned}$$

It is evident from comparing ICH and ICL that, as is common in such incentive problems, these two constraints cannot bind simultaneously as long as the two types are being offered different options. Further, given the fact that the H-type can adopt any

19. Proof is available from the author.

strategy that the L-type has adopted and do strictly better than the L-type, IRH cannot bind. We state this as

LEMMA 1. In any separating equilibrium, ICH and ICL cannot bind simultaneously. IRH never binds in any equilibrium.

The usual analysis of hidden-information models goes on from here to identify the incentive constraint that binds. In our case, however, depending on the values of F and y, either of the incentive constraints may bind. Consider first the case where y is high (higher than L, say). In this case we are in the standard setting where the optimal mechanism is an auction. It both gives the bureaucrat maximal revenue and allocates the slots to the H-types. Therefore, irrespective of the value of F, the chosen mechanism will be an auction, and as is well-known, in the optimal auction the H-type's incentive constraint binds.

The other extreme case is when y is low and F is high. In this setting the bureaucrat's objective is to maximize revenue conditional on every H-type getting a slot. This means that at the optimum the H-types will have a much higher probability of getting the slot than the L-types. If the L-type is to be reconciled to this lower probability of getting the slot, the price he pays must also be significantly lower than the price the H-type pays. Now if y is sufficiently low, the maximal price the L-type can pay is already low, and his participation constraint will not be binding. If this is the case, the bureaucrat will be tempted to raise the price the L-type pays by as much as possible. But there is an obvious tension between this and the need, argued above, to set the L-type's price significantly lower than the H-type's price. As a result, the L-type's incentive constraint will bind in the mechanism chosen by the bureaucrat.

For intermediate values of y and F, either incentive constraint might bind, although from the intuitive discussion in the last paragraph it seems plausible that ICL is more likely to bind when F is high and y is low. Lemma A3 in the Appendix confirms this intuition.

The main analytical goal of this section is to characterize the values of F and y for which there is a high level of red tape. This is complicated by the fact that there are two types of red tape: there is red tape faced by H-types (T_H), and there is red tape faced by L-types (T_L). In principle, depending on which incentive constraint binds, the bureaucrat may want to use either of these

types of red tape (raising T_H relaxes ICL, while raising T_L relaxes ICH). What the next result shows is that the bureaucrat would never want to use red tape against the L-type (the proof is in the Appendix).

CLAIM 2. Self-declared L-types are never subject to any red tape, i.e., there is always an optimum at which $T_L = 0$; and as long as $\nu > 0$, this is the only optimum.

What drives this result is the fact that while more red tape on the L-type does relax ICH, the same effect can be achieved at a lower cost by raising p_L or π. The proof of this result makes use of the fact that the cost of red tape is the same for the two types. Instead, if red tape was much more costly to H-types than it is to L-types, there could be a reason to subject L-types to a little bit of red tape in order to discourage H-types from claiming that they were L-types, and this result would no longer hold.

An obvious consequence of this result is that if red tape is ever used it is used against the H-type. It then follows that if red tape is used at all, it is only used when ICL binds (otherwise there is no reason to use red tape) which happens when F is high and y is low.

To complete the argument, we need to show that when ICL binds the bureaucrat *will* sometimes choose to subject H-types to red tape. This contrasts with the fact that L-types never suffer red tape. The difference between the two cases stems from differences in alternatives to using red tape. In the case of the L-type, the alternative to more red tape was a higher value of π which suits the bureaucrat, since he gets penalized for low values of π. By contrast, in the case of the H-type, the alternative to more red tape was a lower value of π, which hurts the bureaucrat as long as F is positive. As a result, the bureaucrat will be more willing to use red tape.

The final step in the argument is to describe the solution to the bureaucrat's problem. Unfortunately, describing the full solution involves saying what happens in a very large number of different cases. We therefore take the route of describing the full solution in the special case where $\nu = 0$ in the text, while representing the solution to the more general case diagrammatically. The more onerous task of describing the full analytic solution in the more general case is relegated to the Appendix.

CLAIM 3. The solution to the bureaucrat's problem [MB] for the case $v = 0$ is as follows:[20]

(i) If $y \geq H - (H - L)\cdot(1 - N_H)/N_L$: $\pi = 1$, $p_H = H - (H - L)\cdot(1 - N_H)/N_L\}$, $p_L = L(1 - N_H)/N_L$, and $T_H = T_L = 0$.

(ii) If $H - (H - L)\cdot(1 - N_H)/N_L > y \geq L$ and $F \geq L$: $\pi = 1$, $p_H = y$, $p_L = L(1 - N_H)/N_L$, and $T_H = T_L = 0$.

(iii) If $H - (H - L)\cdot(1 - N_H)/N_L > y \geq L/(N_H + N_L)$ and $0 \leq F < L$: $\pi = [N_L y + (H - L)]/[HN_L + (H - L)N_H]$, $p_H = y$, $p_L = L(H - N_H y)/[HN_l + (H - L)N_H]$, and $T_H = T_L = 0$.

(iv) If $L > y \geq L\cdot(1 - N_H)/N_L$ and $L \leq F$: $\pi = 1$, $p_H = p_L = y$, and T_H set to solve the equation $L - y - \delta T_H = 0$.

(v) If $L\cdot(1 - N_H)/N_L \leq y < L\cdot[N_H + N_L]^{-1}$, $L > F \geq 0$: π and T_H set to solve $\pi L - y - \delta T_H = 0$, and $L(1 - N_H\pi)/N_L = y$ and $p_H = p_L = y$.

(vi) If $L\cdot(1 - N_H)/N_L > y$, for any value of F: the outcome is $\pi = 1$, $p_H = p_L = y$, and T_H satisfying $L - y - \delta T_H = L(1 - N_H)/N_L - y$.

Proof. All the statements except the last one follow from Claims A3 and A4 in the Appendix. The last one requires us to extend the argument slightly, but the extension is sufficiently obvious that we feel that it can be excluded.

The essential features of this solution are as follows: (a) Higher values of F are associated with higher values of π and higher levels of T_H. (b) Higher values of y are associated with lower values of T_H for a fixed F. (c) Higher values of y are not necessarily associated with lower values of π—the highest values of π may obtain at very high and very low values of y. (d) An increase in the scarcity of slots represented by an increase in N_H and N_L in the same proportion, while keeping the number of slots fixed, increases the ratio $N_L/(1 - N_H)$ and thereby increases red tape.

The association between high levels of F and high levels of π is hardly surprising since the point of raising F is to force the bureaucrat to raise π. Higher values of π, ceteris paribus, cause ICL to bind more tightly which then gives the bureaucrat a reason to raise T_H as well. An increase in y allows the bureaucrat to charge higher prices. As a result, he does not need to use as much

20. In writing down this solution, we have implicitly assumed that whenever he is indifferent, the bureaucrat always chooses the socially best outcome.

red tape to induce self-selection by the L-type which is why T_H and y will be negatively associated.

A standard intuition from price theory explains one reason why high values of y result in high values of π; the high types value the good more, and therefore it pays more to give it to them as long as they can register their preferences as higher prices. When y is low, the reason why the final allocation is very efficient is that it is essentially costless for the bureaucrat to sort the applicants by using red tape.

Scarcity increases red tape because if the slots are scarce, type-L applicants will be more desperate to get the slots. This makes screening harder.

These broad features of the solution to the bureaucrat's maximization problem all turn out to also hold in the more general case, where ν is positive but small relative to δ (this seems to be the natural case to look at). This solution is depicted in Figures I and II, which are based on Claims A3 and A4 in the Appendix.

What changes when ν is large relative to δ? We show in previous versions of the paper that in this case the outcome is always first-best. This should be intuitive; we have therefore chosen to omit the analysis of this case.

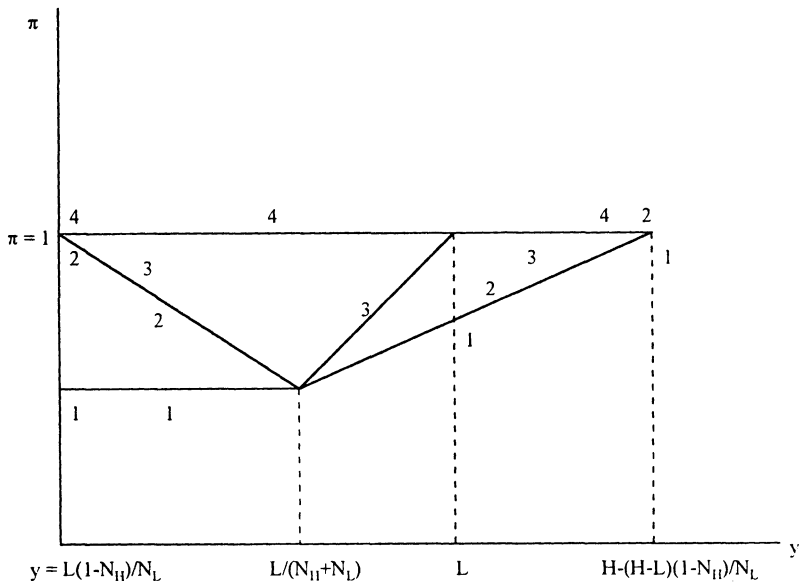
II.B. The Government's Problem

If the government in our model is interested in making money, it will set $F = 0$ and collect the revenue from the bureaucrat as a lump sum fee. When the bureaucrat is welfare-oriented, the choice of F does not matter. The interesting case, therefore, is when the government is welfare-oriented but the bureaucrat is not. The government's maximand in this case will be

$$L + (H - L)N_H\pi(F) - (\delta + \nu)N_H T_H(F),$$

where $\pi(F)$ and $T_H(F)$ are the values of π and T_H that result from the bureaucrat's maximization problem for that particular value of F. In principle, since we have solved the bureaucrat's problem, we can solve the government's problem by comparing the government's maximand for different values of F. In practice, this will involve considering a very large number of cases. We therefore only look at the government's problem in the special case where $\nu = 0$, which makes the problem much more tractable.

It is evident from Claim 3 that in this case the government need only choose between $F = 0$ and $F = L$. Furthermore, for



Curve 1: $F < L(\nu/\delta + \nu N_H/\delta N_L)$

Curve 2: $L(\nu/\delta + \nu N_H/\delta N_L) \leq F < L$

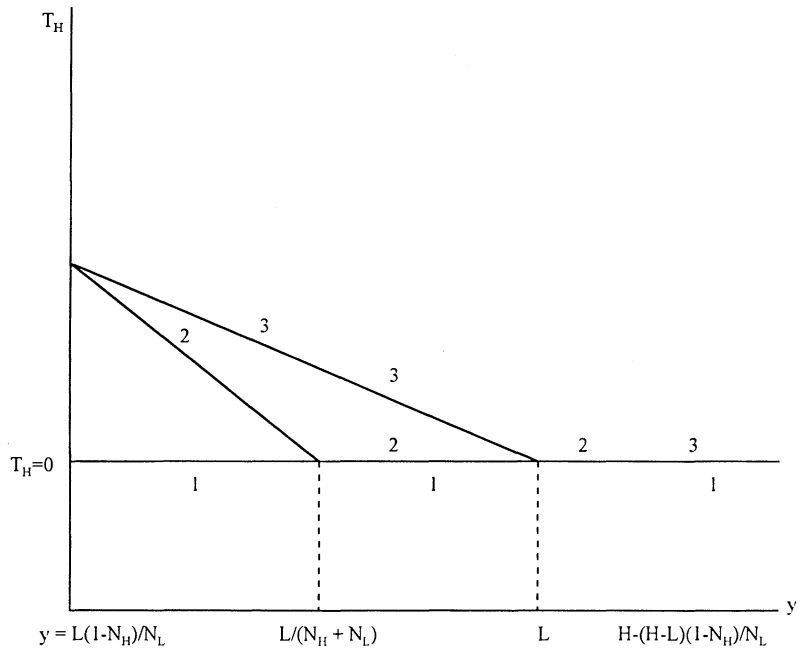
Curve 3: $L \leq F < L\,(1 + \nu/\delta)$

Curve 4: $L\,(1 + \nu/\delta) \leq F$

FIGURE I
π as a Function of y

extreme values of y, i.e., $y \geq H - (H - L)\cdot(1 - N_H)/N_L$ and $y < L(1 - N_H)/N_L$, the value of F does not matter—all values of F result in the same outcome. In both these cases the government will presumably choose $F = 0$; i.e., let the bureaucrat do whatever he wants.

For values of y between $H - (H - L)\cdot(1 - N_H)/N_L$ and L, the solution is also straightforward. It is evident from the comparison of cases (ii) and (iii) in Claim 3 that in this case a higher value of



Curve 1: $F < L(\nu/\delta + \nu N_H/\delta N_L)$

Curve 2: $L(\nu/\delta + \nu N_H/\delta N_L) \leq F < L(1 + \nu/\delta)$

Curve 3: $L(1 + \nu/\delta) \leq F$

FIGURE II
T_H as a Function of y

F is always preferable since it generates a higher value of π without generating any red tape.

The less obvious case is when y is between L and $L(1 - N_H)/N_L$. In this case a simple calculation based on a direct computation of the government's maximand for the two values of F establishes that $H > 2L$ is a sufficient condition for always using $F = L$. However, if $H < 2L$, $F = 0$ will be used as long as y is between $L(1 - N_H)/N_L$ and $L \cdot [N_H + N_L]^{-1}$, but for higher values of y, $F = L$ is still optimal. The chosen value of F is always weakly increas-

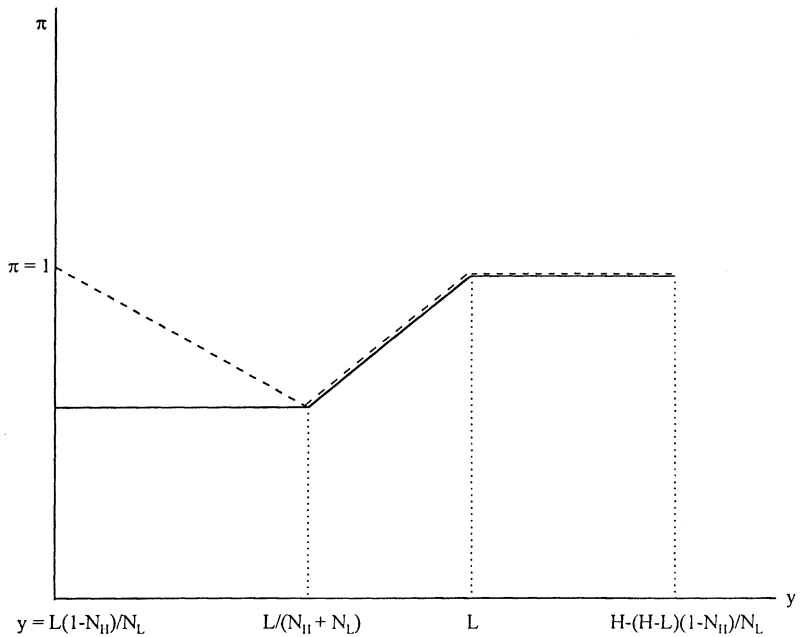


FIGURE III
π as a Function of y When F Is Endogenous

ing in H keeping L fixed: this is because a high H makes it more important for each person of type H to get a slot.

How does the relation between π, T_H, and y look now that F is endogenous and depends on y? These are given in Figures III and IV for two cases: $H > 2L$ and $H = 2L$ (with the interpretation that $H = 2L$ is the limit of the case where $H < 2L$ and represents all such cases). It should be evident from the discussion above that these are essentially the two canonical cases. In the case where $H > 2L$, the pictures are exactly the same as they were when F was exogenously set to be greater than or equal to L. However, in the case where $H = 2L$, endogenizing F does change the picture since, at low levels of y, $F = 0$ is chosen but at higher values the chosen value of F goes up to L. As a result, an increase in y over a certain range causes T_H to go up.

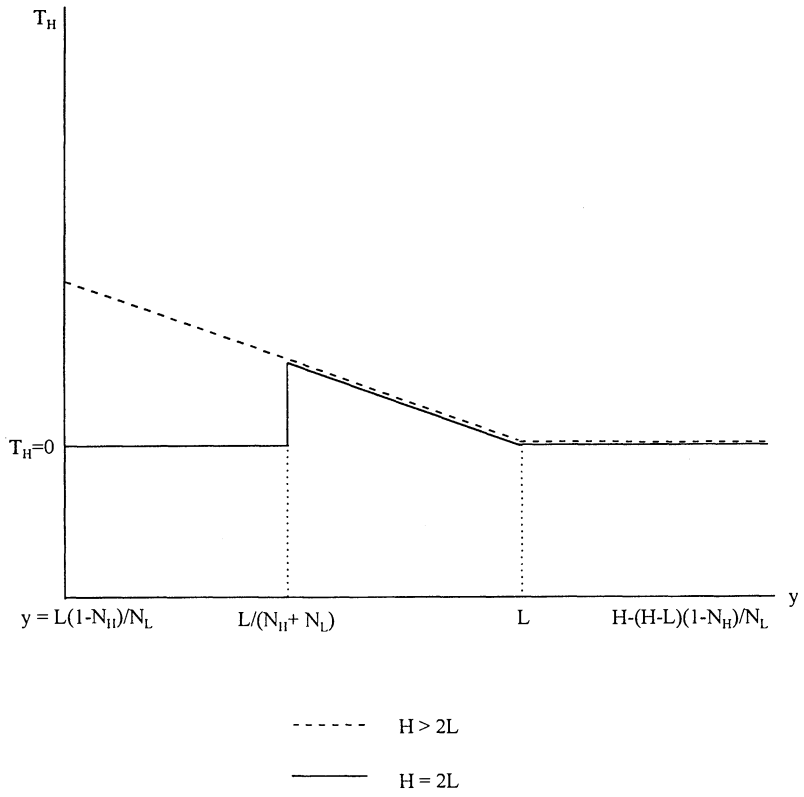


FIGURE IV
T_H as a Function of y When F Is Endogenous

We have not explicitly considered the effect of changes in the scarcity of the slots, but it can be shown that the effect of an increase in the scarcity of the slots is similar to that of a fall in y. It typically leads to a rise in the level of red tape, but it may also cause F to fall, and as a result, for a specific range of parameter values, red tape may be lower even though the slots are scarce.

II.C. What Do We Learn from the Results of the More General Model?

The results here largely confirm what we found in the simpler version of the model analyzed in subsection I.C. As before, for a fixed value of y, an increase in F leads to a higher level of red tape. Combined with Claim 1, this confirms our earlier claim that red tape is maximized when there is a conflict of objectives

between the government and the bureaucrat (with the government being welfare-oriented and the bureaucrat self-serving). It also confirms that there would be no red tape if, instead of the government, a private firm were carrying out the allocation (a private firm would set $F = 0$). Of course, the overall outcome would be worse.

However, note that the effect of an increase in F on the level of red tape in this model is much less dramatic than it was in the model in subsection I.C. There, for $v = 0$, any positive value of F leads the bureaucrat to go immediately to the maximum level of red tape that he would ever use for that level of y. Here, as is evident from Figure II and Claim 3, the response is more gradual. This is because the use of all-pay rather than winner-pay mechanisms allows the bureaucrat to extract more of the surplus from the applicants, which then makes the bureaucrat internalize more of the cost of the red tape he imposes on them. This suggests that a movement toward creating an environment where bureaucrats can use all-pay mechanisms to allocate scarce publicly provided private goods may actually help improve bureaucratic performance.

As in subsection I.C for a fixed value of F, there is a negative relation between y and red tape. The analysis in this section goes beyond the previous analysis in endogenizing F. Endogenizing F does not change the relation between y and red tape as long as H is sufficiently greater than L. However, when H is close to L, the relation between red tape and y may be nonmonotonic, although it will still continue to be true that very low values of y will be associated with very high levels of red tape and red tape will be absent at high levels of y. The relation between red tape and the scarcity of slots is similar to that between red tape and y, with low levels of y corresponding to high levels of scarcity.

The behavior of π as a function of y can be read from Figure I and turns out to be subtler than one would have predicted from the preliminary analysis: except when F is very high (when $\pi = 1$ at all values of y in our range) or very low (when π is constant at low levels of y), π is always U-shaped as a function of y; it is high at high values of y as well as at low values of y and is lower in between. The relationships are more or less the same with F endogenized (see Figure III).

Since we allow the government to choose F, we also find conditions under which a welfare-oriented government will deliberately choose low-powered incentives for the bureaucrat (i.e., set

$F = 0$) in order to avoid generating too much red tape. This will happen when the difference between H and L is not too large (i.e., the misallocation is not too costly), the slots are scarce, and y is relatively small (which imply that the bureaucrat, if pushed, will use high levels of red tape).

Since we have taken the view that lower values of y and greater scarcity go with lower levels of development, this result suggests that at least in situations where the cost of misallocation is small, bureaucrats in less developed countries will tend to have weaker incentives than their counterparts in the developed world. We can also read the model as saying that within the same country, those bureaucracies that deal most with people with low abilities to pay (relative to their willingness to pay), will have the weakest incentives.

III. Toward a Theory of Corruption

To restore efficiency in the economy modeled here, the government will need to be able to control the prices charged by the bureaucrats. We will now modify our model to allow the government some possibility of observing the payments that are made to the bureaucrats.

We introduce the possibility of monitoring payments to bureaucrats by assuming that with some probability $\phi < 1$, the government finds out about *the mechanism being used by the bureaucrat to allocate the slots* (here we are using the word mechanism in its broader sense so that if the bureaucrat uses several different rules to allocate to different people, we will consider them together to be a part of a single mechanism). Recall that we have already assumed and continue to assume that the government knows the fraction of type L applicants who got a slot. What knowing the mechanism tells the government is whether the bureaucrat is charging the recommended prices or whether he is asking for additional bribes.[21]

In this setting, if the government could also inflict arbitrarily large punishments on the bureaucrats, it is easy to see that it could always implement the optimal outcome. All it would have to do is to recommend that the bureaucrat uses the optimal mech-

21. It also tells the government how much red tape is being used, but this is not extra information, since once it knows the prices and the allocation, it can always infer the amount of red tape.

anism and to punish any detected deviation from this mechanism with such severity that no bureaucrat would ever contemplate deviating.

The more interesting case is the one where there is a bound on how much a bureaucrat can be punished. We model this by assuming that there is an institutionally given worst punishment that the government can inflict on any bureaucrat (this may be the loss of his job and a prison stay of several years). Denote the utility level of a bureaucrat who is undergoing this punishment by B, and assume that there is a distribution function $G(B)$, which gives the fraction of the population of bureaucrats whose lower bound is no higher than B.[22] We will assume that B is private information.

There are a number of alternative patterns that can emerge in this setting and investigating all of them is beyond the scope of this paper. Here we confine ourselves to the situation where the government wants to allocate the slots efficiently even at the cost of some red tape (this is the case where H is large relative to L).

To simplify the analysis further, let us revert to the assumption made in subsection I.C limiting the bureaucrat to winner-pay mechanisms. Also, to limit the number of cases, assume that $L \geq y \geq L - L(1 - N_H)/N_L$ and $v = 0$.

Under these assumptions, all mechanisms that achieve the efficient allocation of slots take the form $\{p_H, p_L, T_H\}$, where P_H and T_H are the price and red tape assigned to a type H applicant (who always gets a slot) and p_L, the price paid by a type L, satisfies the incentive compatibility constraint;

$$L - p_H - \delta T_H = (L - p_L)(1 - N_H)/N_L.$$

Of these mechanisms, the one that is least likely to lead to corruption is the one that sets the highest prices for both types (the higher the official price, the less people will want to pay in excess of that price to increase their chances of getting the slot). Therefore, p_H should be set equal to y.

Now suppose that the government announces a mechanism $\{y, p_L^*, T_H^*\}$. In other words, it sets both the prices the bureaucrat is allowed to charge and the maximum amount of red tape that the bureaucrat is permitted to use (an example of a government rule about how much red tape is permitted is the rule recently

22. Those with low levels of B may be thought of as those who especially value their reputation for being honest.

introduced in India requiring all passport applications to be processed within a certain number of days). We assume that the mechanism recommended by the government is incentive compatible from the point of view of the applicants. Once such a mechanism is announced, bureaucrats are required by the government to implement that mechanism, and it is also announced that any bureaucrat who is caught deviating from this mechanism will receive the maximal punishment.[23]

The government also needs to choose F. In deciding on F, the government can take advantage of the fact that if $\nu = 0$ and the bureaucrat only uses winner-pay mechanisms, the bureaucrat will always give every type H applicant a slot for any strictly positive value of F. This was shown in subsection I.C and continues to hold in our current model. Moreover, it holds *irrespective of whether the bureaucrat follows the mechanism the government wants him to follow:* the only difference is that if he chooses to deviate, he will use red tape to screen out L-type applicants instead of relying on prices.

Given the assumption, made above, that H is large relative to L, the government will always set a nonzero level of F. Given that it is indifferent between all nonzero levels of F, assume now that it sets the value of F to be so close to 0 that the expected value of the fines can be ignored while calculating the bureaucrat's utility level (consequently, we do not need to worry how the bureaucrat can be fined in the state of the world where he is already being punished for taking bribes).

Given all these assumptions, the bureaucrat who will be on the margin of deviating and asking for a bribe, will have a B which satisfies

$$N_H y + (1 - N_H)p_L^* = (1 - \phi)y + \phi B.$$

Clearly, the left-hand side of this equation represents the utility of a bureaucrat who follows the rules while the right-hand side represents the utility of a bureaucrat who, instead, charges y for

23. The mechanism used here is an efficiency wage-type mechanism first used in the context of corruption by Becker and Stigler [1974]. It is in principle possible to allow the government to use more sophisticated mechanisms (such as a linear or nonlinear tax on the bureaucrat's income from selling the slots) which may actually work better. We justify not using such mechanisms on the grounds that we do not observe such mechanisms (we also believe that so long as the government has limited ability to observe the bureaucrat's income, the results will not change very much even if we change the model in this direction). For a more detailed discussion of the kinds of incentive schemes used by governments vis-à-vis their bureaucrats, see Banerjee [1995], Kofman and Lawarree [1990], and Tirole [1992].

every slot and gets caught with probability ϕ. Solving for the value of p_L^* using the incentive-compatibility constraint, gives us

$$p_L^* = L - N_L[L - y - \delta T_H^*]/(1 - N_H).$$

Substituting this expression into the above equation gives us

$$N_H y + (1 - N_H)L - N_L[L - y - \delta T_H^*] = (1 - \phi)y + \phi B,$$

which can be written in the form

$$(N - 1)(L - y) - \delta N_L T_H^* = \phi(y - B). \tag{2}$$

Denote the value of B that solves this equation by B^*. Clearly, those and only those with values of B greater than this critical value will choose to break the rules and ask for bribes. In other words, $1 - G(B^*)$ measures the extent of corruption in this economy.

Note that the corruption that arises here is in a very direct sense created by the government. The government creates corruption by imposing a rule on the bureaucrats that some bureaucrats will follow and others disregard: if there were no such rule, there would be no bribes and no corruption. Nevertheless, the reason why the government chooses to impose this rule is that it helps it fight wasteful red tape in the bureaucracy.

This contrasts with the quite common view that corruption arises, at least in part, out of a need to get around the red tape that is endemic in government bureaucracies.[24] In this view, what causes red tape is something that is usually exogenous and explained, if at all, by reference to the sociology of the government. There is therefore little one can do about red tape itself, and anything that helps get around it is probably a good thing. Fighting corruption, in this view, may therefore be a bad thing.

By contrast, our view is that a lot of red tape is deliberately created by the bureaucrats in order to make more money. Fighting corruption, by limiting the amount of money the bureaucrat can make, may therefore also reduce red tape.

A second implication of this analysis of corruption is that corruption only arises when the government has a reason to try to limit moneymaking by bureaucrats. In our model, if the government was indifferent to social welfare and only interested in

24. For a forthright if somewhat dated statement of this view, see Nye [1979] or Leff [1979]. See Waterbury [1979] for a critique of this view on empirical grounds.

making money, there would be no corruption. Like red tape, corruption arises from a conflict of interest.

A number of other conclusions follow from equation (2). First, B^* is increasing in y for any fixed value of T_H^* and the other parameters. In other words, everything else remaining the same, a fall in y *increases* corruption. In other words, somewhat paradoxically, there is more illegal moneymaking precisely when there is *less* money around. This is because an increase in y enables the government to raise the legal price paid by a type L applicant by more than the original increase in y (see the expression for p_L^* given above).

Second, a simple calculation establishes that an equiproportional increase in N_H and N_L, for any fixed value of T_H^* and the other parameters, reduces B^* and therefore increases corruption. In other words, there is more bribery as the good being allocated becomes scarcer.

Third, once again keeping T_H^* fixed, an increase in y or an equiproportional fall in N_H and N_L will lead to a fall in the total amount of red tape. To see this, observe that the average amount of red tape suffered by an H-type applicant is given by

$$G(B^*)T_H^* + (1 - G(B^*))T_H. \tag{3}$$

The first term in this expression is the amount of red tape that is associated with bureaucrats who do not deviate from the recommended mechanism, and the second term comes from those who do deviate. Now, both the increase in y and the fall in N_H and N_L have the effect of reducing the fraction of those who take bribes and therefore lead to a fall in red tape (because $T_H \geq T_H^*$). Also, as shown in subsection I.C, both these changes have the effect of reducing T_H, which goes in the same direction.

However, the above results about the effect of a fall in y, or an increase in N_H and N_L assume that the permitted amount of red tape, T_H^*, is exogenously fixed. This is misleading since in our model the government chooses T_H^* and an increase in T_H^* by itself, increases B^* and therefore reduces bribery.[25] We therefore need to treat T_H^* as an endogenous variable when we do the comparative statics. Since, in the situation considered in this section, all bureaucrats (whether or not they take bribes) allocate the slots in the same way, the government, in choosing T_H^*, needs only to look at the effect on the average amount of red tape. Differentiat-

25. This is because an increase in T_H^* allows p_L^* to be increased.

ing the expression given in equation (3) for the average amount of red tape, with respect to T_H^*, yields the first-order condition:

$$G(B^*)/G'(B^*) = (T_H - T_H^*)\delta N_L/\phi. \tag{4}$$

Equation (2) embodies a very simple trade-off: an increase in T_H^* hurts those already dealing with uncorrupt bureaucrats, but it also increases the fraction of bureaucrats who are not corrupt. Therefore, as the equation makes evident, what matters for the choice of T_H^* is the population of inframarginal (uncorrupt) bureaucrats relative to the population of those who are at the margin of becoming uncorrupt. T_H^* will tend to be high when there are lots of marginal bureaucrats relative to the number of those who are inframarginal.

The effect of an increase in y on T_H^* turns out to be impossible to sign on purely a priori grounds because, while an increase in y increases B^* and therefore increases the number of inframarginal bureaucrats, it also affects the number who are at the margin and the net effect on $G(B^*)/G'(B^*)$ is ambiguous. However, for a large range of distribution functions, $G(\cdot)$ (including, for example, the case where the underlying density is uniform), it can be shown that T_H^* falls when y goes up. Furthermore, it is possible to construct examples where the fall in T_H^* resulting from the increase in y is so large that it swamps the direct effect of the increase y on B^* and the net effect on B^* is negative. In other words, an increase in y *can* lead to an *increase* in corruption because of the endogeneity of T_H^*. For exactly the same reasons, a fall in the scarcity of the good can actually lead to an increase in corruption.

These kinds of "perverse" comparative statics results are less likely to arise if the density function corresponding to the $G(\cdot)$ function has a mass point (or a highly concentrated density) at the lowest point in its support but nowhere else. This kind of density captures the plausible idea that the population of bureaucrats contains a hard core of incorruptible people, but otherwise there is a lot of diversity in how people feel about getting caught taking a bribe. In this case there will always be a large number of inframarginal bureaucrats, and therefore it is costly to raise T_H^* in order to combat corruption. As a result, it is unlikely that when y falls, T_H^* will be raised by so much that there will actually be a fall in corruption.

It is also worth remarking that even with F endogenous,

there will be no corruption in the case where y is higher than L since in this case there is no conflict between making money and furthering social welfare. Thus, the negative relation between y and corruption holds at least when we compare very high and very low levels of y.

As we noted above, the direct effect of an increase in y on red tape is always negative. In addition, we just argued that an increase in y typically leads to a fall in T_H^* which reinforces this effect. However, in the scenario where an increase in y increases corruption, this increase in corruption can increase red tape. However, note that this effect needs to be strong enough to swamp the other two effects if the overall effect of an increase in y is to increase red tape. This seems somewhat implausible.

To summarize, once we endogenize the permitted amount of red tape, we no longer get the simple unambiguous comparative statics results that we got when the permitted amount of red tape was taken as exogenously given. The amount of corruption and somewhat less plausibly, the amount of red tape, may actually go up when the applicants have a higher ability to pay or the slots are less scarce. This is because the government responds to the increase in the ability to pay or the fall in scarcity by severely limiting the amount of red tape the bureaucrat is allowed to use. In a sense, what is going on is that the bureaucrats' effective incentive scheme is becoming much more demanding, and this leads to an outcome where more bureaucrats fail to meet the standard.

We also identify one quite reasonable setting where an increase in the ability to pay or fall in the scarcity of the slots always reduces both corruption and red tape. This is the situation where the population of bureaucrats contains a core of people who are completely incorruptible.

IV. Implications of Inequality in the Ability to Pay

We have so far ignored the possibility that different people may have different abilities to pay. This is an important deficiency since a standard justification of red tape-like procedures is that they protect the poor.[26] The conclusions of this section are as follows. (i) The presence of inequality increases the amount of red

26. See, for example, Weitzman [1977].

tape used by both a profit-minded government and the government in our model, and (ii) it remains true that more red tape is used when the government is welfare-oriented.

There are at least two ways to introduce inequality into this model. The simpler case is where both the bureaucrat and the government can observe each applicant's ability to pay. In this case the government sets an F that depends on the applicant's ability to pay, and the bureaucrat chooses a different mechanism depending on the applicant's ability to pay. The bureaucrat's problem then consists of a number of parallel problems of the type we solve in the previous section. It is easy to see that the outcome of the bureaucrat's maximization problem will be such that those who have less money (smaller y) will face more red tape.

This conclusion gets reinforced if we assume that neither the government nor the bureaucrat can observe the applicant's ability to pay. Assume that the ability to pay takes two values, y_1 and y_2 ($y_1 > y_2$) with probabilities μ and $1 - \mu$, and that a person's valuation of the slot is statistically independent of his ability to pay. Also to make the problem interesting, assume that $1 > \mu (N_H + N_L)$; i.e., there are not enough rich people to fill up the slots (if we do not make this assumption, the poorer people may be irrelevant). In all other respects let the model be exactly the same as the model we introduce in Section I (in other words, we do not allow the government to observe payments to bureaucrats so that the question of corruption does not arise).

This is a two-dimensional screening problem, and these are notoriously difficult to solve. To make it tractable, we make the simplifying assumption we made in the introduction, namely, that the bureaucrat is limited to winner-pay mechanisms. We also assume that $v = 0$ and that $y_1 < L$.

With these simplifying assumptions the problem turns out to be quite simple to solve. Given that we assume that $y_1 < L$ and that only those who get the slot pay for it, the individual rationality constraints will not bind for any of the agents. Therefore, the bureaucrat can impose some extra red tape on the agents without having to cut the price he charges them. Since in addition we have assumed that $v = 0$, extra red tape also costs the bureaucrat nothing. Therefore, a self-serving bureaucrat will always charge the applicants the highest price they can pay and then use red tape to ensure that the mechanism he sets up is incentive compatible.

The problem faced by a profit-minded government with a profit-minded bureaucrat therefore has a simple solution; the bureaucrat will set two prices, y_1 and y_2 (i.e., the maximum possible prices), and offer a slot to each person who pays the higher price and randomly select $1 - \mu(N_H + N_L)$ persons among those who offer to pay the lower price. This will be incentive compatible if[27]

$$(2) \qquad L - y_1 \geq L[(1 - \mu(N_H + N_L))/(N_H + N_L)] - y_2.$$

If not, the bureaucrat will have to threaten those who pay less with some red tape; the exact amount of red tape, T, will be given by

$$(3) \qquad L - y_1 = L[(1 - \mu(N_H + N_L))/(N_H + N_L)] - y_2 - \delta T.$$

In the conflicting objectives model, if the government sets a high enough F, the bureaucrat will want to give a slot to every high type. The mechanism that maximizes the bureaucrat's profits conditional on giving a slot to every high type, will be described by four triplets $(y_1,T_1,1)$, $(y_1,0,\min\{0,(1 - N_H)/\mu N_L)\}$, $(y_2,T_2,1)$ and $(y_2,0,\min\{(1 - N_H - \mu N_L)/(1 - \mu)N_L,0\})$ with T_1 and T_2 satisfying

$$(4) \quad L - y_1 - \delta T_1 = (L - y_1)[\min\{(1 - N_H)/\mu N_L, 1\}]$$

$$(5) \quad L - y_2 - \delta T_1 = (L - y_2)[\min\{(1 - N_H - \mu N_L)/(1 - \mu)N_L, 0\}].$$

The first number of each of these triplets is the price that a person who chooses that option pays. The second number is the amount of red tape he has to go through. The last number is the probability he gets a slot. The first triplet is what a rich high type chooses, the second what a rich low type chooses, the third is what a poor high type chooses, etc. Note that each type is paying the maximum amount he can pay.

The outcome generated by this mechanism is that the rich high types and the poor high types all get slots. If the number of remaining slots is less than the number of rich low types, we assume that the rich low types get all of these slots. If there are slots left over after all the rich low types have chosen, then they will be given to some of the poor low types.

The outcome in the case where both the government and the bureaucrat are welfare-oriented is still going to be socially efficient as long as y_2 satisfies $y_2 \geq L - L \cdot (1 - N_H)/N_L$ since in this

27. It is easily checked that this is the incentive constraint that may bind.

case we can use the mechanism used in the argument for Claim 1, with y_2 substituted for y.

This quite rudimentary analysis yields a number of useful insights.

1. A comparison of equations (4) and (5) with equation (3) establishes that while in the presence of inequality red tape will arise in both the self-serving government model and the conflicting objectives model, there will always be more red tape generated under the latter model. This confirms the results in the previous sections.

2. It is evident from equations (4) and (5) that an increase in inequality in the distribution of y, keeping the mean unchanged, reduces T_1 and increases T_2, but on balance, the social waste due to red tape always goes up. This is shown in the Appendix (See Claim A5). The reason is that the probability that a poor low type gets a slot is lower than the probability that a rich low type gets the slot. As a result, a poor low type has more of an incentive to claim that he is a high type than the rich low type. Moreover, and for the same reason, a change in y has a bigger impact on the poor low type's incentive to misrepresent his type than it has on the corresponding incentive for the rich low type. As a result, the change in the red tape for the poor low type will also have to be larger than the corresponding change for the rich low type. Consequently, the rise in red tape caused by the fall in the poor low types' ability to pay will dominate the fall in red tape resulting from the rise in the rich low types' ability to pay.

3. The poor face more red tape than the rich in the conflicting objectives model. The same result may also be true in the pure self-serving government model, but only if y_2 is sufficiently low. In both cases the bureaucrat uses this extra red tape to threaten the rich, so that the rich are forced to buy their way out of it.

4. The poor of the low type get less access to the slots than the rich of the low type both under the conflicting objectives model and the pure self-serving government model, although the difference in access is greater in the latter case.

V. Conclusions

The model proposed in this paper, while both simple and stylized, makes a number of predictions that broadly fit the pattern of what we know about misgovernance. However, it also has a number of important and obvious limitations.

An implication of this model is that governments in developed countries should use the market more than in LDCs. While this is true in some cases,[28] there are others like health-care where the market is not used. Of course, our model only tells us the efficient outcome and ignores distributional considerations that may explain why the market is not used. It is still a puzzle why, given that the market is not used, there is so little corruption in the health-care bureaucracy in most OECD countries. The explanation suggested by our model is that there is an adequate supply of health-care, i.e., the good is not scarce enough to make corruption worthwhile. Whether this is the right story is an open empirical question.

Our model also does not deal with the issue of whether there are cultural or institutional determinants of government performance. One stereotype we did not take up (because it concerns preferences rather than outcomes) is the characterization of third world societies as being much more casual about corruption in government than first-world governments. It has been pointed out that in this instance what appears to be cultural and exogenous may be endogenous and rational in the sense that there may be multiple equilibria in some of which corruption may be rare and heavily punished and others in which corruption is common and tolerated.[29]

Of course, even if we accept the multiple equilibrium view, it remains to explain why the culture of corruption should emerge principally in LDCs.[30] Two explanations come to mind: one could argue that the culture of corruption is what causes LDCs to be less developed. This we find somewhat implausible given that these LDCs also tended to be poor countries before the recent era of large-scale government interventions in the economy. The other, more convincing (to us) theory holds that development is a process of transforming a large complex of institutions along with increasing the GNP. The culture of corruption in poor countries is at least partly a result of underdeveloped institutions (like a lack of democracy).

28. Few rich countries have licenses for production and imports; and in the United States, for example, oil drilling rights are auctioned off too.

29. See Cadot [1987], Clague [1993], and Sah [1991] for different arguments within this broad category. Tirole [1996] provides a model within which a temporary increase in corruption may become irreversible. Also see Acemoglu [1992] and Murphy, Shleifer, and Vishny [1993] for the related argument that the presence of corruption may actually induce others to become corrupt by reducing the return to the honest activity.

30. Japan and Italy being well-known exceptions.

APPENDIX

Proof of Claim 2

The only interesting case is the one where there is a separating equilibrium. There is no reason to use red tape in a pooling equilibrium.

Now note that if ICH does not bind, then the bureaucrat will always want the value of T_L to be lower. Therefore, $T_L > 0$ implies that ICH binds which in turn implies that ICL does not bind so that $T_H = 0$.

Next observe that if IRL does not bind, we must have $\pi = 1$ because, if not, it is always possible to raise π and relax all the binding constraints. It is also easy to see that if IRL does not bind, we must have $p_L = y$ since otherwise it would be possible to raise p_L and relax all the binding constraints while making the bureaucrat better off.

Consider first the case where IRL does not bind so that $\pi = 1$ and $p_L = y$. Then $H\pi - p_H = H - p_H > H\cdot(1 - N_H)/N_L - y - \delta T_L$ so that ICH does not bind.

Next consider the case where IRL binds. For the reason given in the previous paragraph, we cannot have $\pi = 1$ and $p_L = y$. First, consider the option $p_L < y$. Then an increase in p_L, combined with a reduction in T_L keeping $p_L + \delta T_L$ constant, always improves the outcome.

Finally, consider the possibility that at the optimum $\pi < 1$. In this case increase π while reducing T_L so as to keep the IRL binding. Then $d\pi/dT_L$ will satisfy $(L\cdot N_H/N_L)d\pi/dT_L = -\delta$. Substituting this into the ICH constraint, we find that the left-hand side goes up (because π goes up) and the right-hand side goes down. Therefore, this change relaxes the ICH constraint, and it is always optimal to make such a change. This proves the first part of our claim. The second part follows from the fact that with $v > 0$ a reduction in T_L is strictly in the bureaucrat's interest.

Proved

Solving the Bureaucrat's Maximization Problem [MB]

We solve the bureaucrat's maximization problem [MB] in a number of steps. The first step in solving the bureaucrat's maximization problem is to consider the more limited maximization problem where we drop the constraint ICL. This gives us the problem [mb]:

Choose p_H, p_L, π, T_H, T_L to maximize
$N_H p_H + N_L p_L - N_H \nu T_H - N_L \nu T_L - (1 - \pi) N_H F$,
subject to the constraints

(ICH) $H \cdot \pi - p_H - \delta T_H \geq H \cdot (1 - \pi N_H)/N_L - p_L - \delta T_L$,

(IRH) $H \cdot \pi - p_H - \delta T_H \geq 0$,

(IRL) $L \cdot (1 - \pi N_H)/N_L - p_L - \delta T_L \geq 0$,

$0 \leq p_L \leq y$, $0 \leq p_H \leq y$, $0 \leq \pi \leq 1$, T_H, $T_L \geq 0$.

The solution to this problem is given in Claim A1.

CLAIM A1. The solution to the problem [mb] described above is given below.

If $F \geq L$ and $y \geq H - (H - L) \cdot (1 - N_H)/N_L$, $\pi = 1$, $p_H = H - (H - L) \cdot (1 - N_H)/N_L$, $p_L = L(1 - N_H)/N_L$, and $T_H = T_L = 0$.

If $F \geq L$ and $H - (H - L) \cdot (1 - N_H)/N_L > y > L(1 - N_H)/N_L$, $\pi = 1$, $p_H = y$, $p_L = L(1 - N_H)/N_L$, and $T_H = T_L = 0$.

If $F \geq L$ and $y \leq L(1 - N_H)/N_L$, $\pi = 1$, $p_H = y$, $p_L = y$, and $T_H = T_L = 0$.

If $F < L$ and $y \geq H - (H - L) \cdot (1 - N_H)/N$, $\pi = 1$, $p_H = H - (H - L) \cdot (1 - N_H)/N_L$, $p_L = L(1 - N_H)/N_L$, and $T_H = T_L = 0$.

If $F < L$ and $H - (H - L) \cdot (1 - N_H)\, N_L > y \geq L/(N_H + N_L)$, $\pi = [N_L y + (H - L)]/[HN_L + (H - L)N_H]$, $p_H = y$, $p_L = L(H - N_H y)/[HN_L + (H - L)N_H]$, and $T_H = T_L = 0$.

If $F < L$ and $L/(N_H + N_L) > y \geq L(1 - N_H)/N_L$, $\pi = (1 - N_L y/L)/N_H$, $p_H = y$, $p_L = y$, and $T_H = T_L = 0$.

If $F < L$ and $y < (1 - N_H)/N_L$, $\pi = 1$, $p_H = y$, $p_L = y$, and $T_H = T_L = 0$.

Proof of Claim A1

Observe that at the optimum either the IRL constraint binds or $p_L = y$ (otherwise the bureaucrat would raise p_L). Consider first the case where the IRL constraint binds at the optimum. Assume to start out that the ICH constraint does not bind. Then p_H must be equal to y. What remains to be determined is the value of π. If ICH is not binding, a reduction in π has two effects: it increases $p_L N_L$ by $L \cdot N_H$, and it increases the expected punishment term by $F.N_H$. Therefore, if $L \leq F$, π will be set equal to 1. If $L > F$, π will be reduced until either ICH binds or IRL stops binding so that it ceases to be profitable to reduce π.

This leaves us with four distinct cases we need to consider:

i) $F \geq L$, and IRL binds;

ii) $F \geq L$, and IRL does not bind;

iii) $F < L$, and IRL binds;

iv) $F < L$, and IRL does not bind.

Consider the first two cases together. We know from above that if $F > L$ and IRL binds, π will be set equal to 1; a fortiori this will also be true if IRL does not bind. Then if IRL were to bind, p_L would be $L(1 - N_H)/N_L$. Therefore, IRL binds if and only if $L(1 - N_H)/N_L \leq y$.

Let IRL bind: then from ICH, $H - p_H \geq (H - L)\cdot(1 - N_H)/N_L$ which implies that $p_H \leq H - (H - L)\cdot(1 - N_H)/N_L$. Now either this is an equality or $p_H = y$. Which happens depends on how y compares with $H - (H - L)\cdot(1 - \pi N_H)/N_L$; p_H will be the smaller of the two.

If IRL does not bind, then $p_L = y$. Then ICH cannot bind either since $H(1 - N_H)/N_L - y < H - p_H$. Therefore, $p_H = y$.

Turning now to the case where $F < L$ and both ICH and IRL bind, we substitute IRL in ICH to get

$$H \cdot \pi - p_H = (H - L) \cdot (1 - \pi N_H)/N_L. \tag{A1}$$

If we increase p_H toward y, π has to go up. The rate at which it goes up, $d\pi/dp_H$, is $1/[H + (H - L)N_H/N_L]$. The resulting reduction in p_L will be $L\cdot(N_H/N_L)\cdot[H + (H - L)N_H/N_L]^{-1}$. Therefore, there will be a net gain from the increase in p_H if $N_H > N_L\cdot L\cdot(N_H/N_L)\cdot[H + (H - L)N_H/N_L]^{-1}$ which is always true. So, the outcome in this case is either $p_H = y$ or $\pi = 1$.

Which of these two outcomes obtains at the optimum depends on which binds first as we increase p_H toward y. It can be checked by looking at (A1) that if y is greater than $H - (H - L)\cdot(1 - N_H)/N_L$ then π will hit 1 before p_H hits y. Therefore, this will be the outcome. However, if y is below this critical level, then p_H will hit y with π less than 1.

Of course, these predictions assume that the IRL constraint binds rather than the alternative outcome $p_L = y$. Now so long as y is greater than L, we cannot have $p_L = y$ since this would violate IRL. Therefore, the IRL constraint must bind if y is higher than L. By continuity it will also continue to bind when y is lower than L but not too low. However, as we continue to reduce y, π will fall toward $(1 - \pi N_H)/N_L$, and p_L will rise to close the gap with p_H. This cannot go on indefinitely. y must ultimately reach another

critical value; at this value of y, π must be equal to $(1 - \pi N_H)/N_L$; both p_H and p_L must be equal to y; and any further reduction in y will make p_L greater than y. A simple calculation establishes that the critical value of y must be $L/(N_H + N_L)$ and the corresponding value of π must be $1/(N_H + N_L)$.

Once y falls below $L/(N_H + N_L)$, the constraint $p_L \leq y$ will bind, and therefore there is nothing to be gained by further lowering π. It is easily checked that it is optimal to set $p_L = p_H = y$ and to raise π to meet the IRL constraint (since $\pi > 1/(N_H + N_L)$ and $p_L = p_H$, ICH cannot bind).

The value of π as a function of y in this region of the parameter space will be (from IRL) $\pi = (L - N_L y)/N_H L$. Now as y goes to 0, this value of π goes to a number greater than 1. Therefore, y must hit a critical value beyond which reducing y does not increase π. This value of y is $L(1 - N_H)/N_L$. Below this value of y, $\pi = 1$.

Compiling all the results proved above, we have the claimed result. Proved

We next observe that at the solution to [mb] the suppressed constraint ICL does not always bind.

CLAIM A2. ICL binds at the values of p_H, p_L, π, T_H, and T_L which solve the [mb] iff (a) if $F \geq L$, and $y < L$, and (b) if $F < L$ and $y < L \cdot [N_H + N_L]^{-1}$.

Proof. Immediate from substitution of the solution of [mb] into ICL.

The next step is to note that since when ICL does not bind [MB] is the same as [mb], the solution to [MB] is just the solution to [mb] when conditions (a) and (b) do not hold. We state this as

CLAIM A3. If $F \geq L$, and $y \geq L$, or if $F < L$ and $y \geq L \cdot [N_H + N_L]^{-1}$, the solution to [MB] is the same as the solution to [mb].

Finally, we directly solve the problem for the case where it is known that ICL binds, assuming that ν/δ is not too large. The solution is given below (we only describe the solution for values of y higher than $L(1 - N_H)/N_L$ to prevent the statement from becoming too long—the full statement is given in the previous version of the paper).

CLAIM A4. Let $N_L/N_H > \nu/\delta$ and $\nu/\delta + N_H \nu/N_L \delta < 1$. Then the solution to [MB] for the parameter values $L(1 - N_H)/N_L \leq y < L$

if $F \geq L$ and $L(1 - N_H)/N_L \leq y \leq L \cdot [N_H + N_L]^{-1}$ if $F < L$, is as follows.

If $L > y \geq L \cdot [N_H + N_L]^{-1}$ and $L(1 - \nu/\delta) \leq F$, the outcome is $\pi = 1$, and T_H set to solve the equation $L - y - \delta T_H = 0$.

If $L > y \geq L \cdot [N_H + N_L]^{-1}$, $L \leq F < L(1 + \nu/\delta)$, the outcome is $\pi = y/L$ and $T_H = 0$.

If $L \cdot (1 - N_H)/N_L \leq y < L \cdot [N_H + N_L]^{-1}$, $L(1 + \nu/\delta) \leq F$, the outcome is $\pi = 1$ and T_H set to solve $L - y - \delta T_H = 0$.

If $L \cdot (1 - N_H)/N_L \leq y < L \cdot [N_H + N_L]^{-1}$, $L(1 + \nu/\delta) > F \geq L(\nu/\delta + N_H \nu/N_L \delta)$, the outcome is π and T_H set to solve $\pi L - y - \delta T_H = 0$ and $L(1 - N_H \pi)/N_L = y$.

If $L \cdot (1 - N_H)/N_L \leq y < L \cdot [N_H + N_L]^{-1}$, $F < L(\nu/\delta + N_H \nu/N_L \delta)$, the outcome is $\pi = (N_H + N_L)^{-1}$ and $T_H = 0$.

Proof. Note that since ICH does not bind, raising p_H is always a good thing. Therefore, $p_H = y$. Assume now that $T_H > 0$, and consider the effect of a ΔT_H reduction in T_H on the bureaucrat's objective function. To keep ICL satisfied, we must reduce either p_L or π. In the case when we reduce p_L, the gain is $\nu N_H \Delta T_H$ which is less than the loss that is $N_L \delta \Delta T_H$ by our condition $\nu/\delta < N_L/N_L$. Therefore, it will never pay to reduce p_L. In fact, p_L will be raised until either IRL binds or $p_L = y$.

Assume next that IRL binds. This combined with ICL implies that

$$\pi L - y - \delta T_H = 0. \tag{A2}$$

From (A2) $d\pi/dT_H = \delta/L$. Using this in combination with the formula for $dp_L/d\pi$ derived from IRL, we find that an increase in T_H (weakly) increases the bureaucrat's welfare if $F \geq (1 + \nu/\delta)L$. Therefore, if $F \geq (1 + \nu/\delta)L$, an increase in π accompanied by the corresponding rise in T_H must increase the bureaucrat's welfare. Conversely, as long as $p_L < y$, if $F < (1 + \nu/\delta)L$, a reduction in T_H must raise the bureaucrat's welfare.

Next let IRL not bind. Then from ICH, $dT_H/d\pi = L(1 + N_H/N_L)/\delta$. Therefore, an increase in π accompanied by a rise in T_H (weakly) raises the bureaucrat's welfare iff $F \geq L(\nu/\delta + \nu N_H/N_L \delta)$.

Since $L(\nu/\delta + \nu N_H/N_L \delta) < L(\nu/\delta + 1)$, $F \geq L(\nu/\delta + 1)$ suffices in both cases. Therefore, under this condition π will be set equal to 1 (since an increase in π accompanied by an increase in T_H increases the bureaucrat's welfare). Therefore, $p_L = \min\{(1 - N_H)/N_L, y\}$ which, given our restriction on y, means that $p_L = (1 - N_H)/N_L$.

Next consider the case where $L(v/\delta + vN_H/N_L\delta) \leq F < L(v/\delta + 1)$. In this case it does not pay to increase π once IRL binds but so long as IRL does not bind, π will be increased. Therefore, either $\pi = 1$, or π must be such that IRL just binds. But if IRL does not bind, we must have $p_L = y$ which along with $\pi = 1$ implies that IRL is violated (as long as $y \geq (1 - N_H)/N_l$). Therefore, IRL must bind; i.e., we must have $L(1 - N_H\pi)/N_L = p_L$.

Now we know from above that when IRL binds and $p_L < y$, if $F < (1 + v/\delta)L$ the bureaucrat always wants to reduce T_H. Therefore, at the optimum we will have $T_H = 0$. This implies that the optimal values of π and p_L will be, respectively, y/L and $L(1 - N_H y/L)/N_L$.

By contrast, when $y < L/(N_H + N_L)$, solving IRL and ICL with $T_H = 0$ yields a solution for p_L which is greater than y. Therefore, we must choose $T_H > 0$. Specifically, we will choose $p_L = y$ and π and T_H to satisfy $\pi L - y - \delta T_H = 0$ and $L(1 - N_H\pi)/N_L = y$.

Proved

Claims A3 and A4 between them describe the full solution to the bureaucrat's problem [MB].

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

REFERENCES

Acemoglu, Daron, "Reward Structures and the Allocation of Talent," mimeo, 1992.
Aghion, Philippe, and Robin Burgess, "Financing Development in Eastern Europe and the Former Soviet Union," in A. Giovannini, ed., *Finance and Development: Issues and Experience* (Cambridge, UK: Cambridge University Press, 1993).
Banerjee, Abhijit V., "Eliminating Corruption," mimeo, MIT, 1995.
Becker, Gary, and George Stigler, "Law Enforcement, Malfeasance and the Compensation of Enforcers," *Journal of Legal Studies,* III (1974), 1–19.
Bolton, Patrick, and Gérard Roland, "Privatization Policies in Central and Eastern Europe," *Economic Policy,* XV (1992), 276–309.
Cadot, Olivier, "Corruption as a Gamble," *Journal of Public Economics,* XXXIII (1987), 223–44.
Che, Yeon-Koo, and Ian Gale, "Auctions with Financially Constrained Bidders," mimeo, University of Wisconsin at Madison, 1994.
Clague, Christopher, "Rule Obedience, Organizational Loyalty, and Economic Development," *Journal of Institutional and Theoretical Economics,* CXLIX (1993), 393–414.
Dixit, Avinash, *The Making of Economic Policy* (Cambridge, MA: MIT Press, 1996).
Holmstrom, Bengt, and Paul Milgrom, "Multi-Task Principal Agent Analyses," *Journal of Law, Economics and Organizations,* VII (1991), 24–52.
Kofman, Alfredo, and Jacques Lawarree, "On the Optimality of Allowing Corruption," mimeo, 1990.
Laffont, Jean-Jacques, and Jean Tirole, *A Theory of Incentives in Procurement and Regulation* (Cambridge, MA: MIT Press, 1993).
Leff, Nathaniel H., "Economic Development through Bureaucratic Corruption," in

Monday Ekpo, ed., *Bureaucratic Corruption in Sub-Saharan Africa* (Washington, DC: University Press of America, 1979).
Lewis, Tracy, and David Sappington, "Assigning Productive Assets and Profit-Sharing Opportunities to Wealth Constrained Agents," mimeo, University of Florida, 1996.
Mauro, Paolo, "Corruption, Country Risk, and Growth," *Quarterly Journal of Economics,* CX (1995), 681–713.
Murphy, Kevin M., Andrei Shleifer, and Robert Vishny, "Why Is Rent-Seeking so Costly to Growth?" *American Economic Review Papers and Proceedings,* LXXXIII (1993), 409–14.
Nye, Joseph S., "Corruption and Political Development: A Cost-Benefit Analysis," in Monday Ekpo, ed., *Bureaucratic Corruption in Sub-Saharan Africa* (Washington, DC: University Press of America, 1979).
Olson, Mancur, "Dictatorship, Democracy and Development," *American Political Science Review,* LXXXVIII (1993), 567–76.
Sah, Raaj K., "Social Osmosis and Patterns of Crime," *Journal of Political Economy,* XCIX (1991), 1272–95.
Shleifer, Andrei, and Robert Vishny, "Corruption," *Quarterly Journal of Economics,* CVIII (1993), 599–617.
Tirole, Jean, "Collusion and the Theory of Organisations," in *Advances in Economic Theory: Proceedings of the 6th World Congress of the Econometric Society,* Jean-Jacques Laffont, ed., (Cambridge, UK: Cambridge University Press, 1992).
——. "A Theory of Collective Reputations," *Review of Economic Studies,* LXIII (1996), 1–22.
Waterbury, John S., "Endemic and Planned Corruption in a Monarchical Regime," in Monday Ekpo, ed., *Bureaucratic Corruption in Sub-Saharan Africa* (Washington, DC: University Press of America, 1979).
Weitzman, Martin, "Is the Price System or Rationing More Effective in Getting a Commodity to Those Who Need it Most?" *Bell Journal of Economics,* VIII (1977), 517–24.
Wilson, James Q., *Bureaucracy: What Government Agencies Do and Why They Do It* (New York: Basic Books, 1989).

OBTAINING A DRIVER'S LICENSE IN INDIA: AN EXPERIMENTAL APPROACH TO STUDYING CORRUPTION*

MARIANNE BERTRAND
SIMEON DJANKOV
REMA HANNA
SENDHIL MULLAINATHAN

We study the allocation of driver's licenses in India by randomly assigning applicants to one of three groups: bonus (offered a bonus for obtaining a license quickly), lesson (offered free driving lessons), or comparison. Both the bonus and lesson groups are more likely to obtain licenses. However, bonus group members are more likely to make extralegal payments and to obtain licenses without knowing how to drive. All extralegal payments happen through private intermediaries ("agents"). An audit study of agents reveals that they can circumvent procedures such as the driving test. Overall, our results support the view that corruption does not merely reflect transfers from citizens to bureaucrats but distorts allocation.

I. INTRODUCTION

Public service provision in many developing countries is rife with corruption. A basic question about such corruption is whether it merely represents redistribution between citizens and bureaucrats or results in important distortions in how bureaucrats allocate services. This question underlies the debate on the efficiency implications of corruption, with some arguing that corruption merely "greases the wheels" of the bureaucracy and others arguing that it harms society.[1] In this paper, we use detailed survey data and experimental evidence to study this question in the context of one particular bureaucratic process: the provision of driver's licenses in Delhi, India.

* This project was conducted and funded by the International Finance Corporation. We thank Anup Kumar Roy for outstanding research assistance. We are grateful to Lawrence Katz (the editor), three anonymous referees, Abhijit Banerjee, Gary Becker, Ryan Bubb, Anne Case, Angus Deaton, Luis Garicano, Ed Glaeser, Ben Olken, Sam Peltzman, Andrei Shleifer, and Jakob Svensson and to seminar participants at Harvard, MIT, Princeton, the University of California at Berkeley, the University of Chicago GSB, LSE, Yale University, NYU, Ohio State University, the University of Florida, the University of Toronto, the World Bank, and the ASSA 2006 meeting for helpful comments.

1. For the "grease-the-wheels" view, see Leff (1964), Huntington (1968), and Lui (1985). For example, Huntington (1968) remarked that "[I]n terms of economic growth, the only thing worse than a society with a rigid, overcentralized, dishonest bureaucracy is one with a rigid, overcentralized, and honest bureaucracy." For arguments on how corruption can harm society, see Myrdal (1968), Rose-Ackerman (1978), Klitgaard (1991), Shleifer and Vishny (1992, 1993), and Djankov et al. (2002).

The Quarterly Journal of Economics, November 2007

Specifically, between October 2004 and April 2005, the International Finance Corporation (IFC) followed 822 driver's license candidates, collecting data on whether they obtained licenses, as well as detailed micro data on the specific procedures, time, and expenditures involved.[2] At the end of the process, the IFC administered an independent surprise driving test (simulating the test that is supposed to be given by the bureaucrats) to determine whether individuals who were granted a license could drive.

To understand whether and how corruption affects allocation, license candidates were randomly assigned to one of three groups. The "bonus group" were offered a large financial reward if they were able to obtain a license in 32 days (two days longer than the statutory minimum time of 30 days). The "lesson group" were offered free driving lessons, to be taken immediately after recruitment into the survey.[3] The comparison group were simply tracked through the process. The bonus treatment allows us to assess whether and how the allocation of licenses responds to willingness to pay. Are a group that are willing to pay more for licenses more likely to get them? But also, are there more *unqualified* drivers receiving licenses in such a group? The lesson treatment allows us to assess whether allocation decisions by the bureaucracy are at all responsive to the socially most important component of this regulatory process–one's ability to drive.

The comparison group's experiences already provide evidence of a distorted bureaucratic process. Close to 71% of license getters in the comparison group did not take the licensing exam, and 62% were unqualified to drive (according to the independent test) at the time they obtain a license.[4] The average license getter in this group paid about Rs 1,120, or about 2.5 times the official fee of Rs 450, to obtain a license.

The experimental results highlight how these distortions respond to private willingness to pay. While individuals in the bonus

2. Other microempirical approaches to documenting and measuring corruption are Di Tella and Schargrodsky (2003), Fisman and Wei (2004), and Olken (2005).

3. To ensure that there were no social costs to the study, participants in the comparison and bonus groups were offered free driving lessons upon completion of the final survey and driving test.

4. Why acquire a license without knowing how to drive, especially since licenses are not used as a primary form of identification in India? License getters will likely learn how to drive after they get the license, as we discuss later on. The key point is that their driving skill level is unregulated; they will learn to the level that they find privately useful rather than the socially optimal level.

group are 24 percentage points more likely to obtain a license than those in the comparison group, they are also 13 percentage points more likely to obtain a license without taking the legally required driving exam, as well as 18 percentage points more likely to *both* obtain a license *and* fail the independent driving test.[5] In other words, a higher willingness to pay for a license translates into an increase in the number of license getters who cannot drive. The experimental results regarding the lesson group, however, suggest that social considerations are not totally ignored in the allocation of licenses: the lesson group is 12 percentage points more likely to obtain a license than the comparison group.[6] As a whole, the bonus group pay Rs 178 more in extralegal fees. Individuals in the lesson group continue to make extralegal payments despite being better drivers: the average extralegal payment is about the same in the lesson and comparison groups (albeit with more licensed drivers in the lesson group).

Interestingly, we find no evidence of *direct* bribes to bureaucrats in any of the groups. The extralegal payments are mainly fees to "agents," professionals who "assist" individuals in the process of obtaining their driver's licenses. These agents appear to be more than just time-saving institutions (akin to accountants embodying knowledge of tax regulations). Instead, multiple pieces of evidence suggest that agents institutionalize corruption. We find that 94% of individuals who did not hire agents took the legally required driving test at least once, while only 12% of those who used agents took that test. To investigate this further, we designed a second experiment aimed exclusively at understanding how agents affect the licensing process. Specifically, trained actors were sent to agents to elicit the feasibility of and prices for obtaining a license under different pretexts, which corresponded to bending various official rules. We find that agents can provide services that circumvent official rules. For example, agents were able to procure a license despite someone's lack of driving skills: agents offered to procure licenses for 100% of actors who said they

5. Moreover, the average license getter in the bonus group is more likely to fail our driving test than the average license getter in the comparison group. This suggests that the bonus group's failure rate is higher than one would estimate if one simply added more license getters (but with the same failure rate) to the comparison group.

6. We cannot rule out the possibility that simply being offered lessons also raised the lesson group's desire to get a license and, therefore, the effort they were willing to exert to obtain a license. The lesson group may thus also have a higher private willingness to pay for the license.

did not have the time to learn how to drive. However, they cannot bend all rules as easily: rules that leave a documentary trail (such as place-of-residence restrictions) appear harder for agents to circumvent.

Finally, to understand why good drivers in the lesson group continue to make extralegal payments, we studied nonexperimentally the experiences of those who try to use the formal (i.e., nonagent) channel for getting a license. Examining the subset of participants who began the process by taking the driving test once, we find that a substantial percentage of them (about 35%) failed and must resort to retaking the test or hiring an agent. Most interestingly, this percentage is *unrelated* to actual ability to drive: it is constant across the lesson, bonus, and comparison groups, and it is also constant across scores on the independent driving test. One possible interpretation of these suggestive data is that bureaucrats arbitrarily fail test takers in order to induce them to use agents. This interpretation is consistent with theories of "endogenous red tape," which emphasize that many bureaucratic hurdles might be the *result* of rent-seeking activities by bureaucrats (see for example Myrdal [1968], Shleifer and Vishny [1993], and Banerjee [1997]).

Hence, there appear to be two paths to obtaining a driver's license in New Delhi: the official path and the agent path. While following the agent path involves substantial extra costs, it ensures getting a license even without knowing how to drive, most likely because agents make payments to bureaucrats to bend the rules. While it is possible to obtain a license without hiring an agent, it also appears that bureaucrats may create hurdles (red tape) to encourage the use of agents. Overall, these results support the view that corruption in this particular setting goes beyond simple redistribution from citizens to bureaucrats.

The rest of the paper proceeds as follows. Section II discusses the process of obtaining a driver's license in India, while Section III describes the data collection and lays out the design of the first experiment (comparative experiences of comparison, bonus, and lesson groups). These experimental findings are presented in Section IV. Section V explores the process of getting a license with an agent, relying both on nonexperimental data and also on the findings of the second experiment (audit study of agents); we also investigate the possibility of red tape in the formal process. Section VI discusses alternative interpretations. Section VII concludes.

II. Getting a Driver's License in Delhi, India

The Motor Vehicle Act of 1988 and its subsequent amendments stipulate the official licensing process in India. State governments are responsible for administering this act. In Delhi, the setting for this project, licenses are issued at nine regional transport offices (RTOs). The jurisdiction of each office coincides with the corresponding police district, and individuals can only obtain licenses from their particular RTOs. In 2002, the Delhi Motor Vehicle Department authorized 313,690 licenses.

To be eligible for a license, an individual must be at least 18 years of age. He or she must first obtain a temporary license, which grants the right to practice driving under the supervision of a licensed individual. To obtain the temporary license, proof of residence, proof of age, a passport-sized photo, and a medical certificate must be submitted to the RTO, along with the application form. There is an application fee of Rs 360 ($8). Then the applicant must take a color blindness test and a written examination with 20 multiple choice questions on road signs, traffic rules, and traffic regulations. Upon the applicant's passing these, the temporary license is processed on the same day. If the applicant fails the exam, he or she can reapply after a 7-day waiting period.

After 30 days (and within 180 days) of the issuance of the temporary license, the individual may apply for a permanent license. The applicant must submit proof of age, proof of residence, a recent passport-sized photo, and his or her temporary license. The applicant must also pass a driving road test at the RTO. A Rs 90 fee ($2) is charged for the photograph and lamination of the license. If the applicant fails the road test, he or she can reapply after a 7-day waiting period.

III. Design of the First Field Experiment

In the first experiment, the IFC recruited and observed individuals through the application process for a four-wheeler license. The three main project phases—recruitment, randomization, and follow-up—are described below (see also Figure I).

III.A. Recruitment

Recruitment began in June 2004 and continued through November 2004. Recruiting occurred in a two-week cycle. During each cycle, recruiters intercepted individuals who were entering

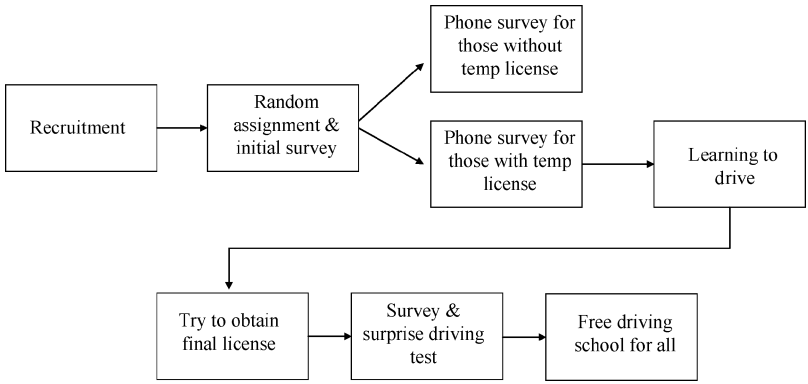


FIGURE I
Project Summary

one of the following four RTOs in Delhi: Southwest, Northwest, South, or New Delhi. The IFC gave recruiters strict guidelines regarding the type of person to approach for the project. First, to reduce attrition, recruiters were instructed to approach only men (in a pilot study, 60% of men remained in the project, while 100% of the women dropped out). Second, they were asked to identify individuals who had not previously had a license, but wanted one. Finally, to comply with government regulations, only individuals over age 18 were allowed to participate.

The recruiters provided each potential participant with a short explanation of the project, offered an information sheet outlining the time frame and payment structure for the project, and invited interested individuals to attend an information session to learn more about the project.

III.B. Initial Session and Randomization

An initial survey session was held at the end of each two-week recruiting cycle near the RTO from which the subjects were recruited. On average, 36 individuals participated in each of the 23 sessions, for a total of 822 project participants (see Figure II). Participation was restricted to individuals who had been officially recruited and up to one of their friends.[7]

To begin, the survey team administered an introduction survey to each participant. In addition to sociodemographic

7. To further limit attrition, the project team rejected any individual whose phone number could not be verified prior to the session and required formal identification (student identification, ration card, etc.).

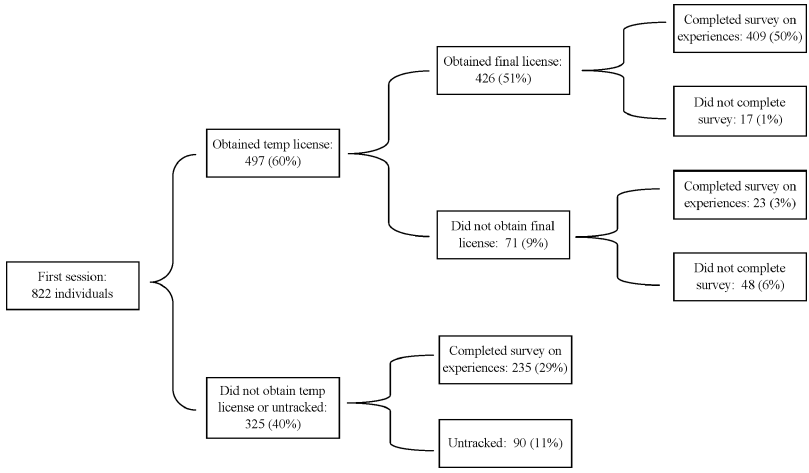


FIGURE II
Final Licesing Status of Participants

Note: Percentage of individuals out of original 822 survey participants reported in parentheses.

information, the survey included questions on previous experiences in obtaining government services and previous driving experience, as well as beliefs about the necessary procedures to obtain a driver's license. The survey concluded with a series of questions regarding driving laws and practices; these questions were drawn from a sample of practice test questions published by the Delhi RTO.[8]

After the survey, each individual was given one of three possible letters. The letters randomly allocated him to one of three groups: a comparison group, a bonus group, and a lesson group. Individuals in the comparison group were simply asked to return for a second survey—documenting their experiences—upon acquiring a permanent license. As an inducement to return, each subject was offered Rs 800 (roughly $17) upon completion of the final survey.[9]

The IFC gave individuals in the bonus group the same set of instructions as those in the comparison group. However, to

8. For example: You are driving in heavy rain. Your steering suddenly becomes very light. You should (1) steer toward the side of the road, (2) brake firmly to reduce speed, (3) apply gentle acceleration, (4) ease off the acceleration, (5) do not know.

9. Since all subjects received a cash payment, their behavior may not be representative of the population as a whole. This does not compromise the internal validity of the differences between treatment and comparison groups.

generate a stronger incentive for obtaining a license, the IFC also offered a bonus of Rs 2,000 (on top of Rs 800 for completing the surveys) if the individuals could obtain their permanent licenses within 32 days of obtaining their temporary licenses (two days over the official minimum wait time). Rs 2,000 was chosen to ensure a large enough treatment effect.[10]

Finally, in addition to being given the same set of instructions as the comparison group, individuals in the lesson group were offered free driving lessons, to be taken immediately. Accredited driving schools were hired to provide up to 15 lessons. Individuals in this group were also promised a payment of Rs 800 upon completion of the surveys.

At the end of this initial session, the project team paid all participants Rs 200 ($4.25). This was done to help alleviate possible credit constraints on acquiring a license. This upfront payment was also made in order to increase the credibility of the final payment. Behavioral studies of this type are not typical in India and participants in the pilot (who did not receive this upfront payment) harbored suspicions about whether the final payment would be made.

While the project team tried to isolate the three groups from each other, we cannot rule out the possibility that individuals in different groups communicated with each other during this process. To increase transparency, each of them was informed that several groups existed in the study, and that some participants were randomly chosen to win additional payments.

III.C. Follow-Up

It may take as few as 30 days or as many as 180 days to obtain a license. During this period, the project team kept in close contact with all participants to remind them about the project and maintain the credibility of the final payments. Extensive phone calls were made (and logged) to ensure that participants understood the instructions and payments schemes, to arrange lessons for subjects in the lesson group, and to remind subjects in the bonus group about the bonus scheme and deadlines.

As shown in Figure II (and, in more detail, in Appendix I), 497 individuals (60%) obtained temporary licenses. The project

10. The monthly gross salary for the 380 employed individuals in our sample is Rs 5,446, and so the bonus is roughly equivalent to one-third of an individual's monthly income.

team administered a phone survey to these individuals regarding the subject's experiences in the bureaucratic process so far. The project team also attempted to administer a phone survey to the 325 individuals who failed to obtain temporary licenses in order to understand the reasons that they did not. Ninety individuals could not be contacted. Since we are unsure whether they obtained any type of license, we exclude them from the rest of the analysis.

Upon earning a permanent license, each subject was invited to a final session. Half of the original set of participants both obtained a final license and returned for the final survey. At this session, the survey team questioned each individual on his experiences in the process, tested his driving skills, gave the final payment, and, for those in the comparison and bonus groups, offered free driving lessons.[11]

Under the supervision of the project team, an accredited driving school administered a surprise practical driving test. The examination was designed to test the skills required to obtain a license. To preserve the integrity of the test, the test-givers were not from any of the schools that provided the instruction to the lesson group and did not know which experimental group a given test-taker belonged to. The driving exam consisted of two parts. First, the test-giver administered an oral examination to judge whether a subject could operate a car.[12] If a subject was unable to answer all of these questions correctly, he was deemed incapable of taking the practical driving test and automatically failed. If the subject adequately answered all questions, the test-giver administered a road test. The test-giver awarded subjects a series of points for satisfactorily illustrating that they could properly start a car, change gears, use indicators, complete turns, and park. The key feature of this test is that it mirrors exactly what the RTO itself is supposed to be testing.

The project team offered Rs 500 to the 71 individuals who obtained temporary licenses, but did not obtain a final license, to also attend a final session. At this session, the project team administered a survey to understand why they did not obtain a license and also administered the surprise driving exam. Twenty-three individuals attended this session (Figure II).

11. Upon earning a permanent license, an individual is required to relinquish his temporary license to the RTO. As proof of date, subjects in the bonus group were required to bring photocopies of their temporary licenses.

12. This oral exam was not a test of technical terms. Instead it tested basic knowledge needed to operate a motor vehicle. For example, individuals were asked, "which pedal would you use to speed up?" "how would you start the car?" etc.

For the rest of the paper, an individual is considered an attritor if he could not be tracked during the study (90 individuals) or if he did not complete the requested final survey (65 individuals); this leaves 667 individuals. Appendix II studies the differences between attritors and nonattritors in terms of socioeconomic characteristics, driving experiences, past bribing experience, and beliefs regarding procedures (as collected in the initial survey). We find very little difference between attritors and nonattritors, with two exceptions: attritors are less likely to be married and more likely to have driven a two-wheeler in the past. If the different treatments caused differential attrition, the comparison of the treatment groups to the comparison group may be less valid. In fact, a few characteristics (mainly age, marital status, and having driven a four-wheeler at one time in the past) are not balanced between attritors and nonattritors across the three groups. Therefore, we control for these characteristics in our empirical specifications.

III.D. Survey Participants' Characteristics

Table I describes the main characteristics of the 667 individuals in the study whom we were able to track and who completed the requested final survey. Column (1) presents means for the full sample, while columns (2)–(4) present means for each group. The stars indicate whether a given group's mean significantly differs from the two other groups', after controlling for session fixed effects. All standard errors are robust.

Panels A and B document the participants' socioeconomic backgrounds and their past driving experience. Individuals tend to be young (24 years of age) and many are high school or college students (49%). Seventy-seven percent are Hindu, while 20% are Muslim; 35% have minority status (Other Backward Castes, Scheduled Castes, or Scheduled Tribes). Many have driven a two-wheeler at least once (88%), yet only 3% report having a two-wheeler license. Close to a quarter report having driven a four-wheeler at least once in the past. As Delhi is India's capital, it is unsurprising that 43% have at least one family member (usually a parent) employed by the government.

The characteristics summarized in Panels A and B appear balanced across the three groups. There are no significant differences across groups in age, education levels (as measured by percentage of people with less than a primary school education), employment status, wealth (as measured by owning a home or owning a car), income, or likelihood of having a two-wheeler

TABLE I
SOCIOECONOMIC CHARACTERISTICS, PAST DRIVING EXPERIENCES, AND BELIEFS ON PROCESS

	Full sample (1)	Comparison (2)	Bonus (3)	Driving lesson (4)
A. Socioeconomic characteristics				
Age	24.28	23.82	24.70	24.11
Married	0.25	0.22	0.27	0.24
Students	0.49	0.50	0.45	0.52
Employed	0.47	0.45	0.50	0.45
Less than primary education	0.08	0.06	0.07	0.09
Owns home	0.61	0.61	0.59	0.63
Owns car	0.11	0.10	0.13	0.09
Minority	0.35	0.43	0.31	0.35
Hindu religion	0.77	0.84**	0.77	0.73
Muslim religion	0.20	0.15	0.19	0.23
Log (salary)	3.90	3.70	4.18	3.73
Family member in government (including self)	0.43	0.38	0.45	0.43
B. Driving experience				
Have 2-wheeler license	0.03	0.03	0.02	0.03
Have driven a 2-wheeler	0.88	0.83**	0.91*	0.86
Have driven a 4-wheeler	0.24	0.24	0.34***	0.11***
Months known how to drive a 4-wheeler (given drive)	3.66	3.38	3.96	3.04
C. You are caught driving without a license. Would you bribe.				
If the fine is 500 and bribe is 300?	0.61	0.64	0.60	0.60
If the fine is 3,000 and bribe is 300?	0.81	0.84	0.79	0.79
D. Ever. . . in the past (conditional on having tried to obtain a public service)				
Paid bribe	0.20	0.18	0.23	0.17
Used agent	0.21	0.19	0.23	0.20
E. Beliefs regarding procedures				
Total trips to obtain license	6.92	7.50	6.87	6.60
Total time at RTO	1,135.35	1,225.15	1,173.69	1,031.52
N	667	155	268	244

Notes.

1. This table reports summary statistics from the initial baseline survey. The mean demographics, driving experiences, and beliefs regarding the license process are presented for the 667 individuals who were tracked during the process and filled out all relevant surveys.

2. Column (1) presents the means for the full sample, while columns (2)–(4) report the means by the three experimental groups: comparison, bonus, and lesson.

3. Stars indicate a significant difference from other two groups, after controlling for session fixed effects. Standard errors are robust. Significance at the 10% level is represented by *, at the 5% level by **, and at the 1% level by ***.

license. There are some exceptions. First, individuals in the comparison group are more likely to be Hindu. Second, a larger fraction of those in the bonus group and a lower fraction of those in the comparison group report having driven a two-wheeler at least once in the past. Third, a larger fraction of those in the bonus group and a smaller fraction of those in the lesson group report having driven a four-wheeler before. However, conditional on having driven a four-wheeler, there are no systematic differences across groups in the tenure of driving a four-wheeler.

Survey participants talk openly about bribes and agent usage. First, to capture attitudes toward bribing, the project team posed the following hypothetical scenario to individuals: "You are driving without a license and are pulled over by a policeman. The policeman offers you a choice of paying a Rs 500 fine or a Rs 300 bribe." Sixty-one percent of the sample indicated that they would pay the bribe, and there were no significant differences in the propensity to bribe across the three groups (Panel C). Participants have some distaste for paying bribes, as evidenced by the fact that when the cost of the fine relative to the bribe increases, more individuals are willing to pay the bribe (for example, 81% of the sample stated that they would pay the bribe if the fine was Rs 3,000 and the bribe remained Rs 300). Second, the project team asked individuals whether they had paid a bribe at least once in the past (Panel D). Conditional on having obtained a service, 20% of individuals paid a bribe and 21% report having hired an illegal agent to help obtain a service (these are not mutually exclusive groups).[13] There are no systematic differences in past bribing behavior or agent usage across the three groups.

The final panel reports the participants' beliefs regarding the process of obtaining a license. Participants think that the entire process will take on average 6.9 trips. As we will see, this is more trips than it will take the average participant in practice. There are no systematic differences in beliefs across the three groups.

In summary, while the precharacteristics are fairly well balanced across the three groups, there are some systematic differences. We directly control for those characteristics in the analysis that follows.

13. The list of services covered in the initial survey was as follows: ration card, passport, land title, building permit, electricity, water, voter's card, personal account number (which is equivalent to a social security number). The highest likelihood of bribe payment was with regard to ration cards, followed by land titles and building permits.

TABLE II
SUMMARY STATISTICS ON THE BUREAUCRATIC PROCESS FOR THE COMPARISON GROUP

Variable	Mean
A. Final license status	
Obtained a final license	0.48
Obtained a license in 32 days or less	0.15
Obtained a final license conditional on trying	0.69
Obtained a license without taking licensing exam	0.34
Obtained license & automatically failed ind. exam	0.29
B. The process by which individuals obtained licenses	
Number of days between temporary and final license	47.99
	(29.14)
Predicted number of trips	6.46
	(4.10)
Number of trips	2.50
	(0.73)
Minutes spent at RTO (across all trips)	206.07
	(111.86)
Number of officials spoken with	4.73
	(2.90)
Lines waited in (final license)	2.51
	(1.09)
Took RTO licensing exam	0.30
	(0.46)

Notes.
1. This table describes the licensing process for the comparison group.
2. Panel A includes all 156 individuals who were both tracked during the course of the study and completed all surveys, while Panel B includes all 74 individuals who obtained a final license and completed all surveys.
3. "Trying" is defined as making at least one trip to the regional transport office after the initial session. "Predicted number of trips" is the number of trips an individual predicted it would entail to obtain a license prior during the initial baseline survey.
4. Standard deviations are in parentheses.

IV. EMPIRICAL RESULTS FROM FIRST EXPERIMENT

How does this bureaucratic system respond to variation in individuals' willingness to pay for a driver's license ("bonus" treatment)? How does it respond to variation in individuals' deservingness of a driver's license ("lesson" treatment)? Before examining the experiment designed to address these questions, we describe some interesting facts about individuals in the comparison group. These are reported in Table II.

Panel A includes all individuals in the comparison group who could be tracked by the survey team and completed the requested surveys, as described in Section III. Only 48% were able to obtain their permanent driver's licenses and only 15% were able to obtain them within 32 days of obtaining their temporary licenses. This low success rate cannot solely be attributed to the difficulty of obtaining a license. Some participants reported that they did not

try to obtain a license (see Appendix I), where trying implies having visited the RTO at least once after the initial session (to talk to either a bureaucrat or an agent). Excluding these individuals, 69% obtained permanent licenses.

Most striking are the statistics in the next two rows of Panel A. We find that 34% of individuals in the comparison group obtained licenses without taking the legally required driving exam at the RTO; given that only 48% obtained licenses, this implies that close to 71% of the license getters did not take the licensing exam. This indicates a large misapplication of the socially most useful component of this regulation—the screening of driving skills. It is possible that bureaucrats use other means, perhaps less time-intensive ones, to assess driving ability. The results of our independent driving test suggest otherwise. Twenty-nine percent of individuals in the comparison group obtained licenses *and* automatically failed our independent driving test, where failing means that the individual knew so little about the workings of the car that the test-giver refused to take him on the road. In other words, 62% of the license getters were unqualified to drive at the time they obtained licenses.[14,15]

In Panel B of Table II, we restrict the sample to the selected set of individuals in the comparison group *who obtained permanent licenses*. On average, it took them 48 days to obtain the licenses. These individuals overestimated what the bureaucratic process would entail: they thought, for example, that the process would take over 6.5 trips to the RTO. In practice, they only spent 3.5 hours (206 minutes) over 2.5 trips. They interacted with about 5 bureaucrats, and waited in 2.5 lines. Few of them (30%) took the required licensing exams at the RTO. Finally, the last row of Panel B shows that individuals in the comparison group on

14. This failure rate reflects a true inability to drive—as defined by the RTO—at the time of the test. As noted above, the test mirrors the RTO exam and checks for basic skills. Of course, these results do not immediately imply that incompetent drivers will be on the road, since we cannot measure investments in driving beyond the study. They do, however, imply that there is no effective regulation of who can drive. People will choose whatever level of driving skill is privately, not socially, optimal. This is especially important since everyone obtains a license for the purpose of driving. Driver's licenses are not used as a primary form of identification in India.

15. One may also ask, though, why individuals who do not know how to drive would go to the RTO to get a license. One explanation might be that it is easier or cheaper to learn how to drive with a permanent license in hand than without one. Learning with a temporary license may be more onerous because of the limited time validity of this license. For example, an unexpected work commitment may arise during the learning process that delays it and necessitates a reapplication for a temporary license. A permanent license (with unlimited validity) provides far more flexibility in timing the learning.

average paid 2.5 times the official fees to obtain their license: the average license getter paid about Rs 1,120, while official fees are only about Rs 450.

In summary, the experience of the comparison group shows distortions in the system, with many individuals obtaining licenses without being screened for driving ability and many paying well above official fees. However, this evidence does not tell us about the forces that generate these outcomes for the comparison group. Do these distortions result from bureaucrats sacrificing social benefits in order to cater to individuals' private willingness to pay? Do these distortions imply that this system does not respond to social considerations (e.g., ability to drive)? The experimental results shed light on these questions.

IV.A. Experimental Results

Our main experimental results are presented in Tables III and IV. Each column reports, for the dependent variable listed in that column, the coefficient estimates on dummy variables for bonus and lesson groups from a regression of the form

$$(1)\quad Outcome_i = \beta_0 + \beta_1 Bonus_i + \beta_2 Lesson_i + \beta_3 Session_i + \beta_4 X_i + e_i.$$

Indicator variables for the initial session the individual attended ($Session_i$) are included to absorb the unobserved heterogeneity in the procedural outcome across the initial sessions. This is important for two reasons. First, the IFC ended the study three months after the last initial session. Thus, individuals who attended the first session in July 2004 had more time to obtain licenses than those who attended the last session in November 2004. Second, because we recruited geographically for each session, all individuals at a given initial session were required to obtain licenses from the same RTO. Controlling for initial session fixed effects therefore also nets out any differences in procedures across RTOs. Demographic variables—age, marital status, religion fixed effects, a dummy variable for having driven a four-wheeler prior to the experiment, and a dummy variable for having driven a two-wheeler prior to the experiment—are used to control for differences in pre-experimental characteristics and differential attrition in the main sample (see Table I and Appendix II).[16] Robust standard errors are

16. The results do not differ significantly if we control for the additional socioeconomic variables from the introduction survey.

TABLE III
OBTAINING A LICENSE

	Obtained license (all tracked) (1)	Obtained license (2)	Obtained license in 32 days or less (3)	Obtained license without taking licensing exam (4)	Obtained license and did not have anyone teach them to drive (5)	Obtained license and attended a driving school (6)	Obtained license and automatically failed ind. exam (7)	Obtained license and exam score <50% (8)
Comp. group mean	0.45	0.48	0.15	0.34	0.23	0.03	0.29	0.32
Bonus group	0.24	0.25	0.42	0.13	0.29	0.03	0.18	0.22
	(0.05)***	(0.05)***	(0.04)***	(0.05)***	(0.04)***	(0.02)	(0.05)***	(0.05)***
Lesson group	0.12	0.15	−0.05	−0.03	−0.12	0.35	−0.22	−0.18
	(0.05)**	(0.05)***	(0.04)	(0.05)	(0.04)***	(0.03)***	(0.04)***	(0.05)***
N	731	666	666	666	666	666	666	666
R^2	0.12	0.14	0.31	0.12	0.26	0.26	0.24	0.20
Fstat	14.24	13.50	87.60	7.48	61.38	52.83	64.48	51.12
p-value	.00	.00	.00	.00	.00	.00	.00	.00

Notes:

1. This table reports on the subjects' ability to obtain a license and their driving ability, by experimental group.

2. Each column gives the results of an OLS regression of the dependent variable listed in that column on indicator variables for belonging to the bonus and lesson group. All regressions include session fixed effects, age, religion fixed effects, an indicator variable for marital status, an indicator variable for whether the individual had ever driven a two-wheeler prior to the project, and an indicator variable for whether the individual had ever driven a four-wheeler prior to the project. For ease of interpretation, the comparison group mean of the dependent variable is listed in the first row. The last two rows report the F-stat and p-value for a test of the joint significance of the bonus and lesson group indicator variables.

3. The sample in column (1) includes all individuals whose final license status was ascertained by the program staff. Columns (2)–(8) include all individuals whose final license status was ascertained and who completed all relevant surveys.

4. All standard errors are robust. Significance at the 10% level is represented by *, at the 5% level by **, and at the 1% level by ***.

reported in parentheses under each estimated coefficient. Below the coefficient estimates, we list the F-statistic and p-value for the joint significance of β_1 and β_2. For ease of interpretation, we also report the mean of the dependent variable for the comparison group in the first row of each column.

Table III focuses on experimental outcomes related to whether or not a given individual obtained a license; Table IV considers payment and process-related outcomes. For ease of exposition, within each table, we first discuss our findings regarding the bonus group and subsequently move to our findings regarding the lesson group.

IV.B. Obtaining a License: The Bonus Group

The first outcome we consider in Table III is whether or not a given individual was able to obtain a license. "Obtained license" is a dummy variable that equals 1 if a given individual obtained a permanent driver's license, and 0 otherwise. In column (1), the sample consists of the 731 individuals for whom we know whether or not they obtained a final license.[17] In column (2), we additionally drop the 65 individuals who indicated their final licensing status to the project team over the phone but refused to attend the final session to take the survey and driving exam. The sample in column (2) will be used for the analysis of all other experimental outcomes, as the only information we have about these 65 individuals is whether or not they obtained licenses. We obtain similar results in these two samples: individuals in the bonus group are about 25 percentage points more likely to obtain final licenses, a difference that is significant at the 1% level.[18] We also consider in column (3) a dummy variable that equals 1 if the individual

17. In the bonus group, the individuals we could not track were more likely to be students and to have known how to drive for a longer period of time (conditional on knowing how to drive), relative to the comparison group. In the lesson group, the individuals we could not track were more likely to be older, married, and employed and to know someone in the government, relative to the comparison group.

18. Since the bonus group has a lower attrition rate (4.4%) than the comparison group (13.4%), one wonders whether selective attrition by the comparison group could generate an apparent difference in success rates even if none existed. This would happen if the dropouts from the comparison group were disproportionately license getters. To quantify the magnitude of this concern, assume conservatively that the license-getting rate among those we cannot track in the comparison group is the same as the license getting rate among those we can track in the bonus group. Assume further that none of those we cannot track in the bonus group obtained licenses. This would imply a license getting rate of 48% in the comparison group, compared to a license getting rate of 65% in the bonus group. This suggests that the attrition is not quantitatively large enough to affect this result.

TABLE IV
PAYMENTS AND PROCESS

	Payment above official fees (1)	Tried to bribe (2)	Hired an agent (3)	Hired an agent and obtained license (4)	Payment to agent above official fees (5)	Obtained license and took more than three trips (6)
Comp. group mean	338.21	0.05	0.39	0.37	313.97	0.05
Bonus group	178.4	0.02	0.19	0.21	142.4	0.03
	(46.33)***	(0.02)	(0.05)***	(0.05)***	(45.54)***	(0.02)
Lesson group	−0.24	−0.02	−0.02	−0.02	−42.22	0.05
	(44.38)	(0.02)	(0.05)	(0.05)	(43.77)	(0.02)**
N	666	666	666	666	666	666
R^2	0.13	0.11	0.12	0.13	0.11	0.09
F-stat	12.06	2.53	14.07	16.45	11.98	2.11
p-value	.00	.08	.00	.00	.00	.12

Notes:

1. This table reports on the subjects' payments and process to obtain a license, by experimental group.

2. Each column gives the results of an OLS regression of the dependent variable listed in that column on indicator variables for belonging to the bonus and lesson group. All regressions include session fixed effects, age, religion fixed effects, an indicator variable for marital status, an indicator variable for whether the individual had ever driven a two-wheeler prior to the project, and an indicator variable for whether the individual had ever driven a four-wheeler prior to the project. For ease of interpretation, the comparison group mean of the dependent variable is listed in the first row. The last two rows report the F-stat and p-value for a test of the joint significance of the bonus and lesson group indicator variables.

3. The sample includes all individuals whose final license status was ascertained by the program staff and who completed all relevant surveys.

4. All standard errors are robust. Significance at the 10% level is represented by *, at the 5% level by **, and at the 1% level by ***.

was able to obtain his permanent license within 32 days of obtaining his temporary license, 0 otherwise. Individuals in the bonus group are 42 percentage points more likely to get their permanent licenses within 32 days or less. Hence, these first findings suggest that individuals who have a greater need to get a license quickly are able to achieve their objective.

Our next findings show that this increased propensity to get a license comes at a social cost: more bad drivers. The dependent variable in column (4) is a dummy variable that equals 1 if the individual obtained a driver's license without taking the legally required RTO driving exam, 0 otherwise. Increasing willingness to pay for a driver's license increases the number of people who obtain a license *without* taking the legally required RTO exam. Columns (5)–(8) of Table III show that this lack of testing is accompanied by an increase in the number of licensed drivers with poor driving skills. Individuals in the bonus group are 29 percentage points more likely to obtain licenses without having anyone teach them how to drive (column (5)) and are not more likely to have attended driving schools (column (6)). They are also much worse drivers than the comparison group: they are 18 percentage points more likely to be licensed drivers who automatically fail the independent driving test (column (7)); they are 22 percentage points more likely to be licensed drivers who score below average on the independent test (column (8)).[19] The interesting finding here is not that the marginal person trying to get a license is of low quality: it is that the bureaucracy allows them to get licenses despite their low quality. In this regard, it is useful to benchmark how bad the marginal drivers actually are. The failure rate on the independent exam is .60 (=.29/.48; see Table II) among the licensed drivers in the comparison group, while it is .75 (=.18/.25) among the marginally new licensed drivers in the bonus group.

In summary, the evidence reported so far in Table III suggests a bureaucratic system where a higher willingness to pay for a license translates not only into an increase in the number of license getters (a socially efficient component of the bureaucratic response) but also into an increase in the number of license getters who do not know how to drive (a socially inefficient component of the bureaucratic response).

19. The score is composed of the individuals' score on the 5 oral questions and on 23 aspects of driving. Thus, the highest possible score is 28.

IV.C. Obtaining a License: The Lesson Group

The motivation for including a "lesson treatment" in our experimental design is to test whether the bureaucrats are at all responsive to the main social consideration in the allocation of licenses: one's ability to drive. Under an extreme view of a corrupt bureaucracy, one might expect the allocation of licenses to be driven only by willingness to pay. This is not the case: randomly helping individuals acquire better driving skills increases the number of license getters among these individuals. Specifically, columns (1) and (2) show that individuals in the lesson group are between 12 and 15 percentage points more likely than the comparison group to obtain permanent licenses.[20]

These findings are, however, difficult to interpret, because we cannot rule out the possibility that offering free driving lessons to these individuals altered their willingness to pay for licenses. Trying harder to get a license could be a justification for the time spent learning how to drive; it could also be that having learned how to drive raises the private value of getting a license, since it can now be used. In support of these points, we found that individuals in the lesson group were about 12 percentage points more likely to "try" to obtain licenses than individuals in the comparison group.[21]

The remaining columns of Table III show that individuals in the lesson group are not more likely than individuals in the comparison group to obtain licenses without taking the exam (column (4)). Thus, while the lesson group has more license getters, it does not have more untested license getters. This suggests that models in which bureaucrats test a fixed fraction of license getters do not fit the data. The lesson group are also more likely to obtain their licenses while having had someone teach them how to drive (column (5)) and especially having attended a driving school (column (6)). These findings are, of course, unsurprising given the nature of the treatment for this group. More generally,

20. Selective attrition could theoretically explain this result if there were more license getters among the dropouts in the comparison group than among the dropouts in the lesson group. Assume that none among those we cannot track in the lesson group obtained licenses. Assume further that the license-getting rate among those we cannot track in the comparison group is the same as the license-getting rate among those we can track in the lesson group. This arguably conservative set of assumptions would (given respective attrition rates of 15.4% in the lesson group and 13.4% in the comparison group) only about equalize the license-getting rate (47%) in these two experimental groups.

21. In comparison, we found that individuals in the bonus group were about 19 percentage points more likely to "try" than individuals in the comparison group.

60% of the individuals in the lesson group who obtained licenses took the free driving lessons; also, conditional on take-up, they attended 12 classes on average. Columns (7) and (8) suggest that these classes did turn these individuals into better drivers.[22] For example, column (8) shows that individuals in the lesson group are 22 percentage points less likely to have obtained licenses and also automatically failed our independent driving test.[23]

In summary, giving a random subset of individuals access to driving lessons did raise their driving skills and also increased the likelihood that they obtained driver's licenses. While this is consistent with the view that bureaucrats do not completely ignore driving ability in the allocation of licenses, this conclusion is somewhat tempered by the fact that giving free access to driving lessons also raised individuals' likelihood of trying to get licenses.

IV.D. Payments and Process: The Bonus Group

Our findings so far show distortions in the application of this regulation, and that the magnitude of these distortions responds to the private willingness to pay for a license. This leads us to question whether bureaucrats receive bribes from misapplying the rules. In Table IV, we study a set of experimental outcomes related to licensing payments and to the process of obtaining a license.

The dependent variable in column (1) of Table IV is the amount paid by an individual *above* the official fees in the process of obtaining a license.[24] The mean of this variable in the comparison group is Rs 338, indicating that the comparison group already incurs substantial payments above the official fees. Column (1) shows that the bonus group makes more of these extralegal payments.

22. Could this be the result of "teaching to the test"? Could the lesson group not be better drivers but merely have been better taught how to take the driving test? The nature of the test, as noted before, makes this an unlikely possibility. Given that general skills are tested, the test likely provides a good approximation to what constitutes a good driver.

23. We also tested driving ability among the set of participants who had only obtained temporary licenses, but agreed to come back for a final survey. As expected, even in that group, driving ability was higher in the lesson group than in the control and bonus groups. Only 26% of the lesson group automatically failed the test, compared to 40% and 50% in the comparison and bonus groups, respectively.

24. Individuals were asked to break down their expenditures for the license. If an individual did not separate his official and unofficial costs, the formal fees of Rs 450 were subtracted from his fees. Note that information on informal fees paid was collected even if the individual did not obtain a license.

In columns (2)–(5), we study the exact nature of these extra payments. While our intuition *ex ante* was that these extra payments were direct bribes paid to bureaucrats, column (2) suggests otherwise. The dependent variable in column (2) is a dummy variable that equals 1 if an individual reported offering to bribe any bureaucrat or being asked for a bribe, 0 otherwise. First, one can see that the mean of this variable in the comparison group is low, with only 5% of individuals having tried to bribe or having been asked for a bribe; this implies that bribes to bureaucrats were only used by 11% of the license getters in the comparison group. More importantly, we do not find a significant (neither economically nor statistically) increase in the use of bribes in the bonus group.

What are these extra payments? Columns (3)–(5) show that most of these payments are payments to agents. Agents are professionals who, for a fee, help individuals through the process of obtaining various services.[25] While illegal, agents are a common institution in India.[26] We find that about 40% of individuals in the comparison group hired agents at some point in the process of getting licenses (column (3)). Nearly as many hired agents and also obtained licenses (column (4)), indicating that hiring an agent pretty much guarantees obtaining a license. The average payment to agents by individuals in the comparison group (Rs 313, column (5)) is about the same as the total average payment above official fees (Rs 338, column (1)); in other words, payments to agents are the bulk of the nonofficial fees paid in the process of getting a license. Individuals in the bonus group report being about 20 percentage points more likely to use an agent (columns (3) and (4)) and spend about Rs 142 more on agent fees (column (5)) than individuals in the comparison group; hence, most of the bonus group's additional payments are agent fees.

One conjecture that emerges from the bonus group's experiences is that agents are the channels of inefficient corruption in this bureaucratic system, and not simply the providers of standard "agency" services (such as standing in line for people). This

25. The existence of agents has been documented before. Rosenn (1984) describes the role of facilitators ("despachantes") in obtaining various public services in Brazil. Fisman, Moustakerski, and Wei (2005) find agents in the arena of international trade in Hong Kong.

26. From the introduction survey, we learned that agent usage is quite prevalent in the procurement of many government services in India. For example, of the 155 participants who obtained ration cards, 54% reported being helped by an agent. Similarly, 47% of the 47 individuals who obtained a land title, 15% of the 104 who obtained a passport, and 20% of the 58 who obtained a personal account number reported hiring an agent.

conjecture is based on the fact that a positive shock to the willingness to pay for a license increases both the number of people that pay for agents (Table IV) and the number of people that obtain licenses despite being unqualified to drive (Table III). However, further evidence will clearly be needed to strengthen this conjecture.

IV.E. Payments and Process: The Lesson Group

The findings in Table IV suggest that the lesson group does not differ much from the comparison group when it comes to average extralegal payments or reliance on agents. How much would we have expected the lesson group to pay? In a model where the extralegal payments are routine payments that have to be made by all license getters, one would have expected the lesson group, who get the license at a higher rate, to also pay more. The fact that the better drivers in the lesson group do not pay more suggests that informal payments are part of an alternative mechanism for acquiring a license, a mechanism that might be used more by those who are attempting to circumvent the driving test.

But the fact that many individuals in the lesson group *continue* to make extralegal payments (and hence use agents) is also intriguing. One possible interpretation is that not everyone in the lesson group knows how to drive. Another interpretation is that the agent route might be an attractive one even for able drivers, possibly because of the many hassles associated with getting a license without an agent. The last column of Table IV gives some credence to the second interpretation. We use as a dependent variable a dummy that equals 1 if an individual obtained a license but also had to make more than three trips in the process of getting that license. This variable may proxy for the hassle in getting a license in that needing more than three visits implies that the individual had to go back either to pick up additional documents or to take additional examinations. We find that individuals in the lesson group were more likely to make more than three trips to the RTO. In other words, it is possible that the formal route involves extralegal hurdles, so that even some of those who know how to drive may choose to hire agents. We return to this possibility in the next section.

V. The Process of Getting a License: Agents and Red Tape

Agents are key players in this bureaucratic process. In fact, more than 70% of the participants *who obtained a license* hired an

TABLE V
OUTCOMES FOR THE COMPARISON GROUP, BY AGENT USAGE

	Hired agent (1)	Did not hire agent (2)	*p*-value of difference in means (3)
	A. Procedures		
Days	46.21	54.44	0.32
No. of trips	2.33	3.19	0.00
No. officials spoken with	3.91	7.69	0.00
Lines	2.41	2.88	0.13
Total minutes spent	178.48	306.06	0.00
Took RTO licensing exam	0.12	0.94	0.00
	B. Expenditures		
Total expenditures	1,282.59	563.13	0.00
	C. Driving ability		
Automatic failure	0.69	0.31	0.01
Driving score	6.60	15.44	0.00

Note: Column (1) presents the mean for the 58 individuals in the comparison group who used an agent and obtained a license, while column (2) provides the mean for the 16 individuals in the comparison group who did not use an agent and obtained a license. Column (3) reports the *p*-value from the test of difference in means between the two groups.

agent. Our experimental results have shown that the greater usage of agents in the bonus group went hand in hand with a greater number of licenses being issued to individuals who had not taken the legally required driving exam at the RTO and did not pass the independent driving test. Based on these results, we conjectured that agents are not simply providing standard "agency" services or greasing the wheels of the bureaucracy but also are a channel for inefficient corruption, facilitating access to licenses among those who are unqualified to drive. Strengthening this conjecture requires further understanding of the role of agents and their relationship to the bureaucrats. This is what we do in the first part of this section, combining nonexperimental descriptive analyses and new experimental data from an audit study. In the second part, we investigate further the possibility that even good drivers may decide to hire agents because of the hurdles, or red tape, bureaucrats are imposing on individuals who attempt to complete the licensing process without an agent.

V.A. Agents: Nonexperimental Analysis

In Table V, we examine processes and outcomes for agent users versus nonagent users in the comparison group. Specifically,

we report the means of a set of variables for individuals in the comparison group who obtained licenses either with (column (1)) or without (column (2)) hiring agents. P-values from t-tests of the difference in means are reported in column (3).

Hiring an agent is associated with a much shorter process. Those who did not use agents spent on average 306 minutes at the RTO, took more than three trips to the RTO, and spoke with close to eight bureaucrats. Agent users spent 130 minutes less time at the RTO, took about one less trip, and spoke on average to only four bureaucrats.

Hiring an agent is also very strongly related to the level of testing at the RTO. While 94% of those who did not hire agents took the legally required RTO practical test at least once, only 12% of those who hired agents took that test. This is consistent with the hypothesis that hiring an agent is the main channel through which bad drivers can end up with licenses, but it is also theoretically possible that only the best drivers, for whom testing would be inessential, hire agents. This hypothesis is rejected in Panel C of Table V. Individuals who hire agents to get their licenses are about 38 percentage points more likely to fail the surprise driving test.

As we had already learned from our experimental results in Table IV, fees paid to agents are nearly the only source of excess payments in this bureaucratic process. Specifically, in Panel B, we compare the average expenditures to obtain a license for those who hired agents and those who did not. For those without agents, the total expenditures were Rs 563. In contrast, those hiring agents paid about Rs 1283, or Rs 720 more, to obtain their licenses.

In summary, this analysis suggests that the role of agents consists of more than simply "standing in line" for their clients. Instead, there is a strong correlation between using an agent and being able to skip the legally required driving exam; there is also a remarkably strong correlation between using an agent and unsafe drivers obtaining licenses.[27] This reinforces our experimental results in Tables III and IV. However, the evidence in Table V is purely correlational. In the next subsection, we move to some new

27. The New Delhi RTO illustrates the correlation between agents and ability to obtain a license. This RTO is situated near the main Federal Buildings. As such, the government has made a special attempt to remove agents from this area, and bureaucrats are more heavily monitored. We find a lower rate of agent usage, a lower rate of license getting, and a higher quality of driving skills among those who received their licenses at the New Delhi RTO. All results in this paper are robust to the exclusion of the New Delhi RTO.

experimental evidence that rules out a noncausal interpretation of these correlations.

V.B. Agents: Experimental Evidence

In January 2006, the IFC performed an audit study of agents involved in the provision of driver's licenses in Delhi. Trained actors were sent to agents under different scripted pretexts. The actor would record whether the agent said a license could be obtained under this pretext and, if so, at what price. The actors were college-aged Hindu men. They were of similar height and weight, and wore similar clothes. In total, six actors had 224 interactions with agents. Appendix III offers more details on the audit design.

Each day, the actors were randomly given one of six scripted pretexts. In the main script of interest, actors stated that they wanted to get a license but did not know how to drive and did not have the time to learn how to drive ("Cannot Drive" script). The five other scripts (in addition to the "Cannot Drive" script) were as follows. First, the actor had to learn what the agent could do for him if he had all the right paperwork and could drive (comparison group). We also focused on what would happen if the actors were missing either residential proof or age proof, two of the documents required to obtain a license. Another script focused on what would happen if the agent could not come back to the RTO to obtain a license. Finally, the last script focused on what would happen if the actor needed a license in less than 30 days, in other words, less than the officially required time between the temporary license and the final license.

After each visit, the actors were asked to fill out surveys describing their experiences with each agent. A series of questions on the work practices of the agents and their relationship with the RTO bureaucrats were also included in the survey. The actors were trained to bring up as many of these questions as possible in casual conversation with the agents (see Appendix III for details).

The results of the audit study are reported in Table VI. The dependent variable in columns (1) and (2) is a dummy variable that equals 1 if the agent says he can procure a license for the actor in a given interaction, and 0 otherwise. Column (1) corresponds to a single regression of this "agent can procure license" dummy on the various pretext dummies; reported in each cell is the estimated coefficient on the pretext in that row, with robust standard errors in parentheses. In column (2), we replicate the

TABLE VI
AUDIT STUDY

	Agent can procure license (Mean = 0.57)		Final price if agent can procure license (Mean = 1,586)	
Group	(1)	(2)	(3)	(4)
Constant	1	1.02	1,277.89	1,303.17
	(0.00)***	(0.04)***	(57.36)***	(83.21)***
Cannot drive	0	−0.01	62.65	110.54
	(0.00)	(0.02)	(81.66)	(85.76)
No residential proof	−0.5	−0.51	1,285.26	1,295.81
	(0.08)***	(0.08)***	(99.34)***	(102.30)***
No age proof	−0.21	−0.23	329	366.85
	(0.07)***	(0.07)***	(87.18)***	(90.96)***
Cannot come back	−0.95	−0.94	317.11	411.55
	(0.04)***	(0.04)***	(256.50)	(263.70)
Need license quick	−0.92	−0.91	855.44	850.51
	(0.05)***	(0.05)***	(212.03)***	(214.55)***
Actor fixed effects		X		X
N	226	226	128	128

Notes:
1. This table reports the audit study results. Each column presents the results of an OLS regression of the dependent variable listed in that column on indicator variables for each script in the audit study.
2. Standard errors are robust. Significance at the 10% level is represented by *, at the 5% level by **, and at the 1% level by ***.

regression in column (1) but further control for actor fixed effects, to net out possible differences across actors in their ability to obtain the service. Columns (3) and (4) follow the same structure as columns (1) and (2), respectively, but focus on the final price quoted by the agent if the agent was able to procure the service.

Several interesting findings emerge. To start, the prices quoted by the agents were of magnitude similar to that of those in the survey data discussed before (see Table V). Second, our finding regarding the "Cannot Drive" script confirms the relationship between agent usage and ability to get a license despite lacking driving skills. Agents saw no problem in helping actors who stated they did not know how to drive and did not have time to learn how to drive. One hundred percent of actors who approached agents with a "Cannot Drive" pretext were told that the agents could help them in getting their licenses. This confirms that the correlation between agent usage and poor driving ability observed in Table V does not simply reflect an omitted third factor. In addition, in cases where the actors managed to ask a few additional

questions of the agents in "casual conversation," the agents openly said that they could get the actors out of the formal driving exam at the RTO. Strikingly, the prices quoted under that script were not statistically different from those quoted to the comparison group.

The remaining rows of Table VI indicate that there are other services that agents can provide, even though these services also imply a deviation from the formal legal requirements. However, not all such services are as easy for the agents to provide as getting a license to someone who cannot drive. For example, only 50% of agents reported that they could procure a license if the actor lacked residential proof (row 3) and 80% if the actor lacked age proof (row 4). Also, in the cases of missing residential proof or age proof, the prices quoted by the agents conditional on being able to help were statistically significantly larger than in the comparison group. However, only 5% of agents could procure a license if the actor stated that he could not come back to hand in forms and take the picture at the RTO (row 5). Finally, only 9% of agents said that they could assist someone who needed a license in less than the official minimum time, and conditional on being able to assist, quoted a much higher price for rendering this service.

How can we explain these findings? Why is assisting someone in getting a driver's license despite his not knowing how to drive easier than assisting someone with some missing pieces of paperwork? One conjecture is that verifiability is an important determinant of which rules can be bent.[28] While it might be easy for the bureaucrat's superiors to crosscheck whether a valid proof of age and proof of residence were submitted by a license candidate and to monitor the dates on which these documents were submitted, it may be harder to cross-check whether the candidate took a road test and how well he did on it. In this view, the audit study suggests that the social inefficiency results would generalize most readily to other contexts where the socially useful part of the regulation is nonverifiable by the bureaucrats' principals. At the same time, the audit findings lead to many more questions. First, is it possible that even verifiable elements of a regulation could be overcome through collusion between the principals and

28. Reinikka and Svensson (2005) illustrate this in the context of Uganda, where a newspaper campaign aimed at reducing corruption in schools by providing parents with information to monitor local officials was highly successful.

the bureaucrats? While we do not have a direct measure of the extent of collusion between the bureaucrats and higher-up officials, the audit results suggest that there was not complete collusion in this particular setting. Second, would bureaucrats still ignore the nonverifiable, but socially useful parts of regulation if the costs to society of breaking the rules were much higher?

V.C. Red Tape

Even the better drivers in our study rely infrequently on the formal channel, which is associated with virtually no extralegal payments. What are the hurdles faced in this channel? The nonexperimental data provide some clues. In particular, our data allow us to examine bureaucrats' behavior when it comes to deciding whether someone has passed or failed the official driving test. Consider an individual entering the RTO and being asked to take the test. What affects the likelihood that this individual will succeed and be awarded a license? One clear determinant of success ought to be that individual's driving ability. However, bureaucrats may strategically manipulate the passing rule in order to extract higher bribe payments, for example, forcing more individuals to go through agents to obtain their licenses. At the extreme, bureaucrats may fail all test takers independent of how well they perform on the test. The fact that a fraction of the participants in our study did manage to obtain their licenses without hiring agents already indicates that such extreme behavior is not taking place. However, the bureaucrats may still be able to manipulate the passing rule in a way that might discourage even some of the good drivers from attempting to get their licenses without agents. This is the possibility we consider in Table VII.

In order to test this red tape hypothesis, we would ideally like to randomly send to the RTO individuals with better and worse driving ability and see how their driving ability affected their success in getting a license. Unfortunately, we do not have such a controlled experiment here and have to rely on descriptive evidence. The evidence in Table VII should, therefore, be interpreted with much more caution than the previous experimental findings in this paper.

We focus on individuals who begin the process without agents and take the driving exam at least once. For this set of individuals, we can define a "success" variable that equals 1 if the individual managed to obtain a license without hiring an agent and without taking the RTO exam twice. This roughly corresponds to

TABLE VII
RED TAPE

	Started without an agent and took exam at least once		Full sample of license getters	
	Success (1)	Used agent in the end (2)	Used agent at start (3)	Used agent in the end (4)
		A. By exam score		
Passed exam	0.62	0.24	0.29	0.61
	[98]	[98]	[219]	[219]
Failed exam	0.74	0.22	0.50	0.84
	[35]	[35]	[186]	[186]
		B. By group		
Comparison	0.65	0.25	0.35	0.78
	[20]	[20]	[76]	[76]
Bonus	0.64	0.27	0.52	0.80
	[46]	[45]	[187]	[187]
Lesson	0.66	0.22	0.22	0.58
	[68]	[68]	[144]	[144]

Notes:
1. This table studies possible red tape in the process of obtaining a driving license. Columns (1) and (2) include the sample of individuals who started without an agent and took the exam at least once. Columns (3) and (4) include the full sample of license getters.
2. "Success" in column (1) is defined as obtaining a license by passing the formal licensing exam, without hiring an agent.
3. Sample sizes are listed below each proportion in square brackets.

individuals who went to the RTO, took the test, and successfully got their licenses. Of course, our objective is to contrast performance on that test based on driving ability. We consider two approaches to identifying heterogeneity in driving ability. First, we can rely on the result of our independent driving test and contrast the mean of this "success" variable for individuals who automatically failed the independent exam and those who passed that exam (Panel A of Table VII). Alternatively, we can go back to our three experimental groups and compare mean "success" across groups, relying on the fact that individuals in the lesson group are better drivers due to the free lessons they were offered (Panel B).

"Success," as defined above, does not appear to vary systematically with driving ability (column (1)). In fact, we find a (statistically insignificant) higher success rate among those individuals we found to be unqualified to drive based on the independent test (74% compared to 62%). The same surprising patterns hold when we contrast success rates across the three experimental groups (Panel B).

With the caveat of a clearly selected sample, this evidence is consistent with the idea that bureaucrats may introduce additional randomness into the application process, or additional red tape, for individuals who plan to use the formal channel, may be to induce them to switch to agents. Interestingly, about 25% of those who started the process at the RTO by taking the driving test eventually resorted to hiring agents to obtain their licenses (column (2)). Similarly, statistics computed for the full sample of license getters also suggest that many of the license getters who used agents did not start the process with agents, but eventually switched to hiring them. Column (3) reports the fraction of license getters who used agents from the start, while column (4) reports the fraction of license getters who ended up using agents. Worse drivers ("failed exam" group; row 2) and drivers in a hurry (bonus group; row 4) are more likely to have used agents from the start. But interestingly, all drivers (good and bad) who start without agents are likely to end with them. For example, we find that while only about 35% of the individuals in the comparison group who obtained a license started the process with agents, 78% of these individuals used agents in the end.

VI. Interpretation

To summarize, there are two main tracks to procuring a driver's license in Delhi. The formal track involves directly applying through the RTO and no bribery. Some of our results, however, suggest that this track might be fraught with extralegal hurdles. The informal channel, on the other hand, is operated by agents, who account for nearly all the extralegal payments in our sample. These agents not only help to secure a license—which they do at nearly a 100% success rate—but also help to circumvent the testing requirement. Applicants with high willingness to pay get their licenses by paying fees to agents and not taking the driving test, resulting in unqualified (yet licensed) drivers. Better drivers are more likely to obtain their licenses through the formal channel, where they get tested but possibly also face extralegal hurdles. The result is a system that fails to regulate the quality of drivers and may force many individuals to make extralegal payments to acquire licenses.

While they reveal a clearly dysfunctional system, do our results imply bureaucratic corruption? One possible alternative

interpretation for these results is that the RTO is unable to test all drivers due to lack of resources and understaffing. It only tests sporadically and many people slip through the cracks; hence the high rates of bad drivers with licenses. At the same time, the understaffing leads to long lines, confusion, and complexity. This generates a demand for agents who provide legal time-saving services, such as waiting in lines and help navigating a confusing system.

While such an "overloaded bureaucrat" model with legal agency services could explain the sporadic testing, it struggles to explain the sharp difference in testing between agent users and nonusers. Specifically, if agents are simply offering time-saving devices, why does the audit study reveal that they can so easily bypass the RTO exam? And why do the survey data show such a strong relationship between agent usage and test-taking at the RTO?

This suggests that the dysfunctional system is not from lack of resources alone. Instead, some form of bureaucratic misbehavior is needed. There are two plausible forms of misbehavior. The first is what we call corruption, where the bureaucrats receive bribes (from agents) in order to both speed up the process, but also skip the test (or ignore the test results). The other form of misbehavior could be lack of effort. Instead of monetary benefits, some "lazy" bureaucrats could be enjoying nonmonetary private benefits by simply not making an effort to test individuals. In this world, agents have knowledge of who to approach at the RTO to both speed up the process and avoid testing (e.g., knowledge of who the rubber-stamping bureaucrats are).

These two explanations are clearly hard to disentangle without direct data on bribery. With this in mind, we attempted to collect more qualitative data from both bureaucrats and agents. First, and as already indicated above, actors involved in the audit study were instructed to engage whenever possible in casual conversations with the agents. When this happened, the agents openly discussed the need for bribing bureaucrats. Of the 208 actor-agent interactions where the actor was able to engage in casual conversation, the agents stated that they would need to pay bribes to the RTO in 81% of the cases. Second, IFC research assistants managed to informally interview three officials in Delhi and one in Chennai. The bureaucrats described weekly to biweekly meetings with agents. At these meetings, the agents pay a fixed fee for each of the agents' clients the bureaucrat granted a license

to. The bureaucrats also indicated that the fee does not vary much based on driving ability.

Beyond these qualitative interviews, our main finding in Table VII also raises doubts about a "lazy bureaucrat" interpretation. Once a person is being tested, the additional effort required to administer the test appropriately is minimal. The bureaucrat is already sitting in the car, and even a small amount of attention to the test-taker would allow far greater differentiation of good and bad drivers than we are finding in Table VII. Thus, while lack of effort could explain the low testing rates, it is harder to understand in this view why the testing that does take place is so poor.

Finally, the prices charged by agents can also be informative, since the agent sector appears quite competitive.[29] Their prices should therefore be somewhat commensurate with their input costs. Our data suggest that an agent saves about two hours of time for the applicants. Assuming agents' opportunity cost of time is about Rs 40 per hour, this would suggest that the marginal cost of assisting an individual in getting a license is only about Rs 80. This is an order of magnitude less than the average agent fee we observe in our data, which is about Rs 700.

As a whole, these qualitative and quantitative considerations lead us to favor a view in which at least some of the failures of this system are generated by corrupt bureaucrats working in collaboration with agents.

VII. Conclusions

Corruption in this study appears to undercut the very rationale for regulation: keeping bad drivers from getting licenses. Agents play a key role in the informal channel, as intermediaries between bureaucrats and applicants. The agent system allows bureaucrats to avoid direct bribery, and the bureaucrats may apply arbitrary failures on the driving exam to entice individuals to use agents. One interpretation of the audit results is that the verifiability of a particular regulatory requirement determines the ease with which corruption can overcome it. This suggests that the social inefficiency results would generalize most readily to other contexts where the socially

29. During the audit, we found at least six agents at each RTO to secure a price from, each vying for business.

useful part of the regulation is unverifiable by the bureaucrats' principals.

The study illustrates two main points for future research in the corruption literature. First, greater efforts to collect micro data are needed to penetrate the black box of corruption. Had we run a survey simply asking individuals who had obtained licenses whether they paid bribes, we might have concluded that there was no corruption in this bureaucratic system. Instead, the detailed questions on payments and the process of obtaining a license allowed us to isolate the central role agents play in this system. Second, this industrial organization of corruption (e.g., around the agent system) is intriguing and has been largely ignored by the theoretical literature. How do agents manage to develop their contacts with the bureaucrats? How do bureaucrats maintain their relationship with agents? Why is the provision of agents apparently so plentiful, rather than their numbers being restricted? Does the agent system limit the ability of the bureaucrat to more finely price discriminate between time-rushed and nonrushed individuals, as seems to be the case here? These are some of the questions we plan to explore in future work.

APPENDIX I: FINAL PROJECT SUMMARY, BY GROUP

	Total (1)	Comparison (2)	Bonus (3)	Lesson (4)
Individuals in initial session	822	202	295	325
Obtained permanent license, completed survey	409	74	189	146
Obtained permanent license, did not complete survey	17	5	3	9
Obtained temp license, completed final survey	23	4	1	18
Obtained temp license, did not complete final survey	48	15	11	22
Tried to get temp license, but failed	105	29	44	32
Did not try to get temp license	130	48	34	48
Unable to track	90	27	13	50

Notes:
1. This table reports the final project status for the 822 individuals present at the initial sessions. Column (1) presents the data for the full sample, while columns (2)–(4) present the data by experimental group.
2. "Trying" is defined as making at least one trip to the regional transport office after the initial session to speak with an agent or an RTO bureaucrat.

APPENDIX II: PATTERNS OF ATTRITION

Panel A	Age (1)	Married (2)	Student (3)	Employed (4)	Less than primary education (5)	Owns home (6)	Owns car (7)	Minority (8)	Hindu (9)	Muslim (10)	Log(salary) (11)
Attritor	−0.74	−0.14	−0.03	−0.01	0.05	−0.06	0.03	0.06	0.02	−0.01	0.13
	(0.67)	(0.05)***	(0.08)	(0.08)	(0.05)	(0.08)	(0.06)	(0.08)	(0.06)	(0.06)	(0.16)
Attritor*	−0.48	0.18	0.18	−0.08	−0.03	0.06	0.04	0.02	0.05	−0.06	−0.28
bonus group	(1.19)	(0.11)	(0.13)	(0.12)	(0.08)	(0.12)	(0.10)	(0.13)	(0.09)	(0.09)	(0.20)
Attritor*	1.87	0.31	−0.09	0.09	−0.02	0.06	−0.07	−0.06	0.12	−0.1	−0.17
lesson group	(1.05)*	(0.08)***	(0.10)	(0.10)	(0.06)	(0.10)	(0.07)	(0.10)	(0.07)	(0.07)	(0.19)

Panel B	Family member in government (including self)	Have a 2-wheeler license	Have driven a 2-wheeler	Have driven a 4-wheeler	Months known how to drive a 4-wheeler (given drive)	Would pay bribe if the fine is 500 and bribe is 300	Would pay bribe if the fine is 3000 and bribe is 300	Ever bribed	Ever used agent	Predicted trips	Predicted time
Attritor	0	0.01	0.12	0.02	0.48	−0.05	−0.01	0.12	0.07	−0.86	−220.85
	(0.09)	(0.03)	(0.04)***	(0.07)	(0.95)	(0.08)	(0.07)	(0.09)	(0.09)	(0.99)	(208.06)
Attritor*	−0.14	0.01	−0.11	0	−1.71	0	0.09	−0.1	−0.13	0.04	−72.85
bonus group	(0.13)	(0.05)	(0.06)*	(0.12)	(1.48)	(0.13)	(0.10)	(0.14)	(0.13)	(1.28)	(252.62)
Attritor*	−0.08	−0.03	−0.12	−0.03	−0.43	0.02	0.02	−0.14	−0.07	0.77	318.72
lesson group	(0.11)	(0.03)	(0.06)*	(0.09)	(1.29)	(0.10)	(0.09)	(0.11)	(0.11)	(1.54)	(281.77)

Notes:

1. This table reports on patterns of attrition. An attritor is defined as an individual whose final licensing status could not be ascertained by the project staff or who did not fill out the relevant surveys.

2. For each panel, a column gives the results of OLS regression of the dependent variable listed in the column on an indicator variable for an attritor, indicator variables for the bonus and lesson group, an indicator variable for belonging to the lesson group and being an attritor, and an indicator variable for belonging to the bonus group and being an attritor.

3. Standard errors are robust. Significance at the 10% level is represented by *, at the 5% level by **, and at the 1% level by ***.

APPENDIX III

The goal of the audit study is to understand whether the agents could obtain licenses under different pretexts, and if so, at what price. Six scripts based on the common barriers individuals face in obtaining a license were written:

Script number	Script
1. Comparison	I have residential proof and proof of age. I know how to drive.
2. Lack of residential proof	I want to get a license but lack residential proof. I am a college student in Delhi and live with friends.
3. Lack of age proof	I know how to drive, but I have no age proof.
4. Lack of ability to drive	I want a driver's license, but cannot learn driving now, as I am extremely busy with my studies.
5. Out of town	Today I will give you all the documents and money. Can you deliver the license to my home, as I cannot come again? Going out of town for some weeks.
6. Need a license fast	Need to get a license as soon as possible. How fast can you get it for me? How much would that cost? [*After the agent asks those questions, ask the following questions*] I need it *X* (answer they give) minus a few days (so you can say, "I need it in two weeks, or a week?). How much would that cost?" [*After the agent asks those questions, ask the following questions*] "What is the fastest you could get it to me? How much would that cost?"

Individuals were recruited through advertisements on a college notice board. Six men from one college were selected. Each was 18–19 years old and Hindu. All were of similar build and height and wore similar clothes.

Of the 9 RTOs in Delhi, eight were chosen for the audit study. The New Delhi RTO was not chosen, as agents were rarely available there. The audit study was conducted over eight days. The evening before the audit, the actors were told which RTO they would have to visit the next day, and which script they needed to use. The actors only visited each RTO once and were randomly assigned scripts and RTO visits in a round-robin fashion.

In total, 224 agents were approached by six different actors. The actors were trained to talk to the agents about their particular problems in obtaining a license and were asked to inquire whether it was possible to obtain a license and how much it cost. In the main experiment, the subjects reported bargaining with the agents on the price, and therefore, all the actors were trained to bargain with the subjects as well.

After visiting the RTO in the morning, all subjects reported back to the project manager to fill out the debriefing survey. The actors filled out one survey per agent to report whether the agent could or could not obtain the service, and, if so, at what price. If the agent could obtain the license despite the hardship, the actors also reported how the agent was able to do this. The actors were also told to ask the name of the agent in order to try to separate out the different pricing schedules of different agents. In 53% of the interactions, agents refused to reveal their names. We were able to identify 52 agents, but we were unable to determine whether some agents simply gave a different name to each actor.

To obtain additional qualitative data on agents and their interactions with bureaucrats, a series of questions on the work characteristics of agents and their relationship with the bureaucrats were included in the surveys. For example:

- How long have the agents worked at the RTO?
- Did they work at more than one RTO?
- Would the agent give a receipt?
- Did they have to bribe a bureaucrat or did the agent do it?
- Can the agent procure other services?

The actors were shown the debriefing survey prior to interacting with the agents, in order to understand what types of information were needed. In particular, the actors were trained on how to bring up these types of questions in casual conversation with the agent, and not to ask the questions if the agent already offered the needed information. Actors practiced these conversation skills with the project managers prior to their visits to the RTO.

UNIVERSITY OF CHICAGO GRADUATE SCHOOL OF BUSINESS, NATIONAL BUREAU OF ECONOMIC RESEARCH CENTER FOR ECONOMIC AND POLICY RESEARCH, AND INSTITUTE FOR THE STUDY OF LABOR
INTERNATIONAL FINANCE CORPORATION
NEW YORK UNIVERSITY WAGNER SCHOOL OF PUBLIC SERVICE
HARVARD UNIVERSITY AND NATIONAL BUREAU OF ECONOMIC RESEARCH

REFERENCES

Banerjee, Abhijit, "A Theory of Misgovernance," *Quarterly Journal of Economics*, 112 (1997), 1289–1332.
Di Tella, Rafael, and Ernesto Schargrodsky, "The Role of Wages and Auditing during a Crackdown on Corruption in the City of Buenos Aires," *Journal of Law and Economics*, 46 (2003), 269–292.
Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer, "The Regulation of Entry," *Quarterly Journal of Economics*, 117 (2002), 1–37.
Fisman, Ray, Peter Moustakerski, and Shang-jin We, "Off-Shoring Tax Evasion: Evidence from Hong Kong as Entrepôt Trader," Mimeo, Columbia University, New York, NY, 2005.
Fisman, Ray, and Shang-jin Wei, "Tax Rates and Tax Evasion: Evidence from 'Missing' Imports in China," *Journal of Political Economy*, 112 (2004), 471–496.
Huntington, Samuel, "Modernization and Corruption," in *Political Order in Changing Societies* (New Haven, CT: Yale University Press, 1968).
Klitgaard, Robert, "Gifts and Bribes," in *Strategy and Choice*, Richard Zeckhauser, ed. (Cambridge, MA: MIT Press, 1991).
Leff, Nathaniel, "Economic Development through Bureaucratic Corruption," *American Behavioral Scientist*, 8 (1964), 8–14.
Lui, Francis, "An Equilibrium Queuing Model of Bribery," *Journal of Political Economy*, 93 (1985), 760–781.
Myrdal, Gunnar, *Asian Drama* (New York, NY: Random House, 1968).
Olken, Benjamin, "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," Mimeo, Massachusetts Institute of Technology, Cambridge, MA, 2005.
Reinikka, Ritva, and Jakob Svensson, "Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda," *Journal of the European Economic Association*, 3 (2005), 259–267.
Rose-Ackerman, Susan, *Corruption: A Study in Political Economy* (New York, NY: Academic Press, 1978).
Rosenn, Keith, "Brazil's Legal Culture: The Jeito Revisited," *Florida International Law Journal*, 1 (1984), 1–43.
Shleifer, Andrei, and Robert Vishny, "Pervasive Shortages under Socialism," *Rand Journal of Economics*, 23 (1992), 237–246.
——, "Corruption," *Quarterly Journal of Economics*, 108 (1993), 599–617.

DO LENDERS FAVOR POLITICALLY CONNECTED FIRMS? RENT PROVISION IN AN EMERGING FINANCIAL MARKET*

ASIM IJAZ KHWAJA AND ATIF MIAN

Corruption by the politically connected is often blamed for economic ills, particularly in less developed economies. Using a loan-level data set of more than 90,000 firms that represents the universe of corporate lending in Pakistan between 1996 and 2002, we investigate rents to politically connected firms in banking. Classifying a firm as "political" if its director participates in an election, we examine the extent, nature, and economic costs of political rent provision. We find that political firms borrow 45 percent more and have 50 percent higher default rates. Such preferential treatment occurs exclusively in government banks—private banks provide no political favors. Using firm fixed effects and exploiting variation for the same firm across lenders or over time allows for cleaner identification of the political preference result. We also find that political rents increase with the strength of the firm's politician and whether he or his party is in power, and fall with the degree of electoral participation in his constituency. We provide direct evidence against alternative explanations such as socially motivated lending by government banks to politicians. The economy-wide costs of the rents identified are estimated to be 0.3 to 1.9 percent of GDP every year.

I. INTRODUCTION

Rent-seeking and corruption are thought to be pervasive around the world, and there is increasing recognition that they impose substantial economic costs. Yet, despite a rich theoretical literature,[1] there is limited empirical work in this area. While cross-country studies are useful in relating subjective measures of corruption to poor economic outcomes, they do not identify the

* We are extremely grateful to the State Bank of Pakistan (SBP) for providing the data used in this paper and clarifying many questions. The results in this paper do not necessarily represent the views of the SBP. Our special thanks to Marina Niessner and Nathan Blecharczyk for excellent research assistance. We also thank Abid Qamar, Tahir Andrabi, Daron Acemoglu, Alberto Alesina, Marianne Bertrand, Ali Cheema, Jishnu Das, Serdar Dinç, Rafael Di Tella, Mara Faccio, Raymond Fisman, Ishrat Hussain, Brian Jacob, Michael Kremer, Erzo Luttmer, Sendhil Mullainathan, Antoinette Schoar, Andrei Shleifer, Christopher Udry, Robert Barro, and Lawrence Katz (the editors), three anonymous referees, and participants at BREAD, the University of Colorado, the University of Chicago, Harvard University, the International Monetary Fund, the Massachusetts Institute of Technology, the National Bureau of Economic Research, and Stanford University for helpful comments and suggestions. All errors are our own.

1. For example, Krueger [1974], Rose-Ackerman [1978], Shleifer and Vishny [1993, 1994], Banerjee [1997], Bliss and Di Tella [1997], Ades and Di Tella [1999], and Acemoglu and Verdier [2000].

The Quarterly Journal of Economics, November 2005

presence of, or channels through which, corruption and rent provision occurs.

This paper uses a unique loan level data set from Pakistan to establish the presence of political rents in banking, identify the means of rent provision by focusing on the role of the public sector, and estimate the economy-wide costs this imposes. The scope and depth of the data used in this study provide several advantages. First, instead of relying on subjective proxies, we have direct measures of a firm's political connections, defined as the firm having a politician on its board. We can therefore test at the individual firm level if political status obtains preferential lending. Second, by using firm fixed-effects and hence only exploiting variation within the *same* firm over time or across lenders, we can account for unobserved firm-specific factors that do not vary over time or across lender types. This allows cleaner identification of the impact of political status on rent provision. Third, using measures of political strength and electoral participation, we can examine the extent to which rents are affected by the local political environment. Finally, given that we have the universe of corporate lending in the country, we can use our micro-level estimates to back out tentative economy-wide costs of political corruption.

Our results show that politically connected firms receive substantial preferential treatment. Not only do such firms receive 45 percent larger loans, but they also have 50 percent higher default rates on these loans. Moreover, this preferential treatment is entirely driven by loans from government banks. Private banks show no such political bias.

The preferential treatment to politically connected firms is not just a result of government banks selecting firms with worse default rates. Using firm fixed-effects and hence exploiting only variation within the *same* firm borrowing from *both* government and private banks, we find that government banks differentially favor politically connected firms by providing them greater access to credit. This preferential access is even higher for politically connected firms that are bigger and have a higher propensity to default.

We also find that the local political environment matters: firms with "stronger" politicians on their boards—as measured by votes obtained, electoral success of the politician, or political party—obtain even greater preferential access to credit from government banks. Also firms whose politicians run from con-

stituencies with greater voter turnout receive lower preferential treatment, hinting at checks imposed by electoral participation and political accountability.

The same politically connected firm also receives greater preferential treatment from government banks when either its politician or his political party wins. Taking advantage of the time dimension of our data, we use firm and quarter level fixed-effects to show that as a politician goes from losing to winning an election, the firm he is affiliated with receives (even) greater access to credit from government banks. We find a similar effect if the politician's political party wins the elections. Both winning or being in the winning party increase preferential treatment, suggesting that our findings indeed reflect the exercise of political power.

These results offer a particular mechanism of political rent seeking consistent with the institutional environment of Pakistan's banking and political system. Politically powerful firms obtain rents from government banks by exercising their political influence on bank employees. The more powerful and successful a politician is, the greater is his ability to influence government banks. This influence stems from the organizational design of government banks that enables politicians to threaten bank officers with transfers and removals, or reward them with appointments and promotions. Government banks survive such high levels of corruption because of the soft-budget constraints that often characterize state institutions [Kornai 1979, 1986].

We argue that our results provide evidence of political corruption and present evidence against alternative interpretations. One such alternative is "social lending" under which government banks lend to socially efficient but high risk projects, and firms with politicians on their boards undertake such socially efficient projects. While it is unlikely that social lending by the government will be carried out through loans to *private* firms (our results exclude loans to government firms), we nevertheless present direct evidence against the social lending view: when we distinguish between government banks that have an explicit social objective[2] versus those meant to run on pure financial profitability, the political preference results *only* appear within the latter "nonsocial" government banks. Social government banks,

2. Examples include banks set up for small and medium enterprises, women's welfare, and agricultural development.

while facing high overall defaults, display no political bias whatsoever. Similarly, the preferential treatment by government banks remains as strong when we examine firms located in a completely different state from their politician's constituency. Such distant firms are unlikely to generate legitimate social value for the politician's constituents.

Since our data form the universe of corporate lending in Pakistan, we can use our estimates to provide a sense of economy-wide costs imposed by these political rents. While there could be a variety of costs, we will only focus on the two for which we can provide estimates. First, as a lower bound, the defaulted amounts due to corrupt lending can be thought of as transfer payments from taxpayers. The deadweight loss from this is estimated between 0.15–0.30 percent of GDP each year. Second, there is an additional direct cost of such lending if the money is poorly invested or not invested at all. The evidence supports this as we find politically connected firms borrowing from government banks have relatively poor real output and productivity. Given the market to book value of investment in Pakistan, we estimate that an additional 1.6 percent of GDP is lost each year due to such investment distortions from corrupt lending.

Our paper broadly relates to the empirical literature on corruption and more specifically, to the role of political actors and state-owned institutions in earning and providing such rents in financial markets. Cross-country or cross-region studies such as Mauro [1995, 1997], Keefer and Knack [1996], Hall and Jones [1999], La Porta et al. [1999], and Glaeser and Saks [2004] study the impact of corruption on aggregate outcomes such as growth and investment rates. Sapienza [2003], Dinc [2004], and Cole [2004] exploit variation across countries or regions within a country and, like our paper, identify how political favors arise through government banks, either in the form of cheaper lending in politically preferred regions or increased lending in election years. Studies such as Fisman [2001], Johnson and Mitton [2003], and Faccio [2004] share our focus in identifying connections between politicians and individual firms and how these connections increase firm value.

Our study both complements and adds to these literatures. Since we link a firm to a politician and directly observe measures of preferential treatment to the firm, we can identify both the precise level, and specific manner in which rents are provided. Moreover, this level of disaggregation enables us to exploit varia-

tion for the same firm across lenders and over time, providing cleaner estimates of these rents. Our results also highlight the role of state institutions in providing political rents but, in addition, suggest that these rents may be checked by political competition and electoral participation. Finally, since we have the universe of corporate loans, we can provide suggestive estimates on some of the significant costs these rents impose on the economy.

The paper is organized as follows. Section II outlines the institutional environment with a focus on political rents. Section III describes the data and methodology. Sections IV–VI present the main results. Section VII provides evidence against alternative explanations such as social lending. Section VIII estimates the economy-wide costs of rent provision, and Section IX concludes.

II. Politics and Lending: The Institutional Environment

II.A. Politicians and Corruption in Pakistan

Politics in Pakistan has been closely linked to clientelism, rent-seeking, and corruption. These factors are often cited as the main problems facing the Pakistani economy. Transparency International, an international nongovernment organization that ranks countries on corruption based on survey data from businesses, has consistently ranked Pakistan very high on their corruption index.

Political events in Pakistan also show a repeated pattern of alleged political corruption leading to political instability. During the past decade and a half, no elected government has completed its five-year tenure, with four prime ministers and their assemblies dissolved by presidents or army generals on accusations of "maladministration, corruption, and nepotism." Pakistan is therefore a good candidate to study the nature and consequences of political corruption.

II.B. A Mechanism for Political Rents

How is political corruption carried out? The National Accountability Bureau (NAB), set up in 2000 with the purpose of prosecuting those involved in large-scale corruption, states that "in terms of the amount of corrupt money changing hands, taxation departments, *state-owned banks and development finance institutions,* power sector utilities, and civil works departments

probably account for the lion's share."[3] *The Guardian,* a British newspaper, reports on the link between politics, corruption and banking in Pakistan:

> Pakistan's state bank . . . moved to freeze the accounts of thousands of politicians . . . The move is seen as the start of a crackdown on the endemic corruption in Pakistan's political system . . . military officials have asked banks to provide lists of anyone who has defaulted on a loan from a state bank—a notorious way of amassing funds by politicians of all parties. (October 1999)

The above quote suggests that one of the means to obtain rents is through the banking sector, with politically connected firms "willfully defaulting on (government bank) loans that are accumulated with the intention of not being returned."[4]

Why are government banks more likely to be the source of political rents? First, they are simply the more dominant domestic player in the banking sector. While financial reforms in 1991 led to a sharp growth in the private sector, the role of the public sector has remained important, constituting 64 percent of domestic lending during our sample period. Second, soft budget constraints—a feature prevalent in government organizations all over the world [Kornai 1979, 1986]—lower the cost of capital for government banks and allow them to remain solvent despite high levels of default. Private banks face harder budget constraints making rent provision more difficult to sustain.

Finally, given the organizational structure of government banks, their lending decisions are particularly prone to political pressures.[5] The Banks Act of 1974 explicitly states that the top hierarchy of government banks—chairman, president, and board members—is to be appointed by the government. The same Act states that the board "determin(es) the credit . . . (and) personnel policies of the bank, including appointment and removal of officers and employees . . . (and) guidelines for entering into any compromise with borrowers and other customers of the bank." In the published words of the current governor of the central bank: "The recruitment, postings and transfers in all government ministries, departments and corporations are largely made either in ex-

3. Quoted from www.nab.gov.pk, June 17, 2004. Emphasis added.
4. National Accountability Bureau report on corruption, December 2000.
5. All banks, government and private alike, face the same regulatory environment which is in-line with international banking practices (Basel accord). Moreover, all banks have access to the same centralized credit information bureau (CIB) database that provides information on each borrower's credit history.

change of outright pecuniary favours or on purely political considerations . . . (with) functionaries who are always trying to please their bosses or political masters."[6]

Thus, the politically appointed top tier bank management not only influences the actions of bank officers through a system of rewards (promotions, sought-after assignments) and punishments (disciplinary action, transfers), but can also play a direct role in how, for example, defaulters are to be dealt with.

Politically connected firms are therefore likely have an advantage over others seeking rents, as they can use their political influence in lieu of monetary bribes which in turn may have larger private costs.[7] However, such political influence is not unbounded. For example, a loan officer may only be willing to expose a certain fraction of his portfolio to political pressures so as not to raise suspicion and inquiry. Similarly, prudential regulations prevent banks from overexposure to a single borrower. Perhaps more importantly, political favors and pressures may act like "gift exchanges," and politicians will be limited in how much and often they can call a friend for favors.

While the mechanism for political rents presented here is stylized, its broad patterns are likely to hold in Pakistan and other countries where state organizations face soft budget constraints and political actors exercise influence on such organizations. We make use of this mechanism to develop our empirical specifications and methodology and generate further testable implications.

III. Data and Methodology

III.A. Data

We use two primary data sets. The first has detailed loan level information for every corporate loan made in Pakistan from 1996 to 2002, while the second has electoral outcomes for the two elections that overlap the loan data period.

The loan-level data are unique both in terms of coverage and detail. They provide quarterly information on *the entire universe*

6. Dr. Ishrat Hussain, "Six Tentacles of Corruption," published in the *Dawn,* a Pakistani newspaper, on November 21, 1998.

7. Nonmonetary bribes are not the exclusive domain of politicians, and other actors such as the army and bureaucrats may also wield similar influence. While links to these actors are not the focus of this paper, their presence in the data only makes our estimates of political rents a *lower* bound of the true rents.

of corporate loans outstanding in Pakistan during a seven-year period from 1996–2002. The data are part of the Credit Information Bureau (CIB) database at the State Bank of Pakistan (SBP) which supervises and regulates all banking activity in the country. The CIB data provide each borrower's credit position by lender and quarter. Included are the amount of the loan outstanding by loan type (fixed, working capital, etc.), default amounts, and any litigation, write-offs or recoveries against the loans. In addition, we have information on the name, location, and directorship of the borrowing firms and lending banks allowing us to construct borrower and bank level attributes.

In terms of data quality, our personal examination of the collection and compilation procedures, as well as consistency checks on the data suggest that it is of very good quality. CIB was part of a large effort by the central bank to set up a reliable information sharing resource that all banks could access. Perhaps the most credible signal of data quality is the fact that all banks refer to information in CIB on a daily basis to verify the credit history of prospective borrowers. For example, we checked with one of the largest and most profitable private banks in Pakistan and found that they use CIB information about prospective borrowers explicitly in their internal credit scoring models. We also ran several internal consistency tests on the data such as aggregation checks, and found the data to be of high quality. As a random check, we also cross-validated the data with the portfolio of a particular branch of a bank.

Given that the loan data cover 1996–2002, there are two relevant national and state elections for this paper—general elections held in 1993 and 1997. We have information on the names and party affiliations for all candidates in these elections including the winner, the number of votes each received, and the total number of registered voters in each constituency. There were around 200 national and 450 state constituencies in each election, with 6–9 candidates per constituency and a total of over 8500 candidates in both election years.

III.B. Matching Politicians to Firms

The CIB data include names and addresses of all directors of a borrowing firm. Since almost all firms are private and closely held, firm directors are typically one of the main owners of the firm. We then use election data to identify firms that have a politician on their board of directors—henceforth referred to as

"politically connected" firms. A politician is defined as any individual who stood in the national or provincial elections. Later on we will also distinguish between whether the politician holds office or not.[8]

A politician is matched to a firm director, if their full (first, middle, and last) names match exactly. Given this literal matching of names, we can have both types of errors—(i) incorrect exclusion (Type I), and (ii) false inclusion (Type II). Type I errors arise when a firm is politically connected but our algorithm is unable to match this firm's directors to a name in the election database. For example, firms that are politically connected because their director is related to or has close links with a politician will not be matched. Type II errors occur when our algorithm matches a firm to a politician, but the match is incorrect.[9] Given that this explanatory variable is binary (i.e., a firm is politically connected or not), the classification error is not classical in that it is correlated with the true value and may not have 0 mean (i.e., we may undermatch more if firms are politically connected through indirect means). Nevertheless, one can show that this nonclassical measurement error still produces a lower estimate of the true effect [Aigner 1973].[10] Thus, given the measurement error in matching politicians to firms, our estimates of political corruption are likely to be *underestimates* of the true effect. One may also be concerned that our measure is correlated with attributes of the firm such as the number of directors a firm has since having more directors may increase the chances of matching. However, our results remain robust to including dummy variables for the number of directors in a firm and more generally, to including firm fixed effects.

8. We define a politician as someone who ran in an election since the institutional setting in Pakistan suggests that it is entry into the political network, and not just whether the individual won the election, that matters. Our subsequent empirical results also bear this out when we separately consider the impact of winning an election from just being a politician.

9. Type I match failures could also be due to different spellings of names (since the data are in English, there are often nonunique spellings of the names). Our algorithm tries to minimize this error by ignoring titles and allowing for common spelling variants. Similarly, as different people may share the same name, Type II errors are also possible. However, since we match on the politician's first, middle (whenever present), *and* last name before classifying a loan as political, such errors are minimized.

10. Suppose that political connectedness (P) is measured with error u ($P = P^* + u$, where P^* is the true classification and $u = -1, 0, 1$ is the error) which is uncorrelated with any controls and the error term in the true specification. Then one can show that $plim\ \hat{\beta}_{OLS} = \beta(1 - [cov(P,u|controls)/var(P|controls)]) < \beta$, where β is the true coefficient.

Since directors in our data almost always reflect one of the primary owners of the firm, politically connected firms should be interpreted as firms that are (partly) owned by a politician with ownership retained over time (i.e., we see little director turnover for a firm over time). As such, the question of when and which types of firms choose to select politicians on their boards is not as relevant in our context. Moreover, our empirical results will primarily use comparisons within a given firm (across different banks or over time), and we are therefore less concerned that our findings are driven by comparing across different types of firms.

III.C. Summary Statistics

Table I presents summary statistics for the variables of interest for the CIB loan database and the matched election data. Since we are interested in analyzing whether *domestic* lenders show preferential treatment to *private* politically connected firms, we exclude loans by foreign banks and loans to all government firms.[11] This leaves us with a panel of 68 private domestic and 23 government banks lending to 93,316 unique firms during the 25 quarters in our data period.[12] The loans are all corporate or business-related loans. While there are fewer government banks in the data, they constitute about 64 percent of overall lending.

As most of our tests exploit cross-sectional variation, we collapse the time component of our panel by "cross-sectionalizing" the data at the firm-bank-level. We do this to avoid issues of autocorrelation over time for a given loan and thus get conservative standard errors. Cross-sectionalizing the data involves converting all values into real 1995 rupees (Rs.) and then taking the time average of each loan, where a "loan" is identified by the borrowing firm and its corresponding bank. The cross-sectionalized data have 112,685 observations or loans. This number is greater than the number of unique firms (93,316) as some firms borrow from more than one bank.

11. Including foreign banks does not change our results as they behave similarly to private domestic banks, i.e., display no political bias. Including lending to government firms, which are backed by government guarantees, may confound the analysis since any preferential treatment they receive is unlikely to reflect private rents, and moreover, government banks may treat such firms differently due to their state ownership.

12. The data set is not a complete panel. The number of loans in any given quarter ranges from 22,361 in the beginning of the sample to 54,554 toward the end, reflecting an overall increase in lending.

TABLE I
SUMMARY STATISTICS

Panel A: Loan-level variables					
Variable	Mean	S.D.	Obs.		
Loan Size ('000s of 1995 Pak Rs.)	6,669	89,298	112,685		
Default Rate (%): Unweighted	16.85	30.22	112,685		
Default Rate (%): Loan size weighted	17.61	31.06	112,685		
Recovery Rate (%): (conditional on default)	8.55	24.50	24,562		
Rate of Return (%)	93.46	35.70	89,223		
Interest Rate (%)	14.05	2.90	89,223		
Loan Type	*Fixed*	*Working Capital*	*Letter of Credit*	*Guarantees*	*Other*
Percent of total lending	32%	49%	7%	7%	5%
Panel B: Borrower/firm attributes					
Politically Connected	*No*	*Yes*			
Percent of total firms	77%	23%			
Percent of total lending (of total loans)	63% (74%)	37% (26%)			
Size (percentile)	*0–50*	*50–75*	*75–95*	*95–99*	*99–100*
Percent of total lending (of total loans)	6% (42%)	3% (21%)	13% (23%)	23% (9%)	55% (5%)
Location (City Size)	*Small*	*Medium*	*Large*	*Unclassified*	
Percent of total lending (of total loans)	8% (17%)	12% (15%)	74% (52%)	6% (16%)	
Foreign Firm	*No*	*Yes*			
Percent of total lending (of total loans)	(99.8%)	4% (0.2%)			
Business Group Size	*Stand Alone*	*Intermediate*	*Conglomerate*	*Unclassified*	
Percent of total lending (of total loans)	20% (54%)	19% (17%)	39% (10%)	22% (19%)	

Panel C: Politician level variables for matched politicians (2073 politicians)		
Variable	Mean	S.D.
Win (%)	9.0	26.0
Percentage Votes	9.83	16.33
Victory Margin	20.53	16.50
Electoral Participation (%)	36.60	10.46

Rate of Return = (1 − Default Rate) * (1 + Interest Rate) + Default Rate * Recovery Rate. Politically Connected = dummy for whether firm has a politician on its board; Other firm level attributes defined in the Appendix; While we report summary statistics for firm location in terms of city size as defined in the Appendix, in the subsequent regressions firm location controls are introduced as separate dummies for each city. Win = politician winning frequency (%); Percentage Votes = percentage votes obtained by politician; Victory Margin = Difference in Percentage Votes between the winner and runner up if politician won, 0 otherwise; Electoral Participation = Registered votes cast (%).

Panel A of Table I gives summary statistics for the loan level variables. These include amount of loan outstanding, rate of default, and the fraction of loan recovered in case of default. Since these data show the *stock* of outstanding loans and defaulted amounts, they also reflect lending activity prior to our data period and, as such, our results, especially on default, should not be construed as driven solely by behavior in the mid-to-late 1990s but also in earlier periods. While we do not have interest rate at the loan level, we are able to proxy this using another data source that contains interest rate information at the bank-branch and loan size category level. For each bank branch we know the average interest rate charged on loans for 40 loan size categories. Using this procedure, a total of 7,518 bank-branch and size-category observations map into 89,223 loans. We cannot match all of the 112,685 loans since some bank branches do not report interest rate information. Using the information above, we can construct the rate of return on a given loan from the bank's perspective. This unit return (η_{ij}), representing earnings of the bank per rupee lent, is given by the following accounting identity:

$$\eta_{ij} \equiv (1 + r_{ij})(1 - \delta_{ij}) + \delta_{ij} * \rho_{ij}, \tag{1}$$

where r_{ij} is the time-averaged interest rate for a loan borrowed by firm i from bank j, δ_{ij} is the time-averaged default rate of the loan, and ρ_{ij} is the recovery rate for loans in case of default. The recovery rate is computed by aggregating all recoveries (against the defaulted principal and interest due) made by bank j from firm i until the end of our sample period.

Given the skewed loan size distribution, there might be a concern that the summary statistics are driven by economically insignificant small loans. For this reason, we also report default rate weighted by loan size. The mean loan size is Rs. 6.7 million, while the mean default rate is 16.9 percent. Notice that the mean net return is 93.46, which means that the average loan actually loses 6.54 cents on every rupee for the bank. As we shall see later, this is driven by the excessively large loan losses of government banks. Private banks have a net loan return of 109.8 percent in the data. Banks recover on average 8.6 percent of default. Panel A also shows the distribution of loans by the type of loan. A loan is classified into one of four different types: fixed (long term), working capital (short term), letter of credit, and guarantees.

Panel B gives various borrowing firm attributes. The main attribute is whether a firm is politically connected. The table shows that while 23 percent of firms are politically connected they receive 37 percent of overall lending. Panel B also presents other firm attributes which will be important to condition on when analyzing whether politically connected firms are treated differently. These variables are the size of a borrowing firm, its location, whether it is a foreign firm, whether it belongs to a business group and how many creditors it has. They are described in more detail in the Appendix.

Panel C uses the matched election data to construct various measures of a politician's strength. *Win* is the percentage of times a politician or his political party wins. *Percentage Votes* is the percentage of total votes a politician obtains, *Victory Margin* is the difference in percentage votes between the winner and runner-up in case the politician won (and 0 otherwise), and *Electoral Participation* is the percentage of registered votes cast in the politician's constituency. Since we have two elections and politicians can run in multiple constituencies, these measures are the average over a politician's individual measures in each election and constituency. We report these statistics for politicians that were matched to the CIB loan data.[13]

While we compare loans to politically connected versus those to unconnected firms in detail later, Table II presents some basic comparisons. Loans to the politically connected firms tend to be given in slightly smaller cities, and to firms that belong to larger business groups. While shorter term working capital loans are the most common type of loans, politically connected firms get greater fixed investment loans. Interestingly, there are sectoral differences in politically connected borrowing, with political loans more likely in sectors such as Textiles. Since these differences may reflect differences in underlying attributes of politically connected firms, we condition on them in our empirical specifications. These differences also hint at rent provision if longer term loans or loans in certain sectors are easier to default on. We will return to these issues toward the end.

13. These summary statistics are similar to those for unmatched politicians suggesting that our matching process did not introduce any selection effects.

TABLE II
POLITICAL LOAN CHARACTERISTICS

	Political loan	Nonpolitical loan	Difference
Average city size	1.28	1.45	−0.17
	(0.11)	(0.10)	(0.03)
Average group size	1.68	1.37	0.31
	(0.11)	(0.07)	(0.04)
Loan type share			
Fixed	37.72	28.80	8.92
Working capital	46.43	49.82	−3.39
Letter of credit	6.45	7.71	−1.26
Guarantees	6.59	7.27	−0.68
Other	2.81	6.40	−3.59

	Political loan		Nonpolitical loan	
	% of total lending	% of Industry type	% of total lending	% of Industry type
Industry Share				
Agriculture	1.4	27.2	2.3	76.0
Chemicals	5.1	53.1	2.6	46.9
Construction	8.3	49.1	5.0	50.9
Engineering/machinery	4.1	20.9	9.0	79.1
Food	11.7	42.8	9.1	57.2
Finance	3.8	23.4	7.3	76.6
Leather	0.5	33.0	0.5	67.0
Paper	2.0	47.4	1.3	52.6
Transport	0.8	19.9	1.9	80.1
Textile	36.6	54.1	18.1	45.9
Energy	1.5	55.8	0.7	44.2
Other	3.1	35.5	3.2	64.5

Firm level attributes (city and group size) are defined in the Appendix. Standard errors are reported in parentheses and clustered at the bank level. The industry shares are percentage of total classified loans and industries classified as follows: Agriculture—Agriculture; Chemical—Ceramics, Foam, Lab, Match, Mineral, Plastic, Rubber, Chemicals, Coating; Construction—Building Material, Construction Metal, Sizing, Storage; Engineering/Machinery—Appliances, Business Machinery, Electronics, Engineering, Fan, Finishing, Mill, IT, Instruments, Power, Telecommunication, Electric, Pump, Capital Goods; Finance—Export/Import, Finance; Leather—Leather; Paper—Books, Packaging, Paper, Photo, Wood, Packages, Printing; Transport—Air transportation, Auto, Aviation, Land transportation, Sea transportation, Tourism, Transportation; Textile—Textile; Energy—Energy, Gas, Petroleum; Other—Cycle, Education, Government, Jewelers, Light, Misc. Service, Medical, Military, Sport, Stationery, Watch, Shopping Mall, Advertising, Entertainment.

III.D. Methodology

The mechanism described in Section II suggests that politically connected firms obtain rents from banks in the form of preferential lending. We examine preference along two margins:

access to credit and the effective price of a loan. Credit access is measured by the amount a firm is able to borrow (logarithm of loan size), a substantial benefit in a credit-constrained economy.[14] The effective loan price is measured as the payments per rupee borrowed that a firm makes (the loan rate of return η_{ij}), as determined by the interest, default, and recovery rates on the loan.

The basic empirical specification employed to test for political preference uses the cross-sectionalized data. For firm i borrowing from bank j, we use OLS to estimate

$$(2) \qquad Y_{ij} = \alpha_j + \beta_1 \cdot Political_i + \gamma_1 \cdot \mathbf{X}_i + \gamma_2 \cdot \mathbf{X}_{ij} + \varepsilon_{ij},$$

where Y_{ij} is one of the measures of preferential treatment mentioned above, and $Political_i$ is an indicator variable for whether a firm is politically connected. $\mathbf{X}_i$ are firm level controls such as firm location, industry, and size, $\mathbf{X}_{ij}$ is a loan type (working capital, fixed investment) control, and α_j is a bank fixed effect. The controls $\mathbf{X}_i$, and $\mathbf{X}_{ij}$ are introduced nonparametrically: we include fixed effects for firm size (5 categories), the number of creditors the firm has (8 categories from 1 to greater than 7), a firm's group size (3 categories), city (134 cities) and industry (21 categories), and the loan type (5 categories). This results in a total of 268 dummy variables (including the 91 bank dummies). β_1 in (2) is our coefficient of interest that captures the preferential treatment a politically connected firm receives, and henceforth will be referred to as the "political preference" effect.

As our unit of analysis is a loan (i.e., firm-bank pair) there may be a concern that the results are driven by the majority of loans which are small in size. Since we are interested in economically significant differences, all regressions (except where loan size is the dependent variable) are weighted by loan size. For example, when default rate is the dependent variable, we can interpret β_1 as the additional default by politicians *per dollar* of borrowed amount. Standard errors are clustered at the bank level.

While (2) includes an extensive set of firm-attribute fixed effects and bank fixed effects, a remaining identification concern

14. Another measure of credit access is whether a firm that applied for a loan received one. Since our data only include firms that receive loans, we can only measure access in terms of how much a firm is lent to, conditional on receiving a loan.

is that β_1 may still be a biased estimate of political preference due to omitted firm level variables correlated with a firm's political status that affect the loan amount or price; i.e., $Political_i$ is correlated with unobserved firm attributes in the error term (γ_i, where $\varepsilon_{ij} = \gamma_i + \nu_{ij}$). For example, more "influential" firms may attract politicians as board members and also use their influence to obtain preferential lending. To the extent that we cannot observe and control for firm influence in (2), β_1 will be an overestimate of the political preference effect.

Given these concerns, a more convincing estimation strategy would be to include firm fixed effects in (2) to account for all time-invariant firm attributes that have a similar (level) effect on a firm's borrowing from all banks; i.e., the firm fixed effect absorbs firm-specific unobservables (γ_i) that enter additively in (2). While including firm fixed effects is not possible in (2) because the fixed effect absorbs our attribute of interest—whether a firm is politically connected or not—there are two ways we can proceed. The first is to define a time-varying measure of political connectedness and use the panel form of our data to exploit variation over time for a given firm. While we will use and describe this approach later, the political rent mechanism outlined earlier suggests that another promising direction, which allows us to retain our original measure of political connectedness, is to exploit differences across lenders, particularly private versus government banks, for a given firm.

We use the following specification to test whether the *same* firm receives (greater) preferential treatment if it is politically connected when it borrows from a government compared with a private bank:

$$(3)\quad Y_{ij} = \alpha_i + \alpha_j + \beta_1 \cdot Political_i * GOV_j + \gamma_1 \cdot \mathbf{X}_{ij} + \gamma_2 \cdot \mathbf{X}_{ij} * GOV_j + \varepsilon_{ij},$$

where in addition to the variables in (2), α_i is a firm fixed effect and GOV_j is an indicator variable for whether the lender is a government bank or not. Our coefficient of interest, β_1, is the "differences-in-differences" estimate of political preference. β_1 captures the extent to which a politically connected firm receives preferential lending from a government bank as compared with a

private bank.[15] In running specification (3), we restrict the data to firms that borrow from both types of banks.[16] The difference-in-difference estimate provides cleaner estimates of the political preference effect and removes the identification concerns mentioned above. The inclusion of both bank and firm fixed effects ensures that our results are not driven by level differences that may arise when comparing across different banks or different firms. For example, bank characteristics such as government banks making larger loans than private banks are captured by the bank fixed effects. Similarly, firm attributes such as political firms having greater loan demand, or different risk classes are subsumed in the firm fixed effects. However, firm fixed effects are not able to eliminate biases that may arise from firm level unobservables that vary over time or across lenders.

In addition to estimating (3), we also run related specifications where we examine whether the relative political preference displayed by government banks differs across different types of firms where firm type is measured by characteristics such as its political strength. Such effects will be introduced as triple interaction terms in (3); i.e., the $Political_i * GOV_j$ term will be interacted with these firm-specific attributes.

Finally, as mentioned above, another strategy to exploit differences within the same firm is to use a time-varying measure of political connectedness and then introduce firm fixed effects in the panel version of the data. We do so by considering changes a firm experiences when its politician or politician's political party wins or loses an election. We use the following specification in the subset of politically connected firms that experience such a change:[17]

15. When we examine preferential treatment in terms of loan size, we aggregate our observations at the firm *X bank-type* level. In particular, we aggregate to firm i and *bank-type* $\bar{j}$ (government or private) since we want to compare how much (more) a politically connected firm is able to borrow from *all* government banks compared with *all* private banks. Therefore, instead of (3) we run

$$\text{Log}(Loan\ Size_{i\bar{j}}) = \alpha_i + \beta_1 \cdot POL_i * GOV_{\bar{j}} + \beta_2 GOV_{\bar{j}} + \varepsilon_{i\bar{j}},$$

where $\bar{j}$ is the bank type index (either government or private bank).

16. We restrict to firms that borrow at least 1 percent of their lending from each type of bank. Firms that borrow from a single bank-type are not included as they do not directly affect our coefficient of interest, β_1.

17. As before, with loan size as the dependent variable for each firm in a given quarter, we aggregate the data at bank-*type* (government or private) level:

$$\text{Log}(Loan\ Size_{i\bar{j}t}) = \alpha_{i\bar{j}} + \beta_1 \cdot WIN_{it} * GOV_{\bar{j}} + \beta_2 \cdot WIN_{it} + \varepsilon_{i\bar{j}}.$$

$$(4) \quad Y_{ijt} = \alpha_{ij} + \alpha_t + \beta_1 \cdot WIN_{it} * GOV_j + \beta_2 \cdot WIN_{it} + \varepsilon_{ijt},$$

where the variables are as before and the additional subscript t specifies the quarter. α_{ij} are bank-lender (i.e., loan-level) fixed effects; WIN_{it} is an indicator for whether the firm's politician holds office during quarter t or not. When we examine changes in electoral success for the politician's political party, we use a similar indicator for whether the politician's political party wins or not, $WIN\text{-}Party_{it}$. The double-difference estimate B_1, captures any (additional) lending preference a politically connected firm receives from a government relative to private bank, when its politician or his political party wins. The bank-lender fixed effects imply that this change is for the *same* loan (i.e., firm-bank pair) over time.

IV. Results—Preferential Treatment for Politically Connected Firms

Table III shows the results of estimating (2) for both margins of preference: loan access and price. The regressions nonparametrically control for firm and loan characteristics by introducing firm attribute, bank and loan type dummies.

Column (1) presents evidence for political preference in terms of credit usage: loans to politically connected firms are 45

TABLE III
Are Politically Connected Firms Given Preferential Treatment?

Dependent variable	Log loan size (1)	Rate of return (2)	Default rate (3)	Recovery rate (4)	Interest rate (5)
Politically connected	0.37	−6.08	6.22	−1.09	0.09
	(0.08)	(2.46)	(1.98)	(1.14)	(0.05)
Controls	YES	YES	YES	YES	YES
R^2	0.26	0.28	0.29	0.24	0.43
No. of Obs.	112,685	89,223	112,685	24,562	89,223

Results are based on cross-sectionalized data. A unit of observation is a loan (bank-firm pair). There are 89,223 observations instead of 112,685 in columns (2) and (5) as interest rate data are not available for all banks. There are 24,562 observations in column (4) because the data are conditional on a firm having defaulted. Rate of Return = (1 − Default Rate) * (1 + Interest Rate) + Default Rate * Recovery Rate. Standard errors reported in parentheses are clustered at bank level. Regressions in columns (2)–(5) are weighted by loan size. Controls in column (1) include dummy for whether borrower is a foreign firm, 91 bank dummies, 134 dummies for each of the city/town of firm. Columns (2)–(5) include column (1) controls plus 8 dummies for the number of creditors the firm has, 5 loan-type dummies and 3 group size dummies, 5 firm size dummies. Firm-level control variables are described in the Appendix.

percent as large as those to unconnected firms. (difference in logs is 0.37). Concerns that this result is biased due to unobserved firm heterogeneity are lessened by the inclusion of firm level controls. Moreover, this will be addressed further in subsequent specifications that allow the inclusion of firm fixed effects.

Columns (2)–(5) show that in addition to better access, politically connected firms also face significantly lower "prices" on their loans: column (2) shows the rate of return on political loans is 6 percentage points lower and is robust to the inclusion of bank fixed effects and firm attribute fixed effects. The difference is both statistically and economically significant.

A breakdown of loan rates of return into its three components specified in (1) in columns (3)–(5) shows that preferential treatment is driven primarily by the higher default rates that the politically connected firms enjoy. Politically connected firms default 6.2 percentage points more than unconnected ones.[18] On a base default rate of 14.8 percent, this implies that the politically connected default 42 percent more. In contrast to default rates, columns (4) and (5) show little difference between politically connected and unconnected firms in the recovery rates on defaulted loans and the interest rates charged.

How do rent-seekers avoid recovery on collateralized loans? The Pakistani setting suggests a couple of answers. First, litigation is a long drawn-out process. Recovering default is not an easy task even for government banks, especially if courts are also subject to political influence. Second, anecdotal evidence suggests that collateral is often overvalued. A common way to create overvalued collateral is through overinvoicing by importing defunct machinery at inflated prices. The political borrower's influence ensures that such overvalued collateral is accepted. Thus, when the firm does default a few years later, preventing recovery or seizure of capital is of little concern.

The results in Table III suggest that politically connected firms receive preferential treatment on two accounts: they are able to borrow larger amounts, and their default rates are higher. For the remainder of the paper we will focus on both these margins of preferential treatment, i.e., receiving larger loans and defaulting more on each rupee lent. For the latter margin we use

18. As we will see later on, since larger political loans are even more likely to default, the unweighted difference in default between political and nonpolitical loans is lower at 3.3 percent (but still significant at the 1 percent level).

default rate instead of the loan return measure because the differences in loan return are entirely driven by differences in default rates and the loan return measure uses interest rate data that are not available for the full data.

We interpret the existence of the political preference effect as evidence of corruption in the form of rents provided to the politically connected. However, the specification presented so far raises plausible concerns regarding both the empirical identification of political preference and in interpreting it as evidence of corruption. In the following sections we present evidence that improves identification and supports our interpretation.

V. Results—Political Rents and Government Banks

Since government banks are more susceptible to political coercion due to their organizational design, we expect them to provide greater rents to politically connected firms. We examine whether this is the case for the two measures of preferential treatment, default rate, and access to credit.

V.A. Default Rate

Columns (1) through (5) in Table IV show that the higher default rates that politically connected firms enjoy arise *entirely* due to loans from government banks. Columns (1)–(2) first run the original specification (2) by restricting the data to loans from government banks only and show that loans to the politically connected firms have 11 percentage points higher default rates. This result remains robust to all of the controls mentioned earlier.

Columns (3)–(4) repeat the same exercise for loans from private banks only. There is hardly any difference in default rates between the politically connected and unconnected firms in private bank loans. Including bank and firm attribute fixed effects (column (4)), shows that politically connected firms have 0.8 percentage points lower default rates on private bank loans.

Column (5) runs specification (3) but with firm attributes controls instead of firm fixed effects and shows the same result. The coefficient of interest is the double interaction term (β_1) that shows politically connected firms default 9.9 percentage points more than the unconnected in loans from government banks relative to loans from private banks. The small negative coefficient on the dummy for political firm shows that if anything,

TABLE IV
ARE POLITICALLY CONNECTED FIRMS FAVORED BY GOVERNMENT BANKS ONLY?
DEFAULT RATE

	Default rate (%)					
	(1)	(2)	(3)	(4)	(5)	(6)
	Government banks only		Private banks only		All banks	Firms borrowing from both government and private banks
Politically connected	10.92	9.13	−0.02	−0.78	−0.78	—
	(4.12)	(1.92)	(0.27)	(0.26)	(0.26)	
Politically connected * government bank					9.91	1.4
					(1.90)	(1.04)
Constant	19.87	—	6.05	—	—	—
	(2.60)		(2.03)			
Controls	NO	YES	NO	YES	YES[a]	Firm fixed effects[b]
R^2	0.02	0.3	0.004	0.15	0.33	0.78
No. of Obs.	61,897	61,897	50,788	50,788	112,685	18,819

Results are based on cross-sectionalized data. Standard errors reported in parentheses are clustered at the bank level. Politically connected = dummy for whether firm has a politician on its board; Government bank = dummy for government banks. Controls include 5 loan-type dummies, 5 firm size dummies, dummy for whether the borrower is a foreign firm, 8 dummies for the number of creditors the firm has, 3 group size dummies, 134 dummies for each of the city/town of borrower, 21 dummies for the industry of the firm, and 91 bank dummies. Firm-level control variables are described in the Appendix.

a. Controls also include government bank dummy and all interactions with the government bank dummy.

b. Regression includes a government bank dummy as well. Data are restricted to firms that borrow from both government and private banks.

politically connected firms have slightly lower defaults suggesting either greater monitoring or better selection for politically connected firms by private banks.

An interesting aside is that while the government banks do treat politically connected firms more favorably, they also face high default rates in general (column (1)). By focusing on political connectedness, we are only capturing one source of "influence." There may be a variety of other avenues such as alternative forms of status (bureaucracy, army, insider networks, familial ties, etc.) and direct bribes that may also contribute to why government banks face higher default rates. In this paper our focus is only on political rents.

Do government banks face higher default rates because they

select worse borrower types—where type is proxied by average default rates—or because they lend greater amounts to the worse types? We will consider the first selection margin here—of choosing whether to lend to a firm—and examine the second margin when we consider credit access.

Note first that if, as one would expect, loans from government and private banks have equal seniority, it is unlikely that a firm will be able to default on one but not on the other. This suggests that the higher default *rate* faced by government banks is because they exclusively deal with worse borrowers, and not that a given firm which borrows from both bank types, defaults more on its government bank loan.

We can check for such selection by including firm fixed effects as in specification (3) and restricting the data to firms that borrow from both types of banks. The firm fixed effect enables us to ask whether the *same* politically connected firm defaults at a higher *rate* on its government versus private bank loan compared with a nonpolitical firm. Column (6) shows that this is not the case, since the default differential reduces to a much smaller and not significant 1.4 percentage points. This decrease is not due to the data restriction since the default differential in this restricted sample is 9 percent without firm fixed effects (regression not shown), similar to that in column (5). It drops only after we have accounted for all selection effects through firm fixed effects. This is not surprising given the cross-default legal stipulations that make it unlikely that a firm can default on one bank and not another when loans have the same seniority.

The mechanism outlined in Section II implies that borrowers are likely to self-select across banks with (the worst) borrowers that have no productive investments but wield (political) influence only borrowing from government banks. Our results also support this. Comparing average default rates for firms that (i) borrow only from government banks, (ii) borrow from both bank types, and (iii) borrow only from private banks, shows that the first have the highest average default rates (25.7 *percent*), followed by the second (16.9 *percent*), and then the last category has the lowest default rates (5.4 *percent*).

V.B. Access to Credit

We next test if the other margin of political preference, access to credit, is also only due to government bank lending. An important concern when comparing credit access for political versus

nonpolitical firms is that the amount borrowed may differ simply due to a firm's different credit needs (a demand effect). In other words, the "preferential treatment" in access to credit identified in Table III earlier, may simply reflect a higher credit demand of political firms and not political preference. To argue that there is political preference, one needs to perfectly condition on a firm's credit demand. The hypothetical comparison would then be between two firms with the *same* credit demand and seeing whether the politically connected firm receives a larger loan from the government bank. Specification (3) allows us to make such a comparison.

Column (1) in Table V shows that while government banks provide larger loans than private banks, they lend even larger amounts—29 percent more—to politically connected firms. The

TABLE V
ARE POLITICAL FIRMS FAVORED BY GOVERNMENT BANKS ONLY?
ACCESS TO CREDIT

	Log loan size		
	(1)	(2)	(3)
Dependent variable	Data restricted to firms that borrow from both government and private banks		
Government bank	0.07	−1.19	−0.2
	(0.03)	(0.14)	(0.03)
Politically connected * government bank	0.29	−0.21	0.13
	(0.05)	(0.22)	(0.05)
Government bank * log firm size		0.14	
		(0.02)	
Politically connected * government bank * log firm size		0.041	
		(0.03)	
Government bank * firm default rate			1.9
			(0.11)
Politically connected * government bank * firm default rate			0.56
			(0.17)
Firm fixed effect	YES	YES	YES
R^2	0.81	0.81	0.83
No. of obs.	10,880	10,880	10,880

Data are restricted to firms that borrow from both government and private banks. Robust standard errors are reported in parentheses. A unit of observation is a firm-bank type (government or private) pair, as all loans of a firm given by the same bank type are summed. There are thus 5,440 firm fixed effects and 10,880 total observations in the regression. Politically Connected = dummy for whether firm has a politician on its board; Government bank = dummy for government banks; Log firm size = Logarithm of a firm's total borrowing from all banks (private and government); Firm default rate = Firm's average default rate across all banks.

use of firm fixed effects strengthens our causal interpretation that the political preference observed is a result of differential treatment and not (level) differences across firms. Moreover, as this preferential treatment stems from government banks, it supports our contention that it arises through the exercise of political power.

We showed above that government banks exclusively lend to the worst type of borrowers in terms of average default rates. Do government banks also perform poorly along the second selection margin, i.e., conditional on choosing to lend to a firm, do they lend greater amounts to the worst firms?

Columns (2)–(3) in Table V check for further selection effects by asking whether certain types of politically connected firms are given greater access to credit. Column (2) (weakly) suggests that government banks lend more to the larger of the politically connected firms. A standard deviation increase in firm size as measured by the logarithm of the total amount it borrows, is associated with 8 percent greater amount that the politically connected borrow from government as compared with private banks. More tellingly, column (3) shows that government banks systematically lend greater amounts to the worst (highest average default rates) of the politically connected firms. The coefficient on the triple interaction term shows that government banks (as compared with private banks) lend 56 percent larger amounts to those politically connected firms that go into default. Finally, one may be worried that by time-averaging each loan, we are no longer guaranteed that a firm is borrowing from private and government banks at the same time. To check for this concern, we also reran the cross-sectional tests of Table V separately for each quarter and found our results to be stable and significant in each quarter.

Tables IV and V paint a stark picture of the political rent-seeking environment and the role of the public sector. It is an environment characterized by politically connected firms that receive greater access to credit and default more, not (only) because they face adverse business shocks but because they *can* default. The worst of such politically connected firms—those that default a lot—exclusively borrow from government banks. Moreover, even after accounting for this poor initial selection, we find that government banks provide greater rents by lending more to the larger politically connected firms and to the worst (in terms of default) of such firms.

VI. Results—Political Strength and Participation

Do political rents vary by the strength of the firm's politician, whether he holds office, and the degree of political participation in the politician's constituency? The mechanism outlined in Section II would suggest so provided that a politician's ability to influence government banks varies by political strength. While we can examine political preference on both margins, greater access to loans, and higher default rates, we found no robust differences in default *rates* and will focus on the margin that does matter, preferential access to credit.

VI.A. Political Strength

Do firms with stronger politicians obtain even greater access to credit from government banks? We use different measures of a politician's strength. These include (i) the percentage of total votes a politician wins; (ii) the fraction of times a politician wins; (iii) the politician's victory margin; and (iv) the fraction of times the politician's political party wins. We aggregate the data to the bank-type and firm level and restrict to firms borrowing from both bank types.

Columns (1)–(3) in Table VI present the results for each of these variables with the logarithm of loan received as the dependent variable.[19] The coefficient of interest is the triple interaction term that reveals whether firms with stronger politicians are able to earn even higher rents from government banks. Table VI shows that along all measures of a politician's strength, firms with stronger politicians borrow even more from government banks.

Column (1) shows that while all politically connected firms are able to borrow more from government banks, a 10 percentage point increase in the number of votes a politician obtains is associated with a further increase of 7 percent in the amount his firm is able to borrow from the government. Columns (2)–(3) similarly show that a 10 percentage point increase in the fraction of times a politician wins and in his victory margin are associated with his firm borrowing 6 and 5 percent more from government

19. Note that since the political strength measures are only defined for politically connected firms ($Political\ Strength_i * Political_i \equiv Political\ Strength_i$) all possible interaction terms are included in these regressions; i.e., they are either subsumed in the firm-fixed effect or the triple interaction term.

TABLE VI
TESTING FOR POLITICAL STRENGTH AND PARTICIPATION

	Log loan size				
	(1)	(2)	(3)	(4)	(5)
Dependent variable	Data restricted to firms that borrow from both government and private banks				
Government bank	0.07	0.07	0.07	0.07	0.07
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Politically connected * government bank	0.25	0.26	0.25	0.23	0.67
	(0.06)	(0.05)	(0.05)	(0.05)	(0.20)
Politically connected * government bank * percentage votes	0.69				
	(0.47)				
Politically connected * government bank * win		0.63			
		(0.32)			
Politically connected * government bank * victory margin			0.53		
			(0.29)		
Politically connected * government bank * winparty				0.29	
				(0.13)	
Politically connected * government bank * electoral participation					−1.04
					(0.53)
Firm fixed effect	YES	YES	YES	YES	YES
R^2	0.81	0.81	0.81	0.81	0.81
No. of Obs.	10,880	10,880	10,880	10,880	10,880

Data are restricted to firms that borrow from both government and private banks. Robust standard errors are reported in parentheses. A unit of observation is a firm-bank-type pair, as all loans of a firm given by the same bank type are summed. There are thus 5,440 firm fixed effects and 10,880 total observations in the regression. Politically Connected = dummy for whether firm has a politician on its board; Government bank = dummy for government banks; Win/WinParty = politician/political party's winning frequency (%); Percentage Votes = percentage votes obtained by politician; Victory Margin = Difference in percentage votes between the winner and runner up if politician won, 0 otherwise; Electoral Participation = Registered votes cast (%).

banks, respectively. Finally, column (4) shows that a 10 percent increase in the fraction of the times a politician's party wins is associated with 3 percent larger loans.[20]

VI.B. Political Participation

Table VI also examines whether there are any constraints to these rents by asking whether a more active electorate is able to monitor and check its politicians. We run a similar specification

20. We restrict the sample to firms that borrow from both government and private banks in order to use firm fixed effects. We get very similar results when we run these regressions (without firm fixed effects) on firms that only borrow from government banks suggesting that our sample restriction is not a concern.

as above using a measure for electoral participation—voter turnout in the politician's constituency—instead of the political strength measures.

Column (5) provides suggestive evidence that electoral checks impose constraints on rent provision. Firms whose politicians run in constituencies with 10 percentage points higher electoral participation receive 10 percent smaller loans from government banks than they would have otherwise. Recall that because we have firm level fixed effects, our result cannot be driven by simple spurious correlations such as firms in less active political constituencies are more likely to default. While other identification concerns remain, this result does suggest that political corruption is higher in weaker political environments, a point that has been highlighted by others at a cross-country level [Shleifer and Vishny 1993].

VI.C. The Impact of Winning

What happens to a politically connected firm's borrowing when its politician or political party wins or loses an election? To what extent does being in power affect the firm's ability to earn rents?

Table VII answers this by exploiting the time series component of our data and estimating specification (4). We use quarterly data and restrict it to quarters where an elected government was in power[21] and to only those politically connected firms that experienced a *change* in whether their politician or political party was in power during our data period. Since we are comparing total firm borrowing from private and government banks, we collapse the data to the firm and bank-*type* level in each quarter.

Table VII shows a significant impact on *access* to credit; i.e., winning or being a member of a winning party affects the ability of a politically connected firm to borrow and hence its *amount* of default.

Column (1) shows that, controlling for firm-bank level time-invariant factors and time trends, when the same political firm wins an election it increases its borrowing from government banks by 20 percent compared with its borrowing from private banks which goes down by 11 percent. Thus, when a firm's poli-

21. We exclude quarters where the new government had not been elected as yet (but the old one had been dissolved) and those during 1999–2002 when there was no elected government due to military rule.

TABLE VII
TIME SERIES TEST OF POLITICAL STRENGTH

Dependent variable	Log loan size			
	Data restricted to politically connected firms that experience change in political status			
	(1)	(2)	(3)	(4)
In power?	−0.120		−0.106	−0.105
	(0.027)		(0.028)	(0.027)
In power * government bank	0.186		0.170	0.168
	(0.032)		(0.032)	(0.033)
Party in power?		−0.132	−0.120	−0.120
		(0.028)	(0.028)	(0.028)
Party in power * government bank		0.170	0.153	0.150
		(0.033)	(0.033)	(0.036)
In power * party in power * government bank				0.008
				(0.040)
Fixed effects	Firm * bank-type, quarter	Firm * bank-type, quarter	Firm * bank-type, quarter	Firm * bank-type, quarter
R^2	0.79	0.79	0.79	0.79
No. of Obs.	29,405	29,405	29,405	29,405

Data are restricted to those politically connected firms that actually experience a change in their "power" status due to elections or their party experiences such a change. There are 2,330 such firms. The data are also restricted to only those quarters when an elected government was actually in power; i.e., we exclude quarters where the old government was disbanded but no new government elected as yet and quarters under military rule. The included quarters are 1996 Quarter 2 and Quarter 3; 1997 Quarter 2 to 1999 Quarter 3. In any given quarter the loans for a given firm from a given bank type (government or private) are summed up. Robust standard errors reported in parentheses. In power = dummy for whether politician is in power (won relevant election) during the given quarter; Party in power = dummy for whether politician's political party forms the government for the given quarter (winning parties were different in the two elections in our data period); Government bank = dummy for government banks.

tician board member wins an election, the firm partly substitutes borrowing from private banks toward government banks. Winning politicians exercise their increased political strength to obtain even greater preferential access to credit from government banks.

Column (2) shows that if a politician's political party wins, the firm connected to him also benefits by getting greater access to credit from government banks (13.2 percent). Since a politician may both win and his party may also be in power, column (3) introduces the two effects together and shows that they both have independent effects. Column (4) interacts the politician winning with his party winning as well, and shows that there is no additional benefit of both winning and being in the winning party.

Thus, a politician is able to obtain (greater) rents for his firm either by being in power himself or through his party.

The effect of a firm's politician or his party being in power is only half of the overall political preferential result (Table V). While winning does matter, what matters equally is whether a firm director is a politician (regardless of whether he or his party is in power). This is not surprising for a couple of reasons. First, a significant number of firms appear to be "politically hedged" as a third have multiple politicians on their board, while 11 *percent* (37 *percent* if weighted by loan size) have politicians from different parties. Second, political lines in Pakistan are quite fluid as politicians frequently switch parties and often have family members in opposing parties. Both firms and families hedge themselves politically. Third, frequent elections with party reversals suggests that a politician may not remain out of power for long. Thus, a politician who is out of the government may still wield substantial influence both because he has links with those currently in power and because he is likely to return to power soon. In terms of rent-seeking, entry into the "political network" has equal importance as the politician's relative position within this network. These results lend further support that our findings reflect political influence as opposed to other forms of influence.[22]

VII. Alternative Explanations

We have interpreted our findings as rents accruing to politically connected firms by virtue of their political influence over government banks and hence indicative of political corruption. Before estimating the economy-wide costs of such corruption, we examine whether there are alternative interpretations that can plausibly explain these findings.[23]

22. One could imagine that an influential individual is both more likely to become a politician and (independently of that) obtain preferential treatment. While we do not take a strong stance on this since our results are also interpretable as rents to such "influence," Table VII does suggest that these results are not due to an individual's unobserved (time-invariant) influence but rather the exercise of political power that increases either by his winning an election or being a member of a winning party.

23. We should emphasize that we are not attempting to explain how these rents are distributed. They may be mostly appropriated by the politician or other firm owners. Even if the politician obtains all the rents, he may have to spend resources on his supporters to retain political influence. From our perspective, these are all forms of rent provision and we do not have the data to be able to distinguish between them.

Note first that omitted variables at the firm or bank level that have time-invariant level effects on outcomes, such as firm "influence" or bank inefficiency, cannot explain our results since they remain robust to firm and bank level fixed effects. Moreover, as discussed before, measurements errors in identifying politically connected firms are likely to underestimate the political preference results. Similarly, while there may be "evergreening" concerns that private banks are better able to hide their poorly performing loans, as Section VIII will show, this is unlikely since firms borrowing from private banks are also more productive in terms of real output. Even if private banks do hide bad loans, this would not explain why government banks treat politically connected firms better than unconnected ones, or why they also do not hide the higher default rates of the politically powerful. Therefore, we only consider alternative explanations that also predict a (correctly identified) political preference effect.

VII.A. Social Lending Explanation

The most likely alternative explanation for our political preference results is "social lending." This explanation relies on two key assumptions: (i) firms with politicians on their boards are more likely to engage in projects with high social but low private returns, and (ii) government banks value social returns more than private banks. Given these two assumptions, one could argue that our political preference results do not reflect corruption but the mutual desires of politicians and government banks to undertake "social" projects.

Such an alternative explanation is unlikely given the institutional details and history of politics and politicians in Pakistan. While certain government banks may have social lending goals, our data set consists of *private* corporate loans and excludes loans to government firms. For the social lending story to hold, one would have to believe that politicians in Pakistan are borrowing money *privately* for achieving social objectives. This is unlikely because social projects are mostly carried out either by directly lending to the targeted social class (such as small farmers), or intermediated through large government-owned firms. To our knowledge, never has a government social scheme been explicitly implemented through loans to private firms. Moreover, politicians generally belong to the richest segment of society, and a recent survey of parliamentarians in Pakistan [Zaidi 2004] suggests that politics enriches individuals, with longer duration in

politics associated with greater wealth. Thus, lending to private political firms with high default rates is unlikely to be socially motivated.

Our empirical results also make it harder to believe the social explanation. First, the preferential treatment results are robust (and in fact hardly change) when conditioning on an extensive set of variables which proxy for social attributes of the loan. These include the location of the loan (lending to small cities), the bank (certain banks may have more social objectives), the size, number of creditors, and group affiliation of the borrower (lending to small borrowers with few creditors), and the type and industry classification of the loan (certain industries generate greater social value).

Second, the social lending explanation is not easily reconciled with further results in Tables VI and VII. For example, to generate the result that firms with stronger politicians receive greater preferential treatment, one would need to assume that the likelihood of a politically connected firm undertaking social projects increases the stronger its politician is, in terms of the votes he obtains, his victory margin, etc. and the lower electoral participation is in his constituency. This is unlikely given that most of these firms are located in the major cities and not necessarily in the politician's constituency.

Nevertheless, regardless of these factors that make it harder to believe the social lending explanation, there is direct empirical evidence against it. Table VIII presents two sets of results that check for the presence of social lending and show that there is no evidence for it.

Our first test of the social lending hypothesis is built on the observation that if mutual social objectives are driving political preference by government banks, then one would expect these results to be stronger for those government banks that have explicit social objectives. These include government banks set up for agricultural development, women's welfare, small and medium enterprises, etc. In total, 25 percent of government bank loans belong to such explicitly social government banks. The remaining government banks are meant to be run on a purely financial basis and have no explicit social goals. Columns (1) and (2) show that on both measures of preferential treatment, i.e., default rates and loan size, there is *no* political preference within the explicitly social government banks, while it is large and

TABLE VIII
TESTING FOR A SOCIAL LENDING EXPLANATION

Dependent variable	Default rate (1)	Log loan size (2)	Default rate (3)	Log loan size (4)
Politically connected * government bank	10.47	0.36	11.68	0.32
	(1.84)	(0.05)	(2.88)	(0.08)
Politically connected * government bank * social government bank	−9.4	−0.21		
	(2.73)	(0.17)		
Politically connected * government bank * local firm			−2.54	−0.042
			(2.09)	(0.08)
Controls	YES		YES	
Firm fixed effects		YES		YES
R^2	0.33	0.56	0.33	0.81
No. of Obs.	112,685	11,549	112,685	10,880

Data are restricted to firms that borrow from both government and private banks in columns (2) and (4). Robust standard errors are reported in parentheses. Errors are clustered at the bank level in columns (2) and (4). In column (2) a unit of observation is a firm-bank-type pair where bank-type is private, social government, or nonsocial government. In column (4) a unit of observation is a firm-bank-type pair where bank-type is private or government. All loans of a firm given by the same bank type are summed. Controls include 5 loan-type dummies, 5 firm size dummies, dummy for whether borrower is a foreign firm, 8 dummies for the number of creditors the firm has, 3 group size dummies, 134 dummies for each of the city/town of firm, 21 dummies for the industry of the firm, and 91 bank dummies. Firm-level control variables are described in the Appendix. Controls also include government dummy and all interactions with the government bank dummy. Politically connected = dummy for whether firm has a politician on its board; Government bank = dummy for lender type; Social government bank = dummy for whether government bank (lender) has explicit social objectives; Local firm = dummy for whether firm is located in same province (state) as politician's electoral constituency.

significant for the nonsocial government banks. This is in stark contrast to what the social lending explanation would predict.[24]

We perform another test of the social explanation based on the observation that if politicians use their firms to generate social returns one would expect that this effect is greater for firms that are located in their own constituency. Columns (3) and (4) separate politicians by whether they own a firm in the same province (state) as their constituency or in a different one.[25] The

24. We should note that the average default rate on the social government banks is indeed higher (41.7 percent) than that on the nonsocial government banks (23.1 percent). This is not surprising if such banks were lending to riskier social projects. Thus, while some government banks may indeed lend for social objectives, such motivations cannot explain the political preference effects.

25. Pakistan is divided into four main provinces. These provinces are different in terms of their ethnic composition and political preferences. A politician's constituency is a strict subset of a province. Given the differences across provinces, it is unlikely that a politician will be interested in increasing the welfare of those in another province.

results show little evidence in support of the social lending explanation as politically connected firms that are not located in the politician's state *also* receive the same degree of preferential treatment as those that are.

VII.B. Efficient Lending Explanation

The results on loan rate of return and default rate in Tables III and IV are based on comparing *averages* for these variables across bank and firm types. However, one could argue that even under efficient lending, it is possible to generate the observed differences in average (as opposed to marginal) loan returns.

To understand this argument, suppose that government banks were lending efficiently without any political bias. In this case government banks would start with the most profitable firm and keep making loans to firms until the *marginal* firm has profitability equal to the marginal cost of deposits for the bank. Suppose further that political firms also happen to be less profitable on average than nonpolitical firms. Then, even though the bank is lending efficiently and without any political preference, we will find differences between political and nonpolitical loans in their average return. Moreover, if government banks have lower cost of funds than private banks, this can also explain why the average loan return for government banks is lower than private banks.

While the above explanation may appear plausible at first, it is unlikely for a number of reasons. First, even the average political loan is losing money for the government bank (a rate of return of -17.5 percent), and so the marginal political loan is likely to be even worse. Such low negative returns are impossible to reconcile with efficient lending given that we know government banks pay positive interest rates on their deposits. Second, if the efficient lending hypothesis were the correct explanation, we should also observe similar differences within private banks, which we do not.[26] Third, our results on preferential access to credit, where politically connected firms receive disproportionately larger loans from government banks than nonpolitical

26. One could make further restrictions on the distribution of average returns for political and nonpolitical firms to generate no differences in average returns for private banks. However, these distributional assumptions are not very plausible since they require the relative density of political firms compared with nonpolitical ones be significantly higher at low return projects, yet be the same for high return projects.

firms, cannot be readily explained by an efficient lending hypothesis. Finally, time series evidence on political firms borrowing (even) more from government banks after winning an election is also hard to reconcile with efficient lending.

VIII. The Costs of Rents

This section estimates the economy-level costs of the rents identified. These cost estimates are admittedly speculative both because we only present the subset of costs that can be inferred from our findings and because, even for this subset, we have to make additional assumptions. We consider costs due to the increased taxation necessary to bail out bad government loans, and from the forgone value when corrupt loans are poorly invested. There are likely to be a variety of other, potentially larger, costs that we ignore due to measurement difficulties. For example, we ignore general equilibrium effects such as distortions in entry and composition of firms, compromised legal institutions, and "wasteful" activities that individuals and firms undertake in seeking rents and getting access to political networks.

VIII.A. The Deadweight Loss of Taxation

Loans that default due to political corruption can be considered a transfer payment to politicians. The transfer is ultimately from taxpayers as the government uses its revenues to bail out government banks. To obtain the taxation deadweight loss from such transfers, we need to estimate the size of this transfer, i.e., the "extra" default due to corrupt lending. Assuming that private banks are lending efficiently,[27] the defaulted amount in government banks over and above the rate of default faced by private banks (6 percent) represents this extra default.

With an average default rate of 30.8 percent on government bank loans to politically connected firms, this suggests that 24.8 percent of such lending is the *incremental* loss due to corruption. Given total government bank lending of Rs. 190 billion ($3.2 billion) in 2002, 38 percent of which was given to politically connected firms, the total additional revenue lost from political corruption is Rs. 17.9 billion annually (0.248 * 0.38 * 190). Alter-

27. This is reasonable since we find no evidence of corruption in private banks. See Mian [2004] for further evidence that private banks are lending efficiently.

natively, given the pervasiveness of corruption in government banks, it is likely that even nonpolitical loans have substantial elements of rents, since such loans also face a high default rate (19.9 percent). If we count nonpolitical loan default on government banks as corruption motivated as well, then the revenue lost from corruption is Rs. 34.3 billion annually (17.9 + 0.139 * 0.62 * 190).

We use conservative DWL estimates that put the marginal costs of taxation at around 40 cents for every dollar raised [Ballard et al. 1985]. Note that others have estimated costs up to a dollar per dollar of revenue raised [Feldstein 1996]. Using the more conservative marginal cost numbers, we get DWL estimates ranging from Rs. 7.2 to 13.7 billion each year, or 0.16–0.3 percent of GDP annually.

VIII.B. Cost of Investment Distortion

It would be unrealistic to assume that wealth transfer is the only distortion resulting from corrupt lending. If influential people like politicians get "cheap" money from government banks, they are unlikely to invest their loans efficiently. This would lead to rates of return to investment that are lower than would have been otherwise. In the extreme, they may not invest at all and simply consume the money or deposit it in offshore accounts. To estimate the cost of such investment distortion, one needs to know the rate of return to corrupt lending.[28]

While one could make different assumptions about this return, it is simpler to present a higher bound where the defaulted amount is assumed to generate zero net returns (i.e., the economy just gains the book value of investment). In this case the cost of investment distortion is losing future streams of income generated if the defaulted amount had been properly invested. Given that the market price of a firm reflects the present value of its underlying assets, we can impute this net present value by subtracting book from market value.

Using this approach and a Market to Book ratio for Pakistan estimated at 2.96 (IFC emerging market database—EMDB), we

28. Note that in well-functioning credit markets, these poor/no investments would not affect aggregate investment since financial markets would compensate for this leakage by lending more (i.e., credit supply would be very elastic). However, in a related paper Khwaja and Mian [2004b] exploit an exogenous shift in credit supply to show that bank credit supply in Pakistan is quite inelastic. Therefore, such perfect market assumptions are unlikely to hold.

get annual costs of Rs. 35–67 billion, or 0.8–1.6 percent of GDP each year.[29] This is estimated as (2.96 − 1) * (Inefficient Government Lending) where the estimates vary depending on whether we only consider the defaulted amount by the politically connected (*Rs. 17.9 billion*) or all government bank default in excess of natural default (*Rs. 34.3 billion*) as inefficient government lending. Note that we are being conservative in only considering the *defaulted* government bank lending as inefficient since, as we show below, it is likely that even the nondefaulted government bank lending is poorly invested.

VIII.C. What Is the Real Rate of Return on Political Loans?

The investment distortion cost only arises if the real return on corrupt lending is less than that on noncorrupt lending. The loan-level financial data used so far do not reveal the real productivity of the loan. For example, it is possible that a politically connected firm defaults because it can, but still invests the loan efficiently.

Table IX shows that this is unlikely, by presenting direct evidence for the lower real quality of government bank lending in the textile industry. We use three measures of firm quality: whether a textile firm exported any amount in the three-year period during 2000–2003, the value of its exports aggregated over the three years, and export "productivity" measured by exports as a fraction of total loans to the firm. These are plausible measures of firm quality since the textile industry in Pakistan is mostly export driven, and it is unlikely that a high quality firm would not be exporting. Moreover, unlike balance sheet information, which for most of these firms is unaudited and hence of highly suspect quality, export information is measured through the banking sector (we obtained the information from the central bank) and therefore harder to manipulate. These data are matched by the name of the textile firm to firm names in our data.

Before presenting the results on government lending quality, columns (1)–(2) first show that our quality measures are indeed related to borrowing performance. Firms in the textile industry with higher default rates are less likely to be exporting. Columns (3)–(8) next present evidence that not only do government banks lend to lower quality firms, but firm quality is even lower for

29. To the extent that private firms have a lower market to book ratio than public firms, we may be overestimating the cost of inefficient investment.

TABLE IX
ARE POLITICALLY CONNECTED FIRMS LESS PRODUCTIVE?

Data restricted to textile firms								
	Exporter?		Exporter?		Log exports		Log export productivity	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Firm default rate	−0.22	−0.17						
	(0.051)	(0.060)						
Government bank borrower			−0.19		−0.79		−0.28	
			(0.08)		(0.44)		(0.18)	
Politically connected				0.05		0.05		−0.02
				(0.06)		(0.20)		(0.09)
Politically connected * government bank borrower				−0.13		−0.64		−0.24
				(0.07)		(0.31)		(0.15)
Constant	0.22							
	(0.029)							
Controls		YES	YES	YES	YES	YES	YES	YES
R^2	0.04	0.27	0.2	0.28	0.1	0.18	0.1	0.21
No. of Obs.	6,313	6,313	6,313	6,313	6,313	6,313	6,313	6,313

All Regressions are run at the firm level. Robust standard errors are reported in parentheses. Exporter is a dummy for whether the firm exports or not; Log exports is the logarithm of export value; Export productivity is export value divided by total firm borrowing (from all bank types). Politically connected = dummy for whether firm has a politician on its board; Government bank borrower = dummy for whether firm borrows from any government bank; Log firm size = Logarithm of a firm's total borrowing from all banks (private, government, foreign); Firm default rate = Firm's average default rate across all banks. Controls include 5 loan-type dummies, 5 firm size dummies, dummy for whether the borrower is a foreign firm, 8 dummies for the number of creditors the firm has, 3 group size dummies, 134 dummies for each of the city/town of firm, and 91 bank dummies. Firm-level control variables are described in the Appendix. When government dummy is reported in columns (3), (5), and (7), the bank dummies are not included in the regression.

politically connected firms. Columns (3)–(4) show that while government bank loans are 19 percentage points more likely to be provided to nonexporting textile firms, within government bank loans, those to the politically connected firms are 13 percentage points more likely to be given to nonexporting textile firms. Columns (5)–(8) illustrate similar findings using the other two measures of firm quality: value of exports and export productivity.

Our cost estimates assumed two investment extremes: normal returns (to the firm) on the corrupt loans or no returns at all. Examining real measures of firm quality suggests that these loans earn below normal rates of returns. The cost of such rent provision is therefore likely to be closer to the upper estimate, giving a total cost (i.e., including DWL) of 1.9 percent of GDP every year. Although this estimate is large, it is comparable to that in cross-country studies [Mauro 1995].

IX. Conclusion

This paper has tried to elaborate on the nature and consequences of political corruption in the form of rents in financial markets by carrying out a detailed micro-level analysis. The techniques used are relatively straightforward and can be replicated in other contexts to examine the role political and other avenues of corruption play in the economies of both developed and developing nations. For example, the rents identified in this paper are likely to have an impact on the structure of industry. Differential access and subsidized credit to the politically connected firm is likely to affect entry and exit of firms and their competitive strategies in general. Firms may devote resources to seek such rents and build political links. Exploring such effects offer promising areas for further research.

A question that arises given our findings is how these rents affect the decision to enter politics and the actions chosen by, and success of politicians. If greater wealth has an impact on political entry and strength, then our results imply a feedback mechanism where influential individuals, particularly the most corrupt, may progressively increase their wealth and influence. There is evidence to suggest that this is indeed the case in Pakistan [Zaidi 2004]. Our results also hint at the importance and robustness of political networks as politicians are able to obtain rents even when not directly in power. They also raise questions on the extent to which political competition imposes checks on rents. Are the excessively corrupt penalized and do rents have to be distributed to retain power? How the nature and extent of rents affect the political and institutional environment presents another interesting direction of future enquiry.

Finally, a positive policy interpretation of our results is that private banks do not provide any political rents, and their low default rates suggest the lack of such concerns in general. Moreover, they show little evidence of related lending [Mian 2004]. This lends credence to the Pakistan government's current push for privatization, with three government banks privatized since 1990. However, we should caution that our results do not suggest that full privatization will eliminate rent provision. If government lending is reduced significantly, those with influence may choose other avenues to seek rents. More generally, our cost estimates are relative to the first best of no corruption. To the extent that constraining the political rents identified in this pa-

per leads to alternative sources of rent extraction, the country may not recover the full cost of corruption identified in this paper. Understanding the importance and costs of alternative sources of rent seeking when more common channels are shut down is an interesting area for future work, especially given that emerging economies are increasingly carrying out such reforms.

APPENDIX: DETAILS ON THE FIRM ATTRIBUTES USED IN THIS ARTICLE

(i) *Size.* The total borrowing by a firm from all the banks in the country (including foreign, domestic, and government banks) is used as a proxy for borrower size. We divide firms into five size categories using 99, 95–99, 75–99, 50–75, and 0–50 percentiles as the cutoff criteria. The cutoff criteria were used given the skewed distribution of lending, with 55 percent of total lending going to the top 1 percent of firms by size.

(ii) *Location.* This variable captures which type of city or town the borrower belongs to. Cities are classified by their population size into three categories: big, medium, and small. Borrowers located in the three largest cities (city population greater than 2 million) are coded as big, while those in cities with population between 0.5–2 and 0–0.5 million are coded as medium and small, respectively.[30] The distribution of lending across city size is also highly skewed with the large cities getting 74 percent of the lending.

(iii) *Foreign.* This variable captures whether the borrower is a foreign firm or not. There are only 212 loans given out to foreign firms in the data, but they represent about 4 percent of the overall domestic lending.

(iv) *Group Size.* Using information on the names and tax identification numbers of all directors of a firm, we can classify firms into "groups" based on their ownership information. In particular, firms are assigned the same group if they have a director in common. Mian and Khwaja [2004a] analyze these group linkages in detail, but for this paper what is important is that forming groups in this way creates three distinct category of firms: (a) Stand-Alone Firms—these are firms whose directors do not sit on the board of any other firm (comprising 20 percent of

30. Karachi, Lahore, and Rawalpindi/Islamabad are coded as "big," Faisalabad, Gujranwala, Multan, Sialkot, Sargodha, Peshawar, Quetta, and Hyderabad are coded as "medium," and the remaining cities and towns are coded as "small."

domestic lending); (b) Intermediate Group Firms—these are firms that belong to intermediate size groups, defined as groups consisting of 2 to 50 firms (20 percent of domestic lending), and (c) Large Conglomerate Firms—these are firms that belong to the large conglomerates, defined as groups consisting of more than 50 firms each (38 percent of domestic lending). Ownership (and hence group) information is missing for 22 percent of domestic lending.

(v) *No. of Creditors.* This variable captures the number of creditors (banks) that a firm borrows from. Loans from foreign banks are also taken into account when constructing this variable.

KENNEDY SCHOOL OF GOVERNMENT, HARVARD UNIVERSITY
GRADUATE SCHOOL OF BUSINESS, UNIVERSITY OF CHICAGO

REFERENCES

Acemoglu, Daron, and Thierry Verdier, "The Choice between Market Failures and Corruption," *American Economic Review,* XC (2000), 194–211.
Ades, Alberto, and Rafael Di Tella, "Competition and Corruption," *American Economic Review,* LXXXIX (1999), 982–994.
Aigner, Dennis, "Regression with a Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics,* I (1973), 49–60.
Ballard, Charles, Don Fullerton, John Shoven, and John Whalley, *A General Equilibrium Model for Tax Policy Evaluation* (Chicago, IL: University of Chicago Press, 1985).
Banerjee, Abhijit, "A Theory of Misgovernance," *Quarterly Journal of Economics,* CXII (1997), 1289–1332.
Bliss, Christopher, and Rafael Di Tella, "Does Competition Kill Corruption?" *Journal of Political Economy,* CV (1997), 1001–1023.
Cole, Shawn, "Fixing Market Failures or Fixing Elections? Agricultural Credit in India," Working Paper, Massachusetts Institute of Technology, 2004.
Dinç, Serdar, "Politicians and Banks: Political Influences on Government-Owned Banks in Emerging Countries," *Journal of Financial Economics,* Forthcoming, 2004.
Faccio, Mara, "Politically connected Firms," Working Paper, Vanderbilt University, 2004.
Feldstein, Martin, "How Big Should Government Be?" NBER Working Paper No. w5868, 1996.
Fisman, Raymond, "Estimating the Value of Political Connections," *American Economic Review,* XCI (2001), 1095–1102.
Glaeser, Edward, and Raven Saks, "Corruption in America," NBER Working Paper No. w10821, 2004.
Hall, Robert, and Charles Jones, "Why Do Some Countries Produce So Much More Output per Worker than Others?" *Quarterly Journal of Economics,* CXIV (1999), 83–116.
Johnson, Simon, and Todd Mitten, "Cronyism and Capital Controls: Evidence from Malaysia," *Journal of Financial Economics,* LXVII (2003), 351–382.
Keefer, Philip, and Stephen Knack, "Institutions and Economic Performance: Cross-Country Tests Using Alternative Institutional Measures," *Economics and Politics,* VII (1996), 207–227.
Khwaja, Asim I., and Atif Mian, "The Value of Business Networks in Emerging Markets," Working Paper, Harvard University, 2004a.
Khwaja, Asim I., and Atif Mian, "Tracing the Impact of Bank Liquidity Shocks," Working Paper, University of Chicago, 2004b.

Kornai, Janos, "Resource-Constrained versus Demand-Constrained Systems," *Econometrica,* XLVII (1979), 801–819.
——, "The Soft Budget Constraint," *Kyklos,* XXXIX (1986), 3–30.
Krueger, Anne, "The Political Economy of the Rent-Seeking Society," *American Economic Review,* LXIV (1974), 291–303.
LaPorta, Raphael, Florencio Lopes-de-Silanes, Andrei Shleifer, and Robert Vishny, "The Quality of Government," *Journal of Law, Economics and Organization,* XV (1999), 222–279.
Mauro, Paolo, "Corruption and Growth," *Quarterly Journal of Economics,* CX (1995), 681–712.
——, "The Effects of Corruption on Growth, Investment, and Government Expenditure: A Cross-Country Analysis," in Kimberly Ann Elliott, ed., *Corruption and the Global Economy* (Washington, DC: Institute for International Economics, 1997), pp. 83–107.
Mian, Atif, "Distance Constraints: The Limits of Foreign Lending in Poor Economies," *Journal of Finance,* 2006, forthcoming.
Rose-Ackerman, Susan, *Corruption: A Study in Political Economy* (New York, NY: Academic Press, 1978).
Sapienza, Paola, "The Effects of Government Ownership on Bank Lending," *Journal of Financial Economics,* LXXII (2004), 357–384.
Shleifer, Andrei, and Robert Vishny, "Corruption," *Quarterly Journal of Economics,* CVIII (1993), 599–617.
Shleifer, Andrei, and Robert Vishny, "Politicians and Firms," *Quarterly Journal of Economics,* CIX (1994), 995–1025.
State Bank of Pakistan, "Pakistan: Financial Sector Assessment 1990–2000," 2002.
Zaidi, Akbar, "Elected Representatives in Pakistan. Socio-economic Background and Awareness of Issues," *Economic and Political Weekly,* November 6, 2004.

Traditional Institutions Meet the Modern World: Caste, Gender, and Schooling Choice in a Globalizing Economy

By KAIVAN MUNSHI AND MARK ROSENZWEIG*

This paper addresses the question of how traditional institutions interact with the forces of globalization to shape the economic mobility and welfare of particular groups of individuals in the new economy. We explore the role of one such traditional institution—the caste system—in shaping career choices by gender in Bombay using new survey data on school enrollment and income over the past 20 years. We find that male working-class—lower-caste—networks continue to channel boys into local language schools that lead to the traditional occupation, despite the fact that returns to nontraditional white-collar occupations rose substantially in the 1990s, suggesting the possibility of a dynamic inefficiency. In contrast, lower-caste girls, who historically had low labor market participation rates and so did not benefit from the network, are taking full advantage of the opportunities that became available in the new economy by switching rapidly to English schools. (*JEL* I21, J16, O15, Z13)

The collapse of the former Soviet Union, followed by the economic and financial liberalization of the 1990s, has restructured and "globalized" many economies throughout the world. One consequence of this restructuring, which has been widely observed, is that some groups have taken advantage of the new benefits afforded by globalization, while others appear to have been left behind. This paper addresses the question of whether and how old institutions clash with the forces of globalization in shaping the response of particular groups of individuals to the new economy. Traditional institutions, such as community networks, are generally believed to play an important role in low-income countries by facilitating economic activity when markets function imperfectly. Less well understood is how traditional institutions affect the transformation of economies undergoing change, affecting in turn the distribution of benefits from macroeconomic structural reform.

We explore the role of one such traditional institution—the caste system—in shaping career choices by gender in a dynamic urban context, using new data on schooling choices and income covering the past 20 years in Bombay city, the industrial and financial center of the Indian economy. Bombay is a useful and important setting in which to study the role of institutional rigidities in a dynamic context, as the Bombay labor market was historically organized along caste lines, with individual subcastes or *jatis* controlling particular occupational niches over the course of many generations.[1] A particularly important feature of these caste networks is that they were most active in

* Munshi: Department of Economics, Brown University, Box B/64, Waterman St., Providence, RI 02912 (e-mail: munshi@brown.edu); Rosenzweig: Department of Economics, Yale University, 27 Hillhouse Ave., New Haven, CT 06520 (e-mail: mark.rosenzweig@yale.edu). Leena Abraham and Padma Velaskar of the Tata Institute of Social Sciences collaborated in the design of the survey instrument, carried out the survey, and provided many valuable insights. We are also very grateful to Suma Chitnis for her support and advice at every stage of the project. We received helpful comments from Abhijit Banerjee, David Card, Jan Eeckhout, Lakshmi Iyer, Duncan Thomas, three anonymous referees, and seminar participants at Columbia University, Harvard University, the Massachusetts Institute of Technology, IZA, University of Pennsylvania, Princeton University, UCLA, the University of Southern California, Washington University, and the World Bank. Research support from the Mellon Foundation at the University of Pennsylvania and the National Science Foundation (grant SES-0431827) is gratefully acknowledged. We are responsible for any errors that may remain.

[1] Rajnarayan Chandavarkar (1994 pp. 122, 223), for instance, describes how "[caste] clusters formed within particular trades and occupations ... [this] occupational distribution reflected neither [traditional rural] caste vocation nor the inheritance of special skills. It was produced partly by exclusionary practices by which social groups, once they

working-class occupations dominated by *lower caste men.* Women historically did not participate in Bombay's labor market and so did not benefit from the caste networks, but both men and women scrupulously adhered to the social rule of endogamous marriage within the *jati.*

Although Bombay was a predominantly industrial city for a hundred years beginning in the last quarter of the nineteenth century, in the early 1990s the liberalization of the Indian economy saw a shift in the city's economy toward the corporate and financial sectors. We study how members of different *jatis,* by gender, responded to these changes in the returns to different occupations, and we will show that the historical pattern of networking within the *jati* continues to shape gender-specific, individual responses to these new opportunities in ways that will importantly affect the future distributions of incomes, independent of pre-schooling human capital effects or liquidity constraints.[2]

Our strategy in this paper is to assess how schooling choice, measured by the language of instruction, varied across *jatis,* across boys and girls within *jatis,* and over time. We focus on schooling choice because most adults were already locked into their occupations when the unexpected economic changes occurred. Schooling choice is an important determinant of future occupational outcomes in the Bombay economy and thus reflects the contemporaneous perceptions of expected occupational returns. University education in Bombay is entirely in English, but children choose between English and Marathi (the local language) as the language of instruction at the time they enter school. Schooling in Marathi channels the child into working-class jobs, while more expensive English education significantly increases the likelihood of obtaining a coveted white-collar job. If the economic liberalization of the 1990s effectively increased white-collar incomes, and by extension the returns to English education, then (future) occupational mobility can be identified from changes in the choice of the language of instruction made by parents of school-age children. Examination of the changing patterns of schooling choice by *jati* and gender thus permits an assessment of the interactions between traditional institutions and the new realities of globalization.

Our empirical analysis is based on a survey of 4,900 households belonging to the Maharashtrian community residing in Bombay's Dadar area and a survey of the schools in that locale that we conducted in 2001–2002. Secondary schools in Bombay run from grade 1 to grade 10. The household survey was based on a stratified random sample of students who entered 28 of the 29 schools in Dadar (in the first grade), over a 20-year period, 1982–2001.[3] English is the language of instruction in ten schools in Dadar, while Marathi is the language of instruction in the remaining 18 schools.

The survey data suggest that the returns to English education, for given years of schooling, increased in the 1990s. Based on retrospective information on the annual earnings of the parents of the sampled children, we estimated the returns to English and the returns to years of schooling at five points in time from 1980 through 2000 for working adults between the age of 30 and 55.[4] Figures 1 and 2 provide the estimated returns to schooling attainment and schooling language, for men and women, respectively, in each time period. As can be seen, the returns to years of schooling increased only mildly over time for both men and women. In contrast, the English premium increased sharply from the 1980s to the 1990s for both sexes, rising from 15 percent in 1980 to 24 percent in 2000 for men and from approximately 0 percent in 1980 to 27 percent in 2000 for women. The returns to English for men increase from the mid-1980s, which is most likely due to the decline around

obtained a foothold in a particular occupation, would not admit an outsider."

[2] A recent literature has shown that historical institutions have long-run consequences for growth in low-income countries (Daron Acemoglu et al., 2001; Abhijit Banerjee and Lakshmi Iyer, 2005). These empirical findings, however, do not provide insight into the mechanisms underlying such persistence.

[3] One school refused to provide us with information on its students and will be ignored in all the discussion that follows.

[4] The details of the estimation procedure and the estimates of the returns to English and the returns to schooling (with standard errors) are provided in Munshi and Rosenzweig (2003).

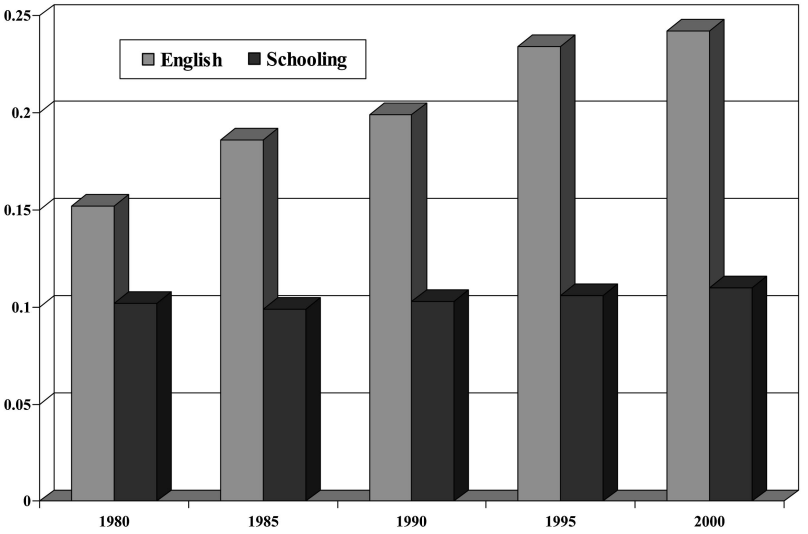


FIGURE 1. RETURNS TO ENGLISH AND SCHOOLING BY YEAR, 1980–2000: MEN AGE 30–55

that time in manufacturing jobs in Bombay (Darryl D'Monte, 2002), but continue to rise through the 1990s.

The survey collected information on schooling choice for 20 cohorts of students who entered the 28 neighborhood schools (in the first grade) over the 1982–2001 period. The time-series data on enrollments in English- and Marathi-medium schools suggest that the changes in the returns to English significantly affected schooling choice for both boys and girls in the sample, across castes and over time. Figure 3 and Figure 4 display the changing proportions of students enrolled in English schools for the 20 entering cohorts from 1982 (cohort = 1) to 2001 (cohort = 20) for three caste group-

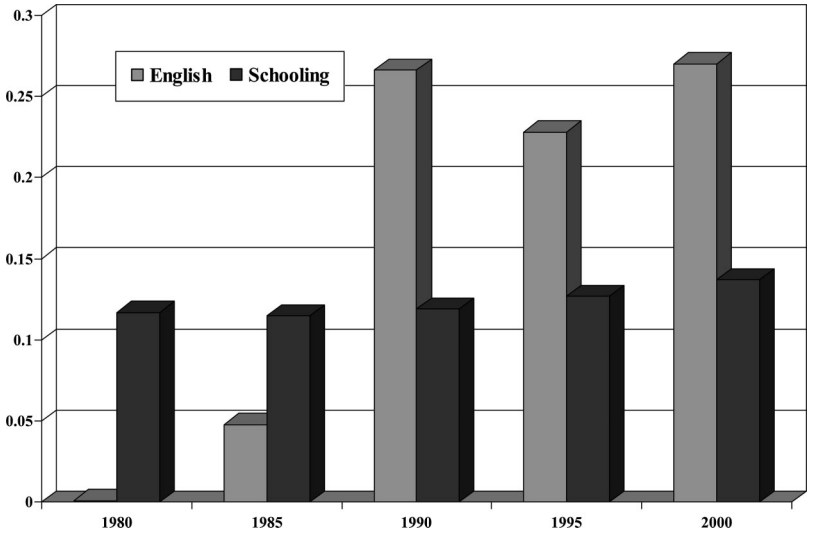


FIGURE 2. RETURNS TO ENGLISH AND SCHOOLING BY YEAR, 1980–2000: WOMEN AGE 30–55

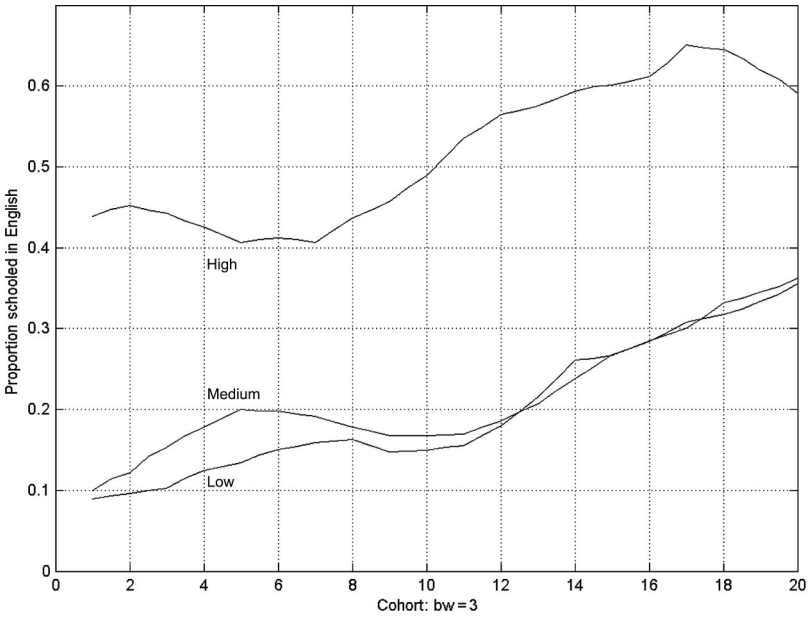


FIGURE 3. ENGLISH SCHOOLING: NET PARENTAL EDUCATION EFFECT—BOYS

ings—low, medium, and high—and by gender.[5] The figures were constructed using the Epanechnikov kernel function to nonparametrically regress schooling choice (1 = English medium; 0 = Marathi medium) on the cohort variable for each caste group, taking into account the strong intergenerational state-dependence with respect to the language of instruction within the family.[6] Although *jatis* define the relevant boundary for the labor-market networks and form the relevant social unit in our analysis,[7] we aggregate the 59 subcastes in our data for expositional convenience in these figures.

Figures 3 and 4 show that enrollment rates in English-medium schools have grown substantially over time for both boys and girls and for all castes.[8] The trajectory is much steeper, however, for the ten most recent cohorts, who would have entered school in the post-reform 1990s. Thus, the increase in the returns to English observed in Figure 1 and Figure 2 appears to have shifted schooling choice toward English education. The figures also indicate substantial differences in English schooling between castes at the beginning of the sample period, reflecting in part the circumstances of the colonial regime. The high castes gained access to clerical and adminis-

[5] Children enter first grade at the age of 6 and complete tenth grade at the age of 15, so the current age of the students in our sample, with only a few exceptions, ranges from 6 to 25. Students in Bombay typically do not change the language of instruction midstream or switch schools after they enter first grade. High castes include all the Brahmin *jatis,* as well as a few other elite *jatis* (CKP and Pathare Prabhus). Low castes include Scheduled Castes, Scheduled Tribes, and Other Backward Castes, as defined by the government of India. Medium castes are drawn mostly from the cultivator *jatis,* such as the Marathas and the Kunbis, as well as other traditional vocations that were not considered to be ritually impure.

[6] If both parents have been schooled in English, it is very unlikely that the child would be sent to a Marathi school, and all the regressions that we later report will also account for such state dependence at the level of the family. Details of the nonparametric estimation procedure and parametric estimates of the schooling regression (with standard errors) are provided in Munshi and Rosenzweig (2003).

[7] As Morris David Morris (1965 p. 76) emphasizes in his historical account of the Bombay labor market, "for any analysis of labor recruitment [in Bombay] ... it is entirely inappropriate to lump into larger groups because of similarity of name, function, social status, or region-of-origin subcastes that are not endogamous."

[8] Details of the nonparametric estimation procedure used to generate these figures are provided in Munshi and Rosenzweig (2003).

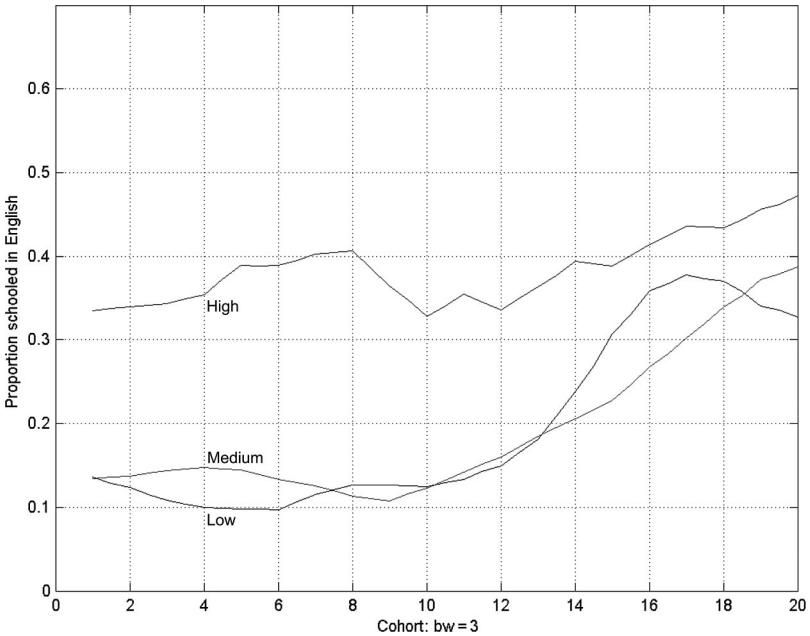


FIGURE 4. ENGLISH SCHOOLING: NET PARENTAL EDUCATION EFFECT—GIRLS

trative jobs under the British, while the lower castes were confined for the most part to working-class jobs. Consistent with the view that Marathi education channels students into working-class jobs, and that English education increases the likelihood of obtaining a white-collar job, we see in Figure 3 and Figure 4 that high-caste boys and girls currently 25 years old (the oldest cohort) were much more likely to have been schooled in English, and that this caste difference in schooling persists over the next ten cohorts. But although the caste gap narrows dramatically for the girls in the 1990s, there is no convergence for the boys. Thus, it appears that caste continues to play a role in shaping schooling choices in the new economy of the 1990s, but only for boys. The key question is why the lower-caste boys seemingly fail to take advantage of the new economic opportunities.

The gender-specific explanation for the observed pattern in Figure 3 that we pursue in this paper is based on network externalities. Numerous studies document higher levels of networking in blue-collar occupations, possibly because the information and enforcement problems that give rise to networks are more acute in those jobs.[9] These studies focus on men, the primary occupants of blue-collar jobs. And among the household heads in Dadar, 68 percent of the men in working-class jobs found employment through a relative or a member of the community, while the corresponding statistic for white-collar workers was 44 percent. Once the (working-class) network is in place, there is a positive externality associated with participation in the network, and hence with the traditional occupational choice in the *jati.* This externality could give rise to intergenerational occupational persistence *at the level of the jati,* with labor market networks channeling boys into particular (traditionally male) occupations and hence toward particular schooling choices.

[9] For example, Albert Rees (1966) found that informal sources accounted for 80 percent of all hires in eight blue-collar occupations versus 50 percent of all hires in four white-collar occupations in an early study set in Chicago. Similarly, 68 percent of blue-collar workers and 38 percent of white-collar workers reported having received help finding a job in M. S. Gore's (1970) study of migrants in Bombay.

Once the returns to the white-collar occupation grow, however, schooling choice must ultimately converge across castes. The explanation for the absence of convergence in Figure 3 that we put forward in this paper is based on the idea that the caste networks might place tacit restrictions on the occupational mobility of their members to preserve the integrity of the network. We will show that although these restrictions might have been welfare-enhancing and indeed equalizing when they were first put in place, such restrictions could result in dynamic inefficiencies when the structure of the economy changes.

The results in this paper provide empirical support for the view that historical occupation patterns kept in place by caste-based networks continue to shape occupational choice, and hence schooling choice, for the boys in the new economy. In contrast, the lower-caste girls who historically kept away from the labor market, and so have no network ties to constrain them, take full advantage of the opportunities that become available in the new economy. The growing disparities in school choices between boys and girls within the traditional *jatis* not only suggest a new balance of economic opportunities by gender, but also could threaten the long-run stability of the caste system, which is based on endogamous marriages within the subcaste. Thus, a complete understanding of the development process must not only take account of the initial conditions and the role of existing institutions in shaping the response to modernization and globalization, but must also consider how these traditional institutions are shaped in turn by the forces of change.

I. The Institutional Setting

A. *Bombay's Labor Market*

Bombay's industrial economy in the late nineteenth century and through the first half of the twentieth century was characterized by wide fluctuations in the demand for labor (Chandavarkar, 1994). It is well known that such frequent job turnover can give rise to labor market networks, particularly when the quality of a freshly hired worker is difficult to assess and performance-contingent wage contracts cannot be implemented. The presence of such recruitment networks has indeed been documented by numerous historians studying Bombay's economy prior to independence in 1947 (Chandavarkar, 1994; Morris, 1965; Alexander R. Burnett-Hurst, 1925). These networks appear to have been organized around the jobber, a foreman who was in charge of a work gang in the mill, factory, dockyard, or construction site, and more importantly also in charge of labor recruitment.

Given the information and enforcement problems associated with the recruitment of short-term labor, it is not surprising that the "jobber had to lean on social connections outside his workplace such as his kinship and neighborhood connections" (Chandavarkar, 1994, p. 107). Here the endogamous subcaste or *jati* served as a natural social unit from which to recruit labor, because marriage ties strengthen information flows and improve enforcement. This widespread use of caste-based networks thus led to a fragmentation of the Bombay labor market along social lines.[10] Other studies also suggest that these patterns tended to persist over many generations. For example, Hemalata C. Dandekar (1986) traces the evolution of a network of Jadhavs (a particular subcaste) belonging to one village in interior Maharashtra. In 1942, 67 percent of the Jadhav migrants from that village were working in the textile mills and 4 percent in other factories. Thirty-five years later, in 1977, 58 percent were still employed in textile mills, while 10 percent were in other manufacturing industries.

A noticeable feature of historical descriptions of caste-based networks in Bombay is that they were restricted to working-class jobs. This is not surprising, because the information and enforcement problems that give rise to such networks tend to be more acute in those occupations. Further, most studies of caste-based networks in Bombay focus on male workers. Women were conspicuously absent from Bombay's labor

[10] The presence of caste clusters has been historically documented in the mills (R. G. Gokhale, 1957), among dock workers (R. P. Cholia, 1941), construction workers, and in the railway workshops (Burnett-Hurst, 1925), in the leather and dyeing industries, and in the Bombay Municipal Corporation and the Bombay Electric Supply and Transportation Company (Chandavarkar, 1994). More recently, Kunj M. Patel (1963) surveyed 500 mill workers in the Parel area, close to the site of our study, in 1961–1962 and found that 81 percent of the workers had relatives or members of their *jati* in the textile industry. Sixty-six percent of the workers got jobs in the mills through the influence of their relatives and friends.

TABLE 1—SCHOOL CHARACTERISTICS AND STUDENT PERFORMANCE

	English medium	Marathi medium
School type	(1)	(2)
Panel A. School characteristics		
Student-teacher ratio	36.71	35.76
	(2.40)	(2.17)
Class size	61.90	62.28
	(3.69)	(3.15)
Students per desk	2.40	2.36
	(0.10)	(0.11)
Proportion of teachers with B.Ed.	0.72	0.70
	(0.07)	(0.05)
Proportion of teachers with higher degree	0.08	0.10
	(0.03)	(0.03)
Computers per student	0.02	0.02
	(0.004)	(0.005)
Student enrollment in secondary section	1528.40	1059.00
	(360.64)	(175.73)
Panel B. School expenses		
Fees	0.48*	0.20*
	(0.01)	(0.01)
Other expenses	1.10*	0.71*
	(0.04)	(0.01)
Panel C. SSC school-leaving exam results (1997–2001)		
Percentage passed	92.59*	51.62*
	(2.04)	(5.95)
Percentage first class among passed	36.2*	24.23*
	(1.69)	(3.35)
Percentage distinction among passed	23.94*	6.90*
	(3.92)	(1.87)
Number of schools	10	18

Notes: Standard errors are in parentheses. Panels A and C use data from the school survey; panel B uses data from the household survey. School characteristics are based on the secondary section (grades 5 through 10). School expenses are measured for 2000–2001 in thousands of 1980 Rupees. To convert to 2000 Rupees, multiply by 4.44. Other school expenses include transportation, coaching classes, textbooks, uniforms, and stationery. Scores above 35 percent are required to pass SSC; scores above 60 percent are required for first class, and above 75 percent for distinction.

* Denotes rejection of the equality of means for the two types of schools with greater than 95-percent confidence.

force, particularly in the working-class jobs (Morris, 1965). These historical patterns of labor force participation by gender will later help explain the schooling choice dynamics, for boys and girls, that we saw in Figure 3 and Figure 4.

B. *The Schools in Dadar*

Our analysis highlights the medium of instruction as the salient feature of schooling choice. It is possible that the choice of the language of instruction merely proxies for school quality. In parallel with the household survey, we carried out a survey of schools based on a questionnaire filled out by school principals. This questionnaire elicited information on a variety of school characteristics, as well as recent student performance on the standardized school leaving examination (common to both Marathi and English schools), which allows us to compare the two types of schools as well as the students across the schools. Table 1, panel A, describes school infrastructure and faculty qualifications in the English and Marathi schools. The average student-teacher ratio, class size, number of students per desk, computers per student, and the proportion of teachers with Bachelor of Education degrees and higher (postgraduate) degrees, are each very similar and statistically indistinguishable for the two types of schools.[11]

[11] A regression of the language of instruction on the set of school characteristics in Table 1, panel A, indicates that

Despite the increase in the demand for English education in the last ten years, as seen in Figures 3 and 4, no new schools were added in this period in Dadar.[12] The English-language schools accommodated this increased demand by adding divisions in each grade, increasing the number of desks in each classroom, and doubling students on each desk. Because the supply of schools was effectively fixed, we would expect the English schools to extract some economic rents from this increased demand through higher fees and schooling costs in general. In contrast, fees in the Marathi schools are subsidized by the state government. Our household survey collected information on school fees and other expenses (transportation, coaching classes, textbooks, uniforms, and stationary) in the last year. Table 1, panel B shows that school fees (in 1980 Rupees) are currently significantly higher in the English schools (480 versus 200 Rupees), as are other expenses (1,100 versus 710 Rupees).

One other difference between the schools is in the performance of the students on the Secondary School Certificate (SSC) school-leaving examination. Table 1, panel C reports student performance on this exam over a five-year period, 1997–2001. Students in the English schools perform much better on this standardized test in terms of the percentage that pass and receive a first class and a distinction.[13] Although these substantial differences in test performance can be explained by differences in school quality, they can also be explained by differential selection by ability into English and Marathi schools, as implied by our network model of school choice, and we will provide evidence supporting this implication of the model below.

In addition, our survey provides direct evidence that the medium of instruction and its implications for children's future role in society, rather than differences in school quality, dominate the schooling choice of parents. The survey elicited from parents the reasons for their choice of school for their child. The percentage of parents reporting that the "quality of education" was a factor in their choice was relatively low and did not differ substantially across parents choosing English-medium schools and Marathi schools—43.7 percent versus 35.2 percent, respectively. In contrast, almost 87 percent of parents who chose English as the medium of instruction for their child reported that career opportunities was a factor in choosing that school. And over 62 percent of parents who chose a Marathi-language school listed closer community ties as a reason.

II. A Simple Model of Schooling Choice

Our first objective in this section is to show how exogenous, historically determined occupational differences across otherwise identical *jatis* can persist when network externalities are present. Because occupational choice translates into schooling choice, this explains the initial caste gap that we observe for the boys in Figure 3 (the model that we lay out in this section applies to the boys, as we will see later that labor market networks are most active among the men). We will show, however, that *jatis* should start to converge once the returns to English grow sufficiently large, which is inconsistent with what we observed in that figure. Our second objective in this section will consequently be to show how network externalities could give rise to endogenous social restrictions on occupational mobility, and by extension schooling choice, preventing convergence across social groups in a changing economic environment.

A. *Population, Community Structure, and Market Structure*

Consider a population with a continuum of individuals. Each individual i is endowed with a level of ability $\omega_i \in \{0, \frac{1}{2}, 1\}$. Note that ability in this section, and throughout the paper, refers to pre-schooling human capital rather than genetic ability. He lives for three periods, studying in the first period and working in the remaining periods. Schooling choice is restricted to instruction in English or Marathi, the local lan-

the joint set of characteristics is not significantly different across the school types.

[12] The average establishment year for the 18 Marathi schools is 1947 and the corresponding year for the 10 English schools is 1959. All schools in the area have now been operating for many decades.

[13] Scores above 35 percent are required to pass, scores above 60 percent are required for a first class, and scores above 75 percent are required for a distinction. The same test is administered to all schools, with the questions translated into English and Marathi.

guage. Occupational choice is restricted to white-collar and working-class jobs. Education in English is required to obtain a white-collar job, but is more expensive than Marathi education, which is assumed for simplicity to be costless. Occupational choice is based on the wage that the individual will receive in the white-collar and the working-class job, net of the pecuniary cost of schooling. Each individual then makes his schooling decision based on the type of job that he (correctly) anticipates he will occupy in the subsequent period. If he prefers to hold a white-collar job, then he will study in English, if not he will study in Marathi, which is less costly.

Each individual is born into a community or *jati.* There is a large number of communities in this economy, and we normalize so that the measure of individuals in each cohort or generation of a *jati* is equal to one. To simplify and highlight the role of network externalities in intergenerational occupational persistence, we assume that the distribution of pre-schooling human capital does not vary over generations or across *jatis.*[14] Within each *jati*-generation there is a measure P_L of low types (with ability $\omega = 0$) and a measure P_M of medium types (with ability $\omega = \frac{1}{2}$).

On the demand side of this labor market, firms operate competitively in both the working-class and the white-collar sectors. We noted earlier that working-class jobs generally tend to be more heavily networked. For the purpose of this simple model, we assume that the white-collar worker's ability, and hence his productivity, can be observed perfectly and so the white-collar wage (net of schooling costs) is specified to be $\theta\omega_i$. Here θ represents the returns to ability in the white-collar job, which in our setup also reflects the returns to English education. In contrast, the nature of the production technology prevents working-class firms from directly observing their employees' ability before they commence work. We take it that the firm is unable to specify a performance-contingent wage contract, and so will use referrals from its incumbent workers to hire new employees, generating a role for the network in the working-class jobs alone. Munshi (2003) provides evidence that experienced workers contribute disproportionately to labor market networks in the United States, and we would expect this pattern to hold up in other economies as well. In our model, the expected working-class wage over the individual's working life is thus specified to be P, the proportion of the previous generation (three-year-olds) in the *jati* that will be employed in the working-class job when he enters the labor force.

[14] We will relax this assumption in the empirical work by allowing for heterogeneity in ability across *jatis.*

B. *The Schooling Equilibrium*

We now proceed to derive the different occupational distributions, and hence schooling equilibria, that can be sustained across *jatis* with the same ability distribution in this setup. Each individual chooses the occupation, and hence the language of instruction, that maximizes his net return. This return depends on his own ability, as well as the proportion of his *jati* in the previous generation employed in the working-class occupation, as described above.

Under conditions that we specify below, with three levels of ability, three distinct schooling equilibria can be sustained within *jatis:* (a) only low types choose Marathi education; (b) low and medium types choose Marathi education; (c) everyone in the *jati* chooses Marathi education.[15]

CONDITION 1: $P_L < \theta/2$.

CONDITION 2: $\theta/2 < P_L + P_M < \theta$.

CONDITION 3: $\theta < 1$.

It is easy to verify that once a *jati* is exogenously assigned a particular occupational distribution, this distribution will persist unchanged over many generations when the conditions above are satisfied.[16] This intergenerational state dependence is a consequence of the network externality associated with the working-class occupation. It implies, in turn, that the probability that any

[15] Munshi and Rosenzweig (2003) consider the general case with N types and N equilibria, without altering the results that we present below.

[16] It is merely necessary to show that no individual wishes to deviate from the occupation, and hence schooling choice, assigned to his type in his *jati* in the previous generation, for each of the schooling equilibria.

individual i drawn randomly from *jati* j will be schooled in English ($E_{ij} = 1$) is related to the proportion of men in the previous generation employed in the working-class job, P_j:

$$\Pr(E_{ij} = 1) = 1 - P_j. \tag{1}$$

This expression will serve as the starting point for the empirical analysis described in Section IV, where we will examine the relationship between schooling choice in the current generation and the occupational distribution in the previous generation, to identify the presence of an underlying network organized around the *jati.*

C. *Schooling Choice as the Returns to English Grow*

The state dependence at the level of the *jati* derived above is obtained under the assumption that the parameters of the model, P_L, P_M, θ, remain stable over time. To explore the effect of the increase in the returns to English (θ) in the 1990s, we now allow for multiple cohorts of unit measure within each generation.

If θ remains constant within a generation, the results derived above follow through without modification for all cohorts. If, however, θ increases across successive cohorts, holding P_j constant, then schooling choice within a *jati* could change over the course of a single generation. When θ just crosses one, high-ability boys belonging to *jatis* that were traditionally in equilibrium 3 switch to English. When θ subsequently reaches $2(P_L + P_M)$, medium-ability boys in *jatis* that were traditionally in equilibrium 2 or equilibrium 3 switch to English, at which point schooling choice across all *jatis* will converge.

Although the network externality described above can explain the persistence of traditional occupational patterns within the *jati* over many generations, and hence the initial caste gap observed in Figure 3, it cannot by itself explain the absence of convergence over the 1990s as the returns to English grew. To explain this absence of convergence, we consider the possibility that heavily networked (working-class) *jatis* might have put restrictions on occupational mobility, and hence schooling choice, in place to preserve the viability of the community network.[17]

To understand why restrictions on mobility might emerge, define a social welfare function that places equal weight on all members of the *jati.* Now the welfare in a *jati* situated in equilibrium 3, in which everyone studies Marathi, is simply the unweighted average of all the payoffs from the working-class occupation, $W = 1$. When θ just crosses one, in a given cohort, all high types in the *jati* can expect to earn more in the white-collar sector than in the *jati's* "traditional" working-class occupation and will thus switch to English schooling. Welfare from that cohort onward is then $W = (P_L + P_M)^2 + (1 - P_L - P_M)$. The new welfare level is a weighted average of $P_L + P_M < 1$ and 1, and so *jati*-level welfare must unambiguously decline when schooling choice, and hence the occupational distribution, shifts. Historically there was intense competition for scarce working-class jobs in Bombay, as noted in Section I. Because larger numbers improve the *jati*'s competitiveness, and increase the working class wage in general, it is easy to see why social restrictions on occupational mobility could emerge endogenously. Moreover, the fact that the lower-caste girls in our sample do not display a similar resistance to change can be attributed to the gender-specific nature of these job networks.

Social restrictions on occupational mobility can be welfare-enhancing for small and medium changes in θ, as noted above. But they could give rise to substantial inefficiencies if they continue to persist when θ grows large. For example, it is easy to verify that the social restrictions described above for equilibrium 3 will be inefficient once θ reaches $1 + (P_L + P_M)$, although a welfare calculation that identifies the presence of such a dynamic inefficiency is beyond the scope of this paper.

[17] Restrictions on mobility do not have to be associated with explicit punishment. Preferences for schooling or future career choices could be determined endogenously, for example, by placing symbolic value on the traditional occupation in the *jati.* Social interactions within the *jati* could also lead individuals to make similar schooling choices and career choices across generations. We do not attempt to distinguish between preferences that are complementary to the network, and the network itself, in this paper.

While we conjecture that restrictions on occupational mobility might be in place in the heavily networked *jatis,* no direct evidence of their presence in Bombay is available. We can, however, test one important implication that is consistent with the presence of these restrictions: the relationship between schooling choice E_{ij} and the occupational distribution within the *jati* in the previous generation P_j must not weaken over successive cohorts in the current generation, even as the returns to English grow. This stability in intergenerational state dependence would then explain the wedge between high-caste and lower-caste schooling choices for boys that was observed through the 1990s in Figure 3.

D. *Selection into Schools*

The model of schooling choice as laid out in this section also has implications for selection, by ability, into English and Marathi schools. Within any *jati,* the average pre-schooling human capital of the English students must be greater than that of the Marathi students. Taking the average across all *jatis,* this implies that average ability must be greater among the English students at any point in time. This observation is consistent with the significantly higher test scores obtained by students in the English schools (Table 1), despite the fact that English and Marathi schools appear to be similar in terms of the resources available per student and the qualifications of the teachers. But how does the ability distribution *within* the English and Marathi schools change across successive cohorts in the current generation as the returns to English grow? Without social restrictions, deviation to English education is ordered by ability, so as θ grows there is a steadily worsening pool of Marathi students. *Jatis* that begin with a greater proportion of their members in working-class jobs have higher ability among the Marathi students, but their shift into English, and hence the decline in ability, must also be more rapid, because all *jatis* ultimately converge. With social restrictions, heavily networked *jatis* continue to begin with a superior ability distribution within Marathi schools, but now there might be no convergence in ability among Marathi students across *jatis.*

Average ability among the English students is greater than average ability among the Marathi students at any point in time, but among the Marathi students it is the group with the highest ability that deviate as θ grows. Thus, while the quality of the Marathi students unambiguously declines over time as the returns to English increase, the change in the quality of the pool of English students is ambiguous.[18]

III. The Household Data

A. *The Survey*

To examine empirically the role of caste networks in shaping mobility during a period of change, we carried out a household survey based on a random sample of students, stratified by caste, who entered the 28 secondary schools in Dadar (in the first grade) over a 20-year period, 1982–2001. This design provides information for the periods before and after the major Indian economic reforms. We obtained a complete list of all students enrolled in grades 1 to 10 in 2001 (the year of the survey), as well as a list of students who were enrolled in grade 10 from 1991 to 2000. Ignoring dropouts, this leaves us with 20 cohorts of students who entered school over the 1982–2001 period. A total of 101,567 students were enrolled in the schools in 2001 or studied in grade 10 over the previous ten years. We drew the roll numbers of 20,596 students randomly from these 20 cohorts, and recovered their names and addresses from the school records. Restricting attention to Maharashtrians residing in Dadar and the immediately adjacent neighborhoods, we were left with 8,092 eligible students to serve as the sampling frame for the survey. The student's name is

[18] For example, in the three-type case with no social restrictions, some *jatis* (in equilibrium 2) have only high-ability children in English schools, while other *jatis* (in equilibrium 1) have both medium- and high-ability children in English schools to begin with. This implies that the quality of the English pool must improve when θ reaches one, because only the high types from *jatis* in equilibrium 3 deviate at that point. But average ability drops below its initial level when θ reaches $2(P_L + P_M)$, because medium and high types in all *jatis* will have switched into English schools by that point. With social restrictions, the change in ability within the English schools becomes even more difficult to characterize.

typically a good indicator of the caste, and we wanted close to 1,000 upper castes in the sample, so all 1,082 students from this population who appeared to be upper castes were selected for the survey. We drew randomly from the remaining students in the sampling frame until the target sample size was reached. The upper castes account for 17.5 percent of the final sample of 4,945 observations, which is slightly higher than the 13.4 percent that we began with in the sampling frame.

The research team interviewed the parents of the selected students at their residences. The survey instrument elicited detailed subcaste information from the respondents and included sections on grandparents' education and occupation, parents' education and occupational and income histories (at five-year intervals from 1980 to 2000), as well as the student's and siblings' subsequent education, occupation, income, and marriage outcomes (where relevant).[19] Information on transfers, assistance in finding jobs, and ties to the community was also collected.

Of the eligible households, 82.5 percent provided completed schedules. This is a relatively high response rate, especially given that some of our addresses were 20 years old. But we might still have obtained a selective sample of households, for a number of different reasons. First, households residing in Dadar who sent their children to study outside the area would be missing from the sample. Second, households who moved out of the area would be among the 17.5 percent of the respondents who did not complete the survey. And third, students from the first ten cohorts who did not reach the tenth grade, and current students who have dropped out, would be missing from the sample. In Section V, we will discuss how our identification strategy is unlikely to be undermined by these potential sources of bias.

[19] The name is usually a good indicator of the individual's community and caste. For example, 98.7 percent of the respondents, whom we had selected on the basis of their names from the school records, said that Marathi was their mother tongue, indicating that they were indeed Maharashtrian. The caste classification is potentially more problematic, however, because lower castes could in some cases change their names or misreport their caste affiliation. Note that such misreporting will not undermine the fixed effects estimation strategy, described below, as long as it does not vary by the gender of the child, within the *jati*.

B. *Descriptive Statistics: Caste, Occupational Networks and Schooling*

The data provide empirical support for three features of the model of schooling choice laid out in Section II. First, the occupational distribution, a product of historical circumstances, varies by caste, and persists across generations, particularly among the men. Second, working-class jobs are associated with a higher level of referrals (networking). And third, working-class jobs are associated with lower levels of English schooling.

The survey elicited information on parental occupations at five-year intervals from 1980 to 2000. For the grandparents, we simply asked for the main occupation over the individual's working life. The 90 occupations in the data were divided by roughly increasing levels of human capital into seven aggregate categories: unskilled manual, skilled manual, organized blue-collar, petty trade, clerical, business, and professional. We further classified unskilled manual, skilled manual, and organized blue-collar as working-class occupations. Clerical, business, and professional were classified as white-collar occupations. Petty trade is treated as an intermediate unclassified occupation.

Table 2, panel A, describes the occupational distribution across broad caste categories (low, medium, high), separately for the employed fathers, based on information in 1995, and the paternal grandfathers of the students in the sample. Columns 1 to 3 of the panel indicate that lower-caste fathers are much more likely to be employed in working-class occupations (54 percent and 43 percent) as compared with high-caste fathers (18 percent).[20] The same cross-caste pattern is obtained for individual occupations within the working-class and white-collar classifications, with the exception of clerical jobs. The comparison of the fathers in columns 1 to 3 with the grandfathers in columns 4 to 6 also indicates that there has been little change in the basic occupational distribution, as well as the percentage of working-class

[20] Note that we use only working-class and white-collar occupations when computing this statistic.

TABLE 2—OCCUPATION, EDUCATION, AND INCOME BY CASTE ACROSS GENERATIONS

Relationship to student	Parent			Grandparent		
Caste	Low	Medium	High	Low	Medium	High
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Fathers and grandfathers						
Employment (%)	97.37	97.31	99.06	98.87	98.86	99.28
Occupational distribution (%)						
Unskilled manual	11.09	7.84	4.41	9.00	3.63	2.10
Skilled manual	17.35	13.70	10.21	11.67	6.72	8.42
Organized blue-collar	22.87	19.22	2.90	22.89	24.23	7.67
Petty trade	4.00	4.51	2.52	3.11	3.20	3.34
Clerical	28.09	36.64	20.81	22.22	23.79	28.84
Business	7.95	8.79	15.51	6.11	4.72	13.00
Professional	8.30	8.79	43.51	5.56	6.18	33.66
Farming	0.35	0.51	0.13	19.44	27.53	2.97
Percent working class	53.64	42.91	18.01	56.24	49.92	19.42
	(1.23)	(1.21)	(1.38)	(1.33)	(1.40)	(1.44)
Years of schooling	9.63	10.22	13.82	—	—	—
	(0.07)	(0.07)	(0.10)			
Monthly income	1.92	1.99	4.61	—	—	—
	(0.04)	(0.04)	(0.25)			
Total number of observations	1,860	1,774	793	1,866	1,934	839
Panel B. Mothers and grandmothers						
Employment (%)	20.56	20.01	51.23	19.31	18.59	15.57
Occupational distribution (%)						
Unskilled manual	29.95	16.94	2.36	24.65	7.18	3.13
Skilled manual	8.82	8.47	6.15	1.70	1.44	3.13
Organized blue-collar	4.01	4.92	0.47	8.50	4.31	0.78
Petty trade	3.74	3.83	1.18	1.13	0.57	0.00
Clerical	31.55	40.71	46.34	4.25	2.30	19.53
Business	4.55	2.46	3.78	2.27	1.44	3.91
Professional	17.38	22.68	39.72	5.10	8.62	67.97
Farming	0.00	0.00	0.00	52.41	74.14	1.56
Percent working class	44.44	31.53	9.09	75.00	51.14	7.14
	(2.62)	(2.48)	(1.41)	(3.39)	(5.36)	(2.30)
Years of schooling	8.03	8.73	13.49	—	—	—
	(0.09)	(0.09)	(0.10)			
Monthly income	0.23	0.30	1.37	—	—	—
	(0.02)	(0.02)	(0.07)			
Total number of observations	1,887	1,954	857	1,885	1,953	854

Notes: Occupational distribution within each caste group is computed using employed individuals only. Employment for fathers and mothers is computed as of 1995. Statistics in columns 4–6 are reported for paternal grandfathers and maternal grandmothers. Working class = 1 if unskilled manual, skilled manual, organized blue-collar; 0 if clerical, business, professional. Standard errors in parentheses. Schooling and income statistics are computed using all parents in the sample, regardless of whether they are employed. Monthly income is measured in thousands of 1980 Rupees in the year closest to the year in which the child entered school.

Occupational categories

Unskilled manual: daily wage labor, deliveryman, servant, hotel worker, helper, cleaner/sweeper, porter, assistant watchman, fisherman, gardener, barber, cobbler (chambhar), unskilled laborer, seaman.

Skilled manual: machine operator, plumber, welder, technician, electrician, mechanic, carpenter, fitter/turner, tailor, painter, film developer, goldsmith, artist, priest, lab assistant, skilled worker, traditional healer (vaidhya), computer operator.

Organized blue collar: mill worker, factory worker, peon, Bombay Port Trust (BPT) worker, Bombay Electric Supply and Transportation (BEST) worker, Bombay Municipal Corporation (BMC) worker.

Petty trade: hawker, storeman (storekeeper), salesman, agent, shopkeeper.

Clerical: supervisor, driver, police, clerk, conductor, stenographer, postmaster, receptionist, foreman/draftsman, secretary.

Business: self business, medical representative, transporter, marketing, consultant, employer, contractor, politician (social worker/leader), merchant.

Professional: tutor, teacher, programmer, engineer, officer, manager, doctor, lawyer, nurse, lecturer, vice-chancellor, librarian, superintendant, director, principal, architect, salaried employee (service), chartered accountant, big businessman.

Farming: farmer, agricultural laborer.

jobs, across the generations within broad caste categories.[21]

Although most men are employed, we see that labor force participation (which includes part-time work) for the women in Table 2, panel B, is relatively low but is growing. Only 15 percent of high-caste grandmothers worked; whereas just over half of high-caste mothers entered the labor force (based on their employment status in 1995). Among the lower castes, the percentage employed remains stable at 20 percent across the generations, but notice that farming is listed as the primary occupation for a large number of working grandmothers. This suggests that urban employment must have increased sharply for the lower-caste women as well.

The occupational distribution across castes for the mothers in panel B, columns 1 to 3, displays a pattern similar to that for the fathers. Lower-caste women are much more likely to be employed in working-class occupations (44 percent and 32 percent) as compared with high-caste women (9 percent). There is an important difference, however, between men and women—although the large difference within the working-class occupations for the men was in access to blue-collar jobs for the lower castes, for women the major difference is in access to unskilled manual jobs; many of the lower-caste women work as sweepers and domestic servants.

Columns 1 to 3 and columns 4 to 6 in panel B suggest that there has been, in contrast to the men, significant intergenerational change in occupational patterns for women within castes. The urban occupations that show the greatest increase are skilled manual, clerical, and professional (with the exception of the high castes).[22] The decline in the percentage of working-class jobs among the lower-caste women, across a single generation, is particularly dramatic. This contrasts with the stability of the occupational distribution for the men, for all castes, that we noted earlier, consistent with the view that labor networks are weak among the women.

Together with the occupational distribution, Table 2 reports the mean years of schooling and monthly income separately by caste for men and women.[23] As expected, high-caste mothers and fathers have significantly more years of schooling and significantly higher incomes. Although the model in Section II assumes that the distribution of pre-schooling human capital is the same across castes (*jatis*), children in a wealthy, educated *jati* that has had access to white-collar jobs for many generations will be nurtured very differently from children in a *jati* that was historically confined to manual jobs. This suggests that pre-schooling human capital could vary in practice across broad caste categories, and across *jatis,* as well. When estimating the effect of the historical occupational distribution on the child's schooling choice, we will consequently take account of the possibility that the occupational distribution could be correlated with the ability distribution in the *jati.*

Table 3 indicates that, as assumed in the model, working-class occupations are associated with higher levels of networking (referrals).[24] Column 1 shows that 68 percent of the working-class men received help from a relative or member of the community in finding their first job (or starting their first business if self-employed), which is significantly higher

[21] The exception is farming, which is listed as the primary occupation for a large proportion of lower-caste grandfathers. This implies, in turn, that roughly one-quarter of the lower-caste fathers are first-generation migrants. Migrants are by definition newcomers in the labor market, and so will be more susceptible to the information problems that generate a need for the caste networks.

[22] The decline in the proportion of high-caste women in professional jobs is most likely because only the highest-ability women of the older generation (grandmothers) entered the labor force.

[23] Recall that income information was collected from each parent at five equal points in time from 1980 to 2000. We use the income (in 1980 Rupees) that coincides as closely as possible with the year in which the child entered school. Thus, the income in 2000 is used for students age 6 to 10, the income in 1995 for students 11 to 15, the income in 1990 for students 16 to 20, and the income in 1985 for students 21 to 25. The same income statistic is used later in the schooling regressions.

[24] The parents of the selected students were asked how they learned about their first job: through a childhood friend, through a college friend, through a relative, through a member of the community (*jati*), or by some other means (which was left open-ended in the questionnaire). This open-ended category included cases in which no help was received, or in which the job was found through newspaper advertisements, campus interviews, and other impersonal information channels. A binary referral variable was then constructed, taking the value of one if the parent learned about the first job from a relative or member of the community, and zero otherwise.

TABLE 3—REFERRALS AND SCHOOLING BY OCCUPATION

Relationship to student	Father		Mother	
Outcomes and choices	Percentage that received referrals	Percentage that studied in English	Percentage that received referrals	Percentage that studied in English
	(1)	(2)	(3)	(4)
Occupation				
Unskilled manual	65.95	0.80	61.29	0.00
Skilled manual	60.13	2.24	45.56	5.56
Organized blue-collar	76.43	0.91	69.44	5.56
All working class	68.44	1.36	57.69	2.24
(standard error)	(1.11)	(0.28)	(2.80)	(0.84)
Petty trade	57.89	1.75	61.76	2.94
Clerical	47.41	2.89	30.56	7.26
Business	49.29	8.53	41.86	9.30
Professional	32.77	11.38	29.25	14.47
All white-collar	43.76	6.20	30.64	10.13
(standard error)	(1.02)	(0.49)	(1.60)	(1.05)
Number of observations	4,515	4,513	1,215	1,215

Notes: Statistics are computed using employed individuals only. Farmers are excluded. A parent is said to have received a referral if a relative or member of the community found him/her a job. A parent is said to have studied in English if he/she studied in that language in secondary school. Occupational categories are defined in Table 2.

than the 44 percent of men in the white-collar jobs who received a referral. The corresponding statistics for the women in column 3 reveal essentially the same pattern, although the level of referrals for the women is generally lower than that for the men, perhaps because the networks for female jobs are less developed.

The model also assumes that Marathi schooling channels the student into a working-class job, while English schooling leads to the white-collar occupation. The survey elicited information on the language of instruction (English versus Marathi) for fathers and mothers, in secondary school. Columns 2 and 4 of Table 3 show that there is a clear distinction between working-class and white-collar jobs with respect to the language of instruction in secondary school. The percentage of men in working-class jobs that attended secondary school in English is just over 1 percent, compared with the 6 percent of men in white-collar jobs. In column 4, a similar pattern is obtained for the women. We have described the relationship between the broad occupational categories (working-class versus white-collar), the level of referrals, and English schooling. But inspection of Table 3 indicates that the level of referrals and English schooling vary systematically within these categories as well. Later, we will take advantage of this finer relationship between particular occupations, the level of referrals, and English schooling, to characterize the occupational distribution in the *jati.*

Table 2 suggests that lower-caste men and women are much more likely to hold working-class jobs. Combining these cross-caste patterns with the results in Table 3, it is not surprising that a much higher proportion of lower-caste men received referrals (60 percent versus 37 percent), and that these men are also much less likely to have been schooled in English (2 percent versus 12 percent). In contrast, although lower-caste women are also much less likely to be schooled in English, the level of referrals is statistically indistinguishable across castes.[25] The level of referrals is low in any case (13 percent for the lower castes and 19 percent for the high castes), especially when compared with the corresponding level for the men, and we will later establish that labor

[25] Although we noted earlier that lower-caste women who work are more likely to hold working-class jobs, which are associated with more referrals, we also saw that lower-caste women are less likely to enter the labor force. These two opposing effects appear to cancel each other, leaving little variation in the level of referrals across castes for the women.

market networks are effectively available for the men only.

IV. Empirical Analysis

A. *Specification and Identification*

The first implication of the model is that the occupational distribution in the *jati* should persist across generations when networks are active. Because schooling choice maps into occupational choice, equation (1) in Section II expressed this implication in terms of schooling choice in the current generation and the occupational distribution in the previous generation:

$$\Pr(E_{ij} = 1) = 1 - P_j.$$

Recall that $E_{ij} = 1$ if individual i belonging to *jati* j is schooled in English; $E_{ij} = 0$ if he is schooled in Marathi; and P_j is the proportion of men in the *jati* in the previous generation who are employed in working-class jobs and so in a position to provide referrals.

The particular relationship between E_{ij} and P_j in the equation above is, of course, a consequence of the modelling assumption that schooling choice maps perfectly into future occupational outcomes. More generally, we would expect to see a negative coefficient, but not necessarily with magnitude one, on P_j. The model laid out in Section II also does not allow for intergenerational state dependence in schooling choice at the level of the *household.* Moreover, we noted above that pre-schooling human capital and family incomes appeared to vary systematically across castes with different occupational backgrounds. The schooling regression that we estimate is consequently specified as

$$\Pr(E_{ij} = 1) = \alpha P_j + X_{ij}\beta + \omega_j, \tag{2}$$

where X_{ij} includes the parents' language of schooling to reflect household-level state dependence in schooling choice, as well as a cohort variable to capture the increase in the returns to English over successive cohorts in the current generation; ω_j measures unobserved or imperfectly observed pre-school human capital and family income in the *jati,* which could independently determine schooling choice.

P_j measures the proportion of men in the previous generation employed in working-class jobs. Although the model assumes that only two types of jobs—working-class and white-collar—are available, as many as 90 occupations are listed in the data. A relatively strong relationship between the level of referrals and the type of occupation was observed earlier in Table 3, and so one convenient statistic that accurately and parsimoniously describes the occupational distribution in the *jati* would be the proportion of fathers (the previous generation) who received a job referral. Working-class jobs were also associated with lower levels of English schooling (Table 3). An alternative measure of the occupational distribution in the previous generation would compute the proportion of fathers who attended English secondary schools. Most of the regressions reported in this paper will use the referrals statistic to measure P_j; English schooling levels were generally low in the previous generation and so there is substantially more variation in the referrals statistic across *jatis.* We will, however, verify that the results hold up with the English-schooling statistic as well.

Following the discussion above, we expect to find $\alpha < 0$ when networks are active and the occupational distribution persists across generations. Recall that α must also remain stable across cohorts in the current generation to explain the absence of convergence in Figure 3. Although much of the analysis treats α as constant, we will later verify that α does indeed remain stable across cohorts.

An identification problem arises when P_j and ω_j are correlated in equation (2). Although *jatis* might have been the same to begin with, we noted in the previous section that their members now have very different characteristics (income and education), depending on the type of occupation that the *jati* has historically been engaged in. A traditionally working-class *jati* could thus be associated with high P_j and low ω_j, in which case a family effect would be erroneously interpreted as a network effect because individuals with lower family resources independently select into Marathi schools.

Our solution to this identification problem exploits the fact, documented in Table 2, that networks are concentrated in working-class jobs

dominated by men. We mentioned earlier that the levels of referrals for women were relatively low, consistent with the significant change in the occupational distribution across generations for women indicated in Table 2. Thus, although the networks might affect schooling choice for the boys, they should have had little or no impact on the girls. The model in Section II then applies to boys only. Instead of using variation in the level of referrals across *jatis* to identify the presence of networks, as in equation (2), we proceed instead to exploit this gender difference in the access to job networks by pooling both sexes in the schooling regression to identify the presence of the network *within* the *jati:*

$$\Pr(E_{ij} = 1) = (\alpha - \tilde{\alpha})P_j \cdot B_{ij} + X_{ij}\tilde{\beta} + X_{ij} \cdot B_{ij}(\beta - \tilde{\beta}) + \gamma B_{ij} + f_j, \quad (3)$$

where $\tilde{\alpha}$, $\tilde{\beta}$ represent the effect of the network and parents' language of schooling on the girls. B_{ij} is a dummy variable that takes a value of one for boys and zero for girls. The advantage of pooling the boys and girls is that the schooling regression can be estimated with *jati* fixed effects, $f_j \equiv \tilde{\alpha}P_j + \omega_j$. Although we can no longer identify α directly, we can obtain a consistent estimate of $\alpha - \tilde{\alpha}$, the coefficient on the $P_j \cdot B_{ij}$ interaction term. For the special case with exclusively male networks, $\tilde{\alpha} = 0$ and the coefficient on the interaction term identifies network-based occupational persistence for the boys directly. More generally, the coefficient on the interaction term provides a conservative estimate of the effect of caste-based networks on schooling choices for the boys.

The identifying assumption in this estimation strategy is that no variable $\phi_j \cdot B_{ij}$ appears in the residual of equation (3), where ϕ_j is correlated with P_j. A sufficient condition for this identifying assumption to be satisfied is that no unobserved determinant of schooling choice should vary by gender or have a differential effect on schooling choice by gender, *within* the *jati.* Later in Section V we will discuss alternative explanations for the negative and significant $\alpha - \tilde{\alpha}$ coefficient we obtain in the schooling regression. These explanations either relax the assumptions of the model, made earlier in Section II, or build on the failure of the identifying assumption. We will argue that none of these explanations fits the data quite as well as the male labor market network explanation we put forward in this paper.

B. *Caste-Based Networks and Schooling Choice*

Table 4, column 1, reports the estimates of the schooling choice regression, equation (2), for the boys. As noted, the sample covers 20 cohorts of students age 6 to 25, who entered school between 1982 (cohort = 1) and 2001 (cohort = 20). The student's cohort (1 to 20), the proportion of fathers in his *jati* who received a referral, and the father's and the mother's language of instruction in secondary school are included as regressors.

The cohort term is included in this regression to account for the increase in the returns to English over time. While the linear cohort effect we specify in Table 4 is clearly restrictive, we verify below that the estimated referral coefficient is unchanged when we allow for more flexible cohort effects. The referral coefficient is also specified to be constant over time in Table 4, and we will subsequently relax this restriction as well. For now, we see that the referral coefficient is negative and significant; children belonging to (historically) working-class and more heavily networked *jatis* are less likely to be schooled in English, consistent with the first implication of the model. The cohort effect is positive and significant, implying a shift into English over time, which is consistent with the increase in the returns to English we saw in Figure 1. Finally, the results imply that a boy is much more likely to be schooled in English if his parents were educated in that language, indicating significant state dependence in schooling choice at the level of the household.

Table 4, column 2, reports estimates from a specification that includes variables that determine the student's pre-schooling human capital as well as the household budget constraint, which could independently determine schooling choices. The parents' years of education, conditional on their language of instruction in secondary school and the level of referrals in the *jati,* are likely significant determinants of children's pre-schooling human capital. The family's access to own resources is measured by the

TABLE 4—CASTE-BASED NETWORKS AND SCHOOLING CHOICE

Dependent variable	English schooling					
Sample	Boys only		Girls only		Boys and girls	
	(1)	(2)	(3)	(4)	(5)	(6)
Referrals	−1.060	−0.377	−0.646	0.124	—	—
	(0.164)	(0.148)	(0.160)	(0.167)		
Referral - boy	—	—	—	—	−0.398	−0.464
					(0.091)	(0.105)
Cohort	0.013	0.009	0.013	0.009	0.017	0.010
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Father studied in English	0.320	0.236	0.388	0.309	—	0.301
	(0.037)	(0.033)	(0.037)	(0.026)		(0.026)
Mother studied in English	0.351	0.220	0.441	0.269	—	0.259
	(0.041)	(0.028)	(0.071)	(0.045)		(0.043)
Father's years of education	—	0.023	—	0.020	—	0.021
		(0.004)		(0.003)		(0.003)
Mother's years of education	—	0.023	—	0.026	—	0.024
		(0.003)		(0.003)		(0.003)
Family income	—	0.005	—	0.009	—	0.007
		(0.005)		(0.003)		(0.003)
Boy	—	—	—	—	0.270	0.297
					(0.049)	(0.077)
Cohort - boy	—	—	—	—	−0.002	−0.001
					(0.002)	(0.002)
Father studied in English - boy	—	—	—	—	—	−0.091
						(0.044)
Mother studied in English - boy	—	—	—	—	—	−0.044
						(0.042)
Father's years of education - boy	—	—	—	—	—	0.002
						(0.005)
Mother's years of education - boy	—	—	—	—	—	−0.001
						(0.004)
Family income - boy	—	—	—	—	—	−0.003
						(0.005)
R^2	0.173	0.274	0.146	0.272	0.163	0.299
Number of observations	2,405	2,286	2,228	2,093	4,635	4,379

Notes: Standard errors in parentheses are robust to heteroskedasticity and clustered residuals within each *jati.* English schooling = 1 if the child is sent to an English school, 0 if the child is sent to a Marathi school. Referrals measures the proportion of fathers in the *jati* who received a referral. Boy = 1 if the student is a boy, 0 if girl. Family income is measured in thousands of 1980 Rupees in the year that is closest to the year in which the child entered school. Columns 1–2: schooling choice for boys. Columns 3–4: schooling choice for girls. Columns 5–6: schooling choice for both boys and girls, including a full set of *jati* dummies.

total income of the father and the mother at the time when the child entered school.[26] Inclusion of these variables results in a substantial decline in the referral coefficient, suggesting that the level of referrals was previously proxying to some extent for unobserved, family-specific determinants of schooling choice, but it remains negative and significant. The coefficient on the cohort variable is quite stable. And the coefficients on the additional regressors all have sensible signs; the boy is more likely to be schooled in a more-expensive English-medium school if his father or mother are more educated, or if the family is wealthier.

The estimates for girls are reported in columns 3 and 4 in Table 4. Column 3 reports the estimates based on equation (2); column 4 reports the estimates from the augmented specification that adds the parents' years of schooling and family income as additional regressors. The estimated cohort effects, and the coefficients on

[26] We use the income in 2000 for students currently age 6 to 10, the income in 1995 for students 11 to 15, the income in 1990 for students 16 to 20, and the income in 1985 for students 21 to 25. All incomes are computed in 1980 Rupees.

parents' language of instruction, parents' education, and family income are similar to those for boys in columns 1 and 2. The referral coefficient, however, becomes negligible for the girls in column 4 once the observed determinants of pre-schooling human capital and access to own family resources are included. One explanation for this result is that girls receive help from the women, not the men, in their *jati.* But we noted that the level of referrals for the women is very low, across all castes. Although not reported, we also found no correlation between referrals and schooling choice, for both boys and girls, when we replaced the level of referrals for the fathers with the level of referrals for the mothers.

The results we have just described are consistent with the view that caste-based networks, net of individual and family characteristics, affect schooling decisions for the boys, but not for the girls. But up to this point, we have controlled only for unobserved ability with a limited number of family characteristics. A more robust identification strategy estimates the schooling regression with *jati* fixed effects, as in equation (3). These estimates are reported in column 5 of Table 4. As noted, only the referral-boy interaction coefficient, and not the linear referral coefficient, can now be identified. The coefficient on this term is negative and significant, and very similar to the referral coefficient for the boys in column 2. Recall from equation (3) that the coefficient on the referral-boy interaction term provides us with a direct estimate of the referral coefficient for the boys if the referral coefficient for the girls is zero. The result we obtained earlier for the girls, in column 4, suggests that this might well be the case.

The regression specification with *jati* fixed effects in Table 4, column 5, did not include family characteristics. Column 6 includes parents' language of schooling, parents' years of schooling and family income, both interacted and uninteracted with the boy dummy, as additional regressors. Including uninteracted family characteristics in the schooling regression has no effect on the estimated referral-boy coefficient by construction, once *jati* fixed effects are included. But we see that the inclusion of the family characteristics, interacted with the boy dummy, has no effect on the estimated referral-boy coefficient as well. Indeed, this coefficient is no smaller than the corresponding coefficient estimated earlier in column 5 without *any* household characteristics. This stability contrasts with the decline in the referral coefficient in Table 4, columns 1 to 4, when family characteristics were included, providing some support for the view that the *jati* fixed effects absorb much of the unobserved heterogeneity in this environment.[27]

Notice also that parents' years of schooling and family income, which had a strong influence on schooling choice for both boys and girls in columns 1 to 4, do not differentially affect schooling choice by gender (column 6). It is only the *jati*-level referral variable that has such a differential effect on schooling choice, as measured by the negative and significant referral-boy coefficient. This observation will be useful later in Section V when we consider alternative explanations for the results presented in this paper.

C. *Schooling Choice over Time*

The second implication of the model laid out in Section II is that the relationship between schooling choice and the occupational distribution in the previous generation will weaken across successive cohorts in the current generation as the returns to English grow, unless restrictions on occupational mobility are in place. To assess empirically the stability of the referral coefficient, we create 4 cohort categories that evenly divide the 20 cohorts, and then estimate the referral coefficient separately for each category.

We begin with a benchmark *jati*-fixed-effects regression, which maintains a constant referral coefficient but relaxes the restriction imposed

[27] A previous version of the paper (Munshi and Rosenzweig, 2003) reported a number of additional robustness tests. First, we accounted for occupational persistence at the level of the family by including a full set of (90) dummies for the student's father's occupation. Second, we allowed for the possibility that the scope of the network was determined by caste and the region of origin (within Maharashtra) by replacing the *jati* by the *jati*-region as the boundary of the network. Third, we dropped very large *jati*-regions (more than 250 observations) and very small *jati*-regions (fewer than 10 observations). The estimated referral-boy coefficient with these alternative specifications was shown to be very similar to what we report in Table 4.

TABLE 5—SCHOOLING CHOICE OVER TIME

Dependent variable	English schooling			
Additional regressors	Without family characteristics	With family characteristics	Without family characteristics	With family characteristics
	(1)	(2)	(3)	(4)
Referral - boy	−0.426	−0.478	—	—
	(0.090)	(0.106)		
Referral - boy - cohort1	—	—	−0.269	−0.416
			(0.168)	(0.167)
Referral - boy - cohort2	—	—	−0.352	−0.333
			(0.100)	(0.112)
Referral - boy - cohort3	—	—	−0.523	−0.540
			(0.145)	(0.143)
Referral - boy - cohort4	—	—	−0.607	−0.665
			(0.256)	(0.238)
Cohort 1	−0.261	−0.161	−0.261	−0.161
	(0.031)	(0.032)	(0.030)	(0.032)
Cohort 2	−0.231	−0.146	−0.231	−0.146
	(0.031)	(0.028)	(0.031)	(0.028)
Cohort 3	−0.161	−0.121	−0.161	−0.121
	(0.030)	(0.023)	(0.030)	(0.023)
Boy	0.236	0.261	0.338	0.364
	(0.065)	(0.091)	(0.156)	(0.149)
Cohort 1 - boy	0.033	0.031	−0.152	−0.106
	(0.038)	(0.037)	(0.209)	(0.169)
Cohort 2 - boy	0.052	0.031	−0.090	−0.153
	(0.042)	(0.035)	(0.174)	(0.151)
Cohort 3 - boy	0.041	0.041	−0.007	−0.030
	(0.032)	(0.024)	(0.117)	(0.114)
R^2	0.164	0.301	0.164	0.301
Number of observations	4,635	4,379	4,635	4,379

Notes: Standard errors in parentheses are robust to heteroskedasticity and clustered residuals within each *jati.* English schooling = 1 if the child is sent to an English school, 0 if the child is sent to a Marathi school. Referrals measures the proportion of fathers in the *jati* who received a referral. Boy = 1 if the student is a boy, 0 if girl. Cohort 1: age 21–25; Cohort 2: age 16–20; Cohort 3: age 11–15; Cohort 4: age 6–10. Column 2 and column 4 include family characteristics, separately and interacted with the boy dummy. Family characteristics include parents' language of schooling and years of education, and total family income. A full set of *jati* dummies is included in all regressions. Sample includes boys and girls.

thus far that cohort effects are linear, by including the cohort categories in Table 5, column 1. The estimated negative referral-boy coefficient is unaffected by the inclusion of the flexible cohort effect and remains very similar to the results shown in Table 4. Inclusion of the family background variables, uninteracted and interacted with the boy dummy, as additional regressors again has no effect on the estimated referral coefficient (column 2).

Table 5, column 3, allows for changes in the referral coefficient across cohort categories. All the referral-boy-cohort coefficients are negative and significant except for the coefficient on the first cohort category, which is slightly less precisely estimated. The referral coefficient is actually increasing for the later cohorts, and we can easily reject the convergence hypothesis which implies a decline in the referral effect over time. Once more, the estimated referral coefficients are robust to the inclusion of the family background variables as regressors (column 4). Although not reported here, the referral coefficient remained stable when the schooling regression was estimated with boys only, including parents' years of education and family income as additional regressors. It is this *jati*-level effect that presumably sustains the gap in schooling choice between broad caste categories observed for the boys in Figure 3.

TABLE 6—ALTERNATIVE MEASURES OF THE OCCUPATIONAL DISTRIBUTION AND SCHOOLING

Dependent variable	English schooling				Test scores		
Occupational distribution measure	Proportion of fathers schooled in English				Proportion of fathers that received a referral		
Sample	Boys only	Girls only	Boys and girls		Boys only	Girls only	Boys and girls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Occupational distribution	0.847	0.083	—	—	−23.151	−23.650	—
	(0.262)	(0.427)			(5.045)	(4.080)	
Occupational distribution - boy	—	—	0.701	0.869	—	—	−0.734
			(0.224)	(0.221)			(5.761)
Cohort	0.008	0.010	0.017	0.010	−0.505	−0.180	−0.190
	(0.002)	(0.002)	(0.002)	(0.002)	(0.134)	(0.204)	(0.223)
Boy	—	—	0.026	−0.006	—	—	3.794
			(0.028)	(0.031)			(4.357)
Father studied in English	0.217	0.301	—	0.313	4.901	2.323	1.847
	(0.032)	(0.027)		(0.026)	(1.397)	(3.711)	(4.028)
Mother studied in English	0.220	0.266	—	0.259	3.312	−2.596	−2.772
	(0.030)	(0.046)		(0.042)	(2.200)	(1.905)	(1.632)
Father's years of education	0.024	0.019	—	0.020	0.929	0.765	0.812
	(0.004)	(0.002)		(0.003)	(0.195)	(0.225)	(0.238)
Mother's years of education	0.023	0.025	—	0.023	0.617	1.074	0.984
	(0.003)	(0.003)		(0.003)	(0.221)	(0.199)	(0.200)
Family income	0.005	0.008	—	0.007	0.260	0.122	0.107
	(0.004)	(0.003)		(0.003)	(0.118)	(0.076)	(0.076)
R^2	0.275	0.272	0.162	0.298	0.322	0.334	0.354
Number of observations	2,286	2,093	4,635	4,379	849	775	1,624

Notes: Standard errors in parentheses are robust to heteroskedasticity and clustered residuals within each *jati.* The sample in columns 5–7 is restricted to children in cohorts 1–10, past the school-leaving age, who passed the SSC exam. Test scores range from 35 to 100. Boy = 1 if the student is a boy, 0 if girl. Family income is measured in thousands of 1980 Rupees in the year closest to the year in which the child entered school. Column 3 also includes cohort interacted with boy. Column 4 and column 7 also include cohort, father/mother studied in English, father's/mother's years of education, and family income, interacted with boy. Regressions pooling boys and girls (columns 3–4 and 7) include a full set of *jati* dummies.

D. *Robustness and Validation: Alternative Measures of the Occupational Distribution and Schooling*

The regression results reported thus far used the proportion of fathers who received a referral for their first job to measure the occupational distribution. We now proceed to verify the robustness of the results by repeating the schooling regressions for boys, girls, and the pooled sample with the proportion of fathers schooled in English as the measure of the occupational distribution.

The coefficients on the cohort variable and the household characteristics in Table 6, columns 1 to 2, are very similar to the estimates reported in Table 4. The coefficient on the English proportion, which can be interpreted as state dependence in schooling choice at the *jati* level, is positive and significant as expected. The coefficient on the English-boy interaction term in Table 6, columns 3 to 4, which include *jati* fixed effects, is very similar to the English coefficient for the boys in column 1, matching the results reported earlier with the referrals variable. Although not reported, once again parents' education and family income do not have a differential effect on schooling choice by gender.

The language of instruction measures the child's future occupation, and the referral statistic measures the occupational distribution in the previous generation, in most of the regressions that we report in this paper. The negative and significant referral-boy coefficient in the fixed effects regressions then reflects the persistence in the occupational distribution across generations, differentially for boys and girls within the *jati.* To validate this interpretation of the results, we proceed to replace the language

of instruction with an alternative schooling outcome as the dependent variable.

We assume that test scores depend on school quality and the pre-schooling human capital of the student. The comparison of English and Marathi schools in Table 1, and the parents' perception of these schools, indicates that school quality does not vary by the language of instruction. Under the maintained assumption that pre-schooling human capital does not vary by gender within the *jati,* this implies that the referral-boy coefficient should be close to zero in the fixed effects regression with test scores as the dependent variable. The fact that referrals have a differential effect by gender on schooling choice should be irrelevant for test scores, if school quality does not vary by the language of instruction.

Columns 5 to 7 of Table 6 replace the language of instruction with performance on the school-leaving SSC examination as the dependent variable. Referrals are once more used to measure the occupational distribution in the *jati,* to be consistent with the specifications used elsewhere in the paper. We restrict attention to the first ten cohorts (age 16–25), which have already attained school-leaving age, in these regressions. Only 17 percent of the students age 16–25 in the sample never passed the SSC examination, so we focus on the test score conditional on having passed the exam in these regressions.[28]

Table 6, column 5, restricts attention to boys, and includes the cohort, family characteristics, and the level of referrals in the *jati* as regressors. The cohort effect is negative and significant, suggesting a decline in the quality of students over time. The referral coefficient is also negative and precisely estimated, which would be the case if students from high-referral *jatis* have lower levels of pre-schooling human capital. Consistent with this interpretation, family characteristics, particularly parents' years of education, have a very large positive effect on test performance. Subsequently we repeat the exercise just described for the girls (Table 6, column 6). The cohort effect is now absent, but the coefficient on referrals remains negative and statistically significant—both boys and girls from high-referral *jatis* do less well on exams. The fixed-effects estimates, reported in Table 6, column 7, of the cohort effect, the cohort-boy interaction, and the boy dummy are not statistically significantly different from zero. More importantly, the coefficient on the referral-boy interaction term is small and statistically insignificant, in contrast to the specifications with language of instruction as the dependent variable. Caste networks affect the language but not the quality of instruction of their members.

E. *Selection into Marathi Schools over Time*

The framework laid out in Section II also has implications for the compositional change in the students who attend Marathi schools over time by *jati:* first, the pre-schooling human capital of boys entering Marathi schools should decline on average as the returns to English grow. Second, when there are no restrictions on mobility put in place to exploit network externalities, the distribution of pre-schooling ability among the boys entering Marathi schools will converge across all *jatis* over time. It is possible that such convergence across *jatis* will be absent when restrictions are in place. Note that the model has no prediction for selection by ability into English schools.

We do not have a direct measure of pre-schooling human capital. The results in Table 6 suggest, however, that, net of income, parental schooling has a positive and significant effect on school performance. In particular, father's schooling has a significant positive effect on test scores for boys and girls, and the effects do not differ significantly by the gender of the child. We thus use the father's schooling level as a proxy for pre-schooling ability.[29] The question we address is whether boys with more educated fathers increasingly exit Marathi schools and whether, and how, the rate of decline in the pre-schooling ability of boys entering Marathi schools varies by *jati.*

[28] Munshi and Rosenzweig (2003) studied the effect of referrals on the probability of success in the SSC exam and obtained results that are qualitatively the same as what we report below with the test score, conditional on success, as the dependent variable.

[29] The results reported below are essentially the same if we replace father's schooling by mother's schooling.

TABLE 7—SELECTION INTO MARATHI SCHOOLS OVER TIME

Dependent variable	Father's years of education									
Cohort	11–20					1–10				
Sample	Boys		Girls		Boys and girls	Boys		Girls		Boys and girls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Cohort	−0.697	−0.577	0.423	0.219	—	0.153	0.240	−0.543	−0.087	—
	(0.230)	(0.155)	(0.230)	(0.189)		(0.191)	(0.146)	(0.326)	(0.156)	
Cohort - boy	—	—	—	—	−0.792	—	—	—	—	0.117
					(0.252)					(0.232)
Referral - cohort	1.430	1.204	−0.596	−0.280	—	−0.147	−0.258	1.054	0.310	—
	(0.400)	(0.260)	(0.376)	(0.311)		(0.323)	(0.250)	(0.554)	(0.270)	
Referral - cohort - boy	—	—	—	—	1.469	—	—	—	—	−0.231
					(0.414)					(0.415)
Referrals	−30.991	—	−3.651	—	—	−10.256	—	−18.624	—	—
	(6.894)		(6.866)			(2.035)		(2.793)		
R^2	0.106	0.205	0.138	0.205	0.215	0.136	0.254	0.184	0.278	0.285
Number of observations	839	839	815	815	1,654	866	866	851	851	1,717

Notes: Standard errors in parentheses are robust to heteroskedasticity and clustered residuals within each *jati.* Referrals measures the proportion of fathers in the *jati* who received a referral. Column 2, column 4, column 7, and column 9 include a full set of *jati* dummies. Column 5 and column 10 include a full set of *jati* dummies, *jati*-boy dummies, and *jati*-cohort dummies. All regressions are restricted to students in Marathi schools.

To test the implications for school selectivity described above, we estimate regressions on the subsample of boys entering Marathi schools of the form

$$(4)\quad E(S_{ij}|E_{ij}=0) = \kappa + \lambda R_j + \mu C_{ij} + \nu R_j \cdot C_{ij} + \psi\omega_j,$$

where S_{ij} is boy i in *jati* j's father's years of schooling; C_{ij} is the boy's cohort; R_j measures the level of referrals in the *jati;* and ω_j measures pre-schooling ability in the entire *jati.* These terms reflect the fact that pre-schooling ability conditional on selection into Marathi is, in general, a function of ability in the *jati* and the level of referrals. The cohort terms reflect the change in this selection process over time. The model, which ignores the variation in ability ω_j across *jatis,* predicts that $\lambda > 0$, $\mu < 0$, $\nu < 0$ without restrictions, and (possibly) $\nu > 0$ with restrictions.

Because the major shift into English schooling occurred in the 1990s, we first estimate equation (4) for the boys in cohorts 11–20. The estimates are reported in Table 7, column 1. The cohort coefficient is negative and significant as predicted, which implies that the pre-schooling human capital of the boys who entered Marathi schools was declining substantially in the 1990s. The coefficient on the referral-cohort interaction term is positive, consistent with the results in Table 5 showing that restrictions on mobility in the high-referral networks were still in place during this period—the more-able boys in high-referral *jatis* were shifting to English-medium schools at lower rates.

The referral coefficient is negative and significant, consistent with the results in Table 6, which indicate that *jatis* with higher referrals R_j have lower ability ω_j; the negative $R_j - \omega_j$ correlation appears to dominate the positive selection $\lambda > 0$ effect in this case. But this tells us that the positive referral-cohort coefficient that we reported above might also be spurious. To assess the robustness of the results in column 1, we add *jati* fixed effects, which subsume $\kappa + \lambda R_j + \psi\omega_j$ (Table 7, column 2). The referral coefficient λ is no longer identified, but the estimated cohort and referral-cohort coefficients are very similar to the results in column 1.

The within-*jati* estimates allow ability to vary across *jatis* but assume that ability is constant over time (both within and across generations). The level of parental schooling could, however, also depend on the access to education, which might have changed over time. If there was convergence in the access to education across *jatis* in the parent generation, that could explain the positive referral-cohort coefficient in columns 1 to 2 without requiring networks to be active. One test to rule out this alternative interpretation of our result would be to estimate the school selectivity regression for girls rather than boys; we have already seen that the network has no effect on schooling choice for the girls, and so both the cohort and the referral-cohort effect should be absent. In contrast, if the referral-cohort term is picking up convergence in (fathers') schooling levels across *jatis,* then this coefficient should be positive and significant for the girls as well.

Table 7, column 3, reports the basic selectivity regression for the girls attending Marathi schools with cohort, referral-cohort, and referrals included as determinants of father's schooling, while Table 7, column 4, repeats this regression with *jati* fixed effects. The referral coefficient in column 3 is again negative (but insignificant), consistent with the lower levels of ability in high-referral *jatis.* The cohort coefficient is positive but insignificant. More importantly, the referral-cohort coefficient is small in magnitude and statistically insignificant—the point estimate is actually negative—and consistent with the results obtained earlier that girls in families belonging to high-referral *jatis* are not restricted in their mobility.

An alternative strategy to control for the confounding effect of changes in access to schooling among the fathers across *jatis* and over time pools boys and girls in the selectivity regression, which can then be estimated with a full set of *jati* dummies interacted with the cohort variable

$$(5)\quad E(S_{ij}|E_{ij}=0) = (\mu-\tilde{\mu})C_{ij}\cdot B_{ij} + (\nu-\tilde{\nu})R_j\cdot C_{ij}\cdot B_{ij} + f_j + g_j\cdot B_{ij} + h_j\cdot C_{ij},$$

where $\tilde{\mu}$, $\tilde{\nu}$ are the coefficients on the cohort variable and the referral-cohort interaction for the girls, and B_{ij} is a boy dummy as before. The fixed effects, f_j, which subsume $\tilde{\kappa} + \tilde{\lambda}R_j + \psi\omega_j$, allow for the possibility that ability varies across *jatis.* The fixed effects interacted with the boy dummy $g_j \cdot B_{ij}$, which subsume $(\kappa - \tilde{\kappa})B_{ij} + (\lambda - \tilde{\lambda})R_j \cdot B_{ij}$, also allow ability to vary by gender across *jatis.* Finally, the fixed effects interacted with the cohort variable $h_j \cdot C_{ij}$, subsume $\tilde{\mu}C_{ij} + \tilde{\nu}R_j \cdot C_{ij}$ and control for changes in access to schooling for the fathers both across *jatis* and over time.

For the special case with $\tilde{\mu} = 0$, $\tilde{\nu} = 0$, as is consistent with the model, the estimated coefficients in equation (5) should match the cohort coefficient and the referral-cohort coefficient when equation (4) is estimated with *jati* fixed effects for boys only. Table 7, column 5, suggests that this is indeed the case: the cohort-boy coefficient is negative and significant, the referral-cohort-boy coefficient is positive and significant, and the point estimates are very similar to the corresponding coefficients in columns 1 and 2. These results confirm that in the most heavily networked *jatis,* high-ability girls were exiting to English-medium schools at significantly faster rates than were boys.[30] The 0.1 quantile–0.9 quantile of the referrals distribution ranges from 0.2 to 0.7. The point estimates in column 5 thus suggest that over the period of the 1990s the gap in father's schooling between boys and girls schooled in Marathi grew by 2.3 years in the highest-referral *jatis* (at the 0.9-quantile level). In contrast, the ability-differential measured by the difference in the father's schooling between boys and girls, declined by as much as five years in the low-referral *jatis* (at the 0.1-quantile level) over the same period. This increasing mismatch in ability levels between the sexes within *jatis* and school types could have important implications for the future stability of the caste system, which relies on endogamous marriage, as discussed below.

Columns 6 to 10 of Table 7 report the estimates of the selectivity equations for the first ten cohorts of students, who entered school in

[30] The negative cohort-boy coefficient implies that the boy-girl pre-schooling human capital differential is declining over time, independent of the influence of the male job network. This may be due to differences in labor force participation or changes in the returns to English by gender (as in Figures 1 and 2).

the 1980s. Schooling choices were stable over this period and thus we do not expect to find changing selectivity effects for the boys or the girls. As before, the referral coefficient, in column 6 and column 8, is negative and significant, reflecting the persistent differences in ability across *jatis.* As expected, however, and in contrast to the cohorts making schooling choices in the post-1990s new economy, the cohort effect and the referral-cohort effect, both uninteracted and interacted with the boy dummy, are insignificant in the pre-reform period.

F. *Alternative Interpretations of the Empirical Results*

The discussion that follows considers alternative explanations for the results we have presented. The identifying assumption in the fixed effects schooling choice regression is that unobserved determinants of schooling choice that are correlated with the occupational distribution should not vary by gender, or have a differential effect by gender on schooling choice, within the *jati.* Some of the alternative explanations we pursue are associated with the failure of this identifying assumption. Other explanations are generated by relaxing the assumptions of the model. We will argue that none of these alternative explanations matches all the results as well as our preferred explanation, based on underlying male labor market networks.

Liquidity Constraints.—The model assumes that schooling choices are based entirely on the individual's ability and the historical occupational distribution in his *jati,* which determines the labor market network that he inherits. When credit markets function imperfectly, liquidity constraints could, in addition, prevent individuals belonging to working class *jatis* from choosing more expensive English schooling. Liquidity does not vary by the gender of the child *within* the *jati,* and so the *jati* fixed-effects regression would appear to rule out this alternative explanation. Boys and girls have different labor market opportunities, however, and it is thus conceivable that liquidity could have a differential effect on schooling choice by gender. The schooling regression accounts for liquidity constraints by including family income at the time the child entered school. Schooling expenses for the children in our sample are relatively low (6.3 percent of family income in English schools and 6.0 percent in Marathi schools), and, not surprisingly, family income has a relatively weak effect on schooling choice for both boys and girls. Moreover, the effect of family income on school choice does not differ by gender at conventional levels of significance (Table 4).

Differences in Ability.—The *jati*-level fixed effects absorb all variation in the *jati* that is not gender specific. But in an economy where men and women historically performed very different roles, the parental and societal inputs that boys and girls received in childhood might have been very different. The results reported earlier, however, provide no evidence of gender distinctions in pre-schooling human capital within households or *jatis.* The estimates reported in Table 4 do not reject the hypothesis that the effects of parental human capital characteristics on school choice are equal for boys and girls. The results reported in Table 6 with test performance as the dependent variable are also consistent with the assumption in the fixed effects schooling regression that pre-schooling human capital does not vary by gender within the *jati.*

Discrimination.—Unless there is a gender-based component to caste discrimination, it will be subsumed entirely by the *jati* fixed effects. But it is possible that firms or schools discriminate against boys from working-class backgrounds, perhaps because they are difficult to discipline, while treating girls from different backgrounds more equally. The referral-boy coefficient would proxy for underlying discrimination in that case.

Recall that household characteristics, such as parental education and family income, had the *same* effect on schooling choice for boys and girls within the *jati.* It was only the *jati*-level referrals statistic that had a gender-specific effect on schooling. If these results are attributed to discrimination, then it implies that firms or schools do not discriminate by family background *within* the *jati,* but by *jati* affiliation alone. It is not obvious why we would expect to see gender discrimination purely along caste lines. Family characteristics, such as parental education and income, were seen to be correlated with pre-schooling human capital and are

at least as easy to observe as caste identity. Historically there does not appear to have been a policy of caste discrimination by employers in any industry in Bombay in any case (Morris, 1965).

Restrictions on High-Caste Women.—We focused in Figure 3 and Figure 4 on the absence of convergence for the boys, which was attributed to restrictions on occupational mobility among the lower castes. An alternative interpretation of these figures is that the ability distribution varies across the population such that it remains optimal for individuals to sort by caste into different careers, even as the returns to English grow. The convergence among the girls with this alternative interpretation is attributed to restrictions on the *high-caste* girls.

There is no evidence that such restrictions are in place, or have been in place historically. High-caste women in Bombay have always had higher labor-force participation rates and more English schooling than lower-caste women, as observed in Table 2. Within the high castes, boys are substantially more likely to be schooled in English than girls, and so the girls could easily switch into English schools without creating a mismatch on the marriage market.

Moreover, although the pooled schooling choice regression with fixed effects cannot distinguish between the alternative explanation, based on female restrictions, and our view that male networks shape schooling choice for the boys alone, recall that we also ran regressions separately for boys and girls. The *jati*-level statistic, measured either by the proportion of referrals or the proportion of fathers with English schooling, affects schooling choice for the boys but not for the girls in these regressions. We thus appear to be picking up restrictions on mobility that are specific to the boys.

Sampling Bias.—We noted three potential sources of sampling bias in Section IIIA. First, particular households might school their children outside the Dadar area. Second, particular households might have moved from Dadar over the past 20 years. Third, children from particular households might have dropped out of school.

The first two sources of sampling bias are easily accommodated in the fixed-effects regression framework. Although school locations and out-migration might vary by *jati,* there is no reason to expect these decisions to vary by the gender of the child *within* the *jati.* The third source of sampling bias is potentially more problematic, because drop-outs could vary by gender within the *jati.* The decision to drop out would depend on the child's pre-schooling human capital and future employment opportunities, both of which determine schooling choice. Selective dropouts, by gender across *jatis,* could consequently violate the identifying assumption underlying the fixed-effects estimation procedure.

However, the sex-ratio of students in the most recent eight cohorts (grades one through eight) in which there would be relatively few dropouts is statistically indistinguishable from the sex-ratio in the older 12 cohorts. Regressions not reported here also reveal that the sex-ratio is uncorrelated with the level of referrals in the *jati,* both in the first 12 cohorts and in the 8 most recent cohorts.[31]

V. Conclusion

As modernization proceeds around the world, there is a perception that indigenous existing institutions may importantly shape the course of the development process across different countries. Yet little is known about how such institutions actually affect the transformation of economies undergoing change, or their impact on the economic mobility of particular groups of individuals. This paper examines the role of one long-standing traditional institution—the Indian caste system—in shaping career choices by gender in a rapidly globalizing economy.

We have found that male working-class networks, organized at the level of the subcaste or *jati,* continue to channel boys into traditional occupations, despite the fact that returns to nontraditional (white-collar) occupations have risen substantially during the post-1990s reform period. In contrast, girls, who have had historically low labor-market participation rates and few network ties to

[31] As an additional test, we also verified that the sex-ratio in the most recent cohort that entered school in 2001 is uncorrelated with the level of referrals in the *jati.* Thus, there does not appear to be selective enrollment by gender and *jati* either.

constrain them, appear to be taking full advantage of the opportunities that have become available in the new economy. It is generally believed that the benefits of globalization have accrued disproportionately to the elites in developing countries. In this setting we find, instead, that a previously disadvantaged group (girls) might surpass boys in educational attainment and employment outcomes in the future in the most heavily networked *jatis.*

Although we have focused on how traditional institutions shape the responses of particular groups of individuals to the new opportunities that accompany globalization, our findings suggest that these institutions are likely to be affected in turn by the forces of change. In our framework, an individual schooled in English no longer needs the traditional caste network; indeed, it has been remarked that "the English educated form a caste by themselves" (M. P. Desai, quoted in Julian Dakin et al., 1968 p. 24). Simple statistics on marriage and migration that we computed for the elder siblings of the students in our sample would appear to support the view that English education will ultimately undermine the caste network. Among the 825 married siblings in our sample, 11.9 percent married outside their *jati.* This contrasts with the parent generation, in which only 3.7 percent of the partners were not members of the same *jati.* Schooling in English appears to be contributing to this increase in inter-caste marriage, as 31.6 percent of the English-educated siblings married outside their *jati,* versus only 9.7 percent of the Marathi-educated siblings. And among the 1,073 siblings who are currently employed, 13.9 percent of the English-educated work outside Maharashtra, versus only 2.1 percent of the Marathi-educated (these differences between the Marathi-educated and the English-educated are statistically significant at the 5-percent level). Both marriage outside the *jati* and out-migration weaken caste ties and the caste network. Increasing exposure to the modern economy through English education, and the mismatch in educational choices and future occupational outcomes between boys and girls in the same *jati* that we have documented, suggest that the forces of modernization could ultimately lead to the disintegration of a system that has remained firmly in place for thousands of years.

REFERENCES

Acemoglu, Daron; Johnson, Simon and Robinson, James A. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review*, 2001, *91*(5), pp. 1369–1401.

Banerjee, Abhijit and Iyer, Lakshmi. "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *American Economic Review*, 2005, *95*(4), pp. 1190–213.

Burnett-Hurst, Alexander R. *Labour and housing in Bombay: A study in the economic conditions of the wage-earning classes of Bombay*. London: P. S. King and Son, 1925.

Chandavarkar, Rajnarayan. *The origins of industrial capitalism in India: Business strategies and the working classes in Bombay, 1900–1940*. Cambridge: Cambridge University Press, 1994.

Cholia, R. P. *Dock labourers in Bombay*. Bombay: Longmans, Green and Co. Ltd., 1941.

Dakin, Julian; Tiffen, Brian and Widdowson, Henry G. *Language in education: The problem in Commonwealth Africa and the Indo-Pakistan sub-continent*. Oxford: Oxford University Press, 1968.

Dandekar, Hemalata C. *Men to Bombay, women at home: Urban influence on Sugao Village, Decan Maharashtra, India, 1942–1982*. Ann Arbor: Center for South and Southeast Asian Studies, University of Michigan Press, 1986.

D'Monte, Darryl. *Ripping the fabric: The decline of Mumbai and its mills*. New Delhi: Oxford University Press, 2002.

Gokhale, R. G. *The Bombay cotton mill worker*. Bombay: Millowners' Association, 1957.

Gore, M. S. *Immigrants and neighborhoods: Two aspects of life in a metropolitan city*. Bombay: Tata Institute of Social Sciences, 1970.

Morris, Morris David. *The emergence of an industrial labor force in India: A study of the Bombay cotton mills, 1854–1947*. Berkeley: University of California Press, 1965.

Munshi, Kaivan. "Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market." *Quarterly Journal of Economics*, 2003, *118*(2), pp. 549–99.

Munshi, Kaivan and Rosenzweig, Mark. "Traditional Institutions Meet the Modern World: Caste, Gender and Schooling Choice in a Globalizing Economy." Bureau for Research and Economic Analysis of Development, Working Paper: No. 038, 2003.

Patel, Kunj M. *Rural labour in industrial Bombay*. Bombay: Popular Prakashan, 1963.

Rees, Albert. "Information Networks in Labor Markets." *American Economic Review*, 1966 (*Papers and Proceedings*), *56*(2), pp. 559–66.

ECONOMIC CONSEQUENCES OF SOCIAL IDENTITY[†]

Discrimination, Social Identity, and Durable Inequalities

By KARLA HOFF AND PRIYANKA PANDEY*

What are the mechanisms by which societal discrimination affects individual achievement, and why do the effects of past discrimination endure once legal barriers are removed? We report the findings of two experiments in village India that suggest that the mechanisms of discrimination operate, in part, within the individuals who are members of the groups who have been discriminated against. We demonstrate that publicly revealing an individual's membership in such a group alters his behavior in ways that make the effects of past discrimination persist over time.

A growing literature in social psychology on *stereotype threat* finds that stereotyped-based expectations affect individual performance in the domain of the stereotype.[1] A study by Jeff Stone et al. (1999) is illustrative. When college students were asked to perform a task described as diagnostic of "natural athletic ability," blacks—stereotyped as better athletes, but worse students than whites—performed better than whites. When the *same* test was presented as diagnostic of "sports intelligence," the performance of blacks declined, that of whites improved, and the racial gap was reversed. Evidence suggests that a mediating factor in stereotype threat is a change in self confidence (Mara Cadinu et al., 2005)

In our studies, we investigated whether the public revelation of social identity (caste) affects cognitive task performance and responses to economic opportunities by young boys in village India. Subjects were sixth and seventh graders drawn from the two ends of the caste hierarchy. We asked subjects to learn and then perform a task under incentives, and we manipulated whether their peers in the experimental session knew their caste. Caste is well-suited to this manipulation because, unlike race, gender, and ethnicity, there are no unambiguous outward markers of caste among young boys. Six subjects, generally from six different villages, participated in each experimental session. In the control condition, the subjects were anonymous within the six-person group. In the experimental conditions, the experimenter publicly revealed subjects' names and caste. In the task—solving mazes—in which performance was studied here,[2] the low-caste subjects in the anonymous condition did not perform significantly differently from high-caste subjects; but when caste identity was publicly revealed in a mixed caste group, a significant caste gap emerged. The caste gap was due to a 20 percent decline in the average number of mazes solved by the low caste. The study shows that publicly revealing the social identity of an individual can change his behavior even when that information is irrelevant to payoffs.

Our results are a generalization of the literature on stereotype threat. Like that literature, we find that individuals' performance is more in accordance with the stereotype of the group when group membership is made salient in some way. Unlike that literature, salience in our experiments depends on the public revelation of social identity and more importantly, we do not argue that the domain of the tasks undertaken by

[†] *Discussants:* Rachel Croson, University of Pennsylvania; Iris Bohnet, Harvard University; Stefano DellaVigna, University of California-Berkeley.

* Hoff: World Bank, 1818 H St., N.W., Washington, DC 20433 (e-mail: khoff@worldbank.org); Pandey: World Bank, 1818 H St., N.W., Washington, DC 20433 (e-mail: ppandey@worldbank.org). Thanks to Rachel Croson, Erika Hoff, and Kenneth Sokoloff. And to Mayuresh Kshetramade, Anaka Narayanan, Ram Pratap, and Sonal Vats for research assistance. We gratefully acknowledge support from the World Bank and the MacArthur Foundation. Data are available at http://www.povertyactionlab.com/data.

[1] A survey is in Claude M. Steele et al. (2002).

[2] Uri Gneezy et al. (2003) showed that mazes are an appropriate task to use to study responses to changes in incentive schemes.

our subjects is one to which a specific stereotype applies. Ninety-five percent of the low-caste subjects and 82 percent of the high-caste subjects had never seen mazes before. If the low caste is stereotyped as inferior in the domain of the task, it is not, or not only, because of a generalization from caste differences in performing other cognitive tasks. The stigma of their caste marks them as unworthy, generally. The ideology intertwined with the discriminatory regime assigns to certain social groups status and social meanings—i.e., social identities. We suggest a broad link between discrimination, social identity, and behavior that can make the effects of past discrimination persist over time for well-identified groups.

I. The Setting in Village India

Indian society is divided into groups called castes which are linked to one or more traditional occupations. The participants in our experiments were drawn from the extreme ends of the caste hierarchy in villages in north India (Uttar Pradesh): the high-caste participants from the traditional landlord, warrior, priestly, and trading castes; and the low-caste participants from a caste that was historically, but is no longer, engaged in leather tanning. Leather tanning is associated with ritual pollution and this caste was subject to the practice of untouchability.

Untouchable castes (*Dalits*) were historically denied political and civil rights and opportunities for economic mobility. In 1947, India ended de jure discrimination against *Dalits.* Discrimination remains a visible part of village India, however, and the caste hierarchy is ritualized in the way many adults interact.[3]

II. Two Experiments

A. *Experimental Paradigm*

The objective of the experiments was to determine whether revealing subjects' social identities publicly would impede the performance of low-caste subjects and their responses to economic incentives.

We brought subjects into a classroom six at a time. Three conditions provided a contrast in the salience of caste. In the control condition termed "anonymous" (A), the experimenter (always a high-caste woman from north India) did not publicly reveal any information about the participants. In the condition termed "caste revealed" (C), the experimenter at the start of the session turned to each participant and stated his name, village, father's name, paternal grandfather's name, and caste.[4] She asked the subject to nod if the information was correct. The final condition, "caste revealed—single caste" (CS), was the same as the preceding condition except, unlike the other conditions, in which a session consisted of three low-caste and three high-caste boys, in CS a session consisted of low-caste boys only or high-caste boys only.

Whereas the extent of *publicly* revealed information varied across conditions, the extent of information *privately* revealed to the experimenters did not. A staff person privately asked every subject (when he boarded the car that brought him from his village) his name, caste, and the names of his father and paternal grandfather. Another staff person verified the subject's name and caste when the subject arrived at the experiment site.

B. *Experiment 1: Solving Mazes*

In our first experiment (Hoff and Pandey, 2005b), 156 subjects participated under condition A, 120 under C, and 60 under CS. A subject participated in only one condition. We asked subjects to solve a packet of 15 mazes in each of two 15-minute rounds. The incentive was one rupee per maze, a significant amount compared to the unskilled adult hourly wage of six rupees. Our dependent variable was the number of mazes solved.

Over the two rounds, the high-caste participants solved 7 percent more mazes than the low-caste participants in the anonymous condition, whereas the high-caste participants solved

[3] In our household survey near the site of the experiment, 56 percent of *Dalit* men reported that they sit on the ground or remain standing when visiting a high-caste household. Likewise, 58 percent of high-caste men said that when a *Dalit* visits their houses, he sits on the ground or remains standing.

[4] We used the names that the children had privately told staff, which for the low-caste subjects never included last names—generally a marker of caste.

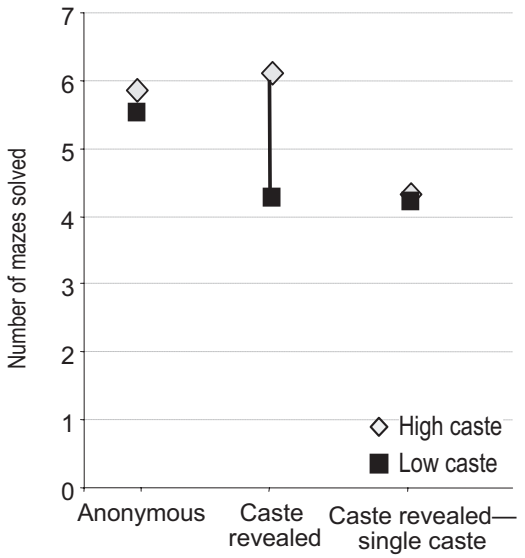


FIGURE 1. AVERAGE NUMBER OF MAZES SOLVED, ROUND 2

34 percent more mazes than the low-caste participants in condition C. To save space, Figure 1 shows the results only for Round 2. The caste differences in performance in both rounds were not significant in the anonymous condition. (The *p*-value of the Mann-Whitney two-sided *U*-test—hereafter MW—was 0.24 in Round 1 and 0.45 in Round 2.) In contrast, in condition C, there was a significant caste difference in both rounds (MW test: $p = 0.04$ in Round 1 and $p = 0.006$ in Round 2). The low-caste participants in condition C solved 20 percent fewer mazes than the low-caste participants in condition A, and the decline was significant in Round 2 (MW: $p = 0.14$ in Round 1 and $p = 0.05$ in Round 2); whereas there were no significant A-C treatment effects for the high caste (MW: $p > 0.80$ in both rounds).

The experimental finding of caste differences in performance in the C condition could be "poor versus rich" effect (in which reminding children of their poor families discourages them), rather than a pure caste effect. We find, however, that our results are robust when we control for the children's class—parents' education, occupation, and land.

The experimental finding that publicly revealing caste caused a significant decrease in the performance of low-caste subjects, compared with that in the anonymous condition, could possibly reflect intimidation of the low-caste subjects by the high-caste subjects, rather than an effect of social identity per se. To check this, we ran condition CS. Condition C was converted to condition CS by constituting experimental groups of low-caste boys only or high-caste boys only. As shown in Figure 1, there was no significant difference between the performance of the low-caste participants in conditions C and CS (MW: $p \geq 0.70$ in both rounds). This supports the conclusion that it is social identity that drives the caste gap in condition C.

An irony uncovered in this condition is that segregation lowers high-caste performance. The high-caste participants in conditions CS solved 21 percent fewer mazes than those in condition C; the treatment effect was significant in Round 2 (MW: $p = 0.73$ in Round 1 and $p = 0.02$ in Round 2). One conjectural explanation is that high-caste segregation in condition CS changed the extent to which subjects anticipated being rewarded because of their social status rather than their effort, while the presence of the low-caste subjects in condition C led high-caste subjects to try to excel in order to distinguish themselves from their low-caste peers.[5]

Figure 2 shows the number of mazes solved on the *x*-axes. The *y*-axes represent the proportion of low-caste and high-caste participants in Round 2 of condition A (the top graph) and condition C (the bottom graph), who solved that many mazes. The modal number of mazes solved by low-caste participants in condition A, Round 2, was seven, whereas it was zero for low-caste participants in condition C. Participants who solved zero mazes did not turn in blank packets; every participant attempted numerous mazes. The proportion of individuals who solved zero mazes in *both* rounds (call it the "drop-out rate") is the proportion who did not learn how to solve a maze over the 10-minute explanation by the experimenter, the 5-minute practice period, and the 30-minute test period. Among the 78 low-caste participants in

[5] Anjini Kochar (2004, p. 16) finds evidence that an increase in schooling by low-caste children increases investment in schooling by other castes. For other experimental evidence that the treatment effect of segregation lowers high-caste performance, see Hoff and Pandey (2004, pp. 24–25).

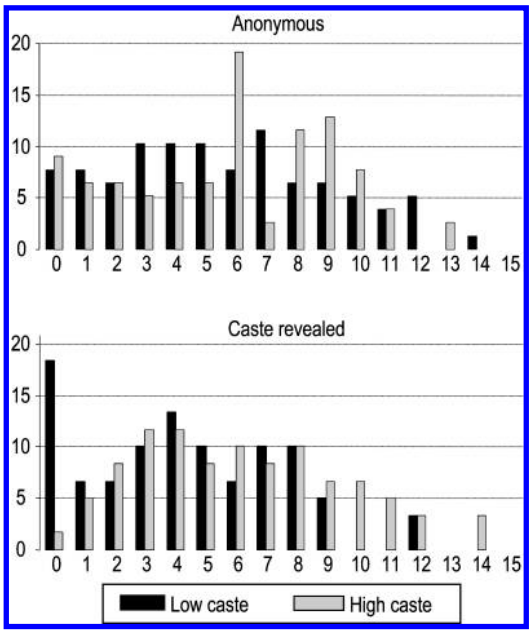


FIGURE 2. NUMBER OF MAZES SOLVED IN THE ANONYMOUS AND THE CASTE REVEALED CONDITIONS, ROUND 2

the anonymous conditions, the drop-out rate was 2.5 percent. Among the 60 low-caste participants in condition C, the drop-out rate was 15 percent. Among the high-caste participants, the drop-out rate was 9 percent and 2 percent in conditions A and C, respectively.

In summary, although the low caste perform as well as the high caste when caste identity is not publicly revealed to their peers in the classroom, publicly revealing caste identity is consistently and robustly associated with a decline in low-caste performance.

C. *Experiment 2: Betting on One's Own Success*

The purpose of our second experiment (Hoff and Pandey, 2005a) was to assess the effect of publicly revealing caste on individuals' willingness to bet on their own success. We manipulated the scope of judgment in evaluating and rewarding success. When subjects were asked to accept or reject a gamble in which there was *no* scope for judgment by an experimenter, making caste salient did *not* produce a caste gap. Instead, it was in the case where there was scope for judgment by others that making caste salient created a caste gap in the proportion of subjects who rejected the gamble.

In this experiment, we showed subjects how to solve a 14-inch square wooden puzzle that we had constructed along the lines of the game Rush Hour Traffic Jam. After practicing one puzzle for eight minutes, a subject chose, in private, to accept or reject a gamble that he could solve a similar puzzle within five minutes. The success payoff was 20 rupees (equal to 2.5 hours' wages for unskilled adult labor), and the failure payoff was one rupee. If he rejected the gamble, he received the safe payoff of 10 rupees.

We manipulated the scope for discretion in awarding the success payoff in the following way. In one gamble, the link between performance and reward was mechanical; by solving the puzzle, a player freed his car from the board and physically extracted his prize money from the underside of the car. This money was visible through the dashboard.

In a second gamble, the link between performance and reward was not mechanical. The prize money was not in the player's car. Moreover, in this condition there was no frame on the game board to keep vehicles within the road grid. We told subjects that another person would give them the puzzle, watch them play, and award them the success payoff only if they did not let any vehicle move off the board. Removing the frame and visible reward changed the gamble in two ways. It made success more difficult because a player could inadvertently push a vehicle off the board as he tried to solve the puzzle, and it created scope for discretion in awarding payoffs.

A total of 360 subjects participated in this experiment: 30 low-caste and 30 high-caste subjects in each treatment (2 gambles $\times$ 3 conditions). We predicted that the interaction with an evaluator would discourage low-caste participants more than high-caste participants when the participants' caste was publicly revealed. We conjectured that in that context, each individual would be more likely to fall into his caste role because he would expect others to treat him according to his caste role. The results bore out our prediction, but only in the A-CS contrast.

The left panel of Figure 3 shows the proportion of subjects who refused the gamble using the game board with the frame and the visible reward. There were no significant differences in the refusal rate between castes, and making

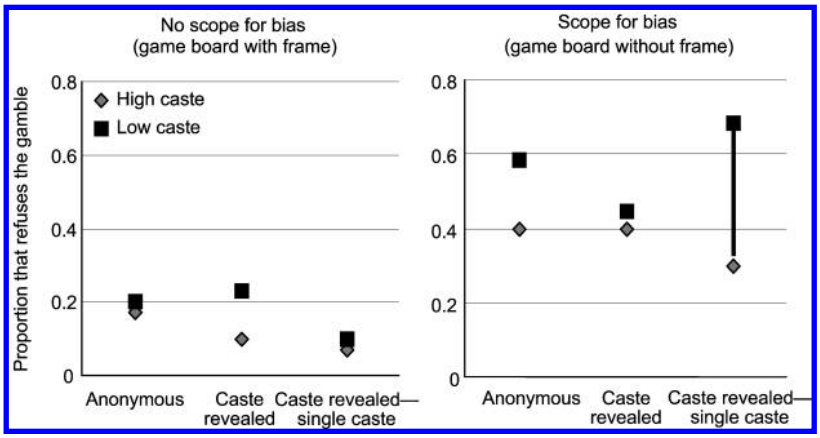


FIGURE 3. PROPORTION THAT REFUSES THE GAMBLE IN THE TRAFFIC JAM GAME

caste identity salient had no significant effects on this rate.

The right panel of Figure 3 shows that when the gamble used the game board without a frame, the proportion of both caste groups that refused the gamble was larger than in the case of the game board with its frame. There was a significant caste gap in the rejection rate *only* in the CS condition.[6] Sixty-seven percent of low-caste participants refused the gamble compared to 30 percent of high-caste participants (z-statistic: p-value $= 0.004$, MW: p-value $= 0.005$). A conjectural explanation of the presence of this caste gap in the CS condition and the absence of a caste gap in the C condition, is that caste segregation implicitly evokes the meaning of caste, which bears on social exclusion and inclusion, which increases the salience of caste roles and/or subjects' concern that the experimenter would be prejudiced against the low caste.

Low- and high-caste children differ from each other in many ways (literacy, parents' income and education, etc.) that might affect their responses to changes in the difficulty of the game and in the scope for experimenter discretion. One way to try to disentangle the influence of these variables from the effect of social identity, per se, is to do a caste-wise test of difference in differences in the refusal rate between the two gambles across two experimental conditions—A *versus* CS. For the *low* caste, the effect of making caste salient was to *raise* the difference in the refusal rate in the two gambles; but for the *high* caste, making caste salient had no effect in this difference (one-sided p-values of 0.09 and 0.50 for low and high castes, respectively). These results suggest that the differential caste behavior, and the gamble with scope for bias, under condition CS, is not driven by differences across castes in ambiguity aversion, since such differences are differences out in this test. Our results provide some evidence that the differential caste behavior is driven by social identity per se.

III. From Discrimination to Social Identity

Where do social identities come from? Having argued that publicly revealing individuals' membership in a discriminated-against group influences their behavior, we now want to emphasize, following Glenn C. Loury (2002), that discriminatory regimes not only categorize individuals and establish category-specific rules; they also invest those categories with social meaning. Discriminatory regimes create a narrative to justify the discrimination. The narrative serves to increase the probability that individuals perceive the regime as legitimate and internalize its values.[7] In the nineteenth

[6] This result mirrors our result in another experiment in which the task was to make a design with colored squares of paper. A subject won the gamble if his design was judged to be "beautiful" (Hoff and Pandey, 2005a).

[7] The same applies generally to polities, as Max Weber emphasized, and to other organizations (e.g., the U.S. Army; see George A. Akerlof and Rachel E. Kranton, 2005).

century, the U.S. South created the race doctrine of biological inequality between whites and Negroes—rationalizations similar to those used to defend discrimination against the low castes. The narrative and the stigma that it creates may outlast the discriminatory regime itself, in which case the legacy of discrimination is "spoiled collective identities" (Loury, 2002, p. 59).

IV. Conclusion

A discriminatory regime affects not only the structure of opportunities open to different social groups, but also the status and social meanings assigned to those groups—their social identities. If these identities influence behavior, then even after opportunities have been equalized across groups, the discriminatory regime will have persistent effects.

The findings of our two experiments suggest that publicly revealing individuals' membership in a group that has been or is being discriminated against impedes the group's ability to respond to economic opportunities. If publicly revealing the social indentity of members of this group increases their negative thoughts about themselves and their distrust, their lack of confidence affects their learing and willingness to bet on their own success, which keeps them from achieving outcomes comparable to those of high castes, which validates the discriminatory ideology and reprodues the effects of discrimination over time. Our experiments used the example of caste discrimination in rural north India, where caste is still "a marker of difference ... [that] harbor[s] the ideologies of pollution and exclusion" (Nicholas B. Dirks, 2001, p. 130). In light of the evidence of stereotype threat in social psychology, our experimental findings suggest that the impact of social identities shaped by discrimination on individuals' responses to economic opportunities may also be very general.

REFERENCES

Akerlof, George A. and Kranton, Rachel E. "Identity and the Economics of Organizations." *Journal of Economic Perspectives*, 2005, *19*(1), pp. 9–32.

Cadinu, Mara; Maass, Anne; Rosabianca, Alessandra and Kiesner, Jeff. "Why Do Women Underperform under Stereotype Threat?" *Psychological Science*, 2005, 16(7), pp. 572–578.

Dirks, Nicholas B. *Castes of mind: Colonialism and the making of modern India*. Princeton University Press, 2001.

Gneezy, Uri; Niederle, Muriel and Rustichini, Aldo. "Performance in Competitive Environments: Gender Differences." *Quarterly Journal of Economics*, 2003, *118*(3), pp. 1049–74.

Hoff, Karla and Pandey, Priyanka. "Belief System and Durable Inequalities: An Experimental Investigation of Indian Caste." World Bank, Policy Research Paper: No. 3351, 2004.

Hoff, Karla and Pandey, Priyanka. "Opportunity Is Not Everything: How Belief Systems and Mistrust Shape Responses to Economic Incentives." *Economics of Transition*, 2005a, *13*(3), pp. 445–72.

Hoff, Karla and Pandey, Priyanka. "The Persistent Effect of Discrimination and the Role of Social Identity." Unpublished paper, 2005b.

Kochar, Anjini. "Reducing Social Gaps in Schooling: Caste and the Differential Effect of School Construction Programs in Rural India." Unpublished Paper, 2004.

Loury, Glenn C. *The anatomy of racial inequality*. Cambridge, MA: Harvard University Press, 2002.

Steele, Claude M; Spencer, Steven J. and Aronson, Joshua. "Contending with Group Image: The Psychology of Stereotype and Social Identity Threat," in Mark Zana, ed., *Advances in experimental social psychology*, *34*. Amsterdam: Elsevier North-Holland, 2002, pp. 379–441.

Stone, Jeff; Lynch, Christian I.; Sjomeling, Mike and Darley, John M. "Stereotype Threat Effects on Black and White Athletic Performance." *Journal of Personality and Social Psychology*, 1999, *77*(6), pp. 1213–27.

This article has been cited by:

1. RYO HORII, MASARU SASAKI. 2012. Dual Poverty Trap: Intra- and Intergenerational Linkages in Frictional Labor Markets. *Journal of Public Economic Theory* **14**:1, 131-160. [CrossRef]
2. Ernst Fehr, Karla Hoff. 2011. Introduction: Tastes, Castes and Culture: the Influence of Society on Preferences*. *The Economic Journal* **121**:556, F396-F412. [CrossRef]
3. Zahra Siddique. 2011. Evidence on Caste Based Discrimination. *Labour Economics* . [CrossRef]
4. Shinji Teraji. 2010. An economic analysis of social exclusion and inequality. *Journal of Socio-Economics* . [CrossRef]
5. Jeffrey Carpenter, Jessica Holmes, Peter Hans Matthews. 2010. Jumping and sniping at the silents: Does it matter for charities?. *Journal of Public Economics* . [CrossRef]
6. Daniel J. Benjamin, , James J. Choi, , A. Joshua Strickland. 2010. Social Identity and Preferences. *American Economic Review* **100**:4, 1913-1928. [Abstract] [View PDF article] [PDF with links]
7. Karla Hoff, , Joseph E. Stiglitz. 2010. Equilibrium Fictions: A Cognitive Approach to Societal Rigidity. *American Economic Review* **100**:2, 141-146. [Citation] [View PDF article] [PDF with links]
8. Michael Woolcock. 2010. The Rise and Routinization of Social Capital, 1988–2008. *Annual Review of Political Science* **13**:1, 469-487. [CrossRef]
9. Jeffrey Carpenter, Jessica Holmes, Peter Hans MatthewsCharity auctions in the experimental lab **13**, 201-249. [CrossRef]
10. MARIA K. HUMLUM, KRISTIN J. KLEINJANS, HELENA S. NIELSEN. 2010. AN ECONOMIC ANALYSIS OF IDENTITY AND CAREER CHOICE*. *Economic Inquiry* no-no. [CrossRef]

Empowering Women: Inheritance Rights and Female Education in India

Sanchari Roy*

February 1, 2012

Abstract

This paper examines the impact of women's property inheritance rights on their education. Using exogenous variation created by state level reforms to the inheritance law in India, I find that mean educational attainment of women who were of primary school-going age at the time of reform increased by 0.5 years in reforming relative to non-reforming states. The impact is present only for women in landowning and "Hindu" households, with no concomitant impact on men. I also provide suggestive evidence that a potential underlying mechanism could be that investment in education was used to compensate daughters for disinheriting them from household property.

JEL Codes: O12, K11, I21
Keywords: Inheritance, education, women, dowry payments

*Centre for Competitive Advantage in the Global Economy (CAGE) and Department of Economics, University of Warwick, Coventry CV4 7AL, UK; email: s.roy@warwick.ac.uk.
I am extremely grateful to Maitreesh Ghatak and Oriana Bandiera for their continuous encouragement and support throughout this project. I also thank Ran Abramitzky, Bina Agarwal, Madhav Aney, Fernando Aragon, Erlend Berg, Sonia Bhalotra, Konrad Burchardi, Robin Burgess, Sayantan Ghosal, Aparajita Goyal, Victor Lavy, Sabyasachi Mukherjee, Jin Wang, Fabian Waldinger, Chris Woodruff and especially Tim Besley and Stefan Dercon for valuable suggestions and feedback. I am also grateful for comments received from seminar participants at ISI Delhi, JNU, LSE, Warwick, the BREAD Summer School (Verona), the Conference on Persistent Poverty (Cornell), the CSAE Conference (Oxford), PAC-DEV (Berkeley), the RES Conference (Royal Holloway) and the Second Riccardo Faini Doctoral Conference (Milan).I gratefully acknowledge financial support from the Bagri Fellowship and the DFID Program on Improving Institutions for Pro-Poor Growth. All errors are my own.

The role of property rights in the process of economic development has been well-emphasized in the economic literature (North, 1990; De Soto, 2000; Besley, 1995; Banerjee, Gertler, and Ghatak, 2002; Field, 2007; DiTella, Galiani, and Schargrodsky, 2007). Property rights, through their impact on distribution of wealth, patterns of production as well as development of markets, especially credit markets, have evolved as one of the prerequisites of economic growth and poverty reduction (Besley and Ghatak, 2009). The primary focus of this literature has been to study the impact of property rights on physical investment, but the role of property rights in the context of human capital investment is relatively under-researched. Moreover, most of the existing research remains gender-neutral, with little attention to the salience of property rights for women. This paper attempts to fill these gaps by studying the impact of property rights, particularly inheritance rights, on the human capital investment of women.

The principal methodological problem faced in estimating the causal impact of property rights at the household level is that of potential endogeneity. There could be unobserved heterogeneity at the household level correlated with both female education and female property rights that may generate spurious results. For example, gender progressive parents may be more likely to invest in their daughters' education as well as give them greater rights to family inheritance. This could lead to the classic omitted variable problem that would bias the estimates of the impact of female property rights. A second complication in this regard may arise due to measurement error as it is often difficult to obtain appropriate measures of female property rights due to the fact that women in many societies lack formal titles to property (Deere and Leon, 2003; Sweetman, 2008). This may introduce further biases in the estimates of the causal impact of female property rights on female

education.

To address these problems, this paper exploits a legislative change to the central inheritance law in India as a source of exogenous variation in female inheritance rights. Like most personal laws in India, inheritance laws too vary by religion. The fundamental law governing present day inheritance rights of four religious communities i.e. Hindus, Buddhists, Jains and Sikhs, called the Hindu Succession Act (HSA) of 1956, was designed to lay down a law of succession whereby sons and daughters would enjoy equal inheritance rights. In fact, however, significant gender inequalities persisted that disadvantaged daughters considerably. The main source of bias came from joint family property, to which sons enjoyed right *by birth* to an independent share but daughters did not. Both had equal rights of inheritance to the separate property that their father accumulated during his lifetime. But, due to the fact that a considerable amount of property, especially land in rural areas, is still jointly owned, such biased rights had a crippling effect on the property ownership of women in India.

The earliest attempts at amending this law were made by five Indian states, namely Andhra Pradesh, Tamil Nadu, Kerala, Karnataka and Maharashtra, between late 1970s and early 1990s. The amendments stated that women who were unmarried at the time the reform was passed in their state would be granted claims equal to that of their brothers in the joint family property, including the right to a share by survivorship (Agarwal, 1994).[1]

The basic identification strategy in this paper uses the fact that exposure to the improved inheritance rights regime following the amendments was jointly determined by state of birth and year of birth. Not only did a woman have to be born in a state that passed the reform, she also had to be of

[1]Details regarding each state amendment is available in "The Hindu Succession Act 1956, with State Amendments (Bare Act)".

school-going age when the reform was passed in her state for it to have any impact on her schooling decisions. Hence, I identify the causal effect of the reform, which I argue is exogenous[2], by using a difference-in-differences methodology that compares mean educational attainment of women who were young enough to be exposed to the reform ("treated" group) to those who were too old ("control" group), between reforming and non-reforming states. The identifying assumption is that in the absence of the reform, the change in female educational attainment across cohorts would not have been systematically different in reforming and non-reforming states. Similar strategies have been used by Duflo (2001), Card and Krueger (1992), Lemieux and Card (2001) etc. to estimate the effect of education on earnings.

However, if there exists unobservable factors that affect female education and are also correlated with the passage of the reform, then the difference-in-differences estimates would be biased. Therefore, in order to address this concern, I employ a triple differences strategy by exploiting another source of variation within each state-cohort, namely land ownership status, religious affiliation and gender. For example, in case of triple differences using land ownership, I estimate the difference in mean educational outcomes between women in the "treated" group relative the "control" group, for land-owning versus non land-owning households in reforming relative to non-reforming states. This would control for two kinds of potentially confounding trends: changes in educational outcomes of women belonging to landed households across states (that have nothing to do with the reform) and changes in educational outcomes of all women in the reforming states (e.g. due to other state policies that affect everyone's education). I do the same for "Hindu"

[2]Concerns regarding the potential endogeneity of the reform process is discussed in Section 2.2.

versus "non-Hindu" households[3], as well as between women and men.

For my analysis on female education, I use household level data from multiple waves of the National Family and Health Survey of India (NFHS), where I focus on the daughters of the head of the household.

The primary finding of this paper is that exposure to the inheritance rights reform was associated with an increase of approximately 0.5 years of education (an improvement of around 0.2 standard deviations) for cohorts of women who were of primary school-going age at the time of the passage of reform. On the other hand, no effect is observed for cohorts that were 16 years or older at the time of the reform, suggesting that the findings are less likely to be driven by correlated unobservables.

Moreover, using triple differencing by land ownership, I find that the entire effect comes from women who belong to land-owning households, and the estimated coefficient is larger at approximately 0.8. Similarly, using triple differencing by religion, I find that the impact to be present only for those women who were either Hindu, Buddhist, Sikh or Jain (to whom the law applied), and the estimated coefficient is even larger at approximately 1.5. Finally, using triple differencing by gender, I find the impact to be present only for daughters with the estimated coefficient being 1.3-1.6 and no concomitant impact on boys, indicating that the reform was successful in narrowing the gap in education between boys and girls.

Due to lack of appropriate data, I cannot provide conclusive evidence using the NFHS on the mechanism through which the inheritance rights reform increased female education. However, suggestive evidence using a separate dataset, the Rural Economic and Demographic Survey 1999, indicates that

[3]I use the term "Hindu" to refer to Hindus, Buddhists, Sikhs and Jains i.e. those to which the HSA 1956 applies, while "non-Hindus" refer to Muslims, Christians, Jews and Parsis.

one potential mechanism could be that investment in education was being used as a form of compensation to daughters for disinheriting them from ancestral property. In a *virilocal* society like India, where married daughters leave their parental household while married sons do not, giving asset bequests to daughters can have poor incentive effect on sons in relation to extending family wealth (Botticini and Siow, 2003), resulting in daughters rarely inheriting property.[4] Instead, daughters have traditionally been given dowries as compensation for relinquishing their rights to property. Post reform, I find no improvement in the situation of women with regard to inheritance, reflecting a weak enforcement of the law. Instead, it appears that daughters are now compensated with larger dowries. However, since larger dowries are costlier to provide, for younger daughters, compensation takes the form of increased investment in education, and thereby, lower dowries. However, it is important to emphasize that this evidence far from conclusive and does not indicate that this is *the* mechanism driving the impact of the reform on female education.

A related paper, Goyal, Deininger, and Nagarajan (2010), also examines the impact of the Hindu Succession Act amendment on women's education and likelihood of inheritance in India. My study differs from Goyal, Deininger, and Nagarajan (2010) in two key ways. Firstly, the results of my paper differ from Goyal, Deininger, and Nagarajan (2010). Although, like Goyal, Deininger, and Nagarajan (2010), I also find a positive impact of the inheritance rights reform on female education, but unlike them, I find

[4]Botticini and Siow (2003) argue that since married daughters in virilocal societies leave their parental home, if they were to share equally in family property, then their brothers, who stay with their parents and work with family property, will not obtain the full benefits of their effort in extending family wealth and hence will supply too little effort. Thus, to mitigate this incentive problem for their sons, parents give dowries to their daughters at the time of marriage (when they cease to contribute to their parents' wealth) but no bequest, and bequests to their sons.

no significant impact on women's propensity to inherit following the reform. To reconcile this difference in the findings of the two papers, however, it is important to note that the reform to the inheritance law applies to ancestral property and not to father's separate property (to which the daughter always had equal inheritance right as the son). Hence the relevant event to consider while identifying the impact of the reform on the likelihood of a woman's inheritance would be the death of her grandfather, since partition of the ancestral property is most likely to happen after the head of the household, in this case the grandfather, dies. Thus, exploiting the variation in the timing of the grandfather's death relative to the reform, I find that the reform had no impact on the likelihood of inheritance of women. Secondly, a unique contribution of my paper is that I also examine the impact of the inheritance rights reform on dowry payment made at the time of a woman's marriage which, coupled with my aforementioned finding on women's propensity to inherit, allows me to shed light on a potential mechanism through which the inheritance rights reform could have impacted female educational attainment in India.

This paper relates to two different strands of literature. The literature on property rights has focused on the role of property rights in enhancing investment incentives in agricultural land (Banerjee, Gertler, and Ghatak, 2002; Besley, 1995), residential investment (Field, 2007), entreprenurial investment of retained earnings (Johnson, McMillan, and Woodruff, 2003) etc. The topic of property rights and human capital investment is relatively under-studied and this is where my paper seeks to make a contribution.

This paper also relates to the literature on dowry and marriage markets. A number of papers focus on the role of dowry as a spot price that clears the marriage market characterized by assortative wealth matching (Becker, 1981;

Anderson, 2003, 2007; Rao, 1993; Edlund, 2001). On the other hand, dowry has also been studied as a "pre-mortem" bequest (Anderson, 2004; Goody, 1973). In this context, it has been argued that change in the environment for producing bridal wealth, in the form of labour market expansion, may lead to reduction in prevalence of dowry (Botticini and Siow, 2003). My paper fits well with such a line of argument as it shows that a legal reform in inheritance rights can have similar consequences on dowry payments through its impact on education, an alternative form of wealth transfer to daughters.

The remainder of the paper is organized as follows: Section 1 describes the institutional background of Hindu inheritance law in India, while Section 2 outlines the data and identification strategy. Section 3 presents results on female education, and Section 4 discusses a potential mechanism underlying the observed effect by looking at the likelihood of inheritance by women and their dowry payments. Section 5 concludes.

1 The Institutional Background

1.1 The Hindu Personal (Inheritance) Law

As mentioned earlier, the laws for inheritance of property in India differ by religion. The inheritance rights of Hindus are governed by the Hindu Succession Act (HSA) of 1956, which also governs the rights of Buddhists, Jains and Sikhs.[5] The Act was built on the foundation of ancient legal doctrines that have prevailed in India since the 12 century A.D., and purported to lay down a law of succession that gave equal rights of inheritance to sons and

[5]These religions are considered to be offshoots of Hinduism and hence are looked upon as being "Hindu-like" religions. For the rest of the paper, I will use the term "Hindu" to denote Hindus, Buddhists, Sikhs and Jain, that is all religions to which the HSA 1956 applied.

daughters. In fact, however, significant gender inequalities remained.

A key feature of the legal structure of "Hindu" inheritance in India is the distinction between "joint family property" and "separate property".[6] Generally speaking, joint family property "consists principally of ancestral property (that is, property inherited from the father, paternal grandfather or paternal great-grandfather), plus any property that was jointly acquired or was acquired separately but merged into the joint property". Separate property, on the other hand, "includes that which was self-acquired (if acquired without detriment to the ancestral estate) and any property inherited from persons other than father, paternal grandfather or paternal great-grandfather" (Agarwal, 1994, p. 85-86).

According to the Hindu Succession Act of 1956, daughters of a "Hindu" male dying intestate (i.e. without leaving a will)[7] were equal inheritors, along with sons, of only their father's separate property and his "notional" portion of joint family property, but had no direct inheritance rights to joint family property itself.[8] [9] Sons, on the other hand, not only inherited their share of the father's own property and his "notional" portion of joint family property, but also had a direct right *by birth* to an independent share of the joint family property. In fact, all persons who acquired interest in the joint family

[6]The joint family here is a legal concept and need not coincide with the joint residence or or any other aspect of a common household economy that may be implied in a sociological use of the term (Agarwal, 1994).

[7]According to Goyal, Deininger, and Nagarajan (2010), the proportion of people who die without making a will in India is very high (around 65%, and probably even higher in rural areas), suggesting that the Hindu Succession Act is what ultimately determines inheritance patterns within the family.

[8]The "notional" portion of the father's share in the joint family property would be ascertained under the assumption of a "notional" or hypothetical partition of that property, as if the partition had taken place just before his death.

[9]In case of a "Hindu" woman dying intestate, all her property devolves equally upon her sons and daughters and husband, if alive. If she has no children or other heirs with first right to her property, then the property devolution takes place according to the source of acquisition.

property by birth were said to belong to the "Hindu coparcenary", which is conceptually similar to an exclusive male membership club in relation to the issue of inheritance to which women had no access.[10]

In order to elaborate, I explain the scenario using a simple example. Let us consider a family consisting of a grandfather and his two sons, Son 1 and Son 2 (see Figure I). Let us assume that the family line begins with the grandfather, such that he has no predecessors. The first son has a son of his own (Grandson 1) as well as a daughter (Granddaughter 1), while for simplicity, I assume the second son is childless. The ancestral/joint family property owned by this family is say 1 acre, and nobody acquires any additional property during his/her lifetime i.e. "separate" property of any individual is zero (for simplicity). Bold letters indicate membership of the "Hindu" coparcenary.

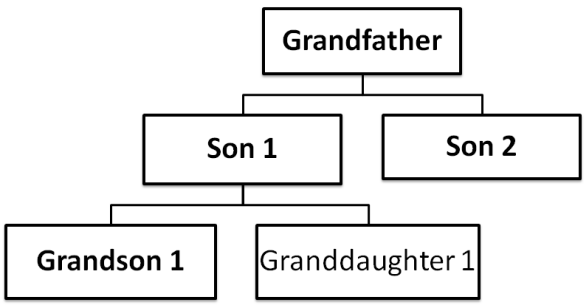


Figure I: Ancestry

The process by which inheritance rights to this ancestral/joint family property will be determined in this family is as follows (see Figure II):

[10]In addition to inheritance, sons could also demand partition of the joint family property while daughters could not. E.g. if the joint family property was a dwelling house, sons (as part of the coparcenary) could demand a partition of the same but daughters were only allowed right of residence but no right of ownership or possession.

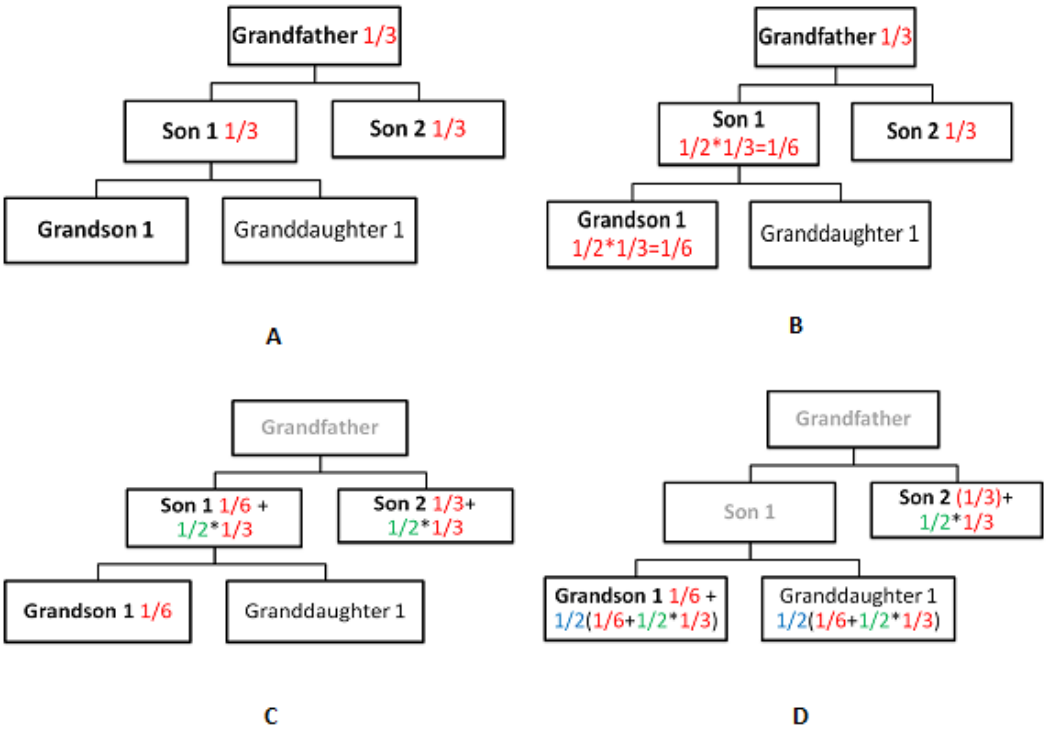


Figure II: Inheritance

During the lifetime of the Grandfather, he himself along with his two sons, Son 1 and Son 2, each have a share of a third in the ancestral property (panel A of Figure II). Moreover, Son 1 shares his third equally with his own son, Grandson 1, since the latter is a member of the male coparcenary (along with his grandfather, his father and his uncle in this example) and hence has a right by birth to an "independent" share to joint family property (panel B). Hence Grandson 1 directly gets a sixth of joint family property as a male coparcener. Granddaughter 1, on the other hand, does not get any share of the joint family property directly.

Now when Grandfather 1 dies (panel C), his share of a third gets split equally between his two living sons, Son 1 and Son 2, such that Son 1's share

now increases to a sixth (his coparcenary share) plus another sixth (inherited from his father), which totals to a third.[11] Next, when Son 1 dies (panel D), his total share is split equally between Grandson 1 and Granddaughter 1, i.e. each get a sixth. So, ultimately, Granddaughter 1 is entitled to a share of one-sixth (inheritance from her father) while her brother, Grandson 1, not only gets that one-sixth (inheritance from his father) but an additional one-sixth which is his coparcenary share. Thus Grandson 1's final share is one-third, which is double that of his sister.

Hence, it is apparent that the daughters suffered from discrimination in terms of inheritance under HSA 1956. Moreover, for the millions living in rural India, the most common form of property is land that is typically family-owned, which makes the gender bias in inheritance rights quite a significant phenomenon. Thus the law, by excluding the daughter from participating in the coparcenary ownership of ancestral property, not only discriminated against her on grounds of gender, but also led to a negation of her fundamental right of equality as guaranteed to her by the Indian Constitution (Ramanujam, 2005).

1.2 State Amendments to the Hindu Succession Act

The topic of inheritance in India is a "concurrent" one, i.e. one over which both the central and the state governments have legislative authority. Thus, although the HSA 1956 is a central law, some of the states have subsequently amended the HSA 1956. In particular, Kerala amended in 1976, Andhra Pradesh in 1986, Tamil Nadu in 1989, Maharashtra and Karnataka in 1994,

[11] Strictly speaking, the grandfather does not have to actually *die* for this so-called "partitioning" to be made: the inheritance shares are decided in a "notional" sense, as described in the earlier footnote. However, in practice, the most common reason behind a split or partition has to do with the death of the household head or patriarch (Foster and Rosenzweig, 2002).

following which daughters were granted *independent* inheritance rights and the right to a share by survivorship in joint family property, equal with their brothers, but only if they were unmarried at the time of the reform.[12],[13] Such a reform opened up the entry of women into what had till now been an exclusively male preserve and sought to, at least partially, redress the concern of gender bias inherent in the original central law. I exploit these legislative amendments as a "natural experiment" to study the impact of a potential improvement in female inheritance rights on female education in India.

2 Data and Identification Strategy

2.1 Data

To estimate the causal impact of the inheritance rights reform on female education, I use household-level data from multiple rounds of the National Family Health Survey of India (NFHS) conducted in 1992, 1998 and 2005.[14] The NFHS is designed along the lines of the Demographic and Health Surveys (DHS) that have been conducted in many developing countries around the world, and are repeated cross-sections.

The NFHS surveys, which are representative at the state level and have an overall response rate of 98 percent, contain detailed information, including

[12]Kerala passed a slightly different amendment in the form of the Kerala Joint Hindu Family System (Abolition) Act that recognized all family members with an interest in the undivided family estate as being independent full owners of their shares from then onwards, i.e. abolished joint family property altogether. But since the spirit of this amendment was similar to those passed by the other reforming states, and could be expected to favourably affect the inheritance of the daughter, I club them together. However, most of the findings of this paper are robust to the exclusion of Kerala.

[13]The inheritance rights of widows were unaffected in these amendments.

[14]The NFHS is carried out by the Ministry of Health and Family Welfare, Government of India.

educational attainment, on all individual members of the household. 29 states of India are covered in the sample.[15] However, the Hindu Succession Act (1956) did not apply to Jammu and Kashmir (Agarwal, 1994). Hence I drop that state from my analysis and are left with 28 states.

I first focus on women who are daughters of the head of the household and who are at least 22 years of age at the time of survey (this ensures that women in the sample have completed their education).[16],[17] There are 18,691 such women in my sample, with year of birth spanning 1943 to 1984.[18] Summary statistics are presented for this sample in Table 1. Mean age for this sample of women is 27 years, while average level of education is 8.14 years of completed education (6 years of education correspond to completion of primary school). Almost half of the sample women belong to households

[15]The 3 newest states of India, i.e. Chattisgarh, Uttarakhand and Jharkhand, were created in 2000, out of Madhya Pradesh, Uttar Pradesh and Bihar respectively. They are part of the NFHS wave of 2005, but not of the waves of 1992 and 1998. Additionally, Sikkim is not a part of the 1992 wave. Smaller Union Territories like Lakshadweep, Andaman and Nicobar Islands, Pondicherry etc. are also excluded.

[16]The findings of this paper are robust to different cut-offs regarding age at time of survey.

[17]Some of the mothers of these women may have been young enough to be exposed to the reform themselves, especially those born after 1965 in 1998 wave and those born after 1972 in 2005 wave. To avoid any confounding impact on outcomes of daughters through their mothers, I calculate the minimum age that mothers need to be in order to be *unexposed* to the reform: this is 44 years at the time of survey, which means that for example, in the 2005 wave, these women had to be born on or before 1961 such that they would be 15 or older at the time of the earliest reform (Kerala in 1976). Hence I restrict my sample to those daughters who were not only themselves at least 22 years old at survey but whose mothers were also at least 44 years old at survey. However, some of the mothers could also be exposed to the reform if they were unmarried and hence had not inherited by the time of the reform was passed. The potential of increased inheritance for them may have independent effects on their daughter's education, if women's inheritance rights are positively correlated with their bargaining power in their husband's family post marriage, that in turn has a positive effect on allocation of resources towards education of the girl children. Since nearly 80 percent of the women in my sample are married by 20, the oldest a mother needs to be to be completely "unexposed" would be to be at least 49 years old at time of survey. The results presented later are robust to restricting my sample to daughters whose mothers are at least 49 at survey.

[18]Of the total 18,691 women, 5,613 obtain from the 1992 round, 6,128 from the 1998 round and 6,950 from the 2005 round.

that own some land while 83 percent are "Hindu".

2.2 Identification Strategy

The basic identification strategy used in this paper exploits the fact that exposure to the inheritance rights reform was jointly determined by a woman's state of birth and year of birth. Not only did a woman have to be born in a state that passed the reform, she also had to be of school-going age when the reform was passed in her state for it to have any impact on her schooling decision. Given that the NFHS is a repeated cross-section, this approach amounts to a difference-in-difference (DD) strategy over cohorts and states.

The NFHS dataset does not contain information on an individual's state of birth but it does collect data on state of residence. Hence my empirical analysis uses state of residence instead of state of birth. If this gives rise to measurement error then my estimates of the reform's impact of female would suffer from attenuation bias. A bigger concern, however is that of systematic variation in migration behaviour in response to the reform. If gender progressive parents marry their daughters to grooms in the reforming states to take advantage of the favourable laws, then too the estimates would be biased. However, using the Rural Economic and Demographic Survey of 1999, I do not find any evidence of significant differences in inter-state female migration between the reforming and non-reforming states in the pre-reform period. In fact, nearly 80 percent of women reside after marriage in the same district or other district of the same state as their parental household. This is also supported by the findings of Rosenzweig and Stark (1989) that in the ICRISAT villages, the mean distance between a woman's original residence place and marital place of residence is around 30 kilometers. Hence, the possibility of systematic migration across states seems relatively remote in

this particular context.

The empirical analysis, as mentioned above, tests for the effect of the reform on "treated" age cohorts. I define the "treated" group as cohorts of women who were of primary school-going age when the reform was passed in their state. In India, children normally attend primary school between the ages of 5 and 10, middle school between the ages of 11 and 13 and high or secondary school between ages of 14 and 15. Hence, my "treated" group consists of cohorts of women who were 10 years or younger at the time of the reform since they were "young" enough for the reform to have affected their education choices. The control group, on the other hand, would consist of women who were already well past school-going age by the time the reform was enacted in their state, i.e. were 21 years or older. The reform ought to have no effect on their educational achievement. Thus, the basic identification strategy is a difference-in-differences between the "treated" or "younger" cohorts and the "control" or "older" cohorts, for reforming relative to non-reforming states. Such a difference-in-differences estimate may be interpreted as the causal impact of the reform, under the assumption that in the absence of the reform, the change in educational attainment of women across cohorts would not have differed systematically between the reforming and non-reforming states.

However, the identification assumption should not be taken for granted. What if the pattern of change in female education across cohorts did vary systematically between the reform and non-reforming states? To address this concern, I test for an implication of the identifying assumption where I compare mean educational attainment of women who were between 16 to 20 years old at the time of reform to that of women who were 21 or older at that time (control group), between reforming and non-reforming states. Since the

former group would also have been out of school by the time the reform was passed in the reforming states, the change in educational attainment for women in this age-group relative to the control group should not, therefore, vary systematically across states.

Within a regression framework, I therefore estimate the following equation:

$$e_{isk} = \alpha_s + \beta_k + \gamma_s k + \delta_1 D_{is,(k \geq k'-5)} + \delta_2 D_{is,(k'-10 \leq k \leq k'-6)}$$
$$+\delta_3 D_{is,(k'-15 \leq k \leq k'-11)} + \delta_4 D_{is,(k'-20 \leq k \leq k'-16)} + X_{isk}\eta + \epsilon_{isk} \quad (1)$$

The dependent variable e_{isk} denotes the educational attainment of woman i in state s belonging to cohort k (i.e. born in year k). Let the reform be passed in year k' in state s. Then $D_{is,(k \geq k'-5)}$ is a dummy indicating whether woman i belonging to cohort k was 5 years old or younger when the reform was passed in her state. Similarly, $D_{is,(k'-10 \leq k \leq k'-6)}$ is a dummy indicating whether she was between 6 and 10 years old, $D_{is,(k'-15 \leq k \leq k'-11)}$ indicating whether she was between 11 and 15 years old and $D_{is,(k'-20 \leq k \leq k'-16)}$ indicating whether she was between 16 and 20 years old respectively. As mentioned earlier, the group consisting of women who were 21 years or older at the time of the reform constitute the omitted category. α_s represents state fixed effect which accounts for state-specific characteristics that do not vary across cohorts, β_k represents cohort of birth fixed effect that accounts for the fact that women born in different years may be exposed to different macro shocks, while $\gamma_s k$ captures state-specific linear trends over cohorts. X_{isk} is a vector of household-level control variables that would affect education: parental age, parental education, land ownership, religion, number of household members and place of residence (urban/rural). ϵ_{isk} is the error term. To address serial

correlation concerns and to allow for heteroscedasticity, the standard errors are clustered at the state level (Bertrand, Duflo, and Mullainathan, 2004).

The coefficients of interest are δ_1 and δ_2, which capture the mean effect on education of being exposed to the inheritance rights reform for the "treated" or "young" cohorts. δ_3 and δ_4, on the other hand, capture the effect of the reform on older cohorts. The oldest cohort category (16 to 20 years) is specifically included as a falsification test - the members of this cohort would have left school by the time the reform was passed in their state and hence would not be expected to experience any impact on their educational attainment.

However, if there existed other unobservable factors affecting female education that were correlated with the passage of the reform itself, then the identification assumption underlying the difference-in-differences approach would be violated. For example, Clots-Figueras (2011) find that election of lower caste women leaders is positively correlated with the probability of passage of female-friendly laws in India, of which the amendments to the HSA 1956 would be an example. If lower caste women leaders also invest more in female education, then the difference-in-difference estimate discussed above could just be picking up the effect of an increase in the presence of such women leaders in state legislatures, who were responsible for both the passage of the reform as well as investment in female education in these states. State policies affecting female education but varying systematically between reforming and non-reforming states would be another example.

To address this concern, I conduct a difference-in-difference-in-difference (DDD) or triple difference analysis by introducing a separate within-state-cohort control group. Three variants of such a control group are explored: women belonging to non-landed (versus landed) households, "non-Hindu"

(versus "Hindu") women, and men (versus women).

Thus, the expanded version of equation 1 that I estimate is:

$$\begin{aligned} e_{isk} = \alpha_s + \beta_k + \gamma_s k + + \delta_1 D_{is,(k \geq k'-5)} + \delta_2 D_{is,(k'-10 \leq k \leq k'-6)} \\ + \delta_3 D_{is,(k'-15 \leq k \leq k'-11)} + \delta_4 D_{is,(k'-20 \leq k \leq k'-16)} + \delta_1' D_{is,(k \geq k'-5)} * C_i \\ + \delta_2' D_{is,(k'-10 \leq k \leq k'-6)} * C_i + \delta_3' D_{is,(k'-15 \leq k \leq k'-11)} * C_i \\ + \delta_4' D'_{is,(k'-20 \leq k \leq k'-16)} * C_i + \mu C_i + X_{isk}\eta + \epsilon_{isk} \end{aligned} \tag{2}$$

where C_i is a dummy variable denoteing either land ownership status, "Hindu" or gender of the individual. The coefficients of interest are δ_1' and δ_2', which capture, e.g. in case of triple differences by land ownership, the differential impact on education for "treated" compared to "control" women that belong to landed relative to non-landed households in reforming versus non-reforming states.

The rationale behind using these groups for triple differencing is as follows. Firstly, the amendment to the Hindu Succession Act 1956 would have a bite in the reforming states only if the parental household of the woman owns any joint family property to begin with. Land is the most commonly held form of joint family/ancestral property in India, hence it makes sense to exploit variation along the dimension of land ownership status of the woman's household to improve identification. Now, since a household's land ownership status obtained at the time of survey, the underlying assumption is that this status has remained unchanged over time. If this assumption does not hold in reality, then measurement error would lead to attenuation bias in the triple differences estimates.[19] A bigger concern, however, is that land own-

[19]To elaborate on this, two possibilities could arise: one, it could be that the woman's family did not own land when she was young but does own land now (at the time of survey) and second, the family owned land when she was young but does not now. In the

ership status maybe correlated with the reform. The identifying assumption of the triple differences strategy is that the difference in educational outcomes between "treated" and "control" women belonging to landed relative to non-landed households in reforming states is on account of the reform. If, however, gender progressive parents had acquired additional land in anticipation of the reform, then this assumption would be violated. But it is important to note here that the reform related to ancestral property, and not to separate property acquired by the father in his lifetime, which allays fears of strategic land procurement by parents that could bias the results.

Secondly, owing to the fact that the HSA 1956 applied differentially across religion in India, only women who were either Hindu, Buddhist, Sikh or Jain should benefit from the reform.

Finally, since the reform to HSA 1956 relates to the issue of inheritance of women, men of similar age group categories may be used as a counterfactual to examine the impact of exposure to the reform on the gender gap in education. However, there may arise concerns that since change in inheritance rights constitute a zero-sum game within the family (more rights for daughters implies less for sons, given a fixed amount of ancestral property), there may occur some compensating impact of the reform on men. This issue is discussed in further detail in section 3.2.3.

Before proceeding to the results, I would like to point out the contribution of each reforming state to the cohort categories constructed above, provided in Table A.1. Since I focus on women who were 22 or older at the time

first case, reform would not have had any impact on the woman's education and including her as being landed introduces downward bias in my estimates. Moreover, the fact that the family did not own land earlier implies that the land was in most probability newly acquired and hence cannot represent ancestral property. In the second case, the reform would have had an impact on the woman's education and excluding her also leads to downward bias.

of survey, the youngest cohort of women were born in 1984 (coming from the 2005 wave).[20] Hence, the variation in $D_{s,(k \geq k'-5)}$ primarily comes from Andhra Pradesh, Kerala and Tamil Nadu, while all five reforming states contribute to the variation in the remaining cohort categories.

3 Impact on Female Education

3.1 Difference-in-Differences Results

Results obtained from estimating equation 1 are presented in Table 2. Focussing on column (2), which includes state and cohort of birth fixed effects, state-specific linear cohort trends and household controls, I find that the suggested impact of exposure to the reform is an increase in mean educational attainment of the 5 or younger group by 0.57 years, and that of the 6-10 group by 0.5 years. This represents an improvement of approximately 0.02 standard deviations for both these "treated" groups relative to the control group. I cannot reject the equality of these two coefficients (δ_1 and δ_2), but can reject (at 1 percent level) the equality of each of them to the coefficient for the 16-20 group (δ_4), which is statistically insignificant as well as small in magnitude. Since women in the 16-20 group in the reforming states were already past school-going age by the time the reform took place, they would not be expected to benefit differentially in terms of education relative to the non-reforming states. This falsification exercise thus increases our confidence that the results are less likely to be driven by correlated unobservables, as well as allows us to rule out the concern that improvements in female education could have led to the passage of the reform in the first place (reverse

[20] A small proportion of interviews in the 2005 wave were carried out in 2006, hence the youngest cohort is that of 1984 rather than 1983.

causality). The coefficient for the 11-15 group (δ_3) is also small in magnitude and statistically insignificant.

3.2 Triple Differences Results

But as outlined earlier, these difference-in-difference estimates do not control for unobservable factors correlated with the passage of the reform that could also affect female education. Hence, I turn to the triple differences approach, using a third source of variation within state-cohorts.

3.2.1 Land Ownership

Table 3 presents the results from estimating the triple differences using equation 2, where the third dimension of variation is in terms of land ownership status of the household the woman belongs to. Column (1) replicates the difference-in-differences result from column (2) of Table 2 for ease of reference, while column (2) provides the triple differences result. In the latter column, the coefficients for the uninteracted cohort groups capture the impact of the reform on women belonging to non-landed households in the reforming states (the δs in equation 2), while those for the cohort groups interacted with the variable "owns land" capture the differential impact of the reform on women who belong to landed households in the reforming states (the δ's in equation 2). Both specifications control for state fixed effects, cohort of birth fixed effects, state-specific linear cohort trends and household controls.

No impact is observed in column (2) of Table 3 on the education levels of non-landed women in any of the age groups in the reforming states, but there exists a positive and significant impact for those who were 5 years or younger and between 6-10 years old at the time of the reform and belonged to landed

households in these states. The suggested effect is that being exposed to the reform increased mean educational attainment of women who were of primary school-going age at the time of reform by 0.7-0.8 years in landed relative to non-landed households in the reforming states, an increase of approximately 0.02 standard deviations. No such differential impact is observed for women who were 16-20 years old (δ_4') at the time of reform, thereby increasing our confidence in the validity of the results. Moreover, the F-test also rejects the equality of the coefficients for 5 or younger and 16-20 groups for the landed (δ_1' and δ_4') at 5 percent and that for the 6-10 and 16-20 groups (δ_2' and δ_4') at 10 percent level.

3.2.2 Hindu

Table 4 presents the results from estimating the triple differences using equation 2, where the third dimension of variation is whether or not the woman belonged to a "Hindu" family. Column (1) replicates the difference-in-difference result from column (2) of Table 2 for ease of reference, while column (2) provides the triple differences result. The coefficients for the uninteracted cohort groups in column (2) capture the impact of the reform on "non-Hindu" women in the reforming states (the δs), while those for the cohort groups interacted with the variable "Hindu" capture the differential impact of the reform on "Hindu" women in the reforming states (the δ's). As in Table 3, I control for state and cohort fixed effects and household controls, but now also add state-religion-specific linear cohort trends to allow for the fact that "Hindus" and "non-Hindus" may have evolved differently across cohorts in different states.

Once again, no impact is observed on the education levels of "non-Hindu" women of any age group in the reforming states, and the coefficients are ac-

tually all negative in sign. For the Hindu women in these states, on the other hand, we find a positive impact on the education level of those who were 5 years or younger and between 6-10 years old at the time of the reform. Note that although the coefficient for the 5 or younger group is significant only at the 10 percent level, its magnitude is quite large. The suggested effect is that being exposed to the reform increased mean educational attainment of "Hindu" women who were of primary school-going age by 1.4-1.5 years compared to "non-Hindu" women in reforming states, an increase of approximately 0.05 standard deviations. No such differential impact is observed for women who were 16-20 years old at the time of reform. The F-test only barely fails to reject the equality of the coefficients for the 5 or younger and 16-20 groups of "Hindus" (δ_1' and δ_4') (p=0.11) but can reject the equality of the coefficients for the 6-10 and 16-20 groups of "Hindus" (δ_2' and δ_4') at 5 percent level.

3.2.3 Gender

Along with the sample of women used for the above analyses, I also observe their brothers in the NFHS dataset, i.e. the sons of the head of the household. Just like in case of the women, I restrict the sample to include only those men who were at least 22 years of age at the time of survey. There are 70,466 such men in my sample, with year of birth spanning 1943 to 1984.[21] For triple differences using gender, I compare "treated" and "control" cohorts between women and men for reforming versus non-reforming states, and the results are presented in Table 5. In each column, the coefficients for the uninteracted cohort groups capture the impact of the reform on men in the reforming states (the δs), while those for the cohort groups interacted with the variable

[21]Of the total 70,466 men, 22,831 obtain from the 1992 round, 23,946 from the 1998 round and 23,689 from the 2005 round.

"daughter" capture the differential impact of the reform on women in the reforming states (the δ's).

Column (1) of Table 5 includes state fixed effects, gender-specific cohort of birth fixed effects (to allow for the fact that education of girls and boys evolved differentially across cohorts) as well as state-specific linear cohort trends and household controls. Focusing on column (1), I find that the impact of the reform on the 5 or younger group of women is positive and highly significant, while that on the corresponding group of men is actually negative and significant. I can reject the equality of these two coefficients (δ_1 and δ_1') at the 1 percent level. The coefficient for the 6-10 group of women is also large but only marginally significant. Nonetheless, I can still reject the equality of this coefficient with the corresponding one for men (δ_2 and δ_2') at the 5 percent level. In addition, both δ_1' and δ_2' are significantly different from δ_4', which passes the falsification test. The suggested impact is women who were exposed to the reform gained approximately 1.1-1.3 additional years of education in the reforming states relative to men, which represents an improvement of approximately 0.03 standard deviations.

The specification in column (1) of Table 5 uses variation in gender across households. Since the sample contains a lot more men than women, it is possible that girls and boys live in different types of households, and the estimates are picking up some of these unobserved differentials that are correlated with education. To address this concern, I introduce household fixed effects in column (2). I restrict the sample to only those households that have at least two children, which reduces the sample size somewhat. The coefficient for the 5 or younger group of women continues to remain highly significant and is indeed slightly larger in magnitude compared to column (1), and I am also able to reject the equality of this coefficient with the

corresponding one for men at the 5 percent level. The coefficient for the 6-10 group of women is also similar in magnitude and level of significance to that in column (1). It is also interesting to note that after controlling for household fixed effects, the coefficients for all age groups of men is positive, although statistically insignificant, which is different compared to what was obtained in column (1). This indicates that there is little evidence of any compensating behaviour on part of the parents towards their sons in response to the inheritance rights reform, which also justifies the use of men as a relevant control group for women.

4 Mechanism

So far, I provide evidence that being exposed to the inheritance rights reform in India was associated with an increase in mean educational attainment for women. But what explains this effect? A key to understanding the mechanism behind this effect would be to first study the impact the reform had on actual likelihood of inheritance by women.

Unfortunately, the NFHS does not contain information on inheritance of women. However, the Rural Economic and Demographic Survey (REDS), which is a representative survey of households from 16 major states of India, contains retrospective information on topics like inheritance and marriage for all members of the household, including for daughters who have married and left the household. In the next section, I use the 1999 wave of REDS to examine the impact of the inheritance reform on likelihood of inheritance by women.[22]

[22]The states that are excluded here but are included in NFHS are Arunachal Pradesh, Chattisgarh, Goa, Jharkhand, Manipur, Meghalaya, Mizoram, Nagaland, New Delhi, Sikkim and Tripura.

4.1 Inheritance

For the inheritance analysis using REDS 99 dataset, I again focus on women who are daughters of the head of the household and at least 22 years of age at the time of survey, such that year of birth ranges from 1946 to 1977. In addition, I restrict the sample to "Hindu" women (since almost 92 percent of the women in this sample belong to these religions) and women belonging to landed households.[23] This leaves me with a dataset of 3,515 women. I also merge the two "treated" groups used in the education analysis above to create a single "treated" group - women who were 10 or younger at the time of reform.

I analyze the impact of exposure to the inheritance rights reform on the probability of inheritance by women by using a triple differences methodology with respect to timing of the death of the paternal grandfather. Inheritance typically occurs when the head of the family dies and the property is partitioned. Since the inheritance rights reform relates to ancestral property, the relevant event in this case is the death of the paternal grandfather in the family. The REDS 99 dataset contains information on the year of death of the paternal grandfather of the women, which allows me to use the variation in the timing of the grandfather's death relative to that of the passage of the reform to identify the impact of the reform on women's likelihood of inheritance. Only if the paternal grandfather died after the reform is passed would there be any impact of the reform on the inheritance of the granddaughters, as in such families bequest is yet to be realized, and ought to follow the new rules.

[23]Restricting the sample to women in landed "Hindu" households is sensible in this context because as results in the previous section showed, this is the group that is most likley to benefit from the reform.

The regression equation takes the following form:

$$\begin{aligned} p_{isk} = \alpha_s + \beta_k + \gamma_s k + + \delta_1 D_{is,(k \geq k'-10)} + \delta_2 D_{is,(k'-15 \leq k \leq k'-11)} \\ + \delta_3 D_{is,(k'-20 \leq k \leq k'-16)} + \delta_1' D_{is,(k \geq k'-10)} * GFA_i \\ + \delta_2' D_{is,(k'-15 \leq k \leq k'-11)} * GFA_i + \delta_3' D'_{is,(k'-20 \leq k \leq k'-16)} * GFA_i \\ + \mu GFA_i + X_{isk}\eta + \epsilon_{isk} \end{aligned} \tag{3}$$

where the dependent variable p_{isk} is a binary variable that takes the value 1 if the daughter inherits any land and 0 otherwise, while GFA_i is a dummy that captures whether or not the grandfather in the household died after the reform was passed. The remaining variables are same as before. According to the amendment of the Hindu Succession Act (1956), *unmarried* women (at the time of the reform) were eligible to inherit ancestral property. Since the mean age at marriage in this sample is approximately 18 years, one would expect that women in the groups 10 or younger and 11-15 would display increased likelihood of inheritance, while those in the group 16-20 would not.

The results from estimating equation 3 are presented in Table 6. Focusing on column (2) that controls for state and cohort fixed effects as well as state-specific trends, I find no differential impact on the likelihood of inheriting land for women who were 10 or younger at the time of reform whose grandfather died after the reform relative to before (equality of the two coefficients cannot be rejected at conventional levels). Similarly for women in the 11-15 group. For 16-20 group, women whose grandfather died after the reform were somewhat less likely to inherit, but I cannot reject the equality of this coefficient with that for 16-20 women whose grandfather died before the reform, which is statistically insignificant and small in magnitude. Hence, the evidence in Table 6 appears to suggest that daughters have continued to be

disinherited from ancestral property even after the reform was passed. This could potentially reflect a weak enforcement of the inheritance rights reform and is consistent with anecdotal evidence obtained in an ethnographic study conducted by Brown, Ananthpur, and Giovarelli (2002) in rural Karnataka, where they find that the amendment to the inheritance law had no impact on the likelihood of inheritance by daughters.[24]

Traditionally, women in India rarely inherited property in their parental family. A potential reason behind such an occurence may be traced to the *virilocal* nature of Indian society, where married daughters leave their parental household while married sons stay and work with family property, e.g. land. In such a societal set-up, giving property bequests to daughters can have poor incentive effect on sons in relation to extending family wealth (Botticini and Siow, 2003). Since married daughters in virilocal societies leave their parental home, if they were to share equally in family property, then their brothers, who stay with their parents and work with family property, will not obtain the full benefits of their effort in extending family wealth and hence will supply too little effort.[25] Thus, to mitigate this incentive problem, parents leave bequests only for their sons, while daughters are traditionally given dowries at the time of their marriage as a form of "pre-mortem" bequest or, if viewed in a different light, as compensation for giving up their rights to household property.[26]

The evidence presented in Table 6 suggests that the situation of women with regard to inheritance did not improve significantly even after the inher-

[24]This would, however, not be the first example of a gender progressive law biting the dust when it comes to practical implementation - the Dowry Prohibition Act (1961) is another law that is regularly flouted in practice.

[25]This is especially true for the rural parts of India where sons typically follow parental occupation.

[26]This was despite the original Hindu Succession Act 1956 granting women equal share, with their brothers, to their father's separate property.

itance rights reform entitled them to a share in ancestral property. But what was the implication of this for dowries? Did the fact that post reform daughters were now being disinherited from a larger share in household property result in greater compensation in the form of larger dowries or were parents substituting away from dowries to other forms of compensation or neither? I turn to examining the impact of the reform on dowry payments in the next section in order to provide a suggestive explanation to these questions.

4.2 Dowry Payments

REDS 99 contains information on nominal dowry payments made by parents at the time of the daughter's marriage. I convert these nominal dowry payments in the dataset to real values using the Indian Consumer Price Index (base: 1966 = 100).[27]

For the dowry analysis, I revert back to using a triple-differencing methodology using variation in land ownership of the daughter's household (as used for the education results in Table 3), but continue to restrict the sample to only "Hindu" households.

Table 7 presents results for the impact of exposure to the inheritance rights reform on dowry payments. The dependent variable is log of real dowry payments made at the time of marriage of the daughter. The two main groups of interest here are 10 or younger and 11-15. For the 16-20 group, a lot of the women are likely to have been married by the time of the reform (mean age at marriage in my sample is 18 years), and hence would be unaffected by the reform, such that one would not expect to see any

[27]I use Consumer Price Index for Agricultural Workers as the deflator since the REDS dataset focuses on a rural sample. I thank Robin Burgess for generously granting me access to his Indian states data for this purpose. Also, over 90 percent of the families in my sample pay dowry and receive nothing, hence I only focus on dowry payments.

differential impact on the average dowry payments of this group relative to the comparison group.

Focusing on column (2) in Table 7, we find that mean dowry payments for women in the 11-15 group increases by 0.45 percentage points in the landowning households relative to the non-landowing ones in reforming states, while that of the 10 or younger group declines by 0.47 percentage points. This latter group is also the same one that enjoys a higher level of educational attainment post reform, while the former did not experience any significant change in their attainment. This could have two potential explanations: parents choose to compensate daughters only along one dimension - the 11-15 group, that were not yet of marriageable age at the time of the reform and hence affected by it, were compensated for giving up right to a larger share of household property following the reform by being given larger dowries at marriage. On the other hand, for the 10 or younger group, the compensation took the form of increased investment in their education, while at the same time, being paid lower dowries. The 11-15 group did not benefit in terms of education since they were already past primary school-going age at the time of reform. This explanation views education and dowry payments as competing channels of compensating daughters for disinheriting them from their rightful share in household property. Alternatively, if dowry is interpreted as a price that clears marriage markets, then higher education could substitute for dowry payments for the 10 or younger group as more educated brides enjoy higher valuation in the marriage market and hence have to pay lower dowries to secure the groom of their choice.

An important question to ask here would be why, in the face of weak enforcement of the reform as captured by unchanged likelihood of female inheritance, would parents increase dowry payments to their daughters at

all? In answering this question, we first need to consider an alternative interpretation of the term "dowry". As Botticini and Siow (2003) point out, "dowry" in contemporary India refers to the goods (including cash) that the groom and his family demand from the bride's family at the time of marriage, over which the bride retains no ownership. Now, even if the inheritance rights reform is weakly enforced such that parents can continue to bypass giving their daughters their rightful share of household property, under the assumption that knowledge of the reform is perfectly diffused among the rest of the population, the potential groom of the daughter and his family would know that the bride is entitled to a share, and in the absence of being granted that share, would demand a higher dowry from the bride's family as compensation for the loss of income that would have accrued to the bride, and thereby to the groom and his family, through that share.

A further question would be why do parents switch from dowry to education as a means of compensating younger daughters after the reform? One potential explanation could be that since post reform dowries are larger and thereby costlier to provide, parents switch to education as an alternative form of compensation for those daughters who are still of school-going age, and also pay them lower dowries. An alternative explanation could be that since the daughter retains hardly any control over her dowry after her marriage, it does not necessarily improve her welfare in reality. Education, on the contrary, is inalienable as an investment in the daughter and hence may be preferred by parents as a means of compensation, especially within a scenario of changing returns to female education in a growing Indian economy. However, the available data does not allow me to provide any evidence on any of these channels or disentangle them.

To sum up, in this section I provide suggestive evidence that a potential

mechanism explaining the positive impact of the inheritance rights reform on female education could be that investment in education constituted a form of compensation to young daughters for disinheriting them from ancestral property. For those daughters who were past school-going age by the time the reform was enacted, compensation consisted of larger dowries. However, it is important to emphasize here that this evidence far from conclusive, and does not prove that this is *the* mechanism driving the education results.

5 Conclusion

Human capital investment is widely considered to be one of the most important drivers of economic growth. This is especially relevant in the case of women as it is well-acknowledged that greater schooling of women enhances the human capital of the next generation and thus makes a unique contribution to economic growth (Behrman, Foster, Rosenzweig, and Vashishtha, 1999). This paper studies the impact of women's property inheritance rights on their human capital attainment by exploiting exogenous variation generated by state level amendments to the central inheritance law in India. I use a difference-in-differences approach that takes advantage of the fact that different states reformed the law at different points in time for identification. In particular, I compare educational outcomes of women who were of primary school-going age at the time of reform (exposed or "treated" group) to those who were too old to go to school ("control" group), between reforming and non-reforming states. I find that being exposed to the reform was associated with a significant improvement in the mean educational attainment of women. In order to improve identification, I also use a triple differences strategy to compare the difference-in-differences estimate by land ownership

status of households, religious affiliation and gender. I find that the positive impact on education exists only for women (compared to men) and that too, only for those women belonging to landed (compared to non-landed) and "Hindu" (compared to "non-Hindu") households.

This paper also attempts to provide suggestive evidence on a potential underlying mechanism explaining this observed effect of the inheritance rights reform on female education. I find that even though the reform entitled daughters to inherit equal shares in joint family property as sons, in reality, this did not happen. In other words, I find no impact of the reform on likelihood of inheritance by women. Instead, parents appear to be compensating their daughters for disinheriting them such by investing in their education as a form of alternative transfer of wealth. For those daughters who were already past school-going age at the time of the reform, the compensation takes the form of higher dowry at the time of their marriage.

Thus, the findings obtained in this paper have policy implications beyond the Indian context with regard to how socio-personal laws can affect economic outcomes. To the extent that inequality in opportunity for women can be traced to legal provisions, changes in inheritance legislation have the potential of addressing gender imbalances and influencing a wide range of outcomes for women, with economy-wide implications.

However, a relevant question to ask in this regard concerns the scalability of such amendments in order to ensure that the benefits can be reaped by a bigger share of the population. Indeed, the amendment to the Hindu Succession Act 1956 as described in this paper was extended to the whole of India in 2005, and it will be interesting to explore if the benefits enjoyed by the women in the first set of reforming states are subsequently shared by the rest of the country's female population.

References

Agarwal, B. (1994): *A Field of One's Own: Gender and Land Rights in South Asia*. Cambridge University Press.

Anderson, S. (2003): "Why Dowry Payments Declined with Modernization in Europe but are Rising in India?," *Journal of Political Economy*, 111(2), 269–310.

——— (2004): "Dowry and Property Rights," BREAD Working Paper No. 080.

——— (2007): "The Economics of Dowry and Brideprice," *Journal of Economic Perspectives*, 21(4), 151174.

Banerjee, A., P. Gertler, and M. Ghatak (2002): "Empowerment and Efficiency: Tenancy Reform in West Bengal," *Journal of Political Economy*, 110(2).

Becker, G. S. (1981): *A Treatise on the Family*. Harvard University Press, Cambridge, MA.

Behrman, J. R., A. D. Foster, M. R. Rosenzweig, and P. Vashishtha (1999): "Women's Schooling, Home Teaching and Economic Growth," *Journal of Political Economy*, 107(4), 682–714.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics*, 119(1), 249–75.

Besley, T. (1995): "Property Rights and Investment Incentives: Theory and Evidence from Ghana," *Journal of Political Economy*, 103(5), 903–937.

Besley, T., and M. Ghatak (2009): "Property Rights and Economic Development," in *Handbook of Development Economics*, ed. by D. Rodrik, and M. Rosenzweig. North Holland.

Botticini, M., and A. Siow (2003): "Why Dowries?," *The American Economic Review*, 93(4), 1385–1398.

Brown, J., K. Ananthpur, and R. Giovarelli (2002): "Women's Access and Rights to Land in Karnataka," Rural Development Institute Report on Foreign Aid and Development No. 114.

Card, D., and A. B. Krueger (1992): "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100(1), 1–40.

De Soto, H. (2000): *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. Basic Books (New York) and Bantam Press/Random House (London).

Deere, C. D., and M. Leon (2003): "The Gender Asset Gap: Land in Latin America," *World Development*, 31(6), 925–947.

DiTella, R., S. Galiani, and E. Schargrodsky (2007): "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters," *Quarterly Journal of Economics*, 122(1), 209–241.

Duflo, E. (2001): "Schooling and Labour Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91(4), 795–813.

Edlund, L. (2001): "Dear Son-Expensive Daughter: Why Do Scarce Women Pay to Marry?," Working paper, Columbia University.

Field, E. (2007): "Entitled to Work: Urban Property Rights and Labor Supply in Peru," *Quarterly Journal of Economics*, 122(4), 1561–1602.

Foster, A., and M. Rosenzweig (2002): "Household Division and Rural Economic Growth," *Review of Economic Studies*, 69(4), 839–869.

Goody, J. (1973): "Bridewealth and Dowry in Africa and Eurasia," in *Bridewealth and Dowry*, ed. by J. Goody, and S. J. Tambiah, pp. 1–58. Cambridge University Press.

Goyal, A., K. Deininger, and H. Nagarajan (2010): "Hindu Inheritance Law, Land Bequests and Educational Attainment of Women in India," in *Innovations in Land Rights Recognition, Administration and Governance*, ed. by S. E. C. Augustinus, K. Deininger, and P. Munro-Faure, pp. 105–116. World Bank.

Johnson, S., J. McMillan, and C. Woodruff (2003): "Property Rights and Finance," *American Economic Review*, 92(5), 1335–1356.

Lemieux, T., and D. Card (2001): "Education, Earnings, and the 'Canadian G.I. Bill'," *The Canadian Journal of Economics*, 34(2), 313–344.

North, D. C. (1990): *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge.

Ramanujam, T. C. A. (2005): "Daughters Better Off than Sons," *The Hindu Business Line*, Saturday 1 October.

Rao, V. (1993): "The Rising Price of Husbands: A Hedonic Analysis of Dowry Increases in Rural India," *Journal of Political Economy*, 101(3), 666–77.

Rosenzweig, M., and O. Stark (1989): "Consumption Smoothing, Migration and Marriage: Evidence from Rural India," *Journal of Political Economy*, 97(4), 905–926.

Sweetman, C. (2008): "How Title Deeds Make Sex Safer: Womens Property Rights in an Era of HIV," *From Poverty to Power: How Active Citizens and Effective States Can Change the World*, Oxfam International.

Table 1: Descriptive Statistics

	Reforming			Non-Reforming	All
	≥ 21	≤ 5	All		
Female	0.31	0.20	0.24	0.19	0.21
	[0.46]	[0.40]	[0.43]	[0.39]	[0.40]
For Females:					
Age	31.05	24.67	27.33	26.82	26.96
	[7.00]	[2.87]	[5.55]	[5.30]	[5.39]
Years of education	6.55	10.78	8.65	7.94	8.14
	[5.58]	[5.58]	[5.40]	[5.85]	[5.74]
Father's age	63.61	58.88	61.19	60.48	60.68
	[9.38]	[7.00]	[8.56]	[8.56]	[8.56]
Mother's age	55.40	50.99	53.12	53.76	53.58
	[7.70]	[5.49]	[6.76]	[7.03]	[6.96]
Father's education	5.35	6.26	5.99	5.96	5.97
	[4.88]	[4.60]	[4.80]	[5.33]	[5.20]
Mother's education	2.92	5.31	4.05	3.25	3.48
	[3.95]	[4.48]	[4.36]	[4.48]	[4.46]
Land ownership	0.44	0.28	0.36	0.52	0.48
	[0.49]	[0.45]	[0.48]	[0.49]	[0.49]
Hindu	0.76	0.76	0.78	0.64	0.83
	[0.42]	[0.43]	[0.41]	[0.47]	[0.36]
HH members	7.76	6.55	7.09	7.63	7.48
	[3.71]	[3.08]	[3.29]	[3.44]	[3.41]
Urban	0.48	0.44	0.50	0.42	0.45
	[0.49]	[0.49]	[0.50]	[0.49]	[0.49]

Notes: * denotes significant at 10 percent, ** denotes significant at 5 percent, *** denotes significant at 1 percent. Numbers in square brackets denote standard deviations."Reforming" denotes states that passed the amendment to the HSA 1956 (i.e. Kerala, Andhra Pradesh, Tamil Nadu, Maharashtra and Karnataka), under which summary statistics are presented separately for groups of women who were 21 or older at reform (denoted by ≥ 21) and those who were 5 or younger at reform (denoted by ≤ 5). "Non-reforming" denotes all the states that did not reform, but a similar split by age at reform is not possible for this category as year of reform varies by state.

Table 2: Impact of Inheritance Reform on Women's Education: Difference-in-Differences

	Years of education	
	(1)	(2)
Aged 5 or less at time of reform	1.10***	0.57**
	(0.36)	(0.26)
Aged 6-10 at time of reform	0.99***	0.50**
	(0.28)	(0.20)
Aged 11-15 at time of reform	0.60**	0.15
	(0.23)	(0.26)
Aged 16-20 at time of reform	0.28	-0.01
	(0.25)	(0.25)
Household controls	Yes	Yes
State fixed effects	Yes	Yes
Cohort of birth fixed effects	Yes	Yes
State-specific linear cohort trends	No	Yes
Adj. R-sq	0.58	0.58
No. of observations	15466	15466

Notes: Standard errors are clustered at the state level and presented in parentheses. * denotes significant at 10 percent, ** denotes significant at 5 percent, *** denotes significant at 1 percent. Household controls include a dummy for "Hindu", which is 1 for Hindus, Buddhists, Sikhs and Jains (to whom the Hindu Succession Act 1956 applied) and 0 otherwise, father's education, father's age, mother's education, mother's age (all in years), a dummy for whether or not the parental household of the woman owns land, a dummy for whether or not the parental household of the woman resides in an urban area and number of household members.

Table 3: Impact of Inheritance Reform on Women's Education: Triple Differences by Land Ownership

	Years of education	
	(1)	(2)
Aged 5 or less at time of reform	0.57**	0.23
	(0.26)	(0.40)
Aged 6-10 at time of reform	0.50**	0.18
	(0.20)	(0.28)
Aged 11-15 at time of reform	0.15	0.00
	(0.26)	(0.31)
Aged 16-20 at time of reform	-0.01	-0.04
	(0.25)	(0.39)
Aged 5 or less at time of reform*owns land		0.83**
		(0.32)
Aged 6-10 at time of reform*owns land		0.74*
		(0.40)
Aged 11-15 at time of reform*owns land		0.28
		(0.39)
Aged 16-20 at time of reform*owns land		-0.01
		(0.44)
Owns land	0.27**	0.18
	(0.12)	(0.16)
Household controls	Yes	Yes
State fixed effects	Yes	Yes
Cohort of birth fixed effects	Yes	Yes
State-specific linear cohort trends	Yes	Yes
Adj. R-sq	0.58	0.58
No. of observations	15466	15466

Notes: Standard errors are clustered at the state level and presented in parentheses. * denotes significant at 10 percent, ** denotes significant at 5 percent, *** denotes significant at 1 percent. Household controls include a dummy for "Hindu", which is 1 for Hindus, Buddhists, Sikhs and Jains (to whom the Hindu Succession Act 1956 applied) and 0 otherwise, father's education, father's age, mother's education, mother's age (all in years), a dummy for whether or not the parental household of the woman resides in an urban area and number of household members.

Table 4: Impact of Inheritance Reform on Women's Education: Triple Differences by Hindu

	Years of education	
	(1)	(2)
Aged 5 or less at time of reform	0.57**	-0.65
	(0.27)	(0.67)
Aged 6-10 at time of reform	0.50**	-0.61
	(0.19)	(0.64)
Aged 11-15 at time of reform	0.15	-0.09
	(0.26)	(0.49)
Aged 16-20 at time of reform	-0.02	-0.45
	(0.25)	(0.62)
Aged 5 or less at time of reform*hindu		1.57*
		(0.83)
Aged 6-10 at time of reform*hindu		1.41**
		(0.64)
Aged 11-15 at time of reform*hindu		0.28
		(0.43)
Aged 16-20 at time of reform*hindu		0.52
		(0.52)
Household controls	Yes	Yes
State fixed effects	Yes	Yes
Cohort of birth fixed effects	Yes	Yes
State*religion linear cohort trends	Yes	Yes
Adj. R-sq	0.58	0.58
No. of observations	15466	15466

Notes: Standard errors are clustered at the state level and presented in parentheses. * denotes significant at 10 percent, ** denotes significant at 5 percent, *** denotes significant at 1 percent. Household controls include a dummy for "Hindu", which is 1 for Hindus, Buddhists, Sikhs and Jains (to whom the Hindu Succession Act 1956 applied) and 0 otherwise, father's education, father's age, mother's education, mother's age (all in years), a dummy for whether or not the parental household of the woman owns land, a dummy for whether or not the parental household of the woman resides in an urban area and number of household members.

Table 5: Impact of Inheritance Reform on Education: Triple Differences by Gender

	Years of education	
	(1)	(2)
Aged 5 or less at time of reform	-0.46**	0.49
	(0.22)	(0.43)
Aged 6-10 at time of reform	-0.26	0.30
	(0.18)	(0.37)
Aged 11-15 at time of reform	-0.25	0.21
	(0.15)	(0.30)
Aged 16-20 at time of reform	-0.07	0.14
	(0.14)	(0.32)
Aged 5 or less at time of reform*daughter	1.33***	1.65***
	(0.42)	(0.47)
Aged 6-10 at time of reform*daughter	1.08*	1.15*
	(0.54)	(0.57)
Aged 11-15 at time of reform*daughter	0.63	0.55
	(0.39)	(0.57)
Aged 16-20 at time of reform*daughter	0.31	0.08
	(0.29)	(0.49)
Daughter	-6.51***	-6.11*
	(1.02)	(3.47)
Household controls	Yes	No
State fixed effects	Yes	No
Gender-cohort of birth fixed effects	Yes	Yes
State-specific linear cohort trends	Yes	No
Household fixed effects	No	Yes
Adj. R-sq	0.43	0.65
No. of observations	73276	56915

Notes: Standard errors are clustered at the state level and presented in parentheses. * denotes significant at 10 percent, ** denotes significant at 5 percent, *** denotes significant at 1 percent. Household controls include a dummy for "Hindu", which is 1 for Hindus, Buddhists, Sikhs and Jains (to whom the Hindu Succession Act 1956 applied) and 0 otherwise, father's education, father's age, mother's education, mother's age (all in years), a dummy for whether or not the parental household of the woman owns land, a dummy for whether or not the parental household of the woman resides in an urban area and number of household members.

Table 6: Impact of Inheritance Reform on Likelihood of Inheritance by Hindu Women

	Only landed households	
	Inheritance	
	(1)	(2)
Aged 10 or less at time of reform	-0.12***	-0.05
	(0.03)	(0.03)
Aged 11-15 at time of reform	-0.03	-0.03*
	(0.03)	(0.02)
Aged 16-20 at time of reform	0.02	0.01
	(0.01)	(0.02)
Aged 10 or less at time of reform*grandfather died after reform	0.01	0.01
	(0.02)	(0.02)
Aged 11-15 at time of reform*grandfather died after reform	-0.05	-0.05
	(0.03)	(0.03)
Aged 16-20 at time of reform*grandfather died after reform	-0.04**	-0.04**
	(0.02)	(0.02)
Grandfather died after reform	-0.01	-0.00
	(0.02)	(0.02)
State fixed effects	Yes	Yes
Cohort of birth fixed effects	Yes	Yes
State-specific linear cohort trends	No	Yes
Adj R-sq	0.08	0.09
No. of observations	3514	3514

Notes: Standard errors are clustered at the state level and presented in parentheses. * denotes significant at 10 percent, ** denotes significant at 5 percent, *** denotes significant at 1 percent. The dependent variable is a dummy which equals 1 if the woman has inherited any land in her parental household and 0 otherwise.

Table 7: Impact of Inheritance Reform on Dowry Payments of Hindu Women

	Log(dowry payments)	
	(1)	(2)
Aged 10 or less at time of reform	0.05	-0.13
	(0.15)	(0.24)
Aged 11-15 at time of reform	-0.51**	-0.53*
	(0.22)	(0.26)
Aged 16-20 at time of reform	-0.27	-0.20
	(0.16)	(0.18)
Aged 10 or less at time of reform*owns land	-0.15	-0.47**
	(0.13)	(0.17)
Aged 11-15 at time of reform*owns land	0.48**	0.45**
	(0.20)	(0.18)
Aged 16-20 at time of reform*owns land	-0.01	0.05
	(0.20)	(0.20)
Owns land	0.44***	0.45***
	(0.09)	(0.09)
State fixed effects	Yes	Yes
Cohort of birth fixed effects	Yes	Yes
State-specific linear cohort trends	No	Yes
Adj R-sq	0.26	0.27
No. of observations	3245	3245

Notes: Standard errors are clustered at the state level and presented in parentheses. * denotes significant at 10 percent, ** denotes significant at 5 percent, *** denotes significant at 1 percent. The dependent variable is log of real dowry payments (in 1966 rupees) made at the time of the woman's marriage.

Table A.1: Distribution of Women by Age at Reform in Reforming States (for education results)

Age at reform	Andhra Pradesh	Karnataka	Kerala	Maharashtra	Tamil Nadu	Total
-8	0	0	82	0	0	82
-7	0	0	79	0	0	79
-6	0	0	97	0	0	97
-5	0	0	92	0	0	92
-4	0	0	89	0	0	89
-3	0	0	84	0	0	84
-2	0	0	73	0	0	73
-1	0	0	169	0	0	169
0	0	0	204	0	0	204
1	0	0	153	0	0	153
2	164	0	172	0	0	336
3	132	0	139	0	0	271
4	131	0	116	0	0	247
5	175	0	236	0	128	539
6	123	0	294	0	120	537
7	99	0	301	0	110	510
8	100	0	235	0	128	463
9	103	0	241	0	92	436
10	209	151	217	286	73	936
11	108	115	176	243	78	720
12	129	127	185	227	190	858
13	119	188	130	312	194	943
14	87	117	135	202	150	691
15	122	88	86	194	152	642
16	161	124	89	188	142	704
17	140	179	71	329	115	834
18	143	269	61	425	115	1,013
19	226	134	76	244	222	902
20	107	248	29	336	208	928
21	70	140	25	213	165	613
22	115	115	35	178	193	636
23	54	328	22	377	137	918
24	166	198	15	244	120	743
25	34	285	18	280	112	729
26	53	193	15	228	66	555
27	25	223	12	219	95	574
28	19	160	8	167	56	410
29	88	115	7	124	58	392
30	23	200	5	154	34	416
31	9	109	3	120	27	268
32	21	97	4	78	66	266
33	7	94	3	93	13	210
34	41	49	0	60	28	178
35	5	81	0	56	22	164
36	8	57	0	50	8	123
37	6	61	0	63	18	148
38	4	32	0	36	9	81
39	12	14	0	23	7	56
40	2	50	0	29	4	85
41	0	18	0	30	4	52
42	0	23	0	27	8	58
43	0	12	0	18	0	30
44	0	6	0	18	2	26
45	0	12	0	9	0	21
46	0	12	0	11	1	24
47	0	17	0	11	0	28
48	0	5	0	4	0	9
49	0	2	0	5	0	7
50	0	10	0	3	0	13
51	0	1	0	2	0	3
Total	3,340	4,459	4,283	5,916	3,470	21,468

Do Property Titles Increase Credit Access Among the Urban Poor? Evidence from a Nationwide Titling Program

ERICA FIELD[†]
Harvard University

MAXIMO TORERO
Group for Development Analysis, and
International Food Policy Research Institute

March 2006

The collateral value of landholdings is generally assumed to increase with ownership rights, thereby improving credit access among landholders. In situations of poverty, however, this outcome is uncertain because of other significant barriers to lending.

To test whether proof of property ownership promotes the use of low-income housing as collateral, we evaluate the impact on credit supply of obtaining a property title through a land-titling program in Peru. By directly observing whether loan applicants are requested to provide collateral, we can isolate the effect of property titles on credit supply from their effect on demand by comparing loan approval rates when titles are requested to rates when they are not. Our results indicate that property titles are associated with approval rates on public sector loans as much as 12% higher when titles are requested by lenders and no relationship between titles and approval decisions otherwise. In contrast, there is no evidence that titles increase the likelihood of receiving credit from private sector banks, although interest rates are significantly lower for titled applicants regardless of whether collateral was requested.

The failure of commercial banks to increase their rate of lending to households that obtain property titles through government programs has important implications for the potential effects of property reform on economic growth and poverty reduction. One explanation for this failure is that titling programs reduce banks' perceptions of their ability to foreclose. This is supported by data from Peru indicating that individuals with title have *less* fear of losing property in cases of default.

1. INTRODUCTION

A large body of work has documented extensive credit rationing in developing countries, whereby low-income households are excluded from the formal banking sector.[1] The widespread inability of small and informal borrowers to provide secure collateral for loans – generally a necessity in formal credit markets – is a critical barrier to access,[2] and

[†] The authors thank Daniel Andaluz of the Committee for the Formalization of Private Property (COFOPRI) office for providing the survey data, and Attila Ambrus, Anne Case, Hank Farber, Jeff Kling, Princeton labor lunch, LACEA and ASSA conference participants for useful comments and suggestions.

[1] In this context, credit rationing refers to non-price rationing, such that asymmetric information and enforcement costs prevent price from serving as the market-clearing mechanism, in turn creating a disequilibirum of excess demand in the credit market (Stiglitz and Weitz (1981). For an overview of this literature, see Hillier and Ibrahimo (1993) and Jaffee and Stiglitz (1990).

[2] According to Berger and Udell (1990), around 70% of all commercial and industrial loans in the US are secured with collateral. Meanwhile, "Lack of collateral satisfactory to banks has almost always been a constraint on disbursement of World Bank SME lines of credit" (Balkenhol *et al.* (1995)).

the large percentage of untitled property in much of the developing world is a frequently cited contributing factor (Holden (1997)). While land is advantageous as collateral because it cannot be removed and does not easily devalue, it is widely believed that many borrowers face credit barriers for lack of formally documented ownership rights.[3]

Consistent with this notion, government land-titling programs are thought to be critical in improving access to credit among the poor, and wide scale land-titling has become a popular policy prescription for reducing credit constraints in developing countries (Binswanger *et al.* (1999)). Nevertheless, property titles are not necessarily sufficient to transform modest landholdings into viable collateral for commercial loans. Use of titles to securitize loans may fail in impoverished settings because transaction costs involved – such as those associated with collateral processing, foreclosure and resale – are sizable compared with the average loan sought. Such costs are even higher when political or legal factors impede repossession of property (Deninger *et al.* 1993). Even when foreclosure is feasible, a high degree of mistrust often exists among lenders as to the validity of ownership documents, and the cost of verification may be prohibitively high even in the context of a formal property system. If poor households are "transactions-cost rationed" in formal credit markets, the lower default risk brought about by collateral provision may be insufficient to facilitate access to loans. Indeed, past research has found the impact of rural titling programs on credit supply and investment demand to be strongly size-differentiated, rationing small producers out of the credit market even when they have titled collateral (See Carter and Olinto (1997)).

Hence, in an era of land-titling reform motivated by credit market improvements, a key question is whether the distribution of property titles in fact enables lenders to profitably use low-income housing as collateral. This paper examines this question by analyzing lender responses to changes in formal ownership rights brought about by a nationwide titling program in Peru, under which more than 1.2 million property titles were distributed to urban households. The staggered timing of the program combined with the collection of cross-section micro-data after partial implementation enables a comparison of households in neighbourhoods already served by the program with households in neighbourhoods yet to be served. In this way, we assess the effect of obtaining a property title on the likelihood of a household receiving a bank loan.

One contribution of the paper is to examine credit market effects of land titling in an urban setting, whereas the existing empirical work on property titles and credit has focused on rural markets (See Feder et al. (1988); Alston et al. (1996); Lopez and Romano (2000); Carter and Olinto (1997); Atwood (1990); Carter and Wiebe (1990); Migot-Adholla et al. (1991); Christensen et al. (1993)). Titling programs arguably have larger potential impact on credit supply in urban settings where geographic barriers play a minimal role in transactions costs, an increasingly relevant distinction as the number of urban titling programs rises.

A second contribution of the paper is to make use of the natural experiment provided by the Peruvian program to address concerns over the endogeneity of tenure

[3] Furthermore, in many countries, including Peru, legal barriers restrict the use of movable property as collateral, such that real estate is the only viable form of security interest (Fleisig and de la Pena (1996)).

status that arises in measuring the collateral value of property titles by comparing titled to untitled households. In particular, any relationship between legal ownership and credit access may reflect spurious correlation between strong property institutions and well-functioning credit markets. Similarly, the decision to title property may be a function of property values or the perceived collateral value of titled land (direct evidence of this is provided by Miceli *et al.* (2001)). For instance, households may have a tendency to seek property titles in communities where loan transactions are less costly based on external factors such as an adequate local property registry that facilitates title verification or a local court known to uphold loan contracts. If tenure status is endogenous to land values or financial markets, the collateral value of titled properties will overstate the gains to titling untitled properties. The Peruvian program, in which all households were "assigned" property titles irrespective of demand, helps isolate the causal effect of property titling on credit market outcomes by reducing these endogeneity concerns.

A second issue complicating empirical work on property rights and credit supply is the fact that land titling efforts have the potential to affect not only the supply but also the demand for credit, which is also a function of ownership rights since tenure security influences incentives to undertake land-related investments (see Besley (1995)). Because of this complexity, the majority of past work has focused on changes in the demand for credit or changes in the total amount of borrowing associated with improved ownership rights. Hence, a final contribution of this paper is to isolate the role of property titles in increasing credit supply from their effect on demand by using extensive micro-data on the loan approval criteria used by banks, including all documentation and information households were asked to provide in loan applications. This allows relatively precise reconstruction of the information set on which banks based approval decisions, legitimizing a selection on observables model to account for potential changes in the composition of titled and untitled loan applicants arising from changes in demand.

More importantly, within the set of reported screening criteria, we can observe whether property titles were used in each loan approval decision. Clearly, improved access to credit resulting from changes in the collateral value of land can only occur among the 60% of applications in which potential borrowers are asked for proof of property ownership. Meanwhile, any difference in approval rates of titled relative to untitled households that is independent of banks' use of property titles to screen prospective borrowers can be attributed to unobservable changes in the applicant pool arising from changes in demand.

Our estimates indicate that titling programs lead to a limited reduction in overall credit rationing and financial market inequalities for the urban poor. In particular, households with no legal claim to property are 9–10 percentage points less likely to secure a loan from a public-sector bank for housing construction materials. Importantly, the effect is concentrated entirely among applicants asked to provide a title as collateral, providing evidence that the observed program effect arises from the increased collateral value of property. Meanwhile, we find no effect of formal property ownership on approval rates of private sector lenders. However, conditional on receiving a loan, titled households face private sector interest rates an average of 9 percentage points lower. Since the measured effect is independent of banks' reported use of titles in loan

transactions, the program effect of titling on private sector interest rates appears to operate through the signalling value of property ownership rather than by increasing the fraction of debt securitized with collateral.

The failure of commercial banks to expand credit to new property owners has important implications for the potential influence of property reform on economic growth and poverty reduction. One explanation for this finding is that titling programs may actually reduce banks' ability to repossess property, which is supported by data from Peru indicating that individuals with title have *less* fear of losing property in cases of default. This suggests that one reason that titling programs may fail to reduce credit constraints is because they unavoidably signal to lenders that a government prioritizes housing for the poor, and hence is more likely to side with borrowers in enforcing credit contracts.

2. THEORETICAL ISSUES

Loans contracts are characterized not only by interest rates but also by non-price elements including collateral (as emphasized by Baltensperger (1976)).[4] Poor borrowers are frequently denied access to loans because they lack adequate collateral to offer the lender as a warranty for their loan and also because of high costs of monitoring and processing relative to the magnitude of loans requested (Hoff and Stiglitz, 1990). Formal treatment of the link between property rights and credit supply is provided by Besley (1995) and Feder (1985). The principal argument is that formal property titles encourage the use of land as collateral by lowering the risk of loss, the costs involved in verifying ownership and the costs of foreclosure in the case of default, thereby reducing the effective leverage ratio and increasing the net collateral value of land. In competitive markets with full information, improved access to collateral reduces the risk premium, and hence the interest rate, on lending. Meanwhile, in the presence of information asymmetries, the use of collateral can eliminate credit rationing by reducing agency problems.

To motivate the empirical question, we consider the model of Bester (1985), in which collateral is used in conjunction with the interest rate to achieve separation of risk types and eliminate credit rationing. Here, a banker faces a heterogeneous distribution of potential borrowers represented by an unobservable risk parameter, θ, such that $\theta \in \{\theta_a, \theta_b\}$, where θ_b is a higher risk type than θ_a The fact that the banker is unable to identify types will lead him to screen borrowers by offering a menu of contracts $\gamma_i = (R_i, C_i)_{i \in I}$ specified with interest rate R_i, and collateral requirement C_i, and constructed such that each type of borrower will choose a specific type of contract.[5] *A priori* two types of Nash equilibria can be obtained: a separating equilibrium in which different types of borrowers choose different types of contracts, and a pooling equilibrium in which both types choose the same contract. In this paper, we test whether

[4] The theoretical function of collateral in lending is discussed extensively by Binswanger et al. (1985), Barro (1976), Benjamin (1978), and Plaut (1985).

[5] As discussed by Bester, the result depends on the correlation of borrowers' preferences and risk type.

a land titling program, under which borrowers shift from not having to having collateral, induces a separating equilibrium (see Bester (1985) for the formal proof of this equilibrium).

In the absence of collateral, quantity rationing will occur if adverse selection inhibits lenders from raising the equilibrium interest rate. However, as shown in Bester's model, if the equilibrium exists, no credit rationing will occur post-reform because property titles enable banks to use contracts with different collateral requirements as a screening mechanism to separate low-risk from high-risk borrowers.[6] Hence, for beneficiaries of the titling program, the shift from an equilibrium with only one set of price characteristics to another in which contracts include non-price components may prevent or reduce rationing. Clearly, this result depends fundamentally on the degree to which land titling promotes the use of property as collateral. Perfect sorting without rationing may be unattainable if titled borrowers still face a binding constraint on the amount of collateral they can provide (*e.g.* if the value of property does not exceed the transaction costs involved in processing loans).

In addition, land titles may have value in loan transactions other than their use as collateral. First, titled property owners may be offered more credit because of the household's higher expected wealth from reduced risk of expropriation. If lenders use wealth as a signal of default risk, titling may give rise to an alternative separating equilibrium in which *all* titled borrowers are considered less risky irrespective of type, lowering the collateral and interest rate requirements on all equilibrium contracts even when loans are not collateralised with property.[7] Secondly, land titles may influence other borrower characteristics that determine credit-worthiness, most notably employment. In particular, if ownership rights increase household labour hours as found in Field (2004), the corresponding increase in wage income could improve applicants' access to credit. In this paper we ignore the indirect influence of ownership rights on credit-worthiness via changes in employment in order to concentrate on the direct effect of titling on banks' use of property as collateral. Hence, the results provide a lower bound estimate of the total effect of the land-titling program on credit access.

3. EMPIRICAL METHODS

This research examines the Peruvian government's recent series of legal, administrative and regulatory reforms aimed at promoting a formal property market in urban squatter settlements. In 1996, under the auspices of the public agency, COFOPRI (Committee for the Formalization of Private Property), and *Decree 424: Law for the Formalization of Informal Properties*, the Peruvian government embarked on an innovative nationwide program whose goal was "rapid conversion of informal property into securely delineated

[6] We are for simplicity ignoring the possibility of equilibrium credit rationing, in which a borrower's demand for credit can be turned down even if the borrower is willing to pay the entire price and non-price elements of the loan contract. For a discussion, see Baltensperger (1978).

[7] Barham et al. (1996) note that lenders may use wealth to assess borrowers' risk, because "repayment capacity under a negative income shock is likely to be lower for [poor] borrowers because of their inability to suppress consumption to meet loan repayments and … inability to establish a diversified asset portfolio."

land holdings by the issuing and registering of property titles" (World Bank, 1992). Implementation involved area-wide titling, in which project teams entered one neighbourhood at a time, moving contiguously within cities until all informal settlements had been reached (World Bank, 1998). While the old process of acquiring a property title was prohibitively slow and expensive, the new process was virtually free and extremely rapid (see Field (2004) for an overview). Eligibility for program participation required title claimants to verify pre-1995 residency on eligible public properties, generally using informal title documents from local registries, post-dated mail, utilities bills or signed sales documents. As a result of the reforms, roughly 80% of the country's eligible residents became nationally registered property owners, affecting approximately 6.3 million individuals.[8] As target households were living in the range from just above to below the poverty line, the value of residences titled through the program was relatively low: In Lima, a comparison of titled and untitled households showed that, on average, untitled lots were roughly 40% smaller than lots titled prior to the intervention.

3.1 Data

To study the effect of titling on credit access we use survey data from March 2000 containing 2,750 randomly sampled households from the program's target population. The survey was modelled after the World Bank's *Living Standards Measurement Surveys (LSMS)*. In addition to capturing detailed information on household and individual characteristics, the survey collected an extensive array of self-reported data on all loan applications submitted by the sample households between 1997 and 1999, including bank requirements and terms of loans provided.

To tackle the question of whether improvements in land rights reduce credit rationing, our empirical analysis employs a quasi-experimental set-up that ideally mimics an experimental design with treatment and control groups. Because the survey was conducted approximately one-third of the way into the program's implementation, roughly 60% of surveyed households belonged to neighbourhoods not yet served by the program. Hence, the treatment group is composed of 536 households that have already participated in the program, and the control group comprises households that have not.[9] The control group is further refined to include only the 1,180 households that eventually received a registered property title through the program (the other 1,034 discarded households already had a title prior to the program).

Table 1 provides descriptive statistics on the sample population, which allow an informal check for random assignment of program timing. As the sample means indicate, there is very little variation in demographic characteristics across program and non-program neighbourhoods. In contrast, households with titles exhibit substantially

[8] By December 2002, 1.64 million lots had been formalized and 1.21 million titles granted, the vast majority of which took place between 1998 and 2000.

[9] In the results presented in the paper, as opposed to an intent-to-treat (ITT) analysis, households in titled neighbourhoods that have not yet received a title are excluded, presenting a potential bias in the comparison of experimental groups if ability to secure a title is related to credit-worthiness. Estimates not presented here reveal that the magnitude and significance of results are robust to an ITT model.

Table 1: Summary statistics

	All households			*Households requesting a formal loan*		
	Untitled	Titled	$\|t_{\Delta}\|$	Untitled	Titled	$\|t_{\Delta}\|$
N:	*1,180*	*536*		*470*	*253*	
Characteristics of household						
Number of working-age members	4.22	4.15	0.64	4.25	4.25	0.00
Number of members	5.30	5.28	0.23	5.45	5.50	0.32
Number of children aged 5 to 11 years	0.87	0.88	0.08	0.99	0.98	0.12
Number of children aged 12 to 16 years	0.64	0.59	1.25	0.67	0.65	0.27
HH head is female	0.23	0.23	0.15	0.18	0.20	0.75
Age of HH head	48.13	48.68	0.83	46.27	46.58	0.34
HH head is literate	0.93	0.93	0.19	0.95	0.95	0.14
HH head had no schooling	0.06	0.05	0.68	0.04	0.04	0.06
HH head attended high school	0.45	0.43	0.71	0.48	0.49	0.35
HH head's attended post-secondary school	0.07	0.06	0.71	0.09	0.08	0.75
HH head's monthly wage	635.21	575.06	0.94	689.52	573.85	0.96
HH employment days per year	460.72	490.56	1.65	458.42	485.66	1.00
Total monthly HH consumption	546.58	548.20	0.10	573.76	574.31	0.03
Bi-monthly HH food expenditures	189.21	190.44	0.23	194.88	195.08	0.03
HH education expenditures, per year	417.38	403.67	0.48	469.03	439.64	0.75
Whether HH has savings	0.08	0.08	0.21	0.10	0.09	0.48
Whether HH is extremely poor	0.27	0.25	1.00	0.20	0.19	0.60
Characteristics of residence						
Whether HH rents part of residence	0.03	0.03	0.02	0.03	0.04	0.49
Years of residence	1982.70	1981.40	1.31	1984.20	1982.70	1.40
Whether HH has a telephone	0.20	0.18	0.78	0.23	0.20	1.10
Whether HH has a home business	0.24	0.26	0.65	0.29	0.32	0.83
Income from home business	332.79	279.00	1.59	335.50	256.40	1.51
Average distance to formal lender	3.82	4.11	1.24	4.65	4.85	0.64
Closest bank two years ago	0.95	0.96	0.50	0.94	0.95	0.52
District number of bank branches per capita	0.03	0.03	0.08	0.03	0.03	0.65
District amount of deposits per capita, soles	1059.97	926.75	0.91	943.27	863.94	0.75
District number of ATMs per capita	0.04	0.04	0.49	0.04	0.04	0.59
HH lending behavior						
Would accept a formal sector loan	0.60	0.73	5.11	-------	-------	-------
Asked for a formal sector for loan	0.40	0.47	2.87	-------	-------	-------
Requested an informal loan	0.12	0.11	0.64	0.17	0.14	1.12
HH housing improvements						
Housing improvements made, 1997–99	0.46	0.56	2.83	0.63	0.76	2.95
HH improvements financed with formal credit	0.18	0.30	4.09	0.42	0.60	3.70
Housing improvements made, ever	0.75	0.83	2.62	0.87	0.95	3.06
Asked for a construction loan, ever	0.37	0.51	4.94	0.60	0.74	3.46

Note: Observations are households. HH indicates household head. Cells contain sample means as reported in the *2000 COFOPRI Baseline Survey*.

different patterns of borrowing and housing investment behaviour than those without titles – presumably a reflection of greater demand for investment associated with higher tenure security. In particular, titled households are 10% more likely to have undertaken housing improvements in the two years prior to the survey and 8% more likely to have

made improvements to the house at some point in the past.[10] Of households that engaged in housing improvements between 1997 and 1999, titled households are 15% more likely to finance improvements through formal loans, and this difference is statistically significant (See Field (2004) for a detailed description of investment responses to the program).

Correspondingly, formal credit demand also increases significantly, measured as the share of households applying for formal loans (18 percentage points higher for households with titles) and the share reporting willingness to accept a loan from a formal lender (14 percentage points higher for households with titles). Regression estimates of the effect of property titles on formal credit applications that control for observable household characteristics are reported in Appendix 1. These estimates indicate that household traits account for more than half the difference in loan application rates between titled and untitled households. Columns 3 and 4 of Table 1 compare observable characteristics of loan applicants with and without titles, the sub-population used in our analysis. While differences exist in the demand for credit associated with property ownership, observable differences between untitled and titled households are even smaller with respect to almost every demographic characteristic. This indicates that marginal applicants (those encouraged to apply for a loan in response to receiving a title) are similar in observable characteristics to unconditional applicants (those who would have applied for a loan in the absence of the program).

3.2 Aggregate level of credit rationing

Before attempting to differentiate supply- and demand-side effects, we first explore whether changes in the demand for loans are accompanied by changes in the aggregate level of credit rationing. If the net change in the level of borrowing exceeds the increase in demand for loans associated with the titling program, we can conclude that credit access has also improved post-reform. Following the definitions of Feder et al. (1991), we classify sample households as fully constrained, partially constrained or unconstrained (price rationed) in formal credit markets. Households with a 100% rejection rate on loan applications between 1997 and 1999 are classified as fully constrained, households who received a lower amount than requested are classified as partially constrained and households with a 100% approval rate on loans at the amount requested are classified as unconstrained.

Rather than inferring demand from loan requests, we follow Barham *et al.* (1996) and construct a measure of latent demand among non-borrowing households. Survey households that do not request loans were asked whether they *would have* accepted credit from several different sources. Households that do not apply for formal loans but report they would accept credit from at least one of these sources are assumed to have self-sorted out of the credit market, and are classified as 'fully quantity rationed.' Households reporting that they would not accept credit from any formal source are further subdivided

[10] This difference persists in regression-controlled means accounting for years of residential tenure, indicating that the difference is not simply a result of treatment group members living in newer neighbourhoods and thus being more likely to have engaged in housing improvements in recent years.

on the basis of their reasons for refusing loans from formal sources. Those whose reason is fear of losing collateral are classified as 'fully risk-rationed,' while those whose reason is anything other than fear of losing collateral are classified as price rationed on the presumption that they have zero demand at the available interest rate.

Interestingly, the pattern of credit rationing indicates that identical shares (34%) of households both with and without titles are fully rationed out of the credit market, meaning they either apply for credit and are rejected or do not apply but would accept a loan if it were offered (Table 2). Meanwhile, the data suggest clear differences in the pattern of credit demand between titled and untitled households. In particular, the share of households that either applies for or would accept a loan rises from 60 to 73 percent, while the share that actually applies rises from 40 to 47 percent. All of the increase in loan applicants is absorbed under the category of partially rationed households (received less than they requested). Since the increase in the share of households that receive loans approximates the increase in the share that apply for loans, we cannot automatically infer that the rise in demand for credit was accompanied by a change in supply. Instead, it is possible that the increased share of loan applicants comes entirely from the population who were previously credit-worthy but unwilling to borrow. In other words, there is either perfect self-sorting among new credit market entrants or corresponding improvements in the availability of credit to households with property titles.

Table 2: Degree of credit rationing

	Did not apply for a loan			Applied for a loan		
	Would not accept		*Would accept*			
	price rationed	**fully quantity rationed:**	**fully quantity rationed:**	**fully quantity rationed:**	**partially quantity rationed**	**price rationed**
	too expensive	*risk-rationed*	*self-sorting*	*rejected*	*received some*	*received all*
Untitled	0.290	0.110	0.202	0.026	0.142	0.231
N:	*342*	*129*	*239*	*31*	*167*	*272*
COFOPRI title	0.203	0.069	0.256	0.015	0.218	0.239
N:	*109*	*37*	*137*	*8*	*117*	*128*
t_{Δ}	3.79	2.62	–2.47	1.46	–3.98	0.38
Pearson $\chi^2 = 36.72$						

Notes: Universe is all hoseholds in sample, and outcomes pertain to formal sector loan applications over the past three years. Among households that applied for at least one loan from a formal source, those that receive no credit are classified as 'fully quantity rationed', those that some but less than all of what they requested are classified as 'partially quantity rationed', and those that received all the credit they requested are classified as price rationed. Households that do not apply for formal loans over the reference period but report they would have accepted credit from at least one source are classified as 'fully quantity rationed.' Among households reporting that they *would not* accept credit from any formal source, those whose reason is fear of losing collateral are classified as 'fully risk-rationed,' while those whose reason is anything other than fear of losing collateral are classified as price rationed.

One notable observation is that, whereas tenure security is generally thought to give rise to an increased fear of losing property, in these data a significantly smaller share of titled households is risk-rationed in the credit market (Table 2). This pattern suggests that perceived risk associated with collateral use is *negatively* related to ownership rights among urban households, which would be the case, for instance, if banks had greater repossession rights over property documents that were not fully registered. This result suggests an important reason the program may have failed to improve credit market functioning: In contrast to standard predictions about gains from property formalization, changes in property institutions accompanied by increased protection of homeowners from collateral loss would *reduce* rather than increase banks' willingness to securitize loans with property. Titling programs will only improve credit markets if rights are strengthened for any formal claim on property – including banks' right to foreclose when a client defaults on a loan. Because of the political climate surrounding land titling, it is very possible that repossession is actually more costly for banks post-reform.

3.3 Household borrowing behavior

Table 3 presents categories of formal credit available to households in the sample and the share of loan applications to each source. The columns on the right-hand side compare the credit sources of households with and households without titles. Here we observe three main categories of banks participating in formal credit markets in urban Peru. According to the survey, the most important source of credit – constituting 35% of all loan applications and 45% of all formal loan applications – is the public-sector Materials Bank (MB), which has historically been one of the largest lending institutions in Peru. Since 1980, MB has targeted in-kind loans of housing construction materials to urban populations living in settlements, housing cooperatives and popular housing associations – the urban titling program's exact target population.[11] The maximum loan amount is roughly $5,000, loans are relatively long-term (up to 15 years) and the bank's official guidelines maintain effective annual interest rates of between 7% and 9% on all loans.[12] Households in this sector are highly dependent on MB for construction materials. Among all households that financed improvements with credit obtained between 1997 and 1999, 73.3% were from MB.

[11] These two government programs, however, operate independently, such that there is no explicit relationship between neighbourhoods targeted for program intervention and MB operations (personal interview with Daniel Andaluz, 14 August 2002, COFOPRI office, Lima, Peru).

[12] While MB functions somewhat as a government relief plan, loan approval is not automatic. To qualify for a loan, the bank's guidelines stipulate that the borrower have a minimum monthly family income equivalent to five times the estimated monthly payment, and borrowers may be asked to provide a co-signer. The bank guidelines also state that all loans involve a lien on the house as collateral for the loan, although a registered mortgage on land is not required. In this sense, all MB loans in theory involve 'inside' collateral, such that, in case of default, control of the construction project and ownership of depreciated assets go to the lender. In cases in which land mortgages backed by a registered property title are used in place of lien mortgages, the loan is additionally securitized with 'outside' collateral. See Chan and Kanatas (1985) for a discussion of these concepts. Official guidelines are reported online at www.banmat.org.pe. As Banerjee and Duflo (2002) point out, it is unclear how closely banks follow guidelines.

Table 3: Allocation of applications across sources of credit

	All	*Untitled*	*COFOPRI*	$\|t_{\Delta}\|$
N:	*1066*	*712*	*354*	
Materials Bank	0.351	0.310	0.435	2.83
Other formal lender	0.233	0.254	0.189	2.07
Commercial supplier	0.209	0.221	0.187	1.09
Informal lender	0.207	0.215	0.189	0.88
Total:	*1.000*	*1.000*	*1.000*	
Composition of other formal lenders				
Commercial bank (fully regulated)	0.544	0.525	0.597	1.01
Savings and loan organization (fully regulated)	0.327	0.343	0.284	0.88
Credit cooperative (fully regulated)	0.057	0.061	0.045	0.48
EDPYME (semi-regulated)	0.012	0.011	0.015	0.25
Nongovernmental organisation/village bank (unregulated)	0.060	0.060	0.059	0.03
Total:	*1.000*	*1.000*	*1.000*	
Composition of informal lenders				
ROSCA	0.329	0.358	0.253	1.54
Local moneylender	0.092	0.079	0.121	0.97
Family	0.222	0.237	0.195	0.67
Friend	0.130	0.135	0.121	0.36
Street vendor	0.227	0.191	0.311	1.98
Total:	*1.000*	*1.000*	*1.000*	

Note: Observations are loan applications, not applicants. Individuals have up to 6 loan applications each, with a sample average of 1.33 loans per applicant. COFOPRI indicates the Committee for the Formalization of Private Property; EDPYMES, 'Entities for the Development of Small and Micro Enterprises'; and ROSCA, 'Rotating Savings and Credit Association'.

Among other creditors, loan applications are fairly evenly divided between supplier or store (hire purchase) credit, credit from other private-sector lenders and informal credit.[13] In-kind loans from retailers or wholesale suppliers (henceforth, supplier credit), which take the form of inputs or merchandise advanced as credit, constitute 21% of loan applications from sample households. Supplier credit is available through stores specializing in selling consumer electronics and home appliances directly to clients on a credit basis, and is generally offered interest free or at very low interest rates, but for short periods of time (Dunn (1999)). In addition, the prices of goods supplied on credit are often considerably higher than the prices for cash purchases in wholesale or retail markets. Thus, the implicit real interest rates are likely to be high.[14] However, because interest costs are built into lease payments, *reported* interest rates on supplier credit are extremely low and often zero. Furthermore, supplier loans have close to a 100% approval rate, likely due to the fact that the good being supplied can easily be repossessed (i.e. full inside collateralization) As a result, property titles are rarely used as collateral to obtain supplier credit, so land titling should have little impact on the supply of inputs or merchandise advanced as credit.[15]

[13] Because of the importance of utilizing data on bank loan requirements, our formal analysis excludes the informal credit market, where unobservable factors are much more likely to determine credit access.

[14] For instance, Barham *et al.* (1996) found store credit in Guatemala provided at a 7% premium.

[15] In the survey data, property titles were used in only six loan applications.

Other private-sector financial institutions include commercial banks and savings and loan organizations, including commercial micro-finance lenders such as MiBanco, credit cooperatives, 'Entities for the Development of Small and Micro Enterprises' (EDPYMEs), village banks, and nongovernmental organisations (NGOs). With the exception of village banks and NGOs, these institutions are regulated by the national bank superintendency.[16] Since our data contain very few applications to semi-regulated or unregulated lenders, all private-sector lenders are grouped together in our empirical analysis and estimates run on the pooled sample along with a dummy indicator for type of institution. While sample size prevents us from studying separately the impact of a title on non-regulated lenders, the results are robust to excluding village banks and NGOs.

Meanwhile, since MB and supplier credit lending practices are much different from those of private-sector financial institutions, we separate formal loan transactions into these three categories throughout the analysis. Most importantly, the nature of credit rationing is likely to be distinct in the market for MB loans for two reasons. First, because MB is designed to reach low- to middle-income households, local branches are positioned and bank administrators accustomed to operating in these neighbourhoods. Therefore, low-income households are less likely to be transaction-cost rationed for MB loans. Secondly, because MB loans are for housing construction, loan amounts are larger on average and have a lower variance than loans from other institutions. The lending practices of MB are also distinct in that they potentially entail substantial project monitoring. Not only are construction materials purchased by the bank, but prospective borrowers must present a certified building plan when applying, and construction projects are at least minimally overseen by bank field representatives.

Significant differences in loan application behaviour exist between titled and untitled households (Table 3). In particular, households with titles are much more likely to request both public- and private-sector loans, while the share of loans sought from stores and informal sources does not vary by ownership status. In terms of differences in credit applicants to each type of lender, on average, MB loan applicants have lower socio-economic status, evidenced by their lower education levels, higher share of female household heads, lower wage incomes, higher education expenditures per year, higher share of extremely poor households and lower income from entrepreneurial activities (Table 4). It is worth noting that, despite having virtually equivalent monthly wage income, applicants to commercial banks have higher monthly spending in all categories of consumption, likely because they spend less of their earned income on housing investment. In addition, the last three rows in Table 4 reveal that mean loan approval and interest rates are distinct across types of lenders.

[16] In Peru, the interest rate on regulated private-sector loans is unconstrained by the government. EDPYMEs represent an intermediate stage between unregulated credit organizations and regulated banks. See Nexus (1998) for a description of the rules for EDPYMEs.

Table 4: Applicant summary statistics, by type of lender

	Commercial bank			*Materials Bank requests*			*Supplier/store credit*		
	Untitled	Titled	$\|t_{\Delta}\|$	Untitled	Titled	$\|t_{\Delta}\|$	Untitled	Titled	$\|t_{\Delta}\|$
N:	*158*	*60*		*220*	*154*		*148*	*62*	
Characteristics of household									
Number working-age members	4.32	4.63	1.09	4.20	4.18	0.11	4.19	4.08	0.39
Number of members	5.35	5.77	1.31	5.49	5.40	0.41	5.43	5.61	0.61
Number children aged 5 to 11 years	0.84	0.95	0.84	1.08	0.95	1.18	1.01	1.19	1.35
Number children aged 12 to 16 years	0.62	0.78	1.27	0.74	0.62	1.32	0.59	0.63	0.25
HH head is female	0.13	0.13	0.01	0.17	0.19	0.43	0.26	0.27	0.17
Age of HH head	47.32	46.58	0.47	45.19	45.85	0.42	46.55	48.52	0.76
Household head is literate	0.97	0.98	0.70	0.95	0.95	0.20	0.93	0.90	0.57
HH head attended primary school only	0.23	0.18	0.86	0.32	0.40	1.49	0.35	0.39	0.44
HH head attended high school only	0.47	0.68	2.92	0.50	0.44	1.19	0.43	0.47	0.55
HH head's attended post-secondary school	0.12	0.13	0.26	0.07	0.07	0.05	0.11	0.06	1.04
HH head's monthly wage	762.96	631.19	1.00	784.79	650.35	0.78	603.10	584.41	0.36
Total monthly HH consumption	630.36	691.14	0.75	519.00	523.64	0.15	604.75	590.03	0.35
Monthly HH food expenditures	212.24	214.75	0.20	179.51	186.95	0.75	207.50	207.17	0.02
HH education expenditures, per year	683.92	573.60	0.66	371.67	400.38	0.62	432.06	455.74	0.36
Whether HH has savings	0.09	0.15	0.87	0.07	0.09	0.53	0.16	0.07	1.78
Whether HH is extremely poor	0.08	0.07	0.23	0.28	0.23	0.89	0.19	0.15	0.15
Characteristics of residence									
Whether HH rents part of residence	0.03	0.02	0.70	0.05	0.02	1.47	0.01	0.10	2.26
Whether HH has a telephone	0.32	0.30	0.26	0.16	0.19	0.56	0.19	0.10	1.57
Whether HH has a home business	0.32	0.38	0.85	0.28	0.29	0.15	0.24	0.34	1.05
Income from home business	434.47	391.33	0.37	280.83	260.80	0.37	293.69	111.03	3.80
Average distance to formal lender	5.23	6.03	1.33	4.19	4.35	0.43	5.33	5.44	0.17
Closest bank two years ago	0.94	1.00	2.62	0.95	0.95	0.16	0.93	0.95	0.36
HH lending behavior									
Number of loans requested	1.34	1.38	0.41	1.21	1.13	1.56	1.37	1.32	0.49
Requested an informal loan	0.11	0.15	0.74	0.10	0.08	0.70	0.32	0.31	0.23
Loan offered, whether any	0.90	0.87	0.66	0.91	0.99	3.40	0.99	1.00	1.00
Average difference in credit amount	–117.7	–242.8	0.72	–184.5	–543	1.02	–0.16	–0.07	1.01
Size of loan, Soles	2,773.30	2,414.30	0.65	3,702.60	3,768.46	0.24	456.60	266.60	1.34
Interest rate, percent	0.32	0.20	3.76	0.09	0.07	1.35	0.03	0.02	1.23

Note: HH indicates household head. Cells contain sample means as reported in the *2000 COFOPRI Baseline Survey.*

Because the sample sizes are small, very few significant differences are observed between titled and untitled households *within* each category of loan. One notable difference is that, among applicants for private-sector loans, households without titles are relatively more educated, while in the pool of MB applications, applicants without property titles are less educated. For all education categories, these differences in differences are statistically significant. With respect to loan application outcomes, the mean differences in approval and interest rates indicate that a higher share of titled applicants receive MB loans and that applicants with titles are offered lower average interest rates on private-sector loans.

3.4 Econometric model

We attempt to measure the collateral value of land titles by modelling the outcomes of individual credit applications. Inference about the impact of a property title on a loan applicant's probability of approval involves speculation about what the applicant would have experienced in the absence of a title. The simplest of such models is:

$$y_{ij} = \alpha_i + \gamma d_j + \varepsilon_i,$$

where y is application outcome, j is the index for the control group ($j=0$) and the treatment group ($j=1$), d_j is 1 if the household has a title and zero otherwise, and γ is the treatment effect of having a title. The no-treatment counterfactual is assumed to obey an additive model, while the treatment effect is constant:

$$y_{i0} = \alpha_i + \varepsilon_i, \text{ where } y_{i1} - y_{i0} = \gamma \text{ and } E[y_{i0}] = \alpha_i. \quad (1)$$

Equation (1), which states that the only reason access to credit changes in the treatment group is titling, is required for identification.

As discussed in Section 1, in non-experimental data, having a land title will generally *not* be independent of potential outcomes since both the decision to obtain a title and the decision to apply for a loan are likely to be correlated with the local lending environment or with property values. The fact that property titles were assigned in our data in a quasi-experimental fashion independent of household demand for tenure security or credit reduces concern over the endogeneity of tenure status. Nonetheless, the large apparent changes in investment demand raise concern over heterogeneity in the pool of loan applicants even if program participation is as good as random. If receiving a title encourages an individual to apply for a loan, titled and untitled applicants will not be comparable in every respect other than the title even if titled and untitled individuals are, so Equation (1) is violated. The direction of bias will depend on whether marginal applicants are more or less credit-worthy than unconditional applicants. A comparison of observables (Table 1) suggests that the two groups are equivalent, lending confidence to our ability to identify treatment effects using untitled applicants as a control group.

To further isolate the effect of changes in the collateral value of land from changes in the pool of applicants, we make use of detailed survey data on the information and documentation used by banks in the screening process of each loan application.[17] This allows us to identify the full set of household characteristics relevant to each particular approval decision. Since loan approval decisions are made by formal lenders on the basis of some finite set of observable characteristics of the applicant, X_i, in theory loan approval outcomes depend only on X_i and treatment (having a land title). In this case, the additive model applies conditional on X, and identification requires:

$$E[y_{i1} - y_{i0} \mid X] = \gamma(X) \text{ and } E[y_{i0} \mid X] = \alpha_i(X). \quad (2)$$

This supports estimating the effect of a title in the approval of loan application i with the following Probit model:[18]

$$Pr(approval)_i = \beta_0 + \beta_1(title) + (\Pi_i' X_i)\alpha + e_i. \quad (3)$$

[17] All households that applied for a loan – regardless of whether the loan was approved – were asked to report the complete set of documents and information, including property titles, requested by the bank.

[18] For households with multiple applications to one type of bank, we use only the most recent application and control for the whether the household applied previously for other formal loans from that category of bank. Robust standard errors are used throughout to account for survey clusters and strata.

where X_i is a k-dimensional vector of applicant characteristics and Π_i is a kXk diagonal matrix containing along the diagonal indicators of whether the bank used each characteristic ($x_1, \ldots, x_k$) in its approval decision for application i. As long as the set of criteria reported by households reasonably captures the information set on which lenders base their approval decisions, the average treatment effect of a property title will be identified by β_1. In other words, even if differences in the demand for credit generate differences between treatment and control applicants, given sufficient information on household characteristics observed by banks at the application stage and the bank's approval algorithm, unconfoundedness is likely to hold conditional on X_i.

Furthermore, even if there are remaining differences in credit-worthiness observable to banks but not captured by the reported criteria, these should be absorbed by the difference in approval rates between titled and untitled households among the approximate 50% of lenders that do not request a title. Hence, we also estimate:

$$Pr(approval)_i = \beta_0 + \beta_1(title) + \beta_2(title*title\ used\ in\ screening) + (\Pi_i X_i)\alpha + e_i. \quad (4)$$

The coefficient estimates on the indicator of whether the applicant acquired a property title through the program, $\hat{\beta}_1$, and the indicator of whether he was asked to provide a title in the loan transaction, $\hat{\beta}_2$, provide inference on the treatment effect of titling. If loan approval rates are higher among the treatment group because property titles are used as collateral, the treatment effect will be fully concentrated among applications in which a title was used, so $\hat{\beta}_1$ will be zero. Conversely, if differences in loan approval rates across experimental groups reflect unobservable differences between treatment and control applicants, approval will be independent of collateral requests, so $\hat{\beta}_2$ will be zero.

Personal identity documents, property titles, wage receipts, co-signer information, reported self-worth and utilities receipts are the most common loan application requirements (Table 5).[19] In total, 51% of loan applicants, comprising roughly equal proportions of applicants with and without titles, are asked to provide a property title prior to the loan approval decision. In less than 10% of applications, banks require other documentation of repayment capacity, including lending group membership, rental contracts, tax receipts, and business registration or professional license documents. Only a handful of banks ask for bonds or collateral in the form of household goods.

To account for these requirements, our empirical estimates control for a number of relevant household characteristics available from the survey. It is important to note that, while several of these variables are potentially endogenous to program participation, in order to isolate the partial derivative of titling on banks' use of collateral it is necessary to account for simultaneous changes in other household characteristics relevant to loan approval decisions. In this sense, we do not measure the net effect of land titling – or total derivative of ownership status on credit access –, which includes indirect channels such

[19] A potentially important source of missing data is the category of loan requirements labelled 'other,' in which the exact requirement was not specified by the household. To correct for this missing information, we include a larger set of potentially relevant household characteristics that might fall under this category.

as employment, and would presumably be larger than the partial derivative. However, (β_2–β_1) sheds light on the size of the difference between total and partial derivatives.

Table 5: Loan requirements

Credit source:	Materials bank	Commercial bank	Supplier credit
N:	*614*	*548*	*266*
Nothing, just reputation	0.003	0.106	0.652
Personal identity document	0.982	0.821	0.303
Property title	0.599	0.429	0.113
Utilities bill	0.503	0.454	0.175
Reported self-worth	0.375	0.299	0.075
Co-signer	0.345	0.285	0.132
Wage receipt	0.269	0.347	0.145
Other	0.246	0.179	0.087
Solidarity group membership	0.083	0.020	0.019
Promisory note	0.072	0.089	0.071
Business registration documents	0.031	0.078	0.011
Household items (collateral)	0.016	0.040	0.009
Bond	0.016	0.038	0.004
Tax receipt	0.015	0.051	0.004
Operating license	0.015	0.041	0.000
Rental contract	0.003	0.003	0.000

Notes: Data come from survey question asked of all loan applicantions reported in the 2000 COFOPRI Baseline Survey (1428) regarding each piece of documentation required by the bank for a given loan application, whether or not loan was approved.

To capture wage income, we control for total monthly household wage income, monthly earnings of the highest wage earner, whether the highest wage earner is self-employed, whether the worker with highest number of working hours is self-employed, monthly earnings of the highest contracted employee and the share of household wages from the contracted employment. We separate self- and contract-employment from non-contract employment because commercial banks may only accept formal wage receipts, although households are likely to report all wage income. To capture the reported self-worth of the loan applicant, we control for the value and age of the property, whether the household is engaged in entrepreneurial activity, monthly income from household entrepreneurial activity, whether the business has a registered tax number, whether the household rents part of their residence and the total amount of other outstanding formal debt incurred between 1997 and 1999.

To account for household utilities bill requirements, we include information on whether or not the household paid an electricity, water, or telephone bill the month before the survey, along with amounts paid for each. To address the remaining loan requirements, we incorporate information on whether the household rents part of its residence, and whether any household member belongs to a community financial group. Capacity to provide a co-signer is proxied by the number of adults in the household and

sex of the household head. Finally, capacity to provide a property document is indicated by whether the household is a member of the treatment group (and therefore has a property title), along with whether the household has an unregistered property document.

We also include in the empirical model basic pieces of household information that are possibly relevant to loan application decisions but not recorded as official screening criteria. These include: sex, age, literacy and education levels of household head; whether the household reports experiencing an economic shock in the past year; and whether the household previously applied for a loan from the same category of institution, the year of the loan application, the intended use of loan funds and the distance from the lender to the household. The last characteristic is relevant for transaction costs of the bank, while loan history could be important if use of collateral decreases with length of relationship with the bank (Berger and Udell (1995) provide evidence of this).

The intended use of loan funds is relevant only for applications to private-sector lenders because loans to MB are uniformly intended for housing construction, while supplier credit is used solely for the purchase of consumer goods. Among applications to private-sector banks, the purpose of the loan is important primarily for identifying risk associated with entrepreneurial credit. Among these loan applications, 34.3% are intended for housing construction, 38.7% for entrepreneurial activity, 8.1% for emergency needs, 2.0% for household goods, 1% for land purchases and 16.1% for other consumption. In the market for private-sector loans, there is little difference between households with and without titles in the composition of loan uses (Table 6).

Table 6: Distribution of loan uses and approval rates by loan use

					Approval rates	
	All loan applications		*Private sector loan applications*		*All loan applications*	
	Untitled	COFOPRI title	Untitled	COFOPRI title	Untitled	COFOPRI title
N:	*526*	*276*	*158*	*60*	*526*	*276*
Household items	0.073	0.076	0.017	0.030	1.00	1.00
Housing construction	0.445	0.545	0.359	0.328	0.89	0.95
Entrepreneurial activity	0.153	0.083	0.381	0.403	0.90	0.89
Emergency	0.071	0.043	0.088	0.060	1.00	1.00
Other	0.260	0.253	0.155	0.179	0.98	0.98

Notes: Observations are loan applications, not applicants. Distribution of credit use across loan applications to all lenders and private sector lenders only (excludes Materials Bank applications only).

By including only application-specific lending criteria (via Π), Equations (3) and (4) assume that the only household information relevant to the approval decision is that which was reported by the household as required documentation. This model is appropriate only under the strong assumption that households report all information used by banks and that banks do not make use of information that was not requested. Another possible specification is to include all potentially relevant information regardless of a particular bank's reported screening criteria by adding a term ϕX_i to the regression equation. In light of the small sample sizes for each type of loan, the regression estimates follow the parsimonious specification.

Given that additional information might be used but not reported and the fact that a non-trivial number of loans involved unspecified 'other' information (Table 5), we also estimate program effects using non-parametric matching techniques, which impose fewer constraints on the total number of covariates included in the model. In particular, titled households are matched to untitled households on the basis of the propensity score, defined as the conditional probability of having a title, $P(X_i, \Pi_i) = \Pr(D_i = 1 | X_i, \Pi_i)$. $P(X_i, \Pi_i)$ is calculated by performing a logistic regression of X_i and $\Pi_i X_i$ on program participation.[21] Propensity scores balances the distributions of covariates in X between program participants and non-participants based on the similarity of their predicted probabilities of participation (Rosenbaum and Rubin 1983).[22] The main advantage to propensity score matching is to capture possible non-linearities in treatment effects and control variables without increasing the dimensionality of the problem. Since lending decisions involve potentially complex interactions among observable borrower characteristics, it is arguably erroneous to impose a parametric functional form linking program participation to outcomes (see Jalan and Ravallion (2003)).

There are several ways to construct estimators based on the propensity score. Kernel matching compares each treated individual with a kernel-weighted average of all comparison observations, with the weights assigned according to the propensity score. In our estimates, the kernel matching estimator is given by a Gaussian kernel function, and standard errors are obtained by bootstrapping. We also present results from random-draw, nearest-neighbour and stratified matching procedures for robustness (see Ichino (2002) for a description of these methods).

Among the pool of approved loans, we also examine differences across treatment and control groups in the interest rate, size of loan obtained, and difference between amount requested and amount received. The corresponding ordinary least squares (OLS) and matching estimates are presented alongside the loan approval estimates. Clearly, loan terms and approval are determined simultaneously; however, data limitations prevent us from estimating a joint model of application outcomes that would give precise estimates of lenders' interest rate and quantity responses to property titles. Instead, we examine differences in average loan terms in order to help distinguish between competing explanations for differences in approval decisions across titled and untitled households.

[21] There are several procedures for matching on the propensity score; see Heckman *et al.* (1998) for a good review. Here we estimate the propensity score with predicted values from a Probit model. We did not find significant differences in the distribution of covariates within strata.

[22] Rosenbaum and Rubin (1983) show that if the D_i's are independent over all i, and outcomes are independent of participation given x_i, then outcomes are also independent of participation given $P(x_i)$, just as they would be if participation was random. In other words, the strong ignorability assumption holds and differences in the outcomes between the control group and the participants can be attributed to the program.

4. RESULTS

4.1 Materials Bank loans

A strong, positive relationship exists between the likelihood that a loan request to MB is approved and the household having received a property title from the program (Table 7). The Probit estimate in column 1 of Table 7 indicates a 4.6 percentage point increase in the likelihood that a loan application to MB is approved, implying a reduction in the rejection rate of nearly 50%. Further, when the treatment effect is combined with whether the document was requested by the bank, in column 2 of Table 7, we observe that the entire treatment effect is concentrated among households that were asked to provide a title. Among the 40% of MB loan applicants that were not asked to provide property titles with loan applications, the treatment effect is close to zero and insignificant – a strong indication that the availability of property titles is indeed responsible for the increase in the loan approval rate associated with having a title.

Table 7: Ordinary least squares (OLS) regressions, outcome of formal credit applications

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent variable:	Offered given applied		Interest rate offered, given received		Amount received, given received		Difference between amount asked and amount received	
Materials Bank loans								
COFOPRI title	0.046 **	0.012	–0.010	–0.015	–216.26	–113.14	399.59	–215.51
	(0.010)	(0.030)	(0.015)	(0.019)	(219.05)	(307.40)	(480.49)	(470.70)
Property documents required COFOPRI title		0.057 **		0.011		–325.03		1,342.86
		(0.024)		(0.022)		(403.03)		(992.63)
Other formal loans								
COFOPRI title	0.002	0.000	–0.085 *	–0.102 *	–25.83	614.71	54.02	247.82
	(0.061)	(0.017)	(0.041)	(0.048)	(684.69)	(715.37)	(119.31)	(192.12)
Property documents required COFOPRI title		–0.062		0.038		–1435.43		–434.31
		(0.082)		(0.063)		(1,107.56)		(260.27)
Supplier loans								
COFOPRI title	0.000	0.000	–0.008	–0.006	110.35	106.50	0.128	0.084
	(0.000)	(0.000)	(0.014)	(0.013)	(100.57)	(102.35)	(0.304)	(0.309)
Property documents required COFOPRI title		0.000		–0.014		192.87		2.210
		(0.000)		(0.147)		(858.88)		(2.59)

Notes: HH indicates household; *, significance of less than 1%; and **, significance of less then 5%. Data in the first two columns are Probit estimates; data in all remaining colunms are OLS regressions. Standard errors are shown in parentheses. Robust standard errors account for sample clustering and stratification. Demographic controls include the following: age, literacy, and education of HH head; whether residence is used as a source of economic activity; total monthly household wage income; monthly earnings of highest wage earner; whether highest wage earner is self-employed; whether the worker with the highest working hours is self-employed; monthly earnings of highest contracted employee; share of household wages from contracted employment; self-reported sale value and age of property; whether household is engaged in entrepreneurial activity and monthly income from HH entrepreneurial activity; whether business has a registered tax number; whether HH rents part of residence; total amount of other outstanding formal debt during 1997–99; whether HH paid for electricity, water or telephone in the previous month and amounts paid for each; whether HH member belongs to local financial group; number of adults; dummy indicating HH has an additional type of unregistered property document; whether an economic shock occurred in the past year; whether HH previously applied for a loan from the same category of institution and year of the loan application; intended use of loan funds; distance from the lender.

The estimated treatment effect of property ownership on MB loan approval rates from propensity score matching suggests an even larger improvement in approval rates of between 9 and 10 percentage points (Table 8). The difference between the estimated treatment effects from Probit and propensity score models is likely related in part to the exclusion of treatment group members who have no well-defined match among the control group (approximately 10% of households fall outside the region of common support). These unmatched households with property titles represent those that would not have applied for a loan in the absence of the program, which suggests that marginal applicants are characterized by below average approval rates based on some non-linear combination of observables.

Table 8: Propensity score estimates

Matching method:	(1) kernel matching	(2) nearest neighbor	(3) stratified matching
Materials bank loans			
Loan application approved	0.094 **	0.104 **	0.093
	(0.028)	(0.036)	(0.029)
Loan amount offered, Peruvian soles	–328.66	–121.67	119.63
	(364.00)	(392.05)	(287.17)
ifference in amount requested and received, Peruvian soles	–656.09	36.75	–127.52
	(1,072.25)	(692.64)	(596.32)
Interest rate, %	–0.017	–0.021	–0.017
	(0.012)	(0.016)	(0.014)
Other formal loans			
Loan application approved	0.047	0.036	0.051
	(0.036)	(0.123)	(0.036)
Loan amount offered, Peruvian soles	1,494.3	789.0	1,654.0
	(1,708.0)	(1,200.5)	(1,744.2)
ifference in amount requested and received, Peruvian soles	108.34	126.53	79.37
	(65.08)	(61.29)	(105.20)
Interest rate, %	–0.087 *	–0.101 *	–0.097
	(0.043)	(0.047)	(0.041)
Supplier loans			
Loan amount offered, Peruvian soles	–258.8	36.29 *	–166.5
	(190.43)	(18.55)	(98.6)
ifference in amount requested and received, Peruvian soles	–0.268	–0.196	–0.092
	(0.22)	(0.50)	(1.02)
Interest rate, %	–0.029 *	–0.005	–0.020
	(0.009)	(0.004)	(0.022)

Notes: ** indicates significance of less than 1%; *, significance of less than 5%. Demographic controls are the same as those listed in the notes to Table 7. Nearest neighbor matching in column 2 is based on a random draw. The kernel matching estimator is given by a Gaussian kernel function, and standard errors are obtained by bootstrapping. Bootstrapped standard errors are shown in parentheses.

In contrast to the loan approval outcome, the provision of a property title appears to have little effect on MB interest rates and loan amounts in both the Probit and propensity score estimates. Although the mean interest rate is nearly two percentage points lower for titled borrowers, the difference is not significant.

4.2 Private-sector loans

Based on the regression estimates in Table 7, the effect of property titles on the market for loans from private-sector lenders is distinct from the market for MB loans. In column 3, the estimated treatment effect from the Probit estimate indicates that the likelihood of loan approval does not change with ownership status. The propensity score results are larger than the regression estimates but insignificant (Table 8). In addition, the average size of private loans to households with titles is around 50% larger, although the point estimate of the difference is not statistically significant. There are no significant or consistent differences between households with and without titles in the difference between amounts requested and received.

Meanwhile, all estimates from Tables 7 and 8 indicate that, conditional on approval, property owners face interest rates that are on average 8–10 percentage points lower than the interest rates offered households without titles. This implies a reduction in the mean private-sector interest rate from 27% to 18%. However, the treatment effect on interest rates does not appear to operate through collateral provision, as indicated by the small and statistically insignificant estimate of β_2, the coefficient on the interaction between treatment and the indicator of title requirement in column 4 of Table 7.

4.3 Supplier credit

As expected, given that the supply of store credit is relatively unconstrained and prices are poorly captured by reported interest rates, we observe little evidence of a treatment effect of property titling on supplier credit. The regression estimates reported in the last two rows of Table 7 find close to zero and insignificant effects of titling on all outcomes. In the propensity score estimates, we observe measurable effects on the interest rate and average loan size with certain matching methods, but neither the sign nor significance of either result is robust to alternative matching techniques.

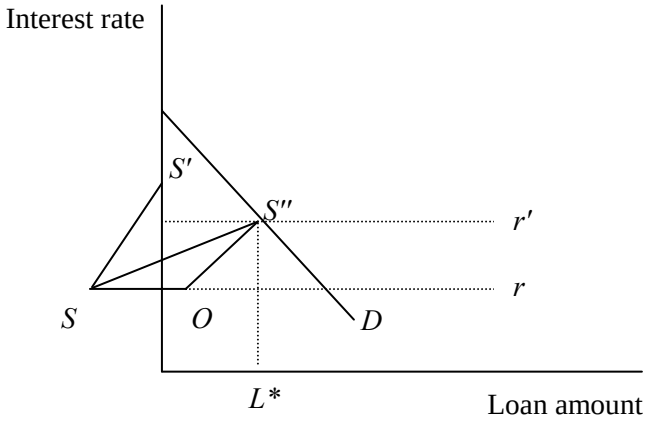
5. DISCUSSION

In the case of MB credit, the absence of a large effect of property titling on interest rates, conditional on receiving a loan, is not surprising because interest rates on MB loans are regulated by the government to fall within a range of two percentage points. The situation of MB is analogous to a credit market model in which the bank is constrained by moral hazard issues from raising the interest rate above a certain level (which may well be the rationale behind the regulation), inducing quantity rationing of MB loans to exclude applicants that cannot provide sufficient collateral or surpass a certain level of default risk. In this market, collateral serves to reduce credit rationing by increasing the share of loans that are free of risk to the lender.

The absence of a strong relationship between loan size and ownership status is also not surprising given that loan amounts are also imprecisely restricted by MB lending rules, which state that amount is limited by the "particular construction needs of the household." Alternatively, differences in risk level inferred by the bank but not captured by the covariates could be responsible for the change in approval probability but not loan

amount. As described by Wette (1983), in the presence of interest rate regulation, increased use of collateral could generate adverse selection effects, depicted in Figure 1.

Figure 1.



Here, the supply curve originates in the negative orthant to reflect the fact that the collateral value of titled property, net of transaction costs, may be negative when land values are low. In this scenario, prior to the titling program, low-risk applicants face the supply curve *SS″*, while high-risk applicants face *SS′*, so only low-risk types are awarded loans of L^* at interest rate *r′*. When both types can provide collateral, the aggregate supply curve becomes *SOS″*. Here, more loans are awarded (since both types enter the market), but the average loan amount remains at L^* and the interest rate stays fixed at *r′*. Hence, if MB applicants are considered riskier, on average, post-reform because the availability of collateral induces high-risk types to enter the market, a higher share of titled applicants will receive loans, but the average interest rate and loan amount could remain unchanged.

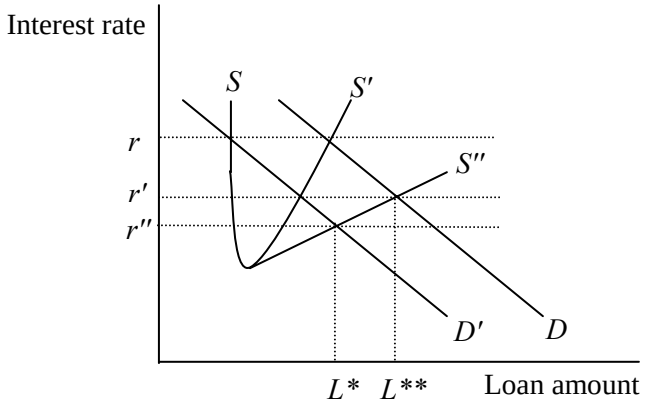
The results for private-sector lenders are more ambiguous because private-sector interest rates do not appear to depend on whether banks use property as collateral. As mentioned in Section 2, one possibility is that private-sector banks do not find it profitable to use land as collateral (*i.e.* the expected effective leverage ratio of capital is non-positive), but that they do infer lower default risk from ownership rights and possession of a property title. This situation is illustrated in Figure 2.

In this figure, no portion of the supply curve is flat, indicating that banks are not using property titles to securitize loans.[23] However, because banks infer lower default risk from the existence of a property title, the supply curve for program participants shifts from *SS′* to *SS″*, lowering the interest rate available to titled borrowers from *r* to *r′*. In this scenario, the average loan size also shifts outward from L^* to L^{**}, which is inconsistent with our findings. However, if the demand curve is steeply sloped, the change in quantity demanded will be small relative to the change in the interest rate. Another possibility is

[23] The downward-sloping portion of the supply curve reflects the fact that effective interest rates are higher for small loans due to costs involved in processing and monitoring loans.

that, given the increased availability of low-interest MB loans for titled households post-reform, the demand curve for more expensive commercial loans shifts inward for program households (D to D′). In this case, the interest rate would fall even further to r'' while the change in quantity demanded at the new interest rate is ambiguous.

Figure 2.



Unfortunately, with available data can cannot distinguish between these two stories, nor can we fully rule out the possibility that results are driven by unobserved heterogeneity in local financial markets. While the result on the private-sector interest rate is robust to controlling for observed heterogeneity among private-sector lender types, clearly the pattern could still be driven by unobservable differences in the lending practices or the level of competition among financial institutions available to households with and without titles. However, measures of aggregate banking activity from 1994 and nation-wide censuses of financial activity from 1996, including the number of automated teller machines (ATMs) and the number of bank employees by district show no differences in local financial sector development (Table 2).

An important issue that emerges from our findings is the discrepancy in lending strategies across public- and private-sector banks. Given the strict lending practices of MB, it is not surprising that MB's interest rate response differs from private-sector banks. Unable to separate the market according to risk, MB must limit the amount of credit it provides to the households without titles, evidenced by the lower approval rates to this population. In particular, since the pool of applicants for MB loans is on average more vulnerable to income shocks, it is likely that MB faces greater moral hazard and enforcement constraints that make it unprofitable to adjust the interest rate, whereas commercial banks that screen out a greater portion of applicants and therefore service a less risky pool of borrowers have more interest rate flexibility. Furthermore, it is unlikely that possessing a property title offers additional information on the default risk of MB borrowers given that they are all borrowing for housing construction, and the bank could reasonably infer the same tenure security and low likelihood of eviction from the household's decision to invest in immobile assets.

It is less clear why private-sector lenders would not make use of property titles as collateral if MB finds it profitable to do so. One explanation is that quantity rationing will

generally be size-biased because the net profit on small loans is lower, making collateral cost-effective only for relatively large loans. Since MB loans are all for housing construction and tend to be fairly large – the mean amount is roughly $1,421 – and the variance in loan size is small, a larger number of loans from non-MB sources will be rejected because of the transaction costs involved. Collateral provision, which only increases loan transaction costs, cannot eliminate this type of quantity rationing.

Another explanation is that commercial banks perceive the transaction costs involved in using land titles as collateral to be higher post-reform. As mentioned in Section 3.2, data on households' fear of losing property suggests that titled applicants perceive foreclosure to be less likely, and commercial lenders may have the same perception in a political climate that prioritizes housing for the poor. Ironically, property titling programs might actually reduce banks' ability to foreclose because they unavoidably send the message that governments will side with poor borrowers in enforcing credit contracts. This could be different for a public-sector bank such as MB if it has inside information regarding the extent to which the government is willing to enforce property collection in the case of default.

A final possibility is that, as a public-sector institution, MB subsidizes its clients and is not, in fact, making profit-maximizing lending decisions. Furthermore, MB may be characterized by higher corruption or misuse of funds for political gain. Indeed, early reports of high default rates among MB borrowers suggest that loans may be distributed according to other criteria.[24] Long-run information on the profitability of MB loans and private-sector lending strategies is needed to disentangle these competing stories.

6. CONCLUSION

Despite the distribution of over 1.2 million property titles to urban squatters, our results indicate that credit rationing is still a key feature of the micro-lending environment in urban Peru. In particular, post-reform, a full 34% of households with titles remain fully rationed out of the formal credit market. These results shed light on the potential impact of titling efforts on financial market integration and development in poor urban communities worldwide. Although property titles are associated with a small reduction in formal-sector credit rationing, it appears that titling efforts will not automatically make collateral-based lending viable for the majority of formal-sector credit applicants.

Most notably, our estimates suggest that the bulk of the reduction in credit rationing associated with the Peruvian titling program can be attributed to one particular lending institution, the publicly funded Materials Bank, which supplies in-kind loans of housing materials. Meanwhile, access to credit from private-sector lenders appears unaltered by titling. While there are a number of possible explanations, one compelling piece of evidence indicates that titling may have reduced banks' ability to foreclose.

The fact that credit access for construction materials improves with ownership rights is important insofar as it helps meet the increased demand for housing investment

[24] In conjunction with the increase in default, the bank's own financial assessment, (Banco de Materiales, "Evaluacion a Junio 2003") suggests operating losses and declining profitability for 2002.

that accompanies improvements in tenure security. In this manner, greater access to MB loans should reduce the dampening effect on other types of investment that will result if demand for construction materials rises while households remain credit constrained in financing home improvements (See Carter and Olinto (2003) for a formal presentation of this relationship). However, given that access to loans for purposes other than housing does not appear to have changed with ownership status, households with titles, post-program, will still be unable to rely on credit as a source of consumption insurance. This is exaggerated by the fact that MB loans are in-kind transfers and hence not fungible in case of unexpected changes in consumption needs.

Perhaps more importantly, property titling does not appear to assist poor households finance micro-enterprise activities. This pattern is clearly illustrated in Table 6, which presents the loan approval rates for households with and without titles according to the designation of credit. Consistent with the regression and matching estimates, we see that the entire improvement in loan approval rates is concentrated among construction loans, while all other categories of credit use have nearly identical approval rates for titled and untitled households. The means in the table indicate that liquidity constraints are still binding on entrepreneurial loans for titled households. Given that collateral-based wealth is an important determinant of small business formation in other settings (Black *et al.* 1996), land titling will have no effect on socially inefficient allocations of entrepreneurial activity across socio-economic groups if post-reform titled property cannot serve as collateral. As a result, the growth implications of titling programs may be greatly overstated.

REFERENCES

ANDALUZ, D. (2002), personal interview with author, 14 August (COFOPRI office, Lima).

ALEEM, I. (1990), "Imperfect Information, Screening and the Costs of Informal Lending", *World Bank Economic Review*, **4** (3), 329–350.

ALSTON, L., G. LIBECAP and R. SCHNEIDER (1996), "The Determinants and Impact of Property Rights: Land Titles on the Brazilian Frontier", *Journal of Law, Economics, & Organization*, **12** (1), 25–61.

ATWOOD, D.A. (1990), "Land Registration in Africa: The Impact on Agricultural Production. *World Development*, **18** (5), 659–671.

BALTENSPERGER, E. (1976), "The Borrower–Lender Relationship, Competitive Equilibrium and the Theory of Hedonic Prices", *American Economic Review*, **66** (3), 401–405.

BARHAM, B. L., S. BOUCHER and M. R. CARTER (1996), "Credit Constraints, Credit Unions, and Small-Scale Producers in Guatemala", *World Development*, **24** (5): 793–806.

BARRO, R. (1976), "The Loan Market, Collateral and the Rate of Interest", *Journal of Money, Credit and Banking*, **8**, 839–856.

BENJAMIN, D. K. (1978), "The Use of Collateral to Enforce Debt Contracts. *Economic Inquiry*, **16** (3), 333–359.

BERGER, A.N., and G.F. UDELL (1990), "Collateral, Loan Quality and Bank Risk", *Journal of Monetary Economics*, **1** (25), 21–42.

BESLEY, T. (1995), Property Rights and Investment Incentives: Theory and Evidence from Ghana", *Journal of Political Economy*, **103** (5), 903–937.

BESTER, H. (1985), "Screening vs. Rationing in Credit Markets with Imperfect Information", *American Economic Review*, **31** (4), 887–899.

BINSWANGER, H. P., and K. DENINGER (1999), "The Evolution of the World Bank's Land Policy: Principles, Experience, and Future Challenges", *World Bank Research Observer*, **14** (2), 247–276.

CARTER, M. R., and P. OLINTO (2003), "Getting Institutions Right for Whom? Credit Constraints and the Impact of Property Rights on the Quantity and Composition of Investment", *American Journal of Agricultural Economics*, **85**: 173-186.

_____ (1997), Wealth, Property Rights and Credit Rationing: Simulated Maximum Likelihood Estimates of a Disequilibrium Credit Market. **Unpublished Manuscript.**

CARTER, M., and K. WEIBE (1990), "Access to Capital and its Impact on Agrarian Structure and Productivity in Kenya", *American Journal of Agricultural Economics*, **December**, 1146–1150.

CHRISTENSEN, S., David DOLLAR, A. SIAMWALLA and P. VICHYANOND (1993), *The Lessons of East Asia. Thailand: The Institutional and Political Underpinnings of Growth.* World Bank Publication No. 12458 (Washington DC: World Bank).

COFOPRI (COMMITTEE FOR THE FORMALIZATION OF PRIVATE PROPERTY) (2000), *Baseline Survey*. March.

DENINGER, K., and G. FEDER (1993), *Land Policy in Developing Countries.* Rural Development Note No. 3, (Washington DC: World Bank).

DUNN, E. (1999), *Micro finance Clients in Lima, Peru: Baseline Report for AIMS Core Impact Assessment.* AIMS Paper. (Washington DC: Management Systems International).

FEDER, G. (1985), "The relation between farm size and farm productivity: The role of family labor, supervision, and credit constraints," *Journal of Development Economics*, **18**, 297–313.

FEDER, G., T. ONCHAN, Y. CHALAMWONG and C. HONGLADARON (1988), *Land Policies and Farm Productivity in Thailand*. (Baltimore: Johns Hopkins University Press).

FEDER, G., and D. FEENY (1991), "Land Tenure and Property Rights: Theory and Implications for Development Policy", *World Bank Economic Review*, **5** (1), 135–153.

FIELD, E. (2004), "Property Rights, Community Public Goods and Household Time Allocation in Urban Squatter Communities" *William and Mary Law Review,* **45** (3): 837-887.

FIELD, E. (2004), "Property Rights and Investment in Urban Slums" *Journal of the European Economic Association Papers and Proceedings,* **3** (2-3): 279-290**.**

FLEISIG, H., and N. DE LA PENA (1996), "Peru: How Problems in the Framework for Secured Transactions Limit Access to Credit", Working paper, Center for the Economic Analysis of Law.

HILLIER, B., and M.B. IBRAHIMO (1993), Asymmetric Information and Models of Credit Rationing", *Bulletin of Economic Research*, **45** (4), 271–304.

HOFF, K., and Joseph STIGLITZ (1990), "Introduction: Imperfect Information and Rural Credit Markets-Puzzles and Policy Perspectives", *World Bank Economic Review,* **4** (3), 235–250.

HOLDEN, P. (1997), "Collateral Without Consequence: Some Causes and Effects of Financial Underdevelopment in Latin America", *The Financier*, **4** (1), 12–21.

ICHINO, A.,(2002), "Estimation of Average Treatment Effects Based on Propensity Scores", *The Stata Journal*, **2** (4), 358–377.

JAFFEE, D.M., and J. STIGLITZ (1990), "Credit Rationing", in *Handbook of Monetary Economics* **91**, 651–666.

JALAN, J., and M. RAVALLION (2003), "Does Piped Water Reduce Diarrhea for Children in Rural India? *Journal of Econometrics,* **112**: 153-173.

LOPEZ, R., and C. ROMANO (2000), "Rural Poverty in Honduras: Asset Distribution and Liquidity Constraints," in *Rural Poverty in Latin America*, eds. R. Lopez and A. Valdez, (Hampshire: Palgrave Macmillan Ltd).

MICELI, T. J., C.F. SIRMANS and Joseph KIEYAH (2001), "The Demand for Land Title Registration: Theory with Evidence from Kenya", *American Law and Economics Review*, **3** (2), 275–287.

MIGOT-ADHOLLA, S., P. HAZELL and F. PLACE (1991a), Indigenous Land Rights Systems in Sub-Saharan Africa: A Constraint on Productivity? *World Bank Economics Review*, **5**, 155–175.

OLÓRTEGUI, I. G. (2001), Informal Settlers in Lima. ESF/N-AERUS International Workshop, Leuven and Brussels, Belgium, 23–26 May 2001.

ROSENBAUM, P. R., and D. B. RUBIN (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, **70** (1), 41–55.

STIGLITZ, J., and A. WEISS (1981), "Credit Rationing in Markets with Imperfect Information", *American Economic Review*, **3** (71), 393–410.

WETTE, H. (1983), "Collateral and Credit Rationing in Markets with Imperfect Information: Note, *American Economic Review*, **3** (73), 442–445.

WORLD BANK DEVELOPMENT NEW ARCHIVES 2000, *Peru's Urban Poor Gain Access to Property Markets*, February 2.

World Bank (1998), *Project Appraisal Document*, Report No.18245PE, Peru, Urban Property Rights Project (Washington DC: World Bank).

_____. (1992), *Project Report No. PID6523*, Peru, Urban Property Rights Project (Washington DC: World Bank).

Appendix 1: Ordinary least squares (OLS) regressions, whether applied for credit from particular source

	(1)	(2)	(3)	(4)	(5)
Universe:	All formal loans	Construction	Materials	Supplier loans	Other formal loans
COFOPRI title	0.067 *	0.099 **	0.104 **	0.060 *	0.034
	(0.032)	(0.027)	(0.028)	(0.028)	(0.027)

Notes: Linear probability estimates, dependent variable in all columns in binary indicator of whether applicant applied for credit from each formal lender type. Similar results are obtained from Probit estimation. COFOPRI indicates whether individual received a property title through the government program, the Committee for the Formalization of Private Property; *, significance of less than 1%; and **, significance of less than 5%.

Empowerment and Efficiency: Tenancy Reform in West Bengal

Abhijit V. Banerjee

Massachusetts Institute of Technology

Paul J. Gertler

University of California, Berkeley

Maitreesh Ghatak

University of Chicago

The paper analyzes the effect of agricultural tenancy laws offering security of tenure to tenants and regulating the share of output that is paid as rent on farm productivity. Theoretically, the net impact of tenancy reform is shown to be a combination of two effects: a bargaining power effect and a security of tenure effect. Analysis of evidence on how contracts and productivity changed after a tenancy reform program was implemented in the Indian state of West Bengal in the late 1970s suggests that tenancy reform had a positive effect on agricultural productivity there.

We are grateful to the editor, Sherwin Rosen, and three anonymous referees for detailed comments, from which the paper benefited significantly. We are indebted to Maitreya Ghatak for his advice and support at all stages of the project, and especially in conducting the survey. Special thanks are due to Debraj Ray and Esther Duflo, whose valuable suggestions greatly influenced our approach. Thanks are also due to D. Bandyopadhyay, Nripen Bandyopadhyay, Pranab Bardhan, Tim Besley, Anne Case, Angus Deaton, Semanti Ghosh, Jonathan Gruber, D. Gale Johnson, Steve Levitt, Eric Maskin, Jonathan Morduch, Sunil Sengupta, and many seminar audiences for helpful feedback. We thank Ajoy Bhowmik, Lipi Ghatak, Arun Ghosh, Swapan Saha, and Nga Vuong for their help in the process of data collection, entry, and analysis. The usual disclaimer applies.

[*Journal of Political Economy,* 2002, vol. 110, no. 2]
 0022-3808/2002/11002-0003$10.00

I. Introduction

While there is widespread support for reforming agricultural property rights (see, e.g., World Bank 1993; Binswanger, Deininger, and Feder 1995), there have been few attempts to evaluate the productivity consequences.[1] Part of the reason is that there are few examples of large-scale changes in property rights that were not accompanied by major social unrest. Moreover, analyzing the impact on efficiency is difficult because of data limitations and the fact that the structure of property rights is itself endogenous.

In this paper we study the effect of a major change in property rights on agricultural productivity in the Indian state of West Bengal. Within a year of being elected in 1977, a left-wing administration launched Operation Barga, a program designed to implement and enforce the long-dormant agricultural tenancy laws that regulated rents and security of tenure of sharecroppers.[2] Under these laws, if tenants registered with the Department of Land Revenue, they would be entitled to permanent and inheritable tenure on the land they sharecropped as long as they paid the landlord at least 25 percent of output as rent. In the decade following the launching of Operation Barga, there was a significant improvement in the terms of tenants' contracts and more secure tenure. Moreover, agricultural productivity grew faster in West Bengal compared to other states in India, earning the administration praise from many, sometimes unexpected, quarters.[3]

An evaluation of the contribution of Operation Barga to the agricultural growth in West Bengal provides a rare opportunity to examine the relationship between property rights and efficiency. It also allows us to reexamine the question of whether there is a necessary trade-off between efficiency and equity in programs that transfer property rights from the rich to the poor. Operation Barga is especially interesting because it involved a limited transfer as opposed to a full transfer of property rights (e.g., redistributing landownership). It gave the incumbent tenant only the right to claim a higher share of the output and permanent tenure. While a full transfer of landownership that would completely eliminate agency costs is likely to have positive effects on productivity, the effect of a more limited transfer such as Operation Barga is less obvious.

Our theoretical analysis shows that the impact on productivity can be decomposed into two effects: a bargaining power effect and a security of tenure effect. The bargaining power effect comes from the fact that

[1] Exceptions include Lin (1992), Besley (1995), and Jeon and Kim (2000).

[2] *Barga* is the local word for sharecropper.

[3] See "Left Gets It Right" (1993), an article in the *Economist* on the Left Front's successful rural reforms in West Bengal.

after the reform the legal contract becomes the tenant's "outside option," which increases his bargaining power vis-à-vis the landlord and forces the landlord to offer him a higher crop share, which translates into stronger incentives.

Security of tenure has two different opposing effects. On one hand, the landlord may use the threat of eviction when output is low to induce the tenant to work harder.[4] Disallowing eviction restricts the use of such incentives and therefore reduces efficiency. On the other hand, greater security of tenure encourages the tenant to invest more since it gives him the confidence that he will stay on the land long enough to enjoy the fruits of his investment. Moreover, his increased bargaining power means that the tenant now expects to get a higher share of the additional output resulting from investment.

We also find empirical support for the hypothesis that the transfer of property rights under Operation Barga positively affected agricultural productivity. We take two approaches to measuring the impact on productivity. The first is a quasi-experimental approach using the neighboring country of Bangladesh, which is similar in many respects to West Bengal but did not implement tenancy reform. The second approach uses sharecropper registration as a measure of program intensity and tests whether productivity is higher in areas in which the program was implemented more intensively. Our results suggest that limited interventions in property rights like Operation Barga, which empower tenants without giving them full landownership, can have a positive effect on productivity.[5] Hence there is no necessary trade-off between efficiency and equity in such programs. Moreover, these strategies of empowerment tend to be politically easier to implement than conventional land reforms. They may therefore offer a real way out of the status quo in the right context.

We have organized the presentation as follows. In Section II, we briefly describe Operation Barga. In Section III, we present our theoretical arguments about how Operation Barga is likely to have affected contracts and incentives, and we discuss results from a survey of sharecroppers that we carried out on how contracts actually changed in response to the reform. In Section IV, we present the analysis of the impact of Operation Barga on productivity using district-level data. Section V presents conclusions.

[4] This observation goes back to Johnson (1950). For a formal analysis, see Dutta, Ray, and Sengupta (1989).

[5] Other examples of strategies that could empower tenants are usury laws, minimum-wage laws, job creation programs, and supply of subsidized credit.

II. Operation Barga

After independence, India sought to improve the living standards of sharecroppers through tenancy reform. The Land Reforms Act of 1955 and its successive amendments have two main clauses: (1) Sharecroppers will have permanent and inheritable incumbency rights to land that is registered in their name provided that they pay the legally stipulated share to the landlords, do not leave the land fallow, and do not sublease the land. Except in such cases, the sharecropper will lose his right to the land only if the landlord wants to use the land for personal cultivation. These rights are inheritable but not transferable. (2) The share that the landlord can demand from a registered tenant will be no greater than 25 percent.[6]

This phase of tenancy reform is widely recognized as a failure (Appu 1975). Loopholes in the law allowed landlords to abuse the personal cultivation exemption and to threaten to evict the tenant whenever he tried to register. Moreover, the tenant was responsible for registering himself, and the government provided little institutional support for him to do so. By virtue of their wealth and superior caste, landlords wielded a lot of power within the village and were therefore able to intimidate tenants. This was compounded by the fact that the government usually took the landlord's side in disputes. As a result, before Operation Barga, very few sharecroppers were registered, crop shares were significantly below the legal minimum, and tenure was widely perceived as being insecure (Bardhan and Rudra 1984).

In 1977, the newly elected government passed the West Bengal Land Reforms Act, which closed most of the loopholes in the 1955 act. Most important, it set very stringent and well-defined conditions under which the landlord could utilize the personal cultivation clause to evict a tenant.

At the same time, the new government launched Operation Barga, a massive and well-publicized village-to-village campaign to register tenants and ensure their rights. Under this program the process used to register tenants was altered to make it easier for the sharecropper to register. Operation Barga officials sought out hesitant sharecroppers, explained the law, and offered them the opportunity to register. Moreover, the new government used its own village political organizations to make sure that landlords did not intimidate tenants, that tenants who registered did not face reprisal from the landlords, and that disputes were handled fairly in the courts. Operation Barga is widely regarded as a success. By 1993, more than 65 percent of an estimated 2.3 million share tenants had been registered.

[6] In cases in which the landlord pays the cost of all nonlabor inputs, the law caps his share at 50 percent. However, this clause rarely applies.

III. Theory: Tenancy Reform, Contractual Change, and Productivity

In this section we develop a simple theoretical model of a landlord-tenant relationship based on moral hazard and limited wealth of tenants. We shall use this model to analyze the potential effects of the reform on the contractual relationship between a given landlord and an incumbent tenant.

There are two ways in which the reform could have altered the set of potential contracts between the landlord and the tenant. First, it changed an incumbent tenant's outside option. The fact that the landlord could no longer evict the tenant meant that the tenant could always hold out for his legal share of the output. The landlord could no longer threaten to replace him with another tenant if he refused to accept a lower share. This does not mean that the contract between them necessarily has to be the legally stipulated contract. Under some conditions there may be a different contract that suits them both better, but the tenant should not be worse off than he would be if he stuck to the letter of the law.

A second potential effect of the reform is directly related to the restrictions on eviction. Under the new law, the tenant could plan to crop the same piece of land for as long as he would like to without fearing eviction. On the other hand, the landlord could no longer expect to use the threat of eviction as a credible incentive device. One would expect the optimal contract to change for both these reasons.[7]

A. The Model

Suppose that there is an infinitely lived landlord who owns a plot of land that he cannot crop himself. In each period he employs exactly one tenant to crop the land. There is a large population of identical infinitely lived tenants who are all willing to work for the landlord as long as the landlord pays them their outside option (or reservation payoff), m, in that period, which is given exogenously. The landlord and the tenants share the same discount factor $\delta < 1$. In each period, output can take on two values, $Y_H = 1$ ("high" or "success") and $Y_L = 0$ ("low"

[7] This is less obvious than it seems because the tenant and the landlord are not bound to honor the letter of the law in their mutual contracting. Thus, in principle, the two parties could continue using threats of eviction as an incentive device even after evictions are made illegal: the tenant can voluntarily agree to let the landlord evict him if he fails to produce enough. This possibility is likely to be limited by commitment problems on both sides. A tenant who is actually facing eviction may want to renege on his promise to leave quietly and may seek the protection of the law. Similarly, a landlord who has been given the right to evict by his tenant may be tempted to abuse his power to his bargaining advantage.

or "failure"), with probability e and $1 - e$, respectively. The realizations of output are independent over time. The tenant chooses e ("effort"), which costs him $c(e)$. For simplicity, we assume that the cost function is quadratic: $c(e) = \frac{1}{2}ce^2$. For reasons that will become apparent later, we assume $c > 1$.

The key assumptions of this model are as follows: (1) Only the tenant's effort matters for output.[8] (2) The tenant's effort choice e is nonobservable and hence noncontractible. (3) Past and present realizations of output are contractible. Specifically, we assume that at the beginning of each period the landlord can commit himself to a one-period contract that maps current and past realizations of output into (a) current payments to each potential tenant and (b) a decision about which tenant will work for him in the next period. (4) The landlord faces a limited liability constraint.[9] In particular, in a given period, each tenant has a limited amount of wealth $w > 0$, so that the least he can get paid is $-w$.[10] (5) Both the tenant and the landlord are risk-neutral.[11]

The fact that both the landlord and the tenants are infinitely lived defines an infinite extensive-form game between the landlord and the tenant that, in principle, can have many equilibria. Here we restrict ourselves to studying equilibria of this game in which the strategies in each period are *history-independent* except for the choice of who is going to be the landlord's current tenant.[12] Furthermore, consistent with the assumption that there are many potential tenants and one landlord, we shall focus on the equilibrium that maximizes the landlord's profits per period.

In this game there is no reason to pay those tenants who are not working for the landlord in the current period, so the contract needs to specify only payments to the tenant who is currently working for the landlord. Likewise, the landlord has no reason to discriminate among those who are not working for him in the current period. Therefore,

[8] Eswaran and Kotwal (1985) have argued that the landlord sometimes contributes to agricultural production by providing managerial inputs. In a previous version of the paper, we argue that our results continue to hold in this case.

[9] There are models of sharecropping based on moral hazard that do not use the hypothesis of limited liability (see Stiglitz 1974).We use it because it provides an analytically simple way of generating rents for the tenant (which is necessary for threats of eviction to be meaningful) as well as the static inefficiency associated with tenancy. See Dutta et al. (1989) and Mookherjee (1997) for similar models of sharecropping based on limited liability.

[10] We are assuming that tenants do not save and nonmonetary punishments are not allowed. Ghatak, Morelli, and Sjöström (2001) and Mookherjee and Ray (2000) study the implications of allowing saving by agents in similar environments.

[11] In a previous version of the paper, we showed that the same results hold when the tenant is risk-averse, as long as the limited liability constraint binds in equilibrium.

[12] Formally we are looking at Markov equilibria in which the state variable is the identity of the current tenant (Fudenberg and Tirole 1991). Dutta et al. (1989) study *history-dependent* Markov equilibria in a similar environment.

if and when he decides to get a new tenant, he can simply choose randomly from among those who are not working for him currently (here we make use of the assumption that there are many potential tenants; otherwise the landlord would randomize only among those who have never worked for him). Furthermore, by the assumption of history independence, the contract for each tenant will just depend on the current realization of output. Therefore, the contract in any given period will just need to specify four numbers: the payment to the tenant and the probability of his continuing in the job when the output is high (denoted, respectively, by h and φ) and the same two numbers when output is low (l and ψ). We shall find it convenient to refer to h and l as success and failure wages. Note that we could have, instead, conducted our analysis in terms of a *linear* contract, $sY - r$, with s denoting the *crop share* of the tenant and r a *fixed-rent component*, with $s = h - l$ and $r = -l$. The reason is that since output takes only two values in this model, all contracts can be expressed as linear contracts.

B. *Optimal Tenancy Contracts without Eviction*

We first solve the landlord's problem under the assumption that incumbent tenants cannot be evicted and will therefore continue to be the tenant in all future periods. In this case the problem reduces to solving the one-period contracting problem. Given the tenant's outside option m and wealth level w, the optimal contract is a solution of maximizing the landlord's expected payoff,

$$\max_{\{e,h,l\}} \pi = e - [eh + (1 - e)l],$$

subject to the following constraints: (i) The limited liability constraint (LLC) requires that the amount of money that could be taken away from the tenant in any state of the world is bounded above by his wealth w and realized output:

$$h \geq -(1 + w), \quad l \geq -w.$$

(ii) The participation constraint of the tenant requires that the contract guarantees an expected payoff to the tenant equal to m:

$$v = eh + (1 - e)l - \tfrac{1}{2}ce^2 \geq m.$$

(iii) The incentive-compatibility constraint (ICC) requires that the tenant chooses the effort level e to maximize his *private* payoff:

$$e = \arg\max_{e \in [0,1]} \{eh + (1 - e)l - \tfrac{1}{2}ce^2\}.$$

Notice that the optimal incentive contract (h, l) must have $h > l$ because if $h \leq l$, then from the incentive-compatibility constraint, $e = 0$ and the landlord gets $-l$; for the same l, if he sets $1 \geq h > l$, he gets $e[1 - (h - l)] - l \geq -l$. This also implies that one of the two LLCs, $h \geq -(1 + w)$, cannot bind. The total expected surplus generated by a project is $S = e - (ce^2/2)$. The first-best level of e, namely the one that maximizes S, is $1/c < 1$ (by our assumption that $c > 1$). Since the constraint that $e \leq 1$ does not bind at the first-best, we can safely ignore it. This also implies that there is no reason to choose $h - l > 1$ since the first level of effort and output is achieved when $h - l = 1$. Hence the ICC can be rewritten as

$$e = \frac{h - l}{c} \in (0, 1).$$

Let us substitute for e using the ICC and rewrite the optimal contracting problem of the landlord as

$$\max_{\{h,l\}} \pi(h, l) = \frac{h - l}{c} - \frac{(h - l)^2}{c} - l$$

subject to

$$\frac{(h - l)^2}{2c} + l \geq m$$

and $l \geq -w$.

Consider first the case in which the participation constraint does not bind. Observe that in this case it is always optimal to reduce l down to $-w$ while keeping $h - l$ unchanged. With l set at $-w$, the value of $h - l$ that maximizes profits is easily determined by differentiating the expression for π with respect to $h - l$. This yields $h - l = \frac{1}{2}$, implying $e = 1/2c$. Substituting these values into the participation constraint, we can write it as $1/8c \geq m + w$. As long as m and w are low enough that this constraint does not bind, we are justified in ignoring the participation constraint. For this case the optimal value of e is therefore $e^* = 1/2c$.

Next consider the case in which $1/8c < m + w$. In this case the participation constraint will bind. Substituting the value of l from the participation constraint into the expression for π gives us

$$\pi(h, l) = \frac{h - l}{c} - \frac{(h - l)^2}{2c} - m.$$

This expression is maximized when $h - l = 1$, which represents a pure

rent contract.[13] Consequently, e will be at its first-best level, $1/c$. Combining $h - l = 1$ with the fact that the participation constraint binds, we get the condition $l = m - (1/2c)$. Since the LLC requires that $l \geq -w$, we conclude that the first-best level of e will be chosen only if $m - (1/2c) \geq -w$ (otherwise the LLC will bind), which is equivalent to $m + w \geq 1/2c$.

Finally, for intermediate values of $m + w$, that is, $1/8c \leq m + w < 1/2c$, both the participation constraint and the LLC will bind. Solving these together, we get $l = -w$ and $h - l = \sqrt{2c(m + w)}$, and using the ICC, we get the optimal value of e,

$$e^* = \sqrt{\frac{2(m + w)}{c}}.$$

There are two results from this analysis that are relevant in studying the effect of the reform.

RESULT 1. The value of e implied by the optimal contract between the landlord and the tenant is

$$e^* = \begin{cases} \dfrac{1}{2c} & \text{if } m + w < \dfrac{1}{8c} \\ \sqrt{\dfrac{2(m + w)}{c}} & \text{if } \dfrac{1}{8c} \leq m + w < \dfrac{1}{2c} \\ \dfrac{1}{c} & \text{if } \dfrac{1}{2c} \leq m + w. \end{cases}$$

Hence an improvement in the incumbent tenant's outside option always (weakly) increases effort.

RESULT 2. The tenant's participation constraint does not bind as long as $m + w < 1/8c$ and hence he earns rents.

These results have simple intuitions. The main trade-off the landlord faces in this model is either to provide incentives or to extract surplus from the tenant. A fixed-rent contract, where the tenant pays the same amount whether or not his output is high, maximizes the tenant's incentives and would always be chosen if the tenant were wealthy enough. However, since the tenant cannot pay more than he has, the fixed rent is bounded above by his wealth, w (this is all he has when his crop fails). Therefore, if w is small, fixed-rent contracts are not in the landlord's interest. The landlord can do better with a contract that makes the tenant pay more when he has more (i.e., when his output is high). However, this clearly taxes success and therefore weakens incentives. This explains why the expected output is less than first-best. However,

[13] It is easy to verify that this contract pays the landlord w in both states of the world, making the tenant a full residual claimant.

as the tenant becomes wealthier, it becomes easier to extract rents from him without sacrificing incentives and expected output approaches the first-best.[14]

An increase in the outside option of the tenant, m, forces the landlord to pay the tenant more. Since the tenant typically has too little incentives, the landlord will want to pay him this extra amount in the form of an extra bonus for success, which will give the tenant stronger incentives to work hard. This result forms the basis of what we call the *bargaining power effect* of the reform: an increase in the tenant's bargaining power, with everything else held constant, leads to an increase in his share and his productivity.

Finally, the tenant may earn rents in this model because if he has very little wealth and his outside options are very low, then the only way the landlord can extract the entire surplus from the tenant (net of m) is to take away almost all of the output when output is high. Since this obviously has adverse incentive effects, the landlord will typically not try to extract the entire surplus when m is very low. Hence the landlord will not reduce the share of the tenant below some minimum level irrespective of m.[15]

The curve $ABCD$ in figure 1 shows equilibrium effort as a function of the tenant's outside option and wealth level when eviction threats are absent.

C. *Optimal Tenancy Contracts with Eviction*

We now turn to the situation in which the landlord can evict the tenant at will. In this case the landlord can typically do better than offering the one-shot contract described above. One feature of the one-shot contract is the fact that the tenant earns rents unless his outside option is sufficiently good: this means that the tenant will strictly prefer to continue being a tenant; therefore, the threat of eviction if output is low can be used as an incentive device.[16]

Let $\bar{V}$ denote the expected equilibrium lifetime utility of an incumbent tenant in the next period. Let M denote the equilibrium lifetime expected utility of someone who is currently not a tenant: $M \equiv m/(1 -$

[14] Laffont and Matoussi (1995), Bandiera (1999), and Ackerberg and Botticini (2002), among others, find evidence for a positive wealth effect on the tenant's share of output.

[15] This problem is similar to that of raising income tax revenue by the government: low taxes will lead to high levels of labor supply and income but will yield low revenue. Result 2 is similar to the idea behind the Laffer curve: higher tax rates may reduce labor supply so much that the government may earn less tax revenue.

[16] The threat of evicting the incumbent tenant is credible from the landlord's point of view because, by assumption, tenants of all types (in terms of wealth and outside options) are available in unlimited numbers. As a result the landlord is indifferent between retaining and firing a given tenant.

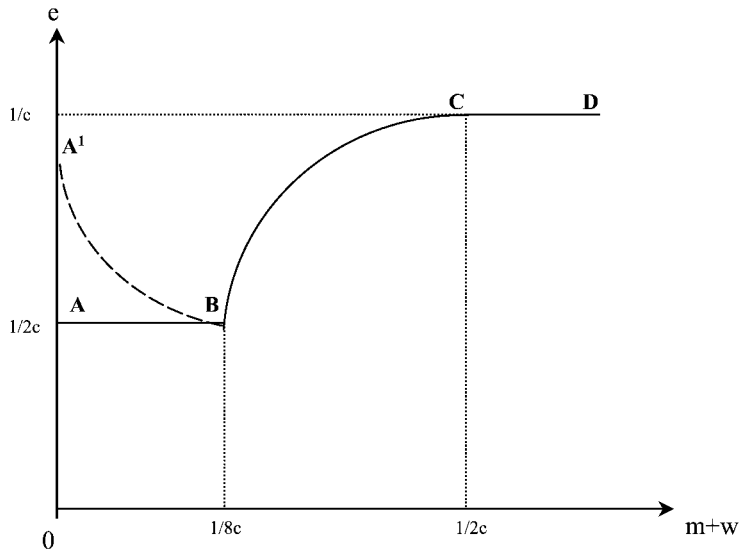


FIG. 1

δ), where, as before, m is the value of the outside option per period. The hypothesis of history independence implies that the landlord cannot precommit anything beyond the current-period incentive contract, (h, l), and the corresponding probabilities of eviction, $(1 - \varphi, 1 - \psi)$. It also implies that the tenant's lifetime utility from next period onward, $\bar{V}$, is taken as exogenous in this period by both players.

Given these assumptions, the tenant's expected lifetime utility in the current period from choosing a level of effort e today, $\bar{V}_0$, must satisfy the Bellman equation:[17]

$$\bar{V}_0 = \max_{\{e \in [0,1]\}} \{eh + \delta[\varphi e + (1 - e)\psi](\bar{V} - M) + \delta M - (1 - e)w - \tfrac{1}{2}ce^2\}. \quad (1)$$

Differentiating this expression with respect to e yields the new ICC:

$$h + w + \delta(\bar{V} - M)(\varphi - \psi) = ce. \quad (2)$$

Comparing this with the ICC in the one-shot game, we see that the

[17] Here we assume that the LLC binds, i.e., $l = -w$. If it does not bind, there will be no rents and the threat of eviction would have no effect.

existence of rents and the tenant's foresight reduce the marginal cost of implementing e by the amount $\delta(\bar{V} - M)(\varphi - \psi)$.

Next we observe that $\varphi = 1$ and $\psi = 0$ in the optimal dynamic contract. As long as the tenant is still getting more than his outside option, raising the probability of eviction is preferred by the landlord rather than raising h for giving more incentives because it is costless from his point of view. Neither φ nor ψ affects the landlord's payoff directly (as long as the participation constraint is not binding), and the only thing they affect is the ICC. Hence from (2) we see that ψ should be set at its minimum possible value, zero, to give the maximum punishment to the tenant for failure. On the other hand, φ should be set at its maximum possible value, one, to maximally reward the tenant for success.

Thus (2) becomes

$$h + w + \delta(\bar{V} - M) = ce. \tag{3}$$

The new participation constraint of the tenant is $\bar{V}_0 \geq M$.[18] In a stationary equilibrium, $\bar{V}_0 = \bar{V}$, and hence from (1) we get

$$\bar{V} - M = \frac{eh - (1 - e)w - \frac{1}{2}ce^2 - m}{1 - \delta e}. \tag{4}$$

Substituting (3) into (4), we get

$$\bar{V} - M = \tfrac{1}{2}ce^2 - w - m. \tag{5}$$

In any equilibrium in which eviction threats are used, $\bar{V} - M$ must be positive. If it is positive, the landlord has to maximize

$$\max_{\{e,h,l\}} e(1 - h) - (1 - e)l$$

subject to the ICC, (3), and the LLC. Using these two constraints, we can rewrite the landlord's objective function as

$$\max_{\{e\}} \{1 - ce + \delta(\bar{V} - M)\}e + w. \tag{6}$$

Maximizing this leads to the first-order condition

$$1 - 2ce + \delta(\bar{V} - M) = 0, \tag{7}$$

[18] This reflects the fact that in this case, in contrast to the case in which evictions are not allowed, the tenant faces a trade-off between current and future rewards. For that reason, the right comparison is made between his lifetime expected utility and his lifetime outside option.

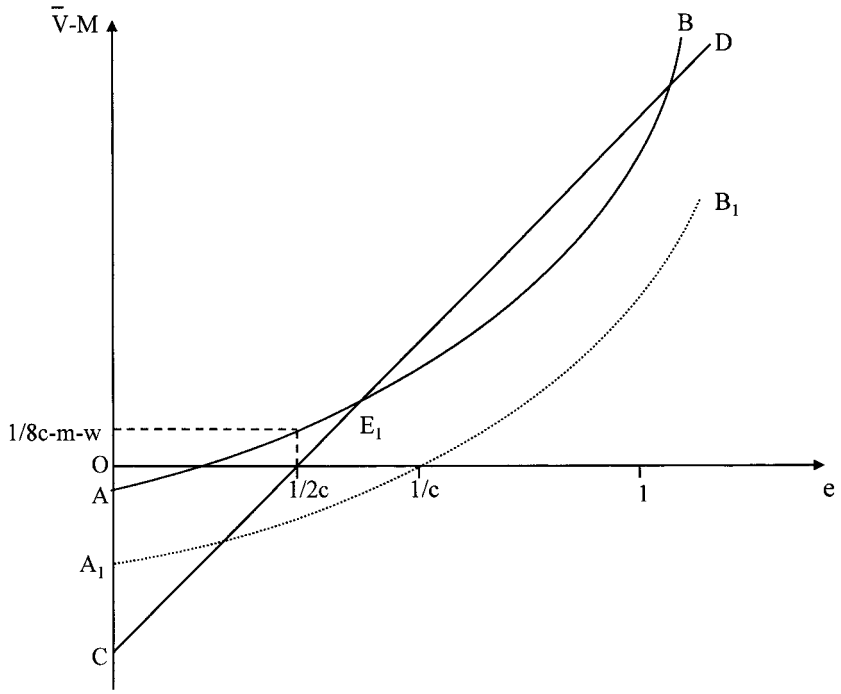


FIG. 2

which can be rewritten in the form

$$e = \frac{1 + \delta(\bar{V} - M)}{2c}. \tag{8}$$

We can find the equilibrium values of e and $\bar{V}$ by solving equations (5) and (8) simultaneously. In figure 2, *AB* and *CD* represent equations (5) and (8). These curves intersect at two points, E_1 and E_2. The curve *AB* is strictly increasing and convex, whereas *CD* is a positively sloped straight line. For $e = 1/2c$, *CD* intersects the horizontal axis. As long as $(1/8c) - m - w > 0$, the curve *AB* lies above *CD* at $e = 1/2c$. Also, for $e = 1$, *CD* lies above *AB*.[19] Hence only the point E_1, which corresponds to a value of $e \in (1/2c, 1/c)$, is an admissible solution since E_2 corresponds to a value of $e > 1$. As $m + w$ increases (but with $[1/8c] - m -$

[19] The relevant condition is $(2c - 1)/\delta > \frac{1}{2}c - w - m$. Since $c > 1$ and $\delta < 1$, $[2 - (\delta/2)]c > 1$, which can be rearranged as $(2c - 1)/\delta > c/2$.

$w > 0$ continuing to hold), the curve AB moves down, and therefore the equilibrium value of e will also go down. This is intuitive since the rents and hence the force of the threat of eviction should be smaller when either m or w is higher.

Let us now turn to the optimal share of the tenant, $h^* - l^*$. Substituting (3) into (7) and using the LLC, we see that

$$h^* - l^* = \tfrac{1}{2} - \delta\tfrac{1}{2}(\bar{V} - M).$$

Since $\bar{V} - M$ goes down when $m + w$ goes up, $h^* - l^*$ must go up.

For the case in which $(1/8c) - m - w < 0$, the curve A_1B_1 represents equation (5) and is obtained by a vertical downward shift of AB. In this case it is clear that the two points at which A_1B_1 and CD intersect involve $e < 1/2c$ and $e > 1$ (the second intersection is not shown in the figure), and none of them are admissible. In this case there is no solution to the optimal contracting problem with eviction where the participation constraint does not bind (i.e., the equilibrium value of $\bar{V} - M$ is positive). Solving the participation equation (5), we get $e^* = \sqrt{2(m + w)/c}$, which is of course exactly the value of e^* we found when eviction was not an option, under the assumption that $1/2c \geq m + w \geq 1/8c$. The optimal choice of the tenant's share, $h^* - l^*$, is also exactly the same as in the no-eviction case. This is as we would expect: when the participation constraint binds, the fact that eviction is an option should be irrelevant.

The rule that $e^* = \sqrt{2(m + w)/c}$ applies only as long as $e^* \leq 1/c$, that is, as long as $m + w \leq 1/2c$. For $m + w > 1/2c$, effort will be set at its first-best level, that is, $e^* = 1/c$, $h^* - l^* = 1$, and the LLC will no longer bind. This, once again, is exactly as in the case without eviction.

The curve A^1BCD in figure 1 shows equilibrium effort as a function of the tenant's outside option when evictions are permitted. It differs from the corresponding curve $ABCD$ for the one-period model only for the range of values of m such that the tenant earns rents ($m + w < 1/8c$). However, for $m + w < 1/8c$, e^* is a declining function of m when eviction is an option, whereas it is constant when eviction is forbidden. Moreover, since the two curves meet at $m + w = 1/8c$, it follows that the supply of effort is strictly higher when eviction threats are possible, for $m + w < 1/8c$. The discussion above is summarized in the following result.

RESULT 3. When evicting the tenant is an option, the optimal choice of e and $h - l$ coincides with that for the no-eviction case as long as $m + w \geq 1/8c$. For $m + w < 1/8c$, the value of e chosen with evictions is strictly higher than the corresponding value without evictions. Moreover, over this range, a higher m is associated with a lower choice of e but a higher value of $h - l$.

This result shows why the effect of Operation Barga on efficiency

could be negative in spite of the bargaining power effect described in Section III*B*. Eviction threats will tend to raise the effort level of very poor tenants, and unless the increase in m is large enough, their effort will fall as a result of the reform, though these tenants will still be better off.[20] However, this analysis applies only to tenants who have a large number of close substitutes so that threats of eviction are credible. This excludes wealthier and more able tenants.

D. Operation Barga and Investment Incentives

The way we have modeled the production technology so far ignores any role of investment. It is often argued that tenurial insecurity discourages investment by the tenant, and this usually forms the strongest efficiency (as opposed to redistributive) argument in favor of tenancy or land reform. This argument typically fails to tell us why the landlord himself cannot undertake such investments directly (given that he is less likely to be credit-constrained than the tenant) or indirectly, by giving incentives to the tenant through suitable contractual means. It is clear that when the investment is contractible (e.g., flattening the land, building soil partitions, planting trees, and digging ponds), the problem cannot be the tenant's unwillingness to invest since the landlord can pay the tenant to invest. The problem is rather that the landlord may not want to invest at the first-best level: given that the tenant's effort is below the first-best level for agency reasons, the value of any investment that is complementary to the tenant's effort will also be below the first-best level, and as a result, the landlord will be reluctant to invest. In this case, Operation Barga can increase investment, but only because it increases the tenant's willingness to put in effort.

Noncontractible investments—such as experimentation with new techniques, the care and maintenance of the land, or the use of manure (the effect of which lasts more than one period)—differ from contractible investments because they create the possibility of a holdup problem unless the landlord can make long-term commitments.[21]

We use a simple two-period extension of our benchmark model of Section III*B* to illustrate the point. Assume that in the first period the model is as before, but now the tenant can make a land-specific investment of amount x, which increases the productivity of the land in

[20] In an eviction equilibrium, h is lower and e is higher than in the no-eviction equilibrium; the tenant's utility per period has to be lower. And since the discount factor of the tenant is lower in an eviction equilibrium than in the no-eviction equilibrium (i.e., $\delta e < \delta$), the tenant's expected lifetime utility is lower as well.

[21] Similar conclusions emerge if the source of noncontractibility of investment is moral hazard (as it is for effort) instead of the landlord's inability or unwillingness to commit to long-term contracts. The analysis is, however, much more complicated (see Banerjee and Ghatak 1996).

the second period in the following way: output is $Y_H = 1 + x$ with probability e and $Y_L = x$ with probability $1 - e$. This investment costs $\frac{1}{2}\gamma x^2$ to the tenant. We assume for simplicity that the second period's payoff is not discounted, $w = 0$, and $m < 1/8c$.

If x was contractible, then the landlord could simply "buy" it from the tenant at the efficient level, $1/\gamma$. Even if x is not contractible, the efficient level of investment can still be achieved as long as the landlord can commit to a two-period contract with the incumbent tenant. Let $r_h \equiv Y_H - h$ and $r_l \equiv Y_L - l$ denote the landlord's payoff when output is high and low, respectively. If in the current period the optimal contract is (r_h, r_l), then by committing to retaining the current tenant next period and increasing the rent by a fixed amount Δr in the next period irrespective of output, the landlord can make the tenant a full residual claimant of the fruits of his investment.

The interesting case of this model is the case in which x is not contractible and it is also not possible for the landlord to commit to rewarding the tenant if he makes the investment. In this case the tenant anticipates that the landlord is going to expropriate the results of his first-period investment by threatening to fire him at the beginning of the second period. As a result the tenant will not invest at all, and so in both periods the outcome will be the same as in the one-period model, that is, $e = 1/2c$, $r_h = \frac{1}{2}$, and $r_l = 0$. Hence if the landlord cannot precommit to future contracts, his total (two-period) expected payoff is $1/2c$ and that of the tenant is $1/4c$.

A possible benefit of Operation Barga in this context is that it rules out all evictions and therefore makes it possible for the landlord to convince the tenant that he will not be evicted. In this respect, both the landlord and the tenant will be better off.

It is also possible for Operation Barga to have beneficial effects on productivity without making the landlord better off. This will be the case if eviction threats were very effective in eliciting extra effort from the tenant before the reform. Let R denote the rents to the tenant from staying in the relationship in the second period. Under our assumptions, $R = 1/8c$. From the analysis of Section III*C*, we know that if eviction threats are used in the first period, then in that period $e = (1 + R)/2c$ and $r_h = (1 + R)/2$. Since the second period is the last period, $e = 1/2c$ and $r_h = \frac{1}{2}$ as in the static model. The landlord's total expected profit is $[(1 + R)^2/4c] + (1/4c)$. Suppose instead that the landlord guarantees tenure to the tenant and precommits to the second-period contract. In this case, the maximum amount by which the landlord can increase the second-period rent ex ante is equal to the net social surplus from investment, $1/2\gamma$. Hence, his total expected profit is $(1/2c) + (1/2\gamma)$. It is readily checked that if c is low and γ is high (which means that it is relatively cheaper to elicit effort from the tenant than invest-

ment), the landlord will prefer to use eviction as a threat. In this case, even though the legal contract raises investment, efficiency, and the tenant's payoff, the landlord will be worse off.

We can summarize our analysis as follows.

RESULT 4. An improvement in the tenant's outside option increases the marginal return on contractible investments that are complementary with effort. Security of tenure and a higher crop share induce the tenant to increase the supply of noncontractible land-specific investments.

E. Operation Barga, Security of Tenure, and Crop Shares

Theoretically we expect Operation Barga to have increased the outside option of tenants and to have made eviction impossible. We know from the aggregate data that the tenants responded positively to the reform: according to official estimates, by 1993, about 65 percent of all sharecroppers were registered, compared to 15 percent before Operation Barga. However, the aggregate data do not indicate whether the reform actually affected contractual terms. In order to fill this gap, we surveyed a stratified random sample of 480 sharecroppers from 48 villages in West Bengal. The survey asked each farmer detailed questions about various aspects of the landlord-tenant contractual relationship before and after the reform.[22]

These data show that the reform greatly improved security of tenure. In the prereform period, tenure was not secure: 74 percent of tenants surveyed said that in the prereform period their leases did not have a specified duration and were subject to arbitrary termination by the landlord, 80 percent reported that landlords in their village had used eviction threats, and 30 percent reported that they or their fathers were actually threatened.[23] The reasons cited for the use of threats of eviction include both low production (in 40 percent of the cases) and disputes with the landlord (in 55 percent of the cases). In other words, eviction was used both as an incentive device and as an instrument for bargaining. After the reform, eviction threats have almost disappeared: 96 percent of all respondents reported that evicting registered tenants is difficult or impossible, and 67 percent also reported that it is difficult or impossible to evict even unregistered tenants—largely because they can register themselves whenever they want. Finally, actual evictions in the postreform period are rare: only 30 percent of respondents said that they know of a tenant who was evicted in the last 10 years.

Since eviction threats were used by the landlord in bargaining in the

[22] See Banerjee and Ghatak (1996) for a more detailed discussion of the survey.

[23] These numbers presumably understate the importance of these threats since in equilibrium the tenants presumably adjust their behavior to avoid the threats.

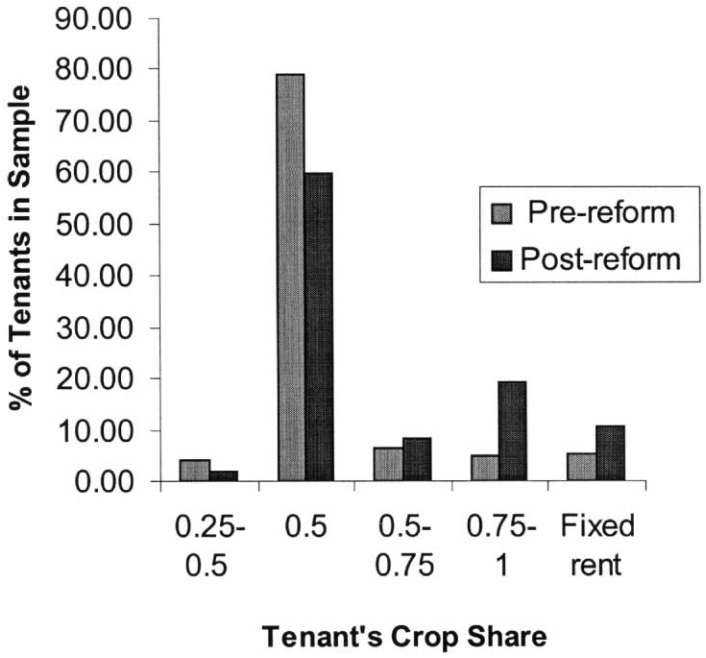


FIG. 3.—Crop share of tenants before and after the reform

prereform period, making eviction difficult or impossible must have strengthened the tenant's bargaining position: in other words, m should have gone up. Our model says that the tenant's share of the crop should go up, or at least not go down, when m goes up. Our survey (as well as smaller surveys by Kohli [1987] and Chadha and Bhaumik [1992]) confirms that crop shares increased after the reform (see fig. 3). For example, the proportion of tenants in our sample getting more than 50 percent of output increased from 17 percent to 39 percent. Evidence from our survey suggests that while shares rose for both registered and unregistered tenants, the increase was greater for registered tenants. To the extent that unregistered tenants faced some insecurity of tenure, their bargaining power presumably increased less, resulting in a smaller increase in the share.[24]

[24] This begs the question of why these tenants did not register. Unregistered tenants in our sample cited two main reasons for not registering: either they had good relations with the landlord or they were dependent on the landlord for credit or other inputs. We might surmise that for both these groups, though for different reasons, the change in m was more limited than it was for those who registered.

This evidence, however, underestimates the extent of change in shares. We were able to survey only those who were still tenants in 1995. This leaves out all those who switched from being sharecroppers to being owner-farmers as a result of the program. This happened for two reasons. First, many landlords, especially those who were absentee, faced with having to deal with a registered sharecropper, preferred to sell out and leave. As a result, land prices fell, allowing erstwhile sharecroppers to buy land. Second, even in cases in which the landlord did not sell out and leave, he often preferred to arrive at an arrangement with his tenant whereby the tenant received ownership of a part of the land in return for giving up his claim on the rest.[25] In a detailed study of the land market in two villages in West Bengal, Rawal (2001) found that between 1977 and 1995, an amount of land constituting over 30 percent of total cultivated area was sold. The major sellers were large or absentee landowners, and the major purchasers were small owner-cultivators and sharecroppers. This is in sharp contrast with other Indian states, where land markets are very thin. To the extent that any land transfer takes place, it occurs from smaller to larger landowners.

F. Discussion

Eliminating the possibility of eviction reduces effort and other noncontractible current inputs as long as m is held fixed in our model. However, once the possibility of eviction is eliminated, a higher m tends to increase the supply of these inputs. Since Operation Barga both eliminated evictions and increased m, its net effect could be positive or negative. There are several other reasons why we might expect the net effect of the reform to be positive. First, investment incentives improve with better security of tenure. Second, our survey indicates that before the reform, eviction threats were not commonly used to punish tenants for producing too little.[26] Third, if the negative incentive effect was indeed significant for some tenant, the landlord could make him a side payment and sell off the land to an owner-farmer (or cultivate it himself). Indeed, such sales were part of the post–Operation Barga scene. Finally, the reform could have had indirect effects that go beyond the contractual relationship between landlords and tenants, something that we have not formally analyzed here. Some commentators (e.g., Gazdar and Sengupta

[25] This is in fact what our model would have predicted if we had allowed the landlord to own several plots of land and self-cultivation by the landlord was an option. Our survey and other studies (Rawal 2001) have found several instances of such land transfers.

[26] Recall that only 40 percent of the 30 percent (i.e., 12 percent of the entire sample) of tenants who indicated that they or their father was threatened with eviction singled out this particular reason. Of course, as pointed out above, these numbers probably underestimate the importance of threats of eviction.

1999) have put particular emphasis on this indirect effect of agrarian reform in West Bengal. It is often argued (see Boyce 1987) that collective action within rural societies (e.g., with respect to management of irrigation water) is severely handicapped by the extreme inequality in the distribution of political and economic power within the society. To the extent that Operation Barga affected this distribution of power, it is likely to have contributed to the alleviation of such collective action problems.[27]

IV. Evidence: The Effect of Operation Barga on Productivity

Our objective in this section is to estimate the effect of the change in property rights brought about by Operation Barga on agricultural productivity in West Bengal. We take two approaches. The first is a quasi-experimental approach that uses Bangladesh as a control. The second uses the number of registered sharecroppers in a district as a measure of program intensity and compares the growth in productivity in districts in which Operation Barga was implemented intensely to districts in which the program was implemented less intensely.

A. Comparison to Bangladesh

Bangladesh, which did not introduce tenancy reform, provides a good comparison to West Bengal. Prior to independence, Bangladesh and West Bengal were parts of the same state in undivided India. Except for religion and political boundaries, the two regions are very similar in most respects. This includes agroclimatic conditions, prevalence of tenancy, and agricultural technology (Boyce 1987). Hence we can expect technological shocks to agricultural yields to be similar between these two regions.

The fact that Operation Barga was implemented in West Bengal but not in Bangladesh can be largely attributed to an exogenous shock. Operation Barga could be implemented only because a left-wing government unexpectedly came into power in West Bengal in 1977. This was a result of a nationwide wave against the Congress Party, which had ruled in most states since independence. In the mid 1970s, a severe political crisis led the Congress-dominated central government to suspend civil liberties. In the subsequent elections in 1977, the voters punished the Congress Party for this: the Left in West Bengal was the beneficiary of this anti-Congress wave. Thus the timing of Operation Barga

[27] See Bardhan, Ghatak, and Karaivanov (2002) for a theoretical analysis of how lower land inequality can improve overall efficiency in the presence of collective action problems ranging from the provision of public goods to the use of common property resources.

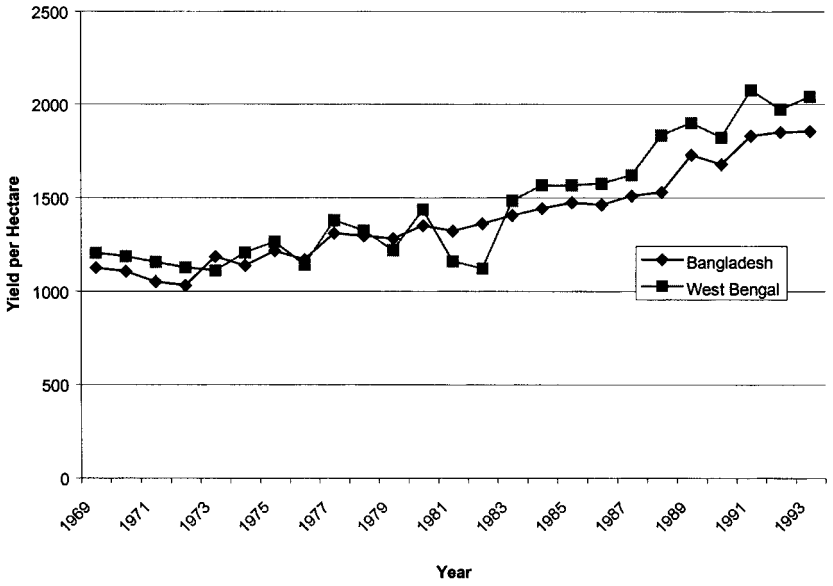


FIG. 4.—Rice yield in West Bengal and Bangladesh, 1969–93

did not reflect what was then happening in West Bengal but rather what was happening in the rest of India.

In the period before Operation Barga, agricultural productivity was growing at almost identical rates in the two states. Rice is the main component of agricultural production in West Bengal and Bangladesh and is planted in over 70 percent of cropped area. Between 1969 and 1978, a period covering the decade before Operation Barga, rice yields increased by 9.3 percent in West Bengal and by 11 percent in Bangladesh. In the period after Operation Barga was introduced (1979–93), rice yields in West Bengal increased by 69 percent compared to 44 percent in Bangladesh.[28] This can be seen more clearly in figure 4, which presents rice yields per hectare over time for West Bengal and Bangladesh. Until 1979, the first real year of Operation Barga, rice yields are approximately the same for the two countries. In the post–Operation Barga period, rice yields in West Bengal are substantially higher in all years except for 1981 and 1982, when West Bengal experienced two

[28] The average exponential rate of growth per year was 4.1 percent in West Bengal and 2.7 percent in Bangladesh during 1979–93. See Saha and Swaminathan (1994) for a detailed analysis of the growth performance of agriculture in West Bengal during this period.

successive years of severe droughts, among the worst experienced in the century (Government of West Bengal *Economic Review,* 1983, pp. 13–14).[29]

During the period of study, agricultural productivity in both regions (and much of eastern India) grew in part as a result of three common factors: the belated arrival of the Green Revolution permitted by the spread of a locally suited high yield variety (HYV) of rice, a fall in the price of fertilizers, and an increase in small-scale private irrigation (Harriss 1993). However, even though the rate of adoption of HYV rice was faster in Bangladesh than in West Bengal, the rate of growth in rice productivity was higher in West Bengal. This difference is what we shall attribute to the implementation of Operation Barga.

1. Methods

We measure the impact of Operation Barga on agricultural rice yields using a difference-in-difference estimator with district-level panel data. The difference-in-difference specification compares the change (before and after Operation Barga) in yields in treatment districts (West Bengal) with the corresponding change in control districts (Bangladesh).

The difference-in-difference model can be specified in regression form as

$$\ln y_{dt} = \alpha_d + \psi_t + \beta \times \text{treatment}_d \times \text{post}_t + \sum \phi_j X_{jdt} + \epsilon_{dt}.$$

The dependent variable is the log of the rice yield per hectare in district d and year t. The right-hand-side variables include a fixed effect for each district, a fixed effect for each year, the interaction of a variable indicating whether the district is a treatment (i.e., in West Bengal), and an indicator of whether it is the postreform period. There are also a series of control variables (the X_j's) that vary over time and across districts. The district fixed effects control for district-specific factors that are fixed over time, and the year fixed effects control for factors that vary over time but are common across all districts—both treatment and control. The coefficient β is the difference-in-difference estimate of the impact of Operation Barga on rice yields.

The difference-in-difference model makes the counterfactual assumption that the treatment districts would grow at the same rate as

[29] Agriculture in South Asia is heavily dependent on the summer monsoon rains, whose distribution over time and across regions tends to be highly variable (see Das 1995, pp. 228–34). While crop yields depend on the total amount and timing of rainfall, we are able to control for only the former. These two years had lower than average total rainfall for both West Bengal and Bangladesh, but crop production in West Bengal (especially the main variety of rice, *aman*) additionally suffered from the erratic timing of the monsoons.

the control districts if there were no intervention. While this assumption is not directly testable, we can test whether the treatment districts and the control districts were growing at the same rate in the preintervention period. If we do find that they were growing at the same rate, it would suggest that our counterfactual assumption is likely to be correct.

This assumption would also be violated if there were some other interventions that were contemporaneous with Operation Barga and were differentially implemented in West Bengal and Bangladesh. To control for the possibility that there were other interventions contemporaneous with Operation Barga that could explain divergence between West Bengal and Bangladesh, we explicitly investigate two important agricultural policies: public irrigation and adoption of HYV grains of rice, which is a measure of the progress of the Green Revolution. These are the interventions that are typically seen as the major technological sources of increased productivity.

2. Simple Difference-in-Difference Results

In this subsection we estimate a simple difference-in-difference model with no time-varying controls on log rice yields for the period 1969–93.[30] The data are district-level data from 14 West Bengal and 15 Bangladesh districts collected from various official government sources.[31] Summary statistics for the log of rice yields for West Bengal and Bangladesh for this period are reported in column 1 of table 1.

We begin by testing the hypothesis that growth in yields in West Bengal districts and Bangladesh districts was the same in the pre–Operation Barga period. This is an indirect test of the difference-in-difference assumption that the change in the control districts is what would have happened in the treatment districts if there were no intervention. To conduct the test we regress changes in log yields over the period 1969–78 against an indicator of whether the district is in West Bengal and year dummies. The hypothesis is rejected if the coefficient on the West Bengal dummy is significantly different from zero. The results are presented in column 1 of table 2. We cannot reject the hypothesis that growth was the same in both control and treatment districts in the pre–Operation Barga period.

The coefficient estimates from the simple difference-in-difference

[30] The data are taken from *Economic Review* (1969–93); the 1990 *Statistical Abstract* (Government of West Bengal); and the 1969–93 *Statistical Yearbook of Bangladesh* (Bangladesh Bureau of Statistics, Statistics Division, Ministry of Planning, Government of the People's Republic of Bangladesh).

[31] From West Bengal we excluded Calcutta, which is almost completely urban, and Purulia, for which data are not available for a considerable number of years. From Bangladesh we excluded eight districts for which data are not available for a large number of years because of changes in the administrative boundaries of these districts.

TABLE 1
SUMMARY STATISTICS

	Log(Rice Yield, kg per Hectare)		HYV Share,[a]	Proportion of Registered Tenants,[b]	Log(Area under Public Irrigation, Hectare),[c]	Log(Road Length, km),[d]	Log(Rainfall, mm),
	1969–93	1977–93	1977–93	1978–92	1977–93	1977–93	1977–93
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	West Bengal (Annual Observations on 14 Districts)						
Grand mean	7.24	7.32	.11	.49	10.01	6.99	7.42
Standard deviation:							
Overall	.31	.31	.09	.23	1.80	.39	.41
Within	.23	.22	.05	.18	.30	.07	.24
Mean in:							
1969	7.06	…	…	…	…	…	…
1977	7.20	7.20	.06	…	9.91	6.93	7.24
1979	7.07	7.07	.06	.15	9.92	6.94	7.17
1993	7.60	7.60	.18	.65	10.13	7.02	7.58
	Bangladesh (Annual Observations on 15 Districts)						
Grand mean	7.22	7.30	.15	0	11.36	…	7.69
Standard deviation:							
Overall	.23	.20	.11	0	.89	…	.35
Within	.19	.15	.07	0	.43	…	.21
Mean in:							
1969	7.05	…	…	…	…	…	…
1977	7.16	7.16	.09	0	11.00	…	7.62
1979	7.14	7.14	.09	0	11.06	…	7.64
1993	7.51	7.51	.25[e]	0	11.76	…	7.84

[a] Fraction of total rice area devoted to the cultivation of the summer crop, *boro*.
[b] Registration data are relevant only for West Bengal and are available for the period 1978–93.
[c] Public minor irrigation schemes include shallow tube wells, deep tube wells, and river lift irrigation.
[d] This information is not available as a continuous series for Bangladesh during the period of analysis.
[e] Information on HYV share for Bangladesh is available up to 1991, so this number pertains to 1991.

TABLE 2
DIFFERENCE-IN-DIFFERENCE MODELS OF LOG OF RICE YIELD PER HECTARE (1969–93)

	DIFFERENCE	LEVEL	
	(1969–78)	1969–93	Excluding 1981–82
	(1)	(2)	(3)
West Bengal (=1)	.004	…	…
	(.17)		
West Bengal × (1979–83)[a]	…	- .09***	−.01
		(3.75)	(.38)
West Bengal × (1984–88)	…	.05**	.05**
		(1.99)	(2.00)
West Bengal × (1988–93)	…	.05*	.05*
		(1.77)	(1.78)
District fixed effects *F*-statistic	…	44.55	42.61
Year fixed effects *F*-statistic	4.26***	29.75***	31.81***
R^2	.12	.80	.81
Sample size	256	717	659

NOTE.—*t*-statistics are in parentheses.
[a] These variables are obtained by interacting a dummy variable that takes the value one if a district is in West Bengal and zero if it is in Bangladesh with another dummy variable that takes the value one if the observation is in the indicated time period (1979–83 in this case) and zero otherwise.
* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

models of log rice yield are presented in columns 2 and 3 of table 2 for 1969–93. The key variables are the interactions of an indicator of whether the district is in the treatment area (West Bengal) with indicators of whether the year was in the postreform period. We split the postreform period into three periods of equal length to accommodate variation in the speed at which registration proceeded, as well as lags in the output response to Operation Barga (e.g., because the effect through increased investment would take time to materialize). The last period reflects the full effect of Operation Barga since registration was mostly complete by then and any resulting investments are likely to have already affected productivity. We reestimated the model excluding 1981 and 1982, when West Bengal experienced two successive years of major droughts.

The first three coefficients in columns 2 and 3 are the difference-in-difference estimates. In the early years of Operation Barga (1979–83), West Bengal grew slower than Bangladesh, but this effect seems to be entirely driven by the presence of the two drought years that disproportionately affected West Bengal. In the next two periods (1984–88 and 1988–93), rice yields were about 5 percent higher. These results are consistent with the hypothesis that Operation Barga had a positive impact on productivity.

3. Adjusted Difference-in-Difference Results

In this subsection, we adjust the simple difference-in-difference estimates for time-varying controls. However, the data on the time-varying controls exist only for the period 1977–91 for both Bangladesh and West Bengal. The data available for both regions include information on rice production and average yield per hectare, a measure of the amount of rice area under HYV cultivation,[32] total annual rainfall, and area covered by public irrigation. Descriptive statistics for these data are reported in table 1. We see that in the postreform period (1979–93), the share of HYV rice in total cultivated area increased from 6 to 18 percent in West Bengal and went up from 9 to 25 percent in Bangladesh. Over the same period, area under public irrigation increased by 23 percent in West Bengal and doubled in Bangladesh.[33]

The availability of data on time-varying controls allowed us to estimate a number of different specifications to test the robustness of the estimates.[34] The results of this exercise are reported in table 3. We estimated a simple unadjusted difference-in-difference model (model 1), another controlling for public irrigation and rainfall (model 2), and another that additionally controlled for HYV share (model 3).

The models show that, except for the first period, rice yields grew faster in West Bengal than in Bangladesh and the differentials grow over time. Model 1, which repeats the simple difference-in-difference analysis for the shorter panel, yields results similar to the simple difference-in-difference results for the long panel reported in the previous subsection. When the drought years are excluded, there is no difference between West Bengal and Bangladesh in the early period, 1979–83. The differential in the later periods (1984–87 and 1988–91) grows when time-varying controls are included in the model. When we control for rainfall, public irrigation, and HYV share, the West Bengal yields are estimated to be 7 percent higher between 1984 and 1987 and 18 percent higher between 1988 and 1991.

These estimated differences are an average of sharecropper and owner-cultivator yields. Assuming that Operation Barga had no effect

[32] Our measure of HYV adoption is the fraction of total rice area devoted to the cultivation of the summer crop, *boro*, which is completely dependent on irrigation and uses HYV seeds and other modern inputs.

[33] Data on private irrigation in West Bengal are unfortunately available for only two years within the period under study.

[34] We repeated the test whether the growth rates of the two regions were the same in the prereform period with the controls for this shorter series (for which the prereform period consists of only 1977 and 1978) and again found the West Bengal dummy to be negative and insignificant.

TABLE 3
DIFFERENCE-IN-DIFFERENCE MODELS OF LOG OF RICE YIELD (1977–91)

	WHOLE SAMPLE			EXCLUDING DROUGHT YEARS 1981–82		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
West Bengal × (1979–83)	− .08*** (−2.43)	−.07** (−2.05)	− .05 (−1.58)	.001 (.01)	.002 (.06)	.015 (.47)
West Bengal × (1984–87)	.04 (1.17)	.05 (1.47)	.07** (2.04)	.04 (1.24)	.04 (1.26)	.06** (1.93)
West Bengal × (1988–91)	.08** (2.20)	.12*** (3.28)	.18*** (5.11)	.07** (2.33)	.11*** (2.97)	.17*** (4.95)
Log(rainfall)	…	.01 (.40)	.007 (.32)	…	.019 (.70)	.01 (.46)
Log(public irrigation)	…	.122*** (7.22)	.07*** (4.27)	…	.103 (5.77)	.04*** (2.69)
HYV share of grain cultivation area	…	…	1.04*** (8.18)	…	…	1.05*** (8.21)
District fixed effects *F*-statistic	40.02***	20.14***	14.76***	41.43***	18.8***	14.64***
Year fixed effects *F*-statistic	20.18***	12.14***	7.73***	21.67***	12.41***	6.04***
R^2	.82	.85	.87	.83	.85	.88
Sample size	424	424	424	367	367	367

NOTE.—*t*-statistics are in parentheses.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

on owner-cultivator productivity, we can estimate the effect of Operation Barga on sharecropper productivity using the formula

$$\frac{1}{A}\frac{dA}{dt} = \frac{s}{1-s}\frac{1}{A^o}\frac{dA^s}{dt},$$

where A is average productivity, A^o is the average productivity of owner-cultivators who are not affected by the reform, A^n is the average productivity of sharecroppers, and s is the average area under sharecropping.[35]

There is unfortunately some controversy about the amount of land under sharecropping in West Bengal. The main reasons are lack of reliable land records, the presence of concealed tenancy to evade tenancy laws, and problems of definition of tenancy. Estimates of total cultivated area under sharecropping in West Bengal before the reform were introduced, provided by various rounds of surveys conducted by

[35] This formula follows from taking logs of the equation $A = sA^n + (1-s)A^o$ using the approximation $\ln(1+x) \simeq x$ when x is small to obtain

$$\log A_d = \frac{sA^n}{(1-s)A^o} + \log[(1-s)A^o]$$

and then differentiating with respect to t. Notice that these percentage changes occur with respect to productivity in owned land (i.e., A^o). Hence the changes with respect to productivity in sharecropped land (i.e., A^n) would be larger.

the National Sample Survey (NSS), are considered to be the most reliable given their large sample base and methodology. These estimates fall within the range of 18–22 percent (see Bardhan 1976, tables 1, 4). The lower bound of this range, which is obtained from the NSS survey of 1970–71, is considered to be an underestimate (Bardhan 1976; Laxminarayan and Tyagi 1977). The upper-bound estimate obtained from the NSS survey of 1953–54 is considered to be more reliable in this respect because it was conducted before tenancy laws were enacted in the country. We take the estimate of 20 percent, which is at the middle of this range.[36] Given that total area under rice cultivation in West Bengal is around 70 percent and sharecropping is observed predominantly with respect to rice cultivation, the proportion of rice area under sharecropping is higher. According to a recent study, over 90 percent of land leased by sharecroppers was under rice cultivation (Bhaumik 1993, table 6.2). This gives us an estimate of about a quarter of rice area under sharecropping.[37]

Under the assumption that there was no differential change in the yields of owner-cultivators between West Bengal and Bangladesh, sharecropper productivity increased by 51 percent during the last period (1988–91).[38] This period gives us the cumulative effect of the reform (i.e., including the effect through investment).

One striking result that can be seen from table 3 is that the estimate of the impact of Operation Barga on productivity increases when we control for public irrigation and the Green Revolution. This suggests that Bangladesh expanded these public programs faster than West Bengal did in the post–Operation Barga period. This hypothesis is consistent with the descriptive statistics reported in table 1. We formally test this hypothesis using the difference-in-difference framework with public irrigation and HYV share as the dependent variables. The results are presented in table 4. The results show that both public irrigation and the share of HYV expanded faster in Bangladesh in the postreform period than they did in West Bengal.

The fact that Bangladesh expanded these public programs designed to improve agricultural productivity faster than West Bengal is important

[36] It is also in the middle of the range provided by Boyce (1987, p. 214) in his authoritative study on agriculture in West Bengal and Bangladesh (namely, one-sixth to one-fourth) based on various sources including the NSS.

[37] Official data suggest that the fraction of land under sharecropping that is formally registered is about 8.2 percent of total cultivated area, or 10.5 percent of area devoted to rice cultivation. This number underestimates the total cultivated area on which Operation Barga had a direct effect since it does not take into account the effect on unregistered sharecroppers and transfer of land from sharecropping to owner cultivation due to land sales and transfers. The size of the total area under sharecropping *before* Operation Barga was introduced is preferred for this reason.

[38] This estimate is obtained by multiplying the coefficient of West Bengal × (1988–91) reported in table 3 by $(1 - s)/s = 3$.

TABLE 4
DIFFERENCE-IN-DIFFERENCE MODELS OF OTHER PUBLIC POLICIES (1977–91)

	Log(Public Irrigation)		HYV Share	
	Whole Sample	Excluding 1981–82	Whole Sample	Excluding 1981–82
West Bengal × (1979–83)	−.24** (−2.28)	−.18 (−1.61)	−.03** (−2.25)	−.022 (−1.45)
West Bengal × (1984–87)	−.27*** (−2.44)	−.24** (−2.18)	−.014* (−1.88)	−.029** (−1.95)
West Bengal × (1988–91)	−.57*** (−4.97)	−.53*** (−4.69)	−.083*** (−5.25)	−.085*** (−5.58)
Log(rainfall)	.06 (.82)	.005 (.06)	.006 (.56)	.007 (.67)
District fixed effects *F*-statistic	250.66***	227.98***	55.32***	49.21***
Year fixed effects *F*-statistic	8.68***	9.51***	29.65***	31.22***
R^2	.96	.96	.85	.85
Sample size	424	367	424	367

NOTE.—*t*-statistics are in parentheses.
* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

for interpreting the results. A concern with this methodology is that there may be unobserved differences in government programs between the two countries. If these unobserved programs behaved like the observed programs and also expanded faster in Bangladesh in the post–Operation Barga period, our difference-in-difference estimates of the impact of Operation Barga on agricultural productivity would give us a lower-bound estimate. However, we cannot completely rule out the possibility that there were unobservable policies that confound the estimated effect. We therefore complement this analysis with an alternative approach in which we estimate the effect of Operation Barga using variation in program intensity across districts within West Bengal.[39]

B. Program Intensity

The approach taken in this subsection uses the district sharecropper registration rate as a measure of program intensity and then examines whether productivity rises faster in areas with greater program intensity.

[39] Some changes were introduced in the methodology of collection of official crop statistics in West Bengal starting in 1986 that, according to some critics, could result in biased estimates of the growth rate. When interdistrict variation in program intensity *within* West Bengal is used to estimate the effect of Operation Barga, our second approach is not subject to any possible bias resulting from this source.

We have data on the number of registered tenants in West Bengal for the period 1978–93, with 1978 being the year in which Operation Barga was launched.[40] Operation Barga was launched in 1978, and at the beginning of 1978, the average level of registration for West Bengal was 15 percent. In 1993, registration stood at 65 percent of the total number of sharecroppers. We augment the set of time-varying controls for West Bengal used in the previous section by data on the length of roads constructed and maintained by the public works department.[41]

We begin by formally deriving the empirical specification, which relates yields to the registration rate and other covariates from the production process. However, we do not have data about the productivity of individual sharecroppers. What we have is district-level yields generated by averaging across registered sharecroppers, unregistered sharecroppers, and owner-cultivators. In order to interpret the coefficients correctly, in the next subsection we aggregate the individual-level model to generate a district-level model.

1. Specification

Individual farm productivity.—Our starting point is a reduced-form productivity equation derived from a structural profit-maximizing model of a tenant farmer. Production depends on the tenant's noncontractible inputs (e.g., effort), contractible inputs (e.g., fertilizer and seeds), publicly provided inputs (e.g., irrigation and roads), and rainfall. Farmers choose effort and contractible inputs to maximize profits subject to the agricultural production function, the parameters of the tenancy contract, prices, public inputs, and rainfall. We assume a Cobb-Douglas specification for farm i's profit-maximizing output per hectare (yield) at time t:

$$Y_{it} = A(\mathbf{c}_{it}, \boldsymbol{\theta}_i)\left(\prod_{j=1}^{n} P_{jt}^{\alpha_j}\right)\left(\prod_{k=1}^{N} X_{kit}^{\beta_k}\right) r_{it}^{\gamma}[\exp(\epsilon_{it})], \tag{9}$$

where A is the X-efficiency of the farm, $\mathbf{c}_{it}$ is a vector of contract parameters (e.g., crop share, probability of eviction for different values of output, etc.), $\boldsymbol{\theta}_i$ represents fixed characteristics of the tenant and the farm (e.g., wealth, ability, and land quality), the P_{jt} are market prices of contractible inputs (we set the output price equal to one), the X_{kit} are publicly available inputs provided by the government (e.g., canal irri-

[40] Data on sharecropper registration were obtained from the Statistical Cell, Department of Land Reforms, Government of West Bengal, and data on districtwise number of sharecroppers from Datta (1981).

[41] These data are not available for Bangladesh after 1984 and hence were not used in the previous subsection.

gation available for the farm and roads for transport of produce to market), r_{it} is the amount of rainfall on the farm during period t, and ϵ_{it} is a zero mean random productivity shock.

The change in the *X*-efficiency parameter A captures the net effect of the two contractual responses to the reforms. The first is the effect of improved crop share of tenants on the supply of noncontractible inputs (e.g., effort). The second is the net effect of the permanency of tenure on the choice of inputs (both current inputs and investments).

In Section III*E,* we found that tenants renegotiated their contracts and obtained better terms after they had the opportunity to register even when they did not register. Therefore, we need to account for both types of tenant farmers in the analysis. Let A^n denote the efficiency of a tenant farm in the prereform period. Further, let A^r and A^u denote the efficiency of tenant farms whose contracts were renegotiated after the reform, with the former referring to a farm cultivated by a registered tenant and the latter to a farm cultivated by a tenant who did not register even though he had the opportunity to do so. As pointed out before, the latter category includes both those who remained tenants and those who became owners. Finally, let A^o be the efficiency of an owner-cultivated farm, which should be unaffected by the reform.

District productivity.—Since the data on total output are at the district level, we have to aggregate the individual farm productivity model to that level. This requires aggregating across registered sharecroppers, unregistered sharecroppers, and owner-cultivators. Over time, the number of registered sharecroppers rose and unregistered sharecroppers fell as Operation Barga was implemented.

The reforms reached tenants in the form of opportunities to register with the land bureaucracy. In order for tenancy laws to be enforced, the tenant had to register his status with the Land Revenue Office. Land revenue officials went village by village to create and update tenancy registration. The government, however, could not make the opportunity to register available to all tenants at the same time within and across districts because of resource constraints and logistical problems. Instead, registration opportunities expanded through districts over time on a village-by-village basis.

Therefore, average district *X*-efficiency at any point in time depends on the proportion of farmers who were tenants, the proportion of tenants who had the opportunity to register, and the proportion of people who chose to register (henceforth, the take-up rate). Because it would take some time for the parties to renegotiate the contracts and for that to have an effect on yields, especially through investment incentives, we use the proportion of tenants who had the opportunity to register lagged

by one period.[42] Formally, let s_d be the share of land that is cultivated by sharecroppers in district d, v_{dt} be the share of sharecroppers who have been offered the opportunity to register in district d at time t, and λ_d be the take-up rate. Then the average X-efficiency of district d in period t is

$$A_{dt} = s_d\{v_{dt-1}[\lambda_d A^r + (1 - \lambda_d)A^u] + (1 - v_{dt-1})A^n\} + (1 - s_d)A^o. \quad (10)$$

In principle, we would like to identify the effect of the reform by examining the effect of registration *opportunities* on district-level productivity. However, there is no information on the proportion of tenants who were offered such opportunities. We shall therefore use time-specific information on the proportion of tenants who actually registered as a proxy for the share of those who were offered registration opportunities. We rewrite (10) in terms of the proportion of tenants who have registered, $b_{dt-1} = \lambda_d v_{dt-1}$, to get

$$A_{dt} = s_d\left[b_{dt-1}\left(A^r + \frac{1 - \lambda_d}{\lambda_d}A^u - \frac{A^n}{\lambda_d}\right) + A^n\right] + (1 - s_d)A^o.$$

Rearranging terms and taking the log, we get

$$\ln A_{dt} = \ln\left\{1 + \frac{s_d}{1 - s_d}b_{dt-1}\left[\frac{\lambda_d A^r + (1 - \lambda_d)A^u - A^n}{\lambda_d A^o}\right] + \frac{s_d}{1 - s_d}\frac{A^n}{A^o}\right\}$$
$$+ \ln (1 - s_d)A^o.$$

Since $\ln (1 + x) \simeq x$ when x is small, we rewrite (9) in log form as

$$\ln y_{dt} = \alpha_d + \gamma b_{dt-1} + \sum_j \alpha_j \ln P_{jdt} + \sum_k \beta_k \ln X_{kdt} + \epsilon_{dt}, \quad (11)$$

where

$$\alpha_d = \frac{s_d}{1 - s_d}\frac{A^n}{A^o} + \ln (1 - s_d)A^o$$

and

$$\gamma = \frac{s_d}{1 - s_d}\frac{\lambda_d A^r + (1 - \lambda_d)A^u - A^n}{\lambda_d A^o}.$$

The coefficient γ measures the effect of the reform on agricultural productivity. The numerator of the coefficient is the average X-efficiency of sharecroppers offered registration opportunities minus the X-efficiency of sharecroppers not offered registration opportunities. This is

[42] This would also partially control for the problem that registration could be driven by current productivity shocks. See the next subsection for a detailed discussion of identification problems.

just the marginal increase in productivity arising from registration opportunities. The marginal increase is measured relative to the *X*-efficiency of owner-cultivated farms. The marginal increase is also weighted by one over the take-up rate. This converts the units from change in productivity due to a change in registration opportunities to change in productivity due to a change in the registration rate.[43]

A limitation on the analysis is that districtwise data on output prices and input prices are available only for a limited number of years. Throughout our analysis we include year dummy variables to capture the common movements of prices over time in the districts. This seems to be a reasonable approximation because the state and federal governments control both input and output prices, and hence their movements over time are not very different across districts.[44] However, to check the robustness of our results, we also estimate the model for the shorter sample for which we have districtwise data on rice prices and real wages. Therefore, the equation to be estimated is

$$\ln y_{dt} = \alpha_d + \psi_t + \gamma b_{dt-1} + \sum_k \beta_k \ln X_{kdt} + \epsilon_{dt}, \tag{12}$$

where ψ_t are the year-specific intercepts. Notice that the year fixed effects also control for any other unobserved time-varying factors that are common to districts such as changes in technology or government policies.

2. Identification

In this subsection we consider issues relating to the identification of the model. The objective of the exercise is to measure the impact of the reform on agricultural productivity using the registration rate as a measure of program intensity. However, the registration rate may be correlated with unobserved productivity shocks for two reasons. First, the registration rate is a combination of the supply of registration opportunities and the demand for such opportunities. Also, the sequence of villages offered registration was not necessarily chosen at random. Second, the progression of registration opportunities could have been correlated with the progression of other (omitted) programs. The ideal response to these problems would be to use an instrumental variables

[43] In principle, γ could vary by district if the take-up rates and the relative importance of sharecropping vary by district. However, we do not have long enough time-series variation within districts to estimate district-specific slopes with much precision. Instead, we can estimate the average effect of the reform across all districts. In this case, our specification could be interpreted as a random-coefficients model.

[44] For example, most inputs (e.g., fertilizer, seeds) are distributed by public-sector agencies and subsidized by the federal government. Also the government through various agencies procures a large part of the crop for public distribution, export, and storage purposes.

approach. However, since anything that affects registration is likely to also affect productivity directly, there are no plausible instruments. We therefore take the approach of controlling for a range of time-varying factors that are likely to have influenced productivity. Below we discuss in detail these important issues and how we handle them.

Sources of variation in the registration rate.—The village-by-village visits by land revenue officials to update tenancy registration were a crucial determinant of the registration rate. While it was possible to register at a time other than during a visit, it was much more difficult (Chattopadhyay et al. [1984] and an interview with D. Bandyopadhyay by Maitreesh Ghatak in April 1997). In fact, 76 percent of the registered respondents to our survey of sharecroppers indicated that the official visits with the Land Revenue Office were the single most important factor leading to their registration decision.

The supply of registration opportunities spread at differential rates across districts (Ghosh 1981).[45] The districts had different bureaucratic resources and physical infrastructures, translating into differential efficiencies of the operation of visits (Ghosh 1981; Chattopadhyay et al. 1984). There were natural shocks to the process of registration such as floods (Lieten 1992). The geographic distribution of sharecroppers within a district varied across districts, and as a result, the marginal cost of making registration opportunities available to tenants varied across districts.

While supply-side frictions explain much of the variation in the registration rate, the distribution of registration opportunities may not have been random. If the government introduced registration opportunities in districts of high or low productivity first, then the registration rate would be correlated with unobserved productivity characteristics and our estimates would be inconsistent. However, if allocations were based on initial productivity, which is a time-invariant factor, the district fixed effects control for this source of bias. On the other hand, if the government dynamically allocated registration opportunities on the basis of current productivity in the district, then the fixed-effects estimate will be biased. A similar problem could occur if the order of villages selected within a district was based on productivity.

While the friction-driven variations in the supply of registration opportunities were clearly important, registration is ultimately a choice. A tenant's decision to register is likely to be affected by his ability, wealth, relations with the landlord, and other characteristics that are associated

[45] Information from our survey supports the hypothesis that the supply of opportunities did not arrive at all the villages at the same time. There is a fair amount of variation among villages in terms of peak year of registration. While 1980 had the highest number of villages experiencing peak registration, some villages peaked as late as 1994, 16 years after the launching of the program.

with his dependence on the landlord (e.g., for loans) or his bargaining power. The wealthier, more able, and more enterprising tenants are likely to be more productive and adopt productivity-enhancing technology. These individuals may also be more likely to register. Therefore, a district with a higher proportion of more productive tenants is likely to have high output as well as high registration. However, as long as these individual characteristics are constant over time, they should not be a problem as long as we allow for district fixed effects.

Finally, a portion of registration decisions could be driven by idiosyncratic shocks that vary across time and district. For example, a drought or flood would affect productivity and therefore the decision to register. While we explicitly control for total annual rainfall, there could still be some other district-specific productivity shocks (such as the timing of rainfall) or introduction of new technology that affects the registration choice.

Omitted programs.—Another potential source of bias comes from the possibility that there were public programs that were implemented or strengthened at the same time and in the same locations as Operation Barga. While Operation Barga itself did not provide any other services other than registration opportunities and the enforcement of the tenancy laws, there were clearly other programs that were part of the government's overall reform package. It is not inconceivable that the implementation of these programs was possibly correlated with the implementation of Operation Barga. Below we discuss various alternative programs and to what extent we control for them.

First, there was some expansion of infrastructure in West Bengal. We partially control for public investment in infrastructure by including measures of the availability of public irrigation and roads within districts.

Second, the use of HYV seeds spread during this period (the Green Revolution) spurred partially by government extension programs. We control for this by including the share of gross cropped area planted with HYV seeds.

Third, it is likely that Operation Barga was better implemented in areas in which the Left Front and its peasant organizations have greater political strength. During this period the role of village-level local governments (*panchayats*) was significantly enhanced in the implementation of various public programs. Operation Barga and other public programs may have been better implemented in districts that had more active local governments, especially those dominated by political parties belonging to the Left Front. To partially control for this, we introduce a Left Front majority district (in 1977) dummy variable interacted with time[46] as additional controls.

[46] We split the postreform period into three time periods of roughly equal length: 1979–83, 1984–87, and 1988–93.

Fourth, the south of West Bengal is closer to the administrative center, Calcutta. For this reason, it may have had better access to a range of government programs including Operation Barga. Calcutta being the metropolitan center and the largest city in eastern India, it is possible that districts closer to it would also experience different patterns of market or technological shocks (e.g., people may be more exposed to new ideas or technologies). To control for these possibilities, we introduce the interaction of a southern district dummy variable with time as additional controls.

Fifth, registration may have been targeted toward areas with a high concentration of sharecroppers. This could lead to a spurious correlation if the evolution of population characteristics and opportunities among sharecroppers was different than in the rest of the population. We control for this by including the initial extent of sharecropping interacted with time dummies as additional explanatory variables.

Sixth, there could be some concern that Operation Barga could be picking up some general equilibrium effects on wages and prices. For the limited number of years (1979–87) for which we have districtwise data of wages and prices, we include them as controls to address this issue.

Finally, there are two other programs that we do not explicitly control for but are unlikely to affect our results. The government started a subsidized loan program for registered sharecroppers. However, the program had very limited impact because of bureaucratic limitations (Kohli 1987). Indeed, 87 percent of the respondents to our survey indicated that they never received a loan from either a government or a commercial lending institution. In addition, the administration also redistributed a limited amount of land to the landless and poor peasants. However, most of the redistribution had been completed before the implementation of Operation Barga (Sengupta and Gazdar 1997). Over the entire sample period (1977–93) of our analysis, the land distributed in this manner constituted around 3 percent of the net cropped area of the state.

3. Results

Table 5 reports the results for the log rice yield models regressed against the registration rate and controls. Column 1 reports the fixed-effects results with no other controls, and in columns 2–5 we successively introduce a number of controls. All these models show that the registration rate was significantly positively associated with yields. The other significant coefficient estimates are as expected: expanding roads increases productivity, and rice yields were higher in districts that planted a greater share of HYV grain. In addition, southern districts and those

TABLE 5
EFFECT OF REGISTRATION ON THE LOG OF RICE YIELD IN WEST BENGAL, 1979–93
(N=210)

	Model 1 (1)	Model 2 (2)	Model 3 (3)	Model 4 (4)	Model 5 (5)	Model 6 (6)
Sharecropper registration (one year lagged)	.43*** (3.46)	.42*** (3.44)	.43*** (3.55)	.35*** (2.69)	.36*** (2.64)	.36*** (2.63)
Log(rainfall)	…	−.07* (−1.67)	−.08* (−1.82)	−.07 (−1.59)	−.08* (−1.74)	−.08* (−1.77)
Log(public irrigation)	…	.02 (1.01)	.01 (.70)	.01 (.60)	.02 (.83)	.02 (.79)
Log(roads)	…	.28*** (2.75)	.25** (2.46)	.21** (1.99)	.19 (1.55)	.22 (1.54)
HYV share of rice area	…	…	.57*** (2.85)	.45** (2.10)	.47** (2.16)	.47** (2.16)
F-statistic:						
South × year[a]	…	…	…	4.73***	4.36***	4.38***
Left Front × year[b]	…	…	…	…	2.64**	2.65**
Sharecropping × year[c]	…	…	…	…	2.64**	.12
District fixed effects	72.23***	15.10***	8.99***	9.01***	8.47***	7.68***
Year fixed effects	28.31***	27.67***	21.60***	17.63***	17.83***	12.17***
R^2	.91	.92	.92	.92	.92	.92

NOTE.—t-statistics are in parentheses.
[a] Represents a set of variables obtained by interacting a dummy variable that takes the value one if that district is in southern West Bengal with each year.
[b] Represents a set of variables obtained by interacting a dummy variable that takes the value one if that district had a Left Front majority at the local-level government in 1977 with each year.
[c] Represents a set of variables obtained by interacting the initial extent of sharecropping in a district with each year.
* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

with a Left Front majority in 1977 grew significantly faster. In table 6 we present the results with wages and prices as additional controls for the shorter sample (1979–87). We find that they do not matter significantly directly, or for the estimated coefficients of other right-hand-side variables including registration as long as we control for year-specific shocks.

The magnitude of the effect of Operation Barga on productivity is estimated by multiplying the coefficient on the registration rate with the change in registration over the period. For the full sample, in the model that includes the full set of controls (model 5 of table 5), the fixed-effects estimate is that Operation Barga raised average productivity of rice in West Bengal by 20 percent.[47] Since rice yields increase by 69

[47] Since at the beginning of the program the take-up rate was 15 percent and at the end of it was 65 percent, the take-up rate due to Operation Barga is $(0.65 - 0.15)/(1 -$

TABLE 6
EFFECT OF REGISTRATION ON THE LOG OF RICE YIELD IN WEST BENGAL, 1979–87
(N=126)

	Model 1a	Model 1b	Model 2a	Model 2b	Model 3a	Model 3b
Sharecropper registration	.44*** (2.71)	.46*** (2.73)	.46*** (2.41)	.48*** (2.89)	.40** (2.34)	.41** (2.29)
Log(real wages)	…	.11 (1.07)	…	.05 (.55)	…	.03 (.31)
Log(price of rice)	…	−.11 (−.98)	…	−.04 (−.40)	…	.001 (.01)
Log(rainfall)	…	…	−.08* (−1.65)	−.08 (−1.52)	−.08 (−1.45)	−.08 (−1.41)
Log(public irrigation)	…	…	.10** (2.34)	.09** (2.30)	.09** (2.19)	.09** (2.14)
Log(roads)	…	…	.10 (.82)	.10 (.78)	.08 (.47)	.08 (.50)
HYV share of rice area	…	…	.66** (2.14)	.59* (1.77)	.49 (1.45)	.47 (1.34)
F-statistic:						
South × year	…	…	…	…	yes	yes
Left Front × year	…	…	…	…	yes	yes
Sharecropping × year	…	…	…	…	yes	yes
District fixed effects	40.93***	29.34***	6.08***	10.20***	4.51**	3.98**
Year fixed effects	24.39***	20.20***	17.71***	4.36**	14.12***	11.29***
R^2	.89	.89	.90	.90	.90	.90

NOTE.—t-statistics are in parentheses.
* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

percent during this period, the share of Operation Barga in this improvement was 28 percent.

The impact on sharecropper productivity is obtained by solving the equation for γ (from eq. [11]) as follows:

$$\frac{\lambda_d A^r + (1 - \lambda_d)A^u - A^n}{A^o} = \frac{1 - s_d}{s_d}\gamma\lambda_d.$$

The left-hand side of this expression is the percentage change in the average productivity of sharecroppers offered registration relative to those not offered registration. Multiplying the point estimate of the effect of Operation Barga (0.36) by the take-up rate due to Operation

0.15) = 0.58. These numbers are obtained by multiplying this number with the point estimate of the coefficient of sharecropper registration.

Barga (0.58) and the relative importance of sharecropping ($[1 - s_d]/s_d = 3$), we get an estimated effect of 62 percent of Operation Barga on sharecropper yields. This is close to the estimate of 51 percent provided by the difference-in-difference approach using Bangladesh.

C. *Discussion*

Let us compare our estimates of the effect of Operation Barga on sharecropper productivity using various approaches with other studies of the impact of changing incentives on agricultural productivity. The two most closely related studies are Shaban (1987) and Laffont and Matoussi (1995). Shaban analyzed farm-level data from eight Indian villages and estimated that controlling for land quality changing the contractual status of a farm from sharecropper-cultivated to owner-cultivated would increase productivity by 16 percent.[48] Laffont and Matoussi use farm-level data from Tunisia to show that a shift from sharecropping to fixed-rent tenancy or owner cultivation raised output by 33 percent and moving from a short-term tenancy contract to a longer-term contract increased output by 27.5 percent.[49] While our estimates are definitely higher, it is worth emphasizing that the 95 percent confidence interval of our estimate goes from 15 percent to 105 percent and therefore includes all the existing estimates. Also, we should probably expect somewhat higher estimates because our measured effect includes the effect of additional investment resulting from the shift in property rights. Shaban's estimate goes up to 32.6 percent when he does not control for land quality. This increase in his estimate should at least partially be interpreted as a measure of the effect of investments in land quality. Finally, as indicated earlier, our estimate is likely to pick up various indirect effects of Operation Barga.

V. Conclusion

We concluded from our theoretical analysis that tenancy laws that lead to improved crop shares and higher security of tenure for tenants can have a positive effect on productivity. Evidence based on aggregate district-level data from the Indian state of West Bengal suggests that the

[48] See Shaban (1987, table 3). Shaban estimated changes with respect to productivity in owned land (i.e., $[A^o - A^n]/A^o$), and hence these numbers are directly comparable with ours.

[49] Laffont and Matoussi (1995, pp. 391–92) obtain estimates of 50 percent and 38 percent, respectively, but in terms of our notation, what they estimate is $(A^o - A^n)/A^n$. To make these numbers directly comparable with our estimates or that of Shaban, we compute $(A^o - A^n)/A^o$ on the basis of their estimates.

tenancy reform program called Operation Barga explains around 28 percent of the subsequent growth of agricultural productivity there. However, given data limitations, we cannot separate the direct and indirect effects of Operation Barga. To get more precise estimates, micro-level data are required, which we leave to future research.

References

Ackerberg, Daniel A., and Botticini, Maristella "Endogenous Matching and the Empirical Determinants of Contract Form." *J.P.E.* 110 (June 2002), in press.

Appu, P. S. "Tenancy Reform in India." *Econ. and Polit. Weekly* 10 (special issue; August 1975): 1339–75.

Bandiera, Oriana. "On the Structure of Tenancy Contracts: Theory and Evidence from 19th Century Rural Sicily." Manuscript. London: London School Econ., 1999.

Banerjee, Abhijit V., and Ghatak, Maitreesh. "Empowerment and Efficiency: The Economics of Tenancy Reform." Manuscript. Cambridge: Massachusetts Inst. Tech. and Harvard Univ., 1996.

Bardhan, Pranab K. "Variations in Extent and Forms of Agricultural Tenancy: Analysis of Indian Data across Regions and over Time." 2 pts. *Econ. and Polit. Weekly* 11 (September 11, 1976): 1505–12; (September 18, 1976): 1541–46.

Bardhan, Pranab K.; Ghatak, Maitreesh; and Karaivanov, Alexander. "Inequality, Market Imperfections, and the Voluntary Provision of Collective Goods." Manuscript. Berkeley: Univ. California; Chicago: Univ. Chicago, 2002.

Bardhan, Pranab K., and Rudra, Ashok. "Terms and Conditions of Sharecropping Contracts: An Analysis of Village Survey Data in India." In *Land, Labor, and Rural Poverty: Essays in Development Economics,* by Pranab K. Bardhan. New York: Columbia Univ. Press, 1984.

Besley, Timothy. "Property Rights and Investment Incentives: Theory and Evidence from Ghana." *J.P.E.* 103 (October 1995): 903–37.

Bhaumik, Sankar Kumar. *Tenancy Relations and Agrarian Development: A Study of West Bengal.* New Delhi: Sage Pubs., 1993.

Binswanger, Hans P.; Deininger, K.; and Feder, G. "Power, Distortions, Revolt and Reform in Agricultural Land Relations." In *Handbook of Development Economics,* vol. 3B, edited by Jere Behrman and T. N. Srinivasan. Amsterdam: North-Holland, 1995.

Boyce, James K. *Agrarian Impasse in Bengal: Institutional Constraints to Technological Change.* Oxford: Oxford Univ. Press, 1987.

Chadha, G. K., and Bhaumik, S. K. "Changing Tenancy Relations in West Bengal: Popular Notions, Grassroot Realities." 2 pts. *Econ. and Polit. Weekly* 27 (May 9, 1992): 1009–17; (May 16, 1992): 1089–98.

Chattopadhyay, B., et al. "Tenancy Reform, the Power Structure and the Role of the Administration: An Evaluation of Operation Barga." *Ecoscience CRESSIDA Transactions* 3, no. 2 (1984): 1–98.

Das, Prosad K. *The Monsoons.* 3d ed. New Delhi: Nat. Book Trust, 1995.

Datta, P. K. "Statistics of Bargadars and Extent of Barga Cultivation in West Bengal: An Analytical Study." West Bengal: Government, Directorate Land Records and Survey, 1981.

Dutta, Bhaskar; Ray, Debraj; and Sengupta, Kunal. "Contracts with Eviction in

Infinitely Repeated Principal-Agent Relationships." In *The Economic Theory of Agrarian Institutions,* edited by Pranab K. Bardhan. Oxford: Clarendon, 1989.

Eswaran, Mukesh, and Kotwal, Ashok. "A Theory of Contractual Structure in Agriculture." *A.E.R.* 75 (June 1985): 352–67.

Fudenberg, Drew, and Tirole, Jean. *Game Theory.* Cambridge, Mass.: MIT Press, 1991.

Gazdar, Haris, and Sengupta, Sunil. "Agricultural Growth and Recent Trends in Well-Being in Rural West Bengal." In *Sonar Bangla? Agricultural Growth and Agrarian Change in West Bengal and Bangladesh,* edited by Ben Rogaly, Barbara Harriss-White, and Sugata Bose. New Delhi: Sage, 1999.

Ghatak, Maitreesh; Morelli, Massimo; and Sjöström, Tomas. "Occupational Choice and Dynamic Incentives." *Rev. Econ. Studies* 68 (October 2001): 781–810.

Ghosh, Ratan. "Agrarian Programme of Left Front Government." *Econ. and Polit. Weekly: Rev. Agriculture* 16 (June 1981): A-49–A-55.

Harriss, John. "What Is Happening in Rural West Bengal? Agrarian Reform, Growth and Distribution." *Econ. and Polit. Weekly* 28 (June 12, 1993): 1237–47.

Jeon, Yoong-Deok, and Kim, Young-Yong. "Land Reform, Income Redistribution, and Agricultural Production in Korea." *Econ. Development and Cultural Change* 48 (January 2000): 253–68.

Johnson, D. Gale. "Resource Allocation under Share Contracts." *J.P.E.* 58 (April 1950): 111–23.

Kohli, Atul. *The State and Poverty in India: The Politics of Reform.* Cambridge: Cambridge Univ. Press, 1987.

Laffont, Jean-Jacques, and Matoussi, Mohamed Salah. "Moral Hazard, Financial Constraints and Sharecropping in El Oulja." *Rev. Econ. Studies* 62 (July 1995): 381–99.

Laxminarayan, H., and Tyagi, S. S. "Tenancy: Extent and Inter-state Variations." *Econ. and Polit. Weekly* 12 (May 28, 1977): 880–83.

"Left Gets It Right." *Economist* 328 (July 31, 1993): 33.

Lieten, Georges K. *Continuity and Change in Rural West Bengal.* New Delhi: Sage, 1992.

Lin, Justin Yifu. "Rural Reforms and Agricultural Growth in China." *A.E.R.* 82 (March 1992): 34–51.

Mookherjee, Dilip. "Informational Rents and Property Rights in Land." In *Property Relations, Incentives, and Welfare,* edited by John E. Roemer. London: Macmillan (for Internat. Econ. Assoc.), 1997.

Mookherjee, Dilip, and Ray, Debraj. "Contractual Structure and Wealth Accumulation." Manuscript. Boston: Boston Univ.; New York: New York Univ., 2000.

Rawal, Vikas. "Agrarian Reform and Land Markets: A Study of Land Transactions in Two Villages of West Bengal, 1977–1995." *Econ. Development and Cultural Change* 49 (April 2001): 611–29.

Saha, Anamitra, and Swaminathan, Madhura. "Agricultural Growth in West Bengal in the 1980s: A Disaggregation by Districts and Crops." *Econ. and Polit. Weekly: Rev. Agriculture* 29 (March 26, 1994): A-2–A-11.

Sengupta, Sunil, and Gazdar, Haris. "Agrarian Politics and Rural Development in West Bengal." In *Indian Development: Selected Regional Perspectives,* edited by Jean Dreze and Amartya Sen. New Delhi: Oxford Univ. Press, 1997.

Shaban, Radwan Ali. "Testing between Competing Models of Sharecropping." *J.P.E.* 95 (October 1987): 893–920.

Stiglitz, Joseph E. "Incentives and Risk Sharing in Sharecropping." *Rev. Econ. Studies* 41 (April 1974): 219–55.

World Bank. *The East Asian Miracle: Economic Growth and Public Policy*. Policy Research Report. New York: Oxford Univ. Press, 1993.