Internal migration and the spread of long-term impacts of historical immigration in Brazil*

Eduardo Cenci† Daniel A. F. Lopes‡ Leonardo M. Monasterio‡

August 10, 2021

Abstract

Immigration shocks in the 1850–1960 period left long-lasting positive impacts in southern Brazil. Yet, little is known about how these benefits spread to other parts of the country in the following decades. We use a surname-based classification of ancestries to identify descendants of immigrants and investigate the spreading of gains from historical immigration in Brazil. We find that the concentration of descendants of historical immigrants in municipalities in northern and central Brazil is positively associated with several indicators of economic development today, in particular with higher wages. Leveraging individual-level information from linked employer-employee data in which we observe both the individual's wage and ancestry, we find a wage premium of approximately 12% for descendants and positive spillovers between ancestry groups. One additional percentage point in the concentration of descendants in a municipality corresponds to a wage increase of 1% for descendants and 2% for non-descendants. Our results are in accord with a model in which descendants and non-descendants have complementary skills in the production function of the firms, particularly those in the agricultural sector.

Keywords: historical immigration, internal migration, descendants, surnames

JEL codes: J15, J61, 015, R23, N36

*The authors thank Bradford Barham, Bruno Barsanetti, Paul Dower, Aguinaldo Maciente, Ana Paula Melo, Laura Schechter, Jeffrey Smith, Emilia Tjernström, and Jeffrey Williamson for their comments at different stages in the development of this paper. We also thank seminar participants at the University of Wisconsin-Madison and conference participants at the 2020 NEUDC (Hanover, NH), the 2020 WEAI Graduate Student Workshop (Denver, CO), the H2D2 Research Day 2020 (Ann Arbor, MI), the 41st SBE Meeting (São Paulo, Brazil), the 2019 LACEA Annual Meeting (Puebla, Mexico), the XVII ABER Meeting (Rio de Janeiro, Brazil), the 2019 AAEA Annual Meeting (Atlanta, GA), and at the 2018 IDEAS Summer School in Development Economics (Prato, Italy). We acknowledge support from Brazil's Institute for Applied Economic Research (IPEA) where a significant part of this project was developed. Any errors are our own.

†University of Wisconsin-Madison

‡National School of Public Administration (ENAP)

1 Introduction

Historical events have lasting effects on present economic development (Nunn, 2009; Spolaore and Wacziarg, 2013). A common theme in the literature is the long-lasting impacts of migration from Eurasia to different parts of the world, from the beginning of the 16th century in various colonizing incursions (Acemoglu, Johnson and Robinson, 2001) to the so-called Age of Mass Migration in the late 19th and early 20th centuries (Hatton and Williamson, 1998). Many studies credit the long-term persistence of economic impacts caused by historical events to human capital, brought to the receiving countries by historical immigrants and transmitted across generations.[1] Indeed, one can understand historical immigration events as human capital transfers from one place to another. Under this view, we can rationalize impacts ensuing from immigration shocks as the effects of changes in the stock and the distribution of human capital in the receiving countries.

Human capital moves. As immigrants move within their new destinations, the human capital brought by them also moves. It spreads over time—transmitted across generations within and between families—and it spreads over space, as immigrants and their descendants move within the receiving country. We study the internal migration of descendants of historical immigrants in Brazil from the South and Southeast regions where they first settled to other regions in the country, and we investigate the effects that arise from the spreading of human capital that followed this movement of people. Specifically, we use a surname-based classification of ancestries to identify descendants of historical immigrants in Brazilian labor markets today and analyze how the concentration of these descendants affects indicators of economic development in these markets.

Migrants played a central role in two events that shaped Brazil's economy. First, about five million immigrants came to Brazil between 1850 and 1960 to work as laborers in the nascent industries of the Southeast and farmers in the fertile lands of the South. Over half of them were of Italian, German, Syrian-Lebanese, Japanese, and other origins (IBGE, 2007). These immigrants changed the profile of the Brazilian population, historically tied to the countries of the Iberian Peninsula (Portugal and Spain), and their colonies. Potentially, the arrival of all these "non-Iberian" immigrants also changed the stock and the distribution of human capital in the locations in which they initially settled (we call these initial locations the "injection points" of historical immigration in Brazil).[2] Second, starting around 1960, millions of internal migrants left the coastal and southern parts of Brazil towards its interior joined in a "March to the West" (Pellegrina and Sotelo,

[1] See, for example, Borjas (1992) and Rocha, Ferraz and Soares (2017).

[2] Carvalho and Monasterio (2012), Ehrl and Monasterio (2017), and Rocha et al. (2017) make similar arguments and show the long-lasting impacts these migrants caused around their injection points.

2019). This internal migration was incentivized by the construction of roads and a new capital city in the interior of the country (Morten and Oliveira, 2016), by land grants and colonization schemes (Jepson, 2006*a*; Hosono and Hongo, 2012), and by the development of new agricultural technologies that enabled grain production in previously unproductive tropical latitudes (Bragança, Assunção and Ferraz, 2021). Many internal migrants involved in this process, particularly those who left the southern regions of Brazil trailing the expansion of the agricultural frontier, were descendants of non-Iberian immigrants. Our study ties these two migration events together, analyzing the spread of descendants of historical immigrants in Brazil and how they impacted the labor markets where they concentrate today. In particular, we look at the relationship between the concentration of descendants and wages.

Focusing on wages allows us to carry out an analysis at the individual level, which is an important advantage of our study. Aggregate outcomes like average wages can mix composition and spillovers effects. For example, if descendants bring higher human capital to a labor market and are paid higher wages, they will raise the average wage of that market by increasing the share of high relative to low wages, not necessarily by raising wages for all workers in that market. On the other hand, if descendants bring human capital that is complementary to the existing human capital of non-descendants, there might be positive spillovers that affect all workers. Using individual-level outcomes like wages, therefore, allows us to separate composition and spillover effects. In other words, we can look at effects on people, not places. Additionally, in an individual-level analysis we can explore several heterogeneities in our results including (but not limited to) different effects the concentration of descendants can have on the non-descendants and on the descendants themselves. Finally, we can leverage the richness of information available in our data to include not only indicators for ancestry in our regressions but also several controls and fixed effects at the individual, firm, and municipality levels.

To circumvent the lack of information on an individual's ancestry, we use her surname to proxy for descendant status. Given the history of the country, any person bearing a non-Iberian (neither Portuguese nor Spanish) surname in Brazil is likely a descendant from immigrants that arrived around 1850–1960 in the South and Southeast regions, not a modern immigrant nor someone descending from earlier colonizers, local indigenous people, or former slaves. Therefore, using the surname-based classification developed by Monasterio (2017) and a linked employer-employee dataset in which we observe the name of every worker formally employed in Brazil in the years between 2004 and 2017, we are able to assign an ancestry to every individual observed in our data.

We find a wage premium of approximately 12% for descendants, which suggests they are either more productive than non-descendants or are favorably discriminated. We

also calculate the concentration of descendants in the workforce of Brazilian municipalities (our analogs for labor markets) and find evidence in favor of positive spillovers between ancestry groups. In our preferred specification, one additional percentage point in the concentration of descendants corresponds to a wage increase of 1% to descendants and of 2% for non-descendants. Similar results hold in different specifications, including one that uses an instrumental variables strategy that combines distance to the injection points with terrain ruggedness to predict the concentration of descendants in the municipalities. When exploring heterogeneity, we find that our results are stronger for low-skilled men and those working in the agricultural sector.

We use a simple model, borrowed from studies that investigate imperfect substitutions between domestic and foreign-born workers in the US (Borjas, Grogger and Hanson, 2008; Ottaviano and Peri, 2012) to discuss the mechanisms behind our results. Firms in this model can combine labor from descendants and non-descendants in CES production functions with some degree of substitutability (or complementarity). We assume that descendants and non-descendants carry the skills, knowledge, and cultural traits of those who were once foreign-born or domestic workers, respectively. With that, we can rationalize the wage premium of descendants (descendant labor may be more productive) and the wage spillovers between ancestry groups (there can be complementarities between descendant and non-descendant labor). The model also allows for alternative but not exclusive explanations that operate via capital and firm technology.

We show that the persistent economic impacts of historical events—of an immigration shock, in particular—can spread due to the mobility of human capital. Other studies have documented a link between historical events and persistent economic impacts in Latin America (Dell, 2010; Droller, 2017; Valencia Caicedo, 2018), and, in the case of historical immigration, in Brazil (Carvalho and Monasterio, 2012; Ehrl and Monasterio, 2017; Rocha et al., 2017; Vigna and Rocha, 2019). Existing evidence on long-lasting positive impacts of historical immigration, however, is usually constrained to locations around immigrants' injection points.[3] In focusing on how the descendants of historical immigrants spread their human capital in a receiving country and affect its labor markets, our study broadens the current knowledge on the long-term impacts of historical events.

Our paper makes several additional contributions. First, we add to the literature on the long-term impacts of the Age of Mass Migration. Hatton and Williamson (1998) note that most work on the consequences of this global historical event focus on the US.

[3]Droller (2017) investigates the spread of Europeans in the Argentinian pampas but does not focus on how the immigrants and their descendants moved to and affected other parts of Argentina. Ehrl and Monasterio (2017) looks at descendants of historical immigrants in Brazil but their analysis is restricted to the state of Rio Grande do Sul and, therefore, to municipalities near the injection points of historical immigration there.

Less is known about the consequences of the Age of Mass Migration in Latin America, the destination of millions of those immigrants (Sánchez-Alonso, 2007). Second, because the internal migration flows that spread the descendants of historical immigrants in Brazil are closely tied to the expansion of modern mechanized agriculture in the country, we indirectly contribute to the literature on the causes and consequences of the expansion of the Brazilian agricultural frontier (Bustos, Caprettini and Ponticelli, 2016; Bustos, Garber and Ponticelli, 2017; Bragança, 2018; Pellegrina, 2020; Bragança et al., 2021). Anecdotal accounts often mention how instrumental the descendants were in shaping the modern profile of agriculture and the economies along the frontier (Wagner and Bernardi, 1995; Rezende, 2002; Alves, 2005, 2016). However, no study directly identifies the descendants of historical migrants and attempts to analyze the impacts of their human capital in these regions as we do. Third, we add to the literature that investigates imperfect substitution—the converse of complementarity—between domestic and immigrant work (Borjas and Katz, 2007; Borjas et al., 2008; Ottaviano and Peri, 2012), extending the analysis to the descendants of these workers. In doing so, we connect also to the literature on immigrants' assimilation in other countries (Abramitzky, Boustan and Eriksson, 2014; Pérez, 2019).

The rest of the paper is organized as follows. Section 2 provides background information on immigration, ancestries, and surnames in Brazil. This section explains the surname-based classification of ancestries, discusses our samples of municipalities, and presents a brief account of the expansion of the agricultural frontier in Brazil. Section 3 presents the data used in the study and some descriptive statistics. Section 4 presents our theoretical framework and discusses potential outcomes of our analysis. Section 5 presents our empirical strategy, discusses the identification concerns, and proposes an instrument to address these concerns. Section 6 presents the results of our main and complementary analysis. Section 7 builds on our theoretical framework to discuss the possible mechanisms behind our results. We close the paper with a series of robustness checks in Section 8 and the conclusion in Section 9.

2 Background information

2.1 Immigration, surnames, and ancestries in Brazil

The colonial ties of Brazil to Portugal and its proximity to the Spanish colonies in South America resulted in a regular flow of colonizers and immigrants coming from the Iberian Peninsula. This process gave Portuguese and/or Spanish ancestries—and surnames—to

most of Brazilian whites.[4] At the same time, Brazil's historical (and many times forceful) integration of former slaves and Amerindians into its national population, left the descendants of those groups with Iberian surnames as well. As a result, not only Brazilian whites but also virtually all Brazilian blacks, mixed, and natives have Portuguese and/or Spanish surnames today.

In the late 19th and early 20th centuries, state-sponsored settlements (*colônias*) attracted a large number of non-Iberian immigrants to the South and Southeast regions of Brazil.[5] There are historical records of non-Iberian settlements in the states of Minas Gerais (Monteiro, 1973), Espírito Santo (Franceschetto, 2014), and São Paulo (Rocha et al., 2017), in the Southeast region. There are also records for all the states in the South region: Paraná (Nishikawa, 2015), Santa Catarina (Piazza, 1983), and Rio Grande do Sul (Carvalho and Monasterio, 2012).

Non-Iberian immigrants came to Brazil first from Germany, then from Italy, and finally from Japan. Smaller groups came also from Syria, Lebanon, Turkey, Poland, Russia and other countries. Immigration from Portugal and Spain, which happened throughout the history of Brazil and is more widespread on its territory, continued during that period. International immigration was encouraged by the Brazilian government in the belief that bringing in Europeans and other foreign settlers was an efficient way to develop the interior of the country and to replace the slave labor force after slavery was abolished in Brazil.[6] Immigration intensified from 1850 through the late 1950s (with an expected decline during World War II) but declined sharply after 1960. The absence of new substantial inflows and the natural aging of the immigrant population combined to make the current share of foreign-born people in Brazil negligible (around 0.23% in 2010, according to the national census). Figure 1 below illustrates this sharp increase and later decline of international immigration in Brazil, from 1850 to 1960.

For some periods in 1850–1960, we have information on immigrants' country of origin.[7] Table 1 shows the breakdown of immigrants by country of origin for 1884–1939, one of the periods of highest intensity of international immigration in Brazil. More than half of those for which we can identify a country of origin are of non-Iberian ancestry. If we exclude those in the "others" category, immigrants from non-Iberian countries account for 51% of the total (if we group "others" with the rest of the non-Iberian countries, their share rises to 57%).

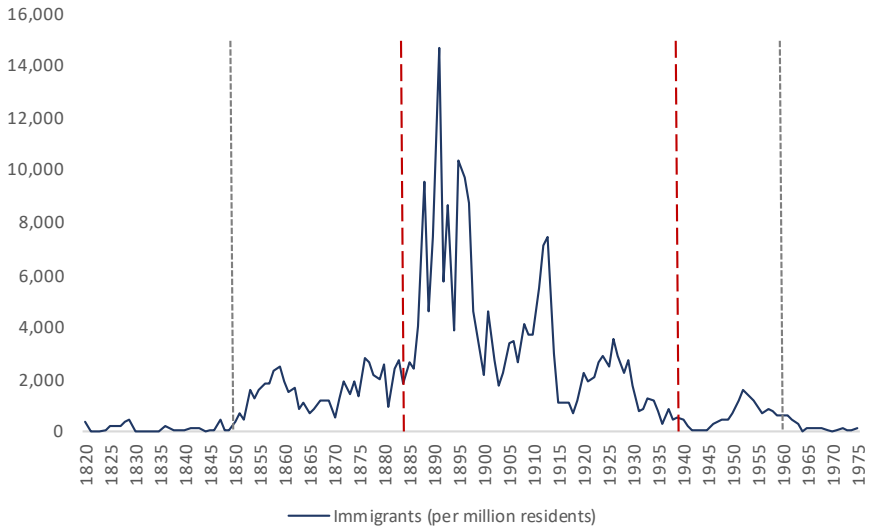
[4]For the purposes of this study, we define "ancestry" as the country of origin of one's ancestors.

[5]Informative discussions of the causes, context, and consequences of the state-sponsored immigration in Brazil can be found in Carvalho and Monasterio (2012), Ehrl and Monasterio (2017), and Rocha et al. (2017).

[6]The abolition of slavery in Brazil was a gradual process started in 1850 and finalized only in 1888.

[7]Estimates of the resident population, number of immigrants, and their country of origin breakdown are all obtained from IBGE (2007).

Figure 1: International immigration in Brazil, 1820–1975



Note: Immigration and resident numbers from IBGE (2007). The gray lines delimit the 1850–1960 period, whereas the red lines delimit the period for which we have more information on immigrants' country of origin (1884–1939).

Table 1: Immigration to Brazil by country of origin, 1884–1839

Country of origin	Immigrants (1,000)	Share	Ancestry group	Group share
Italian	1,412	34%	Non-Iberian	45%
Japanese	186	4%		
German	171	4%		
Syrian-Lebanese	99	2%		
Portuguese	1,204	29%	Iberian	43%
Spanish	582	14%		
Others	505	12%	Undefined	12%
Total	4,159	100%		100%

Notes: Data from IBGE (2007).

Brazil's historical background, combined with this intense (but later interrupted) experience of international immigration in its post-slavery period, generated a unique landscape of surnames and ancestries in the country. Because the fraction of foreign-born in the country today is close to zero and because most Brazilians have Iberian surnames, a person bearing a non-Iberian surname in Brazil has a high probability of having descended from immigrants who arrived in the country between 1850 and 1960. Therefore, the surname-based classification of ancestries employed in our analysis serves well as a proxy for the presence of descendants of historical immigrants in the current

population of Brazilian municipalities.

We close this sub-section with a couple of clarifications regarding the terminology we use in this paper. First, we note that state-sponsored settlements were not the only points of attraction of non-Iberian immigrants in Brazil. Many immigrants came on their own or following private enterprises. However, the state-sponsored settlements were the destination for a significant portion of immigrants and were often located in the same regions where private settlements formed. Therefore, throughout the paper, we use the terms "injection points" and "settlements" interchangeably.

Second, throughout the paper, we refer to the "concentration of descendants" in the population of a region or municipality when, in fact, what we observe and measure is the concentration of non-Iberian surnames in that population. The concentration of non-Iberians surnames (observed variable) is the proxy we use for the concentration of descendants (variable of interest). Also, in this paper, for "descendants" we mean "the descendants of non-Iberian immigrants from the 1850–1960 period." Likewise, by referring to the complement of this group in the population as "locals", we mean "the local population, their descendants, and the descendants of immigrants from Portugal, Spain, and other countries in Latin America."[8] We use the terms "locals" and "non-descendants" interchangeably.

2.2 Surname-based classification of ancestries

Since our classification of ancestries is based on surnames, a brief discussion of the social and legal norms surrounding surnames in Brazil follows. Children in Brazil usually receive two surnames: first the mother's second surname, followed by the father's second surname. Because only the second surname of each parent is passed on, and because the father's surname comes last, effectively, only the father's surname survives. As for name changes after marriage, Brazilian civil law required a married woman to adopt her husband's second surname until 1977. After that, adoption became optional, and in 2002, adoption of the spouse's surname became optional for both men and women. In most cases, when adopted, the husband's surname becomes the second.[9] In our study, we consider only the second surname of each person. Therefore, our way of tracking the

[8]Those coming from Africa to Brazil voluntarily, as free individuals in the past and today, most likely kept their surnames. They and their descendants count as Iberians or non-Iberians. Those who came forcibly in the past, as enslaved individuals, had to adopt Iberian surnames and are not considered immigrants in this study. Their descendants count as Iberians (unless they gained a non-Iberian surname through marriage).

[9]Some people in Brazil have three or more surnames. Nevertheless, here we use the term "second" to refer to the surname that comes last in a person's full name to avoid confusion with the term "last name", commonly used in English to denote one's surname.

offspring of historical immigrants effectively ignores maternal lineage.

The algorithm used in the classification allows for considering more than one surname, creating classifications of mixed ancestries (when each surname comes from a different group) or homogeneous ancestries (when both surnames are from the same group). However, there are challenges to implementing a surname classification that uses two or more surnames with our data.[10] Moreover, we do not expect such refined classifications to improve the approximation of our measure to the true concentration of individuals of non-Iberian ancestry in the municipalities of our study samples for two reasons. First, the group of non-Iberian ancestries is large enough to accommodate many cases of mixed ancestries (e.g., German-Italian). Second, we expect neither a consistent pattern in the order of Iberian and non-Iberian surnames when a person has both, nor a correlation between this order and the concentration of descendants in a given municipality. Therefore, the measurement error arising from assigning Iberian or non-Iberian ancestry to workers with mixed surnames is probably classical in our setting (any measurement error arising from the choice of using only the second surname in our classification will bias our results towards zero).

Simply put, the surname-based classification of ancestries used in this study follows four steps: (1) collect the second surname of all workers in the sample in a given municipality for a given year; (2) match these surnames to historical sources where surnames are accompanied by countries of origin; (3) link each unique surname to a country of origin (e.g., Italy) or region of origin (e.g., Eastern Europe); and (4) attribute ancestry of the historical source to current observations based on this surname-origin matching process.[11] In a small number of cases, we refine the classification using information on race (native Brazilian surnames, for example, can be misclassified as Japanese).[12]

The classification yields, for each paired municipality and year in our sample, the total number of workers of each ancestry. We group all ancestries into two groups: Iberians and non-Iberians. To obtain a proxy for the concentration of descendants in the population of the municipalities m in our sample in each year t (denoted by C_{mt}^{Desc}), we simply divide the number of workers of non-Iberian ancestry by the total number of workers in the

[10]Such a refined classification is used by Lopes, Silva and Monasterio (2017), who, unlike us, have detailed information on an individual's parents. We cannot trace maternal and paternal lineages n the RAIS data and correctly identify cases of homogeneous or heterogeneous ancestries.

[11]Refer to Monasterio (2017) for a thorough explanation of the algorithm and the data requirements, and refer to Ehrl and Monasterio (2017) and Lopes et al. (2017) for a description of the updated versions of the algorithm, which are similar to the one used in this study.

[12]IBGE, the Brazilian Statistical Office, uses the term "color/race", usually divided into five categories: black, white, mixed ("*pardo*"), yellow (East Asian), and indigenous. The "yellow" category is seldom chosen by Asian-Brazilians, who often choose the mixed category instead. In this study, we use the term "race" as a way to follow the standard in the literature.

sample for that municipality-year pair. The resulting measure is then multiplied by 100 to facilitate interpretation of coefficients in descriptive statistics and regressions.

We do not use race to define ancestries nor do we aim to investigate racial wage disparities in Brazil.[13] We acknowledge, however, that the share of whites is higher in the group of descendants: 61% of the descendants in our sample in 2010 are identified as whites in the data, while this proportion is only 39% for the locals. Therefore, in most of our analyses we control for race, in addition to other relevant controls discussed in detail in Section 5.1.

2.3 Study regions (municipality samples)

Several states in Brazil have documented injection points of historical non-Iberian immigration. All the states in the South and the Southeast regions, with the exception of Rio de Janeiro, had a significant number of state-sponsored settlements that served as points of attraction and destination to non-Iberian immigrants coming from different parts of the world to Brazil between 1850 and 1960.[14]

In this study, we investigate how internal migration spread the gains from historical immigration from the southern to the northern and central regions of Brazil.[15] Thus, it is convenient that we separate the municipalities in our sample into two samples. The first, which we call the "injection sample," includes all municipalities in states with documented injection points of historical immigration in Brazil. (We also add to this sample the municipalities in the state of Rio de Janeiro due to its location in the Southeast region and its past as home of the national capital until 1960.) Municipalities in this sample may have been impacted by the arrival of immigrants in the past, by accommodations made to receive them (e.g., investments in infrastructure and land redistribution), and by the spreading of their descendants around the injection points. The municipalities in the states of the other three regions form what we call the "spread sample." Many of these municipalities were impacted by the March to the West that brought millions of internal migrants (many of them descendants) to the interior of the country. Municipalities in the spread sample may have been impacted by the spreading of descendants of immigrants in Brazil (but not by the injection points directly).

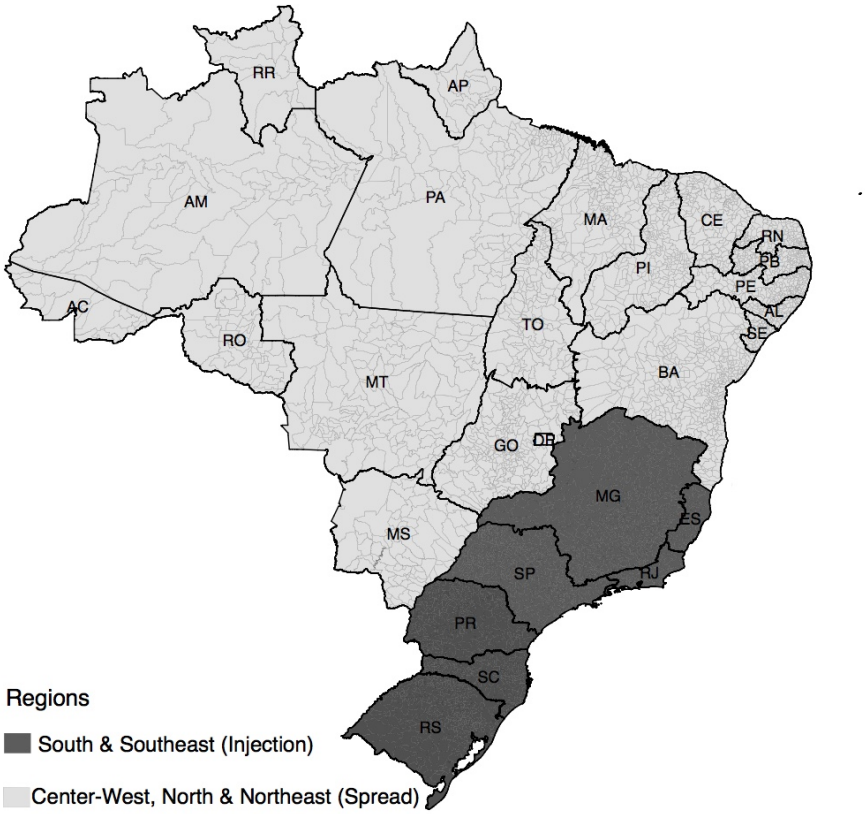
[13] Gerard, Lagos, Severnini and Card (2018) does such an investigation, using the same data we use in this study (RAIS), while many works in the sociology literature, like Andrews (1991) and dos Santos (2002), discuss the connection between racial wage differences and historical immigration in post-slavery Brazil.

[14] There are occasional historical records of non-Iberian settlements in states of the other regions in Brazil (e.g. Japanese settlements in the states of Amazonas and Pará in the North region). However, the records of settlements in these regions are rare compared to the other regions, and the settlements are smaller.

[15] Precisely, from the South and Southeast to the North, Northeast, and Center-West regions.

Figure 2 shows the two resulting samples in the map of Brazil. We use the union of these two samples of municipalities in some exercises and refer to it as the "national sample". In all samples used in our study, we remove state capitals thus excluding also the Federal District (DF).[16]

Figure 2: Samples of municipalities



Notes: Black lines denote state boundaries and thin gray lines denote municipality boundaries. The Spread sample encompasses all municipalities in the states of Rondônia (RO), Acre (AC), Amazonas (AM), Roraima (RR), Pará (PA), Amapá (AP), and Tocantins (TO) in the North region; the states of Maranhão (MA), Piauí (PI), Ceará (CE), Rio Grande do Norte (RN), Paraíba (PB), Pernambuco (PE), Alagoas (AL), Sergipe (SE), and Bahia (BA) in the Northeast region; and the states of Mato Grosso do Sul (MS), Mato Grosso (MT), and Goiás (GO) in the Center-West region. The municipalities in the states of Rio Grande do Sul (RS), Santa Catarina, and Paraná (PR) in the South region, and the states of São Paulo (SP), Espírito Santo (ES), and Minas Gerais (MG) in the Southeast region—all of which have a significant and well documented presence of injection points of non-Iberian immigration—plus the state of Rio de Janeiro (RJ), form the Injection sample.

For descriptive statistics and empirical analyses, we exclude municipalities with less than five individual observations or missing values for the concentration of descendants. Many municipalities in our sample have a small number of individual observations.[17] This is due to the small size of these municipalities and to the fact that our study

[16]We also leave out Fernando de Noronha, a small municipality on an island far off the Brazilian coast.

[17]Like the concentration of descendants, the distribution of the number of individual observations is skewed to the left.

region includes some of the poorest regions of Brazil, where informality is higher than the national average (the degree of formalization in the Spread region is 31.3% compared to 43.4% in the National sample—both excluding state capitals). Because our main dataset covers only individuals employed in the formal sector, municipalities with a small population and a large share of workers in the informal sector are bound to have few observations in our data. In Section 8 we show that our results are robust to different exclusion criteria based on the number of individual observations in a given municipality.

In the robustness checks in Section 8 we also show results for different samples of municipalities, including one that excludes the Northeast region from the Spread sample. We also, show results for a version of the Spread sample that only includes municipalities that match the definition of agricultural frontier in Bustos et al. (2016).

2.4 The agricultural frontier in Brazil

We close this section presenting a brief account of the expansion of the agricultural frontier in Brazil. This account is useful to understand the migration of descendants from the South and Southeast to the other regions of Brazil and the rationale behind our instrument discussed in Section 5.

Following the adaptation of soybean varieties to tropical climates, migrants from the South and Southeast of Brazil started to settle in the Center-West around 1960–1970 and in parts of the North and Northeast after 1990, triggering the expansion of the agricultural frontier in the country. This expansion of agricultural production had implications that go beyond the development of the agricultural sector in Brazil. The development of the frontier integrated local markets, spread modern agricultural technologies, induced migration, and changed the land use and the economic structure of the region and the country (Bustos et al., 2016, 2017; Bragança et al., 2021).

The low population density and the abundance of (mostly flat) farmland in the Brazilian Cerrado, the savannah-like biome that dominates the agricultural frontier, combined to result in low land prices that attracted farmers from the South and Southeast regions of the country (Rezende, 2002). This process was further stimulated by private colonization companies, farmers cooperatives, land reform initiatives, and rural development programs implemented by the national government (Jepson, 2006*a*,*b*; Hosono and Hongo, 2012). The frontier continued to expand in recent decades, stimulated by the arrival of new technologies that impact agricultural production, labor markets, and internal migration (Bustos et al., 2016; Bragança, 2018).

Because internal migrants often came to the frontier from the same regions that had received non-Iberian immigrants in the late 19th and early 20th centuries, many of them

were direct descendants of those historical immigrants (Alves, 2005). Plenty of anecdotal accounts (Wagner and Bernardi, 1995; Santos, 2008) and the high incidence of non-Iberian surnames in a region so distant from the original injection points (Monasterio, 2017) suggest that this was, indeed, the case. Farmers who settled on the agricultural frontier many times came from former non-Iberian colonies in the South and the Southeast, where the extant tradition in soybean cultivation and association in cooperatives matched the definition of modern agriculture desired by the Brazilian government for the region at the time (Hosono and Hongo, 2012; Alves, 2016).[18]

The expansion of the agricultural frontier in Brazil, thus, helps explain why we can find descendants of non-Iberians immigrants so far from the original injection points today. Moreover, this process of agricultural development accompanied by strong internal migration motivates the construction of an instrument that relies on the interplay between the availability of flat farmland—well suited for the mechanized modern agriculture sought after by the descendants coming from the South region, in particular—and the distance to the non-Iberian settlements. We explain this instrument and its rationale in more detail in Section 5.2.

3 Data

3.1 Data sources

The main data source used in this study is the *Relação Anual de Informações Sociais* (RAIS), a report of all labor contracts that employers in Brazil are required to file every December in order to comply with labor regulations. These reports form a database used by the national government to administer unemployment benefits and allowances for low-income workers and produce statistics on the formal sector. This makes RAIS a high-quality annual census of all formally employed workers in Brazil.[19] Stacked over the years, RAIS becomes a linked employer-employee dataset, a type of data that is increasingly popular in economic studies (Card, Cardoso, Heining and Kline, 2018).

The RAIS data include demographic characteristics of employees, their remuneration, and some characteristics of their jobs. The data also include characteristics of the employers like industry, size (number of employees), and the municipality in which the

[18]In fact, Alves (2016) observes that migrants "sold their smallholdings, bought larger areas, and settled them using modern agricultural techniques" and concludes that the Brazilian Cerrado is a "typical case of agricultural development promoted by farmers from more advanced agronomic culture."

[19]For examples of papers using RAIS, see Dix-Carneiro and Kovak (2017) and Gerard et al. (2018). For detailed information on RAIS data, its variables, and structure, refer to these papers' data appendices.

firm/establishment is located.[20] Using the employer's municipality and the link they have with workers, we can assign workers to municipalities. The full name of workers, from which we extract the second surname to be used in our surname-based classification of ancestries, is available in the RAIS data, as is the information on worker's race, which we use to improve the classification as discussed in section 2.2.

We complement our data with multiple sources. Municipality-level socioeconomic characteristics come from Ipeadata, the Atlas Brasil project, and IBGE. Unless noted otherwise, these variables use information from the 2010 population census. The approximate location and the year of establishment of the historical non-Iberian settlements in São Paulo and Rio Grande do Sul come from Rocha et al. (2017) and Carvalho and Monasterio (2012), respectively. Rainfall variables are constructed using data from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS). Temperature variables use data from the Global Meteorological Forcing Dataset for land surface modeling. Average elevation and the Terrain Ruggedness Index are calculated using data from the Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) from the US Geological Survey. The distance to state capitals and non-Iberian settlements uses municipalities' economic centers obtained using Google Maps. Variables on soil types use data from Embrapa Solos, while the dummies and shares for biomes use information from the MapBiomas project. We also use potential agricultural yields constructed with data from the Food and Agriculture Organization Global Agro-Ecological Zones (FAO/GAEZ) database in robustness checks.

One concern with the use of RAIS data to classify surnames is that it covers only a share of the population in a given municipality-year: the formally employed. In a country like Brazil, where informality is high, this share can be particularly small and not representative of the population of the municipalities in our sample. To amend this deficiency, we resort to a couple of other data sources that also contain information on respondents' names in Brazil: the *Cadastro Único*, the unified registry of beneficiaries of the Brazilian cash transfer program (*Bolsa Família*), and the *Base Sócios*, a record of business owners maintained by the national tax authority. For at least one year (2010), we can compare the concentration of descendants obtained using only RAIS or using these three data sources combined, which cover a larger portion of the population in the municipalities we study. The correlation between the two measures in the National sample in 2010 is positive and high ($r = 0.96$).

In figure A1 in the appendix we plot the concentration of descendants using only RAIS data on the horizontal axis, the same concentration using the extended data sources

[20] We treat each establishment as a unique firm: different establishments of the same firm, either in different municipalities or within the same municipality, are counted and treated as different firms.

on the vertical axis, and the 45° line. The results of this exercise reassure us that the concentration of descendants we calculate using only surnames of formally employed workers is a good representation of the true concentration of descendants in the population of the municipalities in our samples.

3.2 Sample restrictions (worker-year data)

The RAIS data starts in 1986 but workers' names are only available in the dataset starting in 2004. Therefore, 2004 is the initial period of our data. The final period of our study is 2017, currently the last year for which the data is available.

After restricting the sample of municipalities in our study regions, we make some restrictions to the individual-level data used in regressions and descriptive statistics (the same restrictions apply to the data used in the surname-based classification). In every year, we exclude all foreign-born, all public servants, and the military.[21] We keep only workers between 16 and 70 years old who have a valid identifier (PIS number). We also exclude those who work less than 10 hours per month and those without a positive remuneration. Because workers may change jobs in any given year, the same worker can appear more than once in each annual RAIS dataset. We keep only the last occurrence of a worker in any given year.[22] After adjusting remunerations for inflation and calculating hourly wages, we drop those workers with hourly wages above the 99.9th percentile.

Some of our analyses use only data from 2010, a census year for which several economic indicators are available at the municipality level. In other analyses, including the main wage regressions, we use the whole 2004–2017 period as repeated cross-sections to take advantage of the richness and size of our data, and to explore all the existent variation in the concentration of descendants to identify our parameter of interest. In fact, some of our regressions use the 2004–2017 data as a pseudo-panel of municipalities and also as a panel of workers, which allows us to perform different fixed-effects regressions (details in Section 5.1 and results in Section 6.1). The instrumental variables strategy, however, uses the 2004-2017 sample as pooled data (details in Section 5.2 and results in Section 6.2).

Because our main interest in this study is the spreading of the impacts of historical

[21] The exclusion of public servants and the military is done because many of those workers are registered at the capital of the state regardless of their actual workplace. Also, their remuneration follows legally established norms and is, therefore, less likely to be impacted by municipality-level characteristics like the concentration of descendants.

[22] In robustness checks, we show that our results hold in a sample that keeps multiple occurrences of the same worker, and also in a sample that keeps only those employed on December 31st—the date on which employers file their RAIS reports.

immigration in Brazil, we focus most of our analysis on the municipalities in the regions Center-West, North, and Northeast of Brazil (the Spread sample). Therefore, we extend the period of analysis from 2010 to the 14 years in the 2004–2017 period only for that sample.[23] Table 2 summarizes information regarding the size of each of our samples in terms of states, municipalities, firms, workers, and worker-year observations (for the 2004–2017 period).

Table 2: Main study samples

Sample:	Injection (South & Southeast)	Spread (Center-West, North, & Northeast)	
Period:	2010	2010	2004–2017
States	7	19	19
Municipalities	2,849	2,624	2,680
Firms	1,714,578	537,143	1,548,719
Workers	20,209,660	6,036,142	19,217,877
Worker-Year			85,324,069

Notes: We keep only one observation per worker per year and treat each establishment as a unique firm (different establishments of the same firm are counted and treated as different firms). Municipalities excluded from the Spread sample in 2010 for having less than five individual observations that cross this cutoff in any year during 2004-2017 are counted in the sample for that period. Therefore, the total number of municipalities in 2004–2017 is higher than in 2010.

3.3 Descriptive statistics

We close this section by presenting some descriptive statistics that highlight differences between our samples, and differences between descendants and locals within each sample. We also discuss the distinction between averaging some of our variables at the municipality level and at the individual level and the implication of this distinction to the interpretation of our results. We start by showing descriptive statistics for variables of interest at the municipality-level for both the Injection and the Spread samples. We then show the distribution of the concentration of descendants on the map of Brazil and also as histograms for both samples. We then repeat some descriptive statistics and the histogram of the concentration of descendants for averages over the number of individuals, which ends up placing more weight on larger municipalities. Finally, we show a table of differences in means for descendants and locals in each sample, which we will use to motivate the theoretical frameworks that comes next in Section 4.

Table 3 below shows the mean, median, and standard deviation for selected variables aggregated at the municipality level to highlight differences between our samples.

[23]When using data from 2004–2017, we adjust municipality borders to accommodate changes over the period, thus using the municipality boundaries and codes of 2004 as our main unit.

Table 3: Descriptive statistics for the municipalities in the study samples, 2010

Variable	Injection sample			Spread sample		
	Mean	Median	S.D.	Mean	Median	S.D.
Panel A: Observations, population, descendants, and income						
Number of individual observations	7,094	1,134	23,057	2,300	308	7,907
Population (1,000 residents)	29.45	8.94	75.34	23.27	13.42	38.32
Concentration of descendants (%)	16.45	9.89	17.68	2.88	1.75	3.82
Average income per capita (R$)	635.41	616.82	210.47	339.67	285.79	154.69
Average wage in the formal sector (R$)	8.33	7.90	2.28	7.71	7.01	3.11
Panel B: Demographic (shares)						
Female (%)	32.86	33.34	10.23	24.94	24.32	12.20
Age 25–54 (%)	69.92	69.96	4.81	73.66	73.53	7.46
White & Asian (%)	73.64	79.68	20.09	35.25	31.96	20.48
High school graduates (%)	42.36	42.86	14.81	44.33	43.71	19.31
College graduates (%)	4.64	4.19	2.90	4.48	3.21	5.55
Panel C: Labor characteristics (shares)						
White collar occupations (%)	17.83	17.01	8.10	20.90	18.80	12.50
Agriculture (%)	22.20	14.48	22.03	22.52	10.93	26.09
Manufacturing & construction (%)	32.89	29.72	21.87	26.38	19.58	24.31
Large firms: 100 or more employees (%)	24.13	20.56	23.94	17.57	0.00	25.45
N (municipalities)		2,849			2,624	

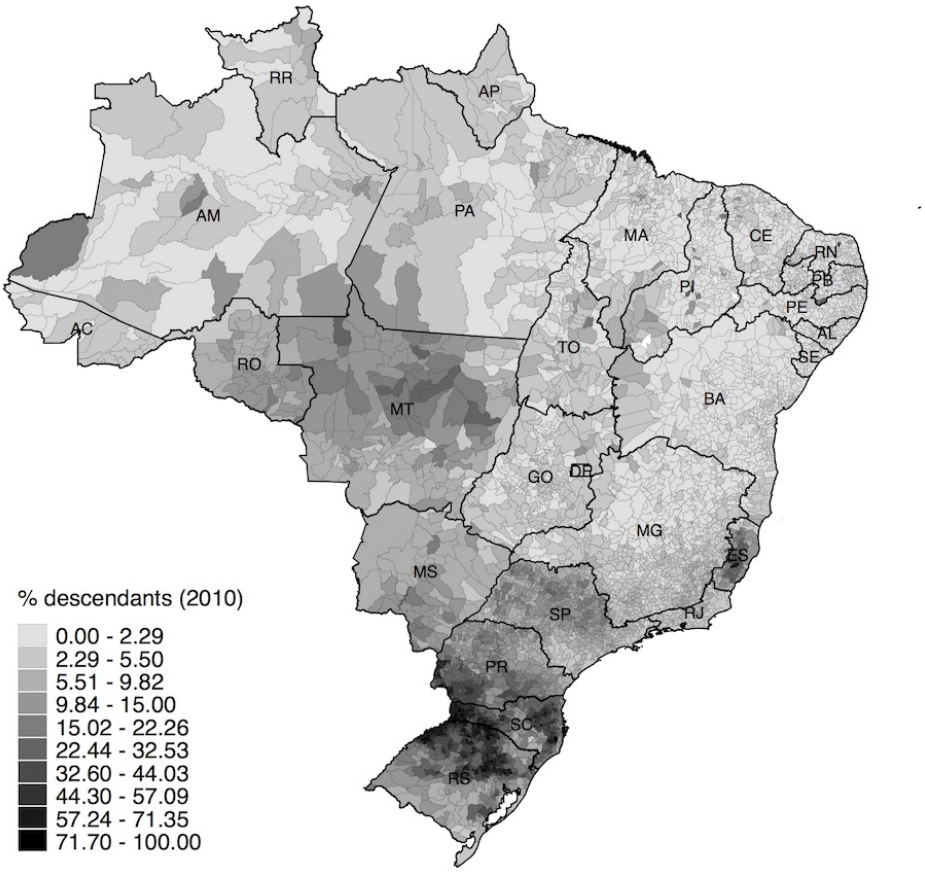
Notes: The Injection sample considers municipalities in all states of the regions South and Southeast. The Spread sample considers municipalities in the states of the regions Center-West, North, and Northeast. Both samples exclude state capitals and municipalities with less than five individual observations in the RAIS data in 2010 or with a missing value for the concentration of descendants. The number of observations refers to the number of individuals in our selected RAIS data. The concentration of descendants (% of workers with a non-Iberian surname in each municipality) and the average hourly wage (in R$ of 2017) also come from RAIS data. The figures for the total population and average income per capita (income from all sources, in R$ of 2010) come from the 2010 Brazilian population census. All other variables use data from RAIS in 2010 and come from aggregations, at the municipality level, of our selected sample of individual-level data.

We observe that while the average population size is similar (29.45 thousand residents in the Injection sample versus 23.27 thousand in the Spread sample), the average number of individual observations we have in our selected RAIS data per municipality is thrice as large in the Injection sample. This suggest that informality, unemployment, and self-employment are larger in the Spread sample, since only the formally employed appear in our data. We also observe that the average concentration of descendants is much higher in the Injection sample, which is expected since this is the region of Brazil that concentrated most of the arrivals and settlements of non-Iberian immigrants during the 1850–1960 period. We also observe much higher average income per capita and wage in the Injection sample and, going to panel B, also a much higher proportion of whites and Asians.[24]

[24]When grouping race into a binary white/non-white category, we follow Firpo and de Pieri (2018) and

The map in Figure 3 below shows the distribution of descendants in Brazil in 2010.

Figure 3: Concentration of descendants in Brazilian municipalities, 2010 (%)



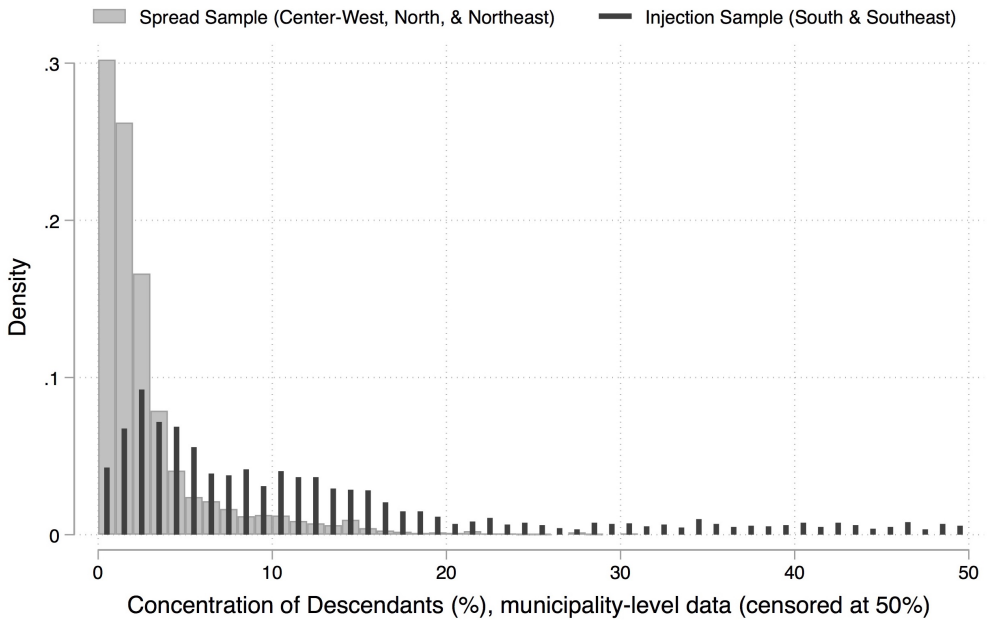
Notes: The concentration of descendants is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality in 2010. The scale in the map uses k-means smoothing to highlight variation in the concentration across space, not its intensity in each municipality.

We notice a high concentration in parts of the states of the South, in São Paulo (SP) and Espírito Santo (ES), all places in Brazil that are known for having received significant numbers of historical non-Iberian immigrants. We also notice higher concentrations in the Center-West region, particularly in Mato Grosso, a state that symbolizes the expansion and modernization of the agricultural frontier in Brazil. The scale in the map uses k-means smoothing to highlight variation in the concentration of descendants across space, not its intensity in each municipality. Therefore, we present a histogram of the distribution of the concentration of descendants in each region to clarify how different they are and how many fewer descendants we have in the Spread region.

include workers of Asian descent in the group of whites.

Figure 4 shows the distribution of the concentration of descendants across the municipalities in our two samples. The figure shows the distribution up to 50% only to improve visualization, since few municipalities in the Injection sample and no municipality in the Spread sample go beyond that level. Both distributions are skewed to the left, as suggested by the averages higher than medians seen in the third row of Table 3.

Figure 4: Distributions of the concentration of descendants in the municipalities, 2010



Notes: The concentration of descendants is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality in 2010. The Injection sample considers municipalities in all states of the regions South and Southeast. The Spread sample considers municipalities in the states of the regions Center-West, North, and Northeast. Both samples exclude state capitals and municipalities with less than five individual observations in the RAIS data in 2010 or with a missing value for the concentration of descendants. For the Injection sample, the average concentration is 16.45%, the median is 9.89%, and the standard deviation is 17.68%. For the Spread sample, the average concentration is 2.88%, the median is 1.75%, and the standard deviation is 3.82%.

All numbers presented in Table 3 and Figures 3 and 4 are aggregated at the municipality level. Our main analyses, however, are conducted at the individual level. Because the number of individual observations in each municipality varies considerably—as shown by the standard deviations of the first row in Table 3—any statistic aggregated at the individual level may differ considerably as well. For example, if the concentration of descendants is generally higher in larger municipalities (in terms of their number of individual observations), then the concentration of descendants in the municipalities, when averaged across all individuals in our sample, will be higher than the average concentration when averaged across all municipalities. This is precisely what happens in the Spread sample, where the average across individuals is higher than we had before: 3.83% versus 2.88%. The opposite happens for the Injection sample, in which small municipalities with

high concentration of descendants make the average across municipalities (16.45%) higher than the average across individuals (11.94%). These differences are meaningful when we consider marginal effects in Section 6.2.

Figure A2 in the appendix shows how the distribution of the concentration of descendants in the municipalities where each individual in our data works is distributed across these individuals. As before, we show the distribution up to 50% only to improve visualization. Like the distributions weighted by municipality shown before, these distributions weighted by individuals are skewed to the left. The mass around zero, however, is smaller, particularly for the Spread sample.

Figures 3 and 4 show a strong presence of descendants in municipalities to the west and north of Brazil and away from the sites of historical settlements and points of arrival of immigrants in 1850–1960. This corroborates the story we tell in section 2.4, that descendants of historical immigrants in Brazil spread over the country following the expansion of the agricultural frontier.

We close this section by presenting descriptive statistics that highlight the differences between descendants and non-descendants (locals) in our two main samples.[25] Table 4 shows the mean for descendants and locals and their difference for variables we use as outcomes, explanatory variables, and controls. Due to the large size of both samples, traditional t-tests show that all differences are statistically significant. Therefore, we present the normalized difference next to the differences in parentheses (Imbens and Wooldridge, 2009). A common rule of thumb is to consider a normalized difference greater than one quarter as meaningful.

In both samples, we observe in panel A that descendants work in smaller municipalities than locals, but the normalized difference suggests this difference is not as sizeable as it may appear. Not surprisingly, we also observe that descendants work in municipalities where the concentration of descendants is higher. Less obviously, is the fact that the wage of descendants is substantially higher than the wage of locals in both samples and that this difference is more pronounced in the Spread sample. We do not find differences in hours worked, which indicates that using hourly wages or total labor income as the outcome of interest should produce similar results.

In panel B we note that descendants and locals employed in the formal sector differ in several demographic characteristics. In particular, we note that descendants are more likely to be female, slightly older, much more likely to be white or Asian, and more educated whether we look at the share with a high school degree or higher, or with a college degree or higher. The differences hold with similar normalized magnitudes and

[25]Most statistics for the National sample resemble the figures we have for the Injection sample, as the latter contains more than three quarters of the individual observations in the former.

Table 4: Differences in means by ancestry, individual-level, 2010

Variable	Injection sample			Spread sample		
	Descendants	Locals	Difference	Descendants	Locals	Difference
Panel A: Labor market size, descendants, and income						
Municipality size (1,000 workers)	66.84	84.27	-17.43 (-0.13)	25.09	29.65	-4.56 (-0.10)
Concentration of descendants (%)	21.35	10.53	10.82 (0.62)	8.30	3.65	4.65 (0.61)
Log hourly wage (formal sector)	2.27	2.13	0.14 (0.16)	2.15	1.92	0.22 (0.26)
Hours worked	42.46	42.94	-0.48 (-0.08)	43.05	43.34	-0.29 (-0.06)
Panel B: Demographic characteristics						
Female	0.42	0.36	0.06 (0.08)	0.34	0.28	0.05 (0.08)
Age (years)	33.79	33.24	0.55 (0.03)	32.99	32.59	0.40 (0.03)
White & Asian	0.87	0.69	0.18 (0.31)	0.54	0.32	0.22 (0.32)
High school graduate	0.62	0.52	0.10 (0.15)	0.58	0.47	0.11 (0.15)
College graduate	0.13	0.06	0.07 (0.17)	0.12	0.04	0.07 (0.20)
Panel C: Job characteristics						
Firm tenure (months)	40.81	32.07	8.74 (0.11)	29.24	28.73	0.51 (0.01)
White collar occupation	0.20	0.13	0.07 (0.14)	0.20	0.11	0.09 (0.18)
Agriculture	0.04	0.07	-0.03 (-0.09)	0.17	0.13	0.03 (0.06)
Manufacturing & construction	0.38	0.36	0.02 (0.03)	0.28	0.36	-0.08 (-0.13)
Large firm (100 or more workers)	0.40	0.47	-0.06 (-0.09)	0.34	0.44	-0.09 (-0.13)
Number of individuals (millions)	2.62	17.59		0.23	5.80	

Notes: Normalized differences are shown in parentheses next to the differences in means. All variables come from RAIS in 2010. The number of observations, both per municipality (first row) and total (bottom row) refers to the number of individuals in our data after sample restrictions. The Injection sample considers municipalities in all states of the regions South and Southeast. The Spread sample considers municipalities in the states of the regions Center-West, North, and Northeast. Both samples exclude state capitals and municipalities with less than five individual observations in the RAIS data in 2010 or with a missing value for the concentration of descendants.

the same directions in both samples.

Finally, in panel C we look at characteristics of the jobs descendants and locals hold in our data. Descendants have longer job tenure than locals, especially in the Injection sample. They are also more likely to hold white-collar occupations in both samples, but they are less likely to work in the agricultural sector in the Injection sample, whereas they are more likely to work in that sector in the Spread sample. Descendants and locals concentrate differently in the manufacturing and construction industries in both samples. Descendants are more likely to work in manufacturing and construction in the Injection sample, but less likely to do so in the Spread sample (and the difference there is larger). The differences in Table 4 suggest that descendants have a wage premium over locals, and that, at least in the Spread sample, they have a stronger presence in the agricultural sector.

4 Theoretical framework

4.1 Motivation

Three empirical regularities arise in the descriptive statistics of Section 3.3 and are confirmed later by the regression results of Section 6. First, municipalities with a strong presence of descendants have a higher average income per capita and, in particular, higher average wages in the formal sector. Second, descendants of historical immigrants in Brazil earn a premium over the formal sector wage of locals. Third, these differences in aggregate income and individual wages are larger in the Spread sample, where the concentration of descendants in the municipalities is rather small. The individual-level analysis we use in this paper adds a fourth empirical regularity. The association between the concentration of descendants and higher wages is larger for locals than for the descendants themselves. And this association is also stronger in the Spread sample.

These empirical regularities motivate the theoretical framework we develop in this section. We first introduce a basic set-up with two types of workers—descendants and locals—and a representative firm that hires labor from these workers in each municipality. We then discuss the possibility of complementarity between the two types of labor in the production function of firms, potential constraints that limit the firm's optimization, and the worker's labor choices. We conclude with a list of possible outcomes from this stylized framework, which serve both as predictions for results and as explanations for the mechanisms behind them.

Our framework borrows elements from Borjas et al. (2008) and Ottaviano and Peri (2012), who estimate the impacts of immigrants on the wage of native workers in the US. The basic model in these papers allows both types of workers—immigrants and natives—to differ and to imperfectly substitute each other in the production process of firms. We extend this thinking to allow the descendants of immigrants and the (descendants of) locals to differ in the type of labor they provide to firms. Like Borjas et al. (2008) and Ottaviano and Peri (2012), we allow for the possibility of imperfect substitution—or, conversely, complementarity—between the two types of labor. Unlike these studies, however, we do not assume a specific functional form for the production function. Their models feature a Cobb-Douglas function with capital and labor, where the labor aggregate has a nested CES form. This way, the elasticity of substitution between immigrants and natives can be estimated within different education and experience groups. In contrast, we do not aim to estimate the elasticity of substitution in our setting. In our framework, the possibility of imperfect substitution between descendants and locals is enough to help us think about how the concentration of descendants of historical immigrants can affect

wages in Brazilian municipalities. We do not mathematize our theoretical framework beyond what is helpful for intuition nor try to estimate it. Our goal in this section is to provide intuition for our results and guidelines for our discussion of mechanisms.

4.2 Environment: labor markets, firms, and workers

Our environment is a municipality (indexed by m), which we interpret as the empirical counterpart of a labor market. Each municipality has a representative firm (also indexed by m), and a pool of workers divided into two ancestry types: descendants (D) and locals (L). Firms are characterized by their production technologies and workers are characterized by their ancestries (a factor common to all workers of an ancestry type) and outside options (varying by individual, indexed by i).

Firms are competitive and utilize only labor to produce a single consumption good $Y_m = A_m f\left(D_m, L_m; \theta, \rho\right)$. We do not impose a particular form for the function f.[26] Firms have different technology multipliers A_m and combine labor from descendants and locals with some degree of substitutability ρ. We allow one ancestry type to be more or less productive than the other. The term θ>0 in the production function captures this possibility by multiplying the descendant labor term (the equivalent term for locals is normalized to 1, and any level adjustment is subsumed in the firm's technology multiplier). A value of $\theta \geq 1$ means that descendants are more productive than locals.

Workers of both ancestry types in this economy are endowed with one unit of labor, which they supply to the local firms if the wage offered to their type is greater than their outside option. It follows from our exposition on firms that a worker's ancestry defines not only her type but also her average labor productivity (the θ multiplier). The utility function of a worker when formally employed is $u_i(w^a) = w^a - \gamma_i$ for ancestry types $a = D, L$ and value of outside option $\gamma_i > 0$, which may be unemployment, self-employment, or employment in the informal sector.

We assume that firms can observe workers' names and infer their ancestries in the same way we do when employing our surname-based classification of ancestries. We do not assume that firms necessarily treat workers of different ancestries as different

[26] One example of a production function with these characteristics and flexibility with respect to the degree of substitutability between the two types of workers is the CES function. In our setting, it would take the following form:

$$Y_m = A_m \left[\theta D_m^{\rho} + L_m^{\rho}\right]^{1/\rho}, \qquad \rho \in (-\infty, 1], \qquad \theta > 0.$$

Such a production function can be extended to include capital. In that case, we could write the production function in a Cobb-Douglas format and let the labor aggregate follow the CES form. Moreover, the labor aggregate admits a great deal of flexibility and the division of workers into many other types if we re-write it as a nested CES.

types, only that they can do so. As for the reasons why ancestry types may, in fact, represent different types of labor, there are many. First, workers of different ancestries may have inherited ethnic capital from their ancestors, who came from countries that differ in several dimensions from the origins of locals.[27] Second, once in the country, the immigrants and their descendants concentrated in particular regions where they developed institutions, skills, and knowledge specific to that region that may differ from those developed by the local population in other places. Third, ancestry in Brazil may correlate with factors that affect human capital development, like access to good quality education and discrimination in the labor market.

4.3 Equilibrium

If firms are unconstrained on how much labor of each type they can use, their profit maximization yields optimal quantities of descendant and local labor (D_m^* and L_m^*, respectively), and a corresponding wage schedule $w^a\left(D_m^*, L_m^*; \theta, \rho, A_m\right)$ for $a = D, L$. The exact form of the wage function is purposefully left undefined.[28]

In equilibrium, every worker of ancestry a with $w^a > \gamma_i$ in municipality m is formally employed by one of the firms j operating in that municipality. The worker receives a wage w^a, and the competitive firms produce Y_j^*. Free entry ensures that all firms make zero economic profits.

There can be a wage premium for one ancestry group in this equilibrium if $w^D \neq w^L$. For example, if $w^D/w^L > 1$, descendants earn a premium over the wage of locals. The exact format of the wage premium and how it reacts to disturbances in the equilibrium depend on assumptions we make on the production function. In most cases, the wage premium will vary with the ancestry-specific productivity term θ. Moreover, whenever there is no perfect substitution, we may have a wage premium even in the absence of productivity differentials (when $\theta = 1$). In this scenario, the premium may vary with the proportion of descendant to local labor in the municipalities. This variation, in turn, would be governed by the degree of substitutability between the two types of labor (the parameter ρ).

[27] We follow Borjas (1992) and define ethnic capital as "the average quality of the ethnic environment in which parents make their investments" (p.124), which combine with parental inputs to determine the skills of the next generations in the families of immigrants of a particular ethnicity. In our setting, the ancestry types descendants and locals correspond to two large ethnic groups in Brazil: non-Iberians and Iberians.

[28] Using the same CES structure we gave as an example before, the wage schedule would be:

$$w^D = \theta A_j^\rho \left(Y_j/D_j\right)^{1-\rho}, \qquad w^L = A_j^\rho \left(Y_j/L_j\right)^{1-\rho}.$$

Finally, with imperfect substitution, we can expect fluctuations in the relative scarcity between types of labor to affect their price (wage). In other words, changes in the proportion of descendants and locals in each labor market may make one type less scarce (and the other, less abundant). Their wages, then, may adjust accordingly, becoming lower and higher, respectively.

4.4 Labor market frictions and technology choices

We assume that labor markets in Brazil are not perfectly integrated. Firms and workers are partially constrained to the characteristics of their current labor markets. Specifically, we assume that firms can hire workers only within their labor market (the municipality) and that workers do not migrate between labor markets for work. The representative firm in each municipality is, therefore, constrained to a mix of descendants and locals in its labor aggregate that reflects the concentration of descendants in the municipality. Workers, on the other hand, are constrained by the wage schedule offered by the firms to each ancestry type in their municipalities.[29]

Firms adjust their technology choices to this constraint on the labor mix. Different concentrations of descendants may allow different technology and labor choices in the municipalities. The wage offered to each ancestry type will reflect these differences in technology and labor optimization accordingly.

When comparing two municipalities, the representative firm in the municipality with a higher concentration will be less constrained on its choice of technology and labor mix than the representative firm in the municipality with a lower concentration. Starting from low levels, a higher concentration of descendants allows firms to use a different mix of labor in their production, one in which descendants become relatively more abundant and, therefore, receive lower wages. However, a higher concentration in the municipality can also allow the representative firm to choose a different technology. If the new technology better exploits complementarities between the labor of the two ancestry types making both more productive, wages could rise not only for locals but for both ancestry types.

The concentration of descendants in most Brazilian municipalities is quite small. The median is 3.70% in the national sample of municipalities and 1.74% in our main sample of interest (the Spread sample, formed by the Center-West, North, and Northeast regions). In this context, increases in the concentration of descendants in a municipality may allow

[29] One way to rationalize these assumptions is to assume that it is difficult for firms to hire more descendants when their concentration in the municipality is small (due to search costs, for example), and that it is costly for workers to migrate to another labor market. There is, in fact, evidence that Brazilian labor markets are not well-integrated (Dix-Carneiro and Kovak, 2017) and that migration costs significantly hinder integration (Morten and Oliveira, 2016).

the representative firm to access a larger pool of technologies, including technologies in which both types of workers are more productive. The central idea in this argument is one of diversity. There may be a menu of technologies that the firms can choose from, each corresponding to a given proportion of descendant and local labor. When constrained by a very small concentration of descendants, firms can choose amongst a small subset of technologies. When the concentration increases from very small levels, so does the diversity in the labor force and the number of possible technologies the firm can choose from. Potential increases in overall firm productivity are higher when constraints bind strongly. As the concentration increases and approaches 50% (maximum diversity), potential gains are gradually exhausted.[30]

4.5 Possible outcomes

Considering the setting and assumptions presented here, the following outcomes may follow from an increase in the concentration of descendants in a municipality or a comparison between municipalities with high and low concentration levels. This list is not exhaustive and focuses on the possibilities we judge more pertinent to our setting. We return to these outcomes in Section 7 in which we discuss our results and the possible mechanisms behind them.

First, a higher concentration of descendants in a given municipality may relax optimization constraints for the representative firm when there is imperfect substitution (complementarity) between descendant and local labor. In this case, an increase in the concentration would correspond to higher productivity for both workers if the firms can access technologies that capitalize on labor complementarities, and higher wages for both ancestry groups. Effects could be stronger in municipalities and firms where the concentration is small (e.g., below the median). It follows that effects can also be non-monotonic, showing decreasing returns that approach zero when the concentration reaches some close-to-optimal level.

Second, an increase in the concentration of descendants in a given municipality may correspond to increases in the wage of locals, who become relatively more scarce, and decreases in the wages of descendants, who become relatively less scarce. It follows that the wage premium of descendants when it exists would decrease. We could have stronger effects where the concentration is very small and a non-linear pattern for these effects. Also, the descendant wage premium may decrease faster where the concentration

[30] We note that the concentration of descendants in the municipalities of the Spread sample averages less than 4% and is always less than 33%. That compares to averages between 12% and 16% (depending on the weighting) and a range that can reach 100% in the Injection sample.

is lower, but it can also start from higher levels.

Third, when the previous two points combine, we may see wage increases that happen for both ancestry types but are stronger for locals. Finally, a higher concentration of descendants may correspond to higher average wages in the municipalities via a composition effect if the reduction in the descendant wage premium does not outweigh the increase in the proportion of descendants in the municipality.

The first three possibilities represent wage spillovers between ancestries. They are cases in which the concentration of descendants may affect the wages of each ancestry group, not only the average wage in the municipality. The last possibility, on the other hand, exemplifies a situation where the average wage in the municipality can change even when the wages of each type remain the same, via changes in the relative share of each ancestry group, i.e., a composition effect.

5 Empirical framework

5.1 Estimation strategy

5.1.1 Regressions at the municipality level

Our main explanatory variable is the concentration of descendants in the municipalities proxied by the proportion of workers in the formal sector with non-Iberian surnames. We want to know if the concentration of descendants affects income measures in the municipalities of our study regions, particularly wages in the formal sector. Municipalities in our Spread sample are far from the injection points of historical immigration in Brazil. Therefore, a positive association between the presence of descendants and income in those municipalities indicates that internal migration spread the positive impacts of historical immigration in Brazil.

Our empirical analysis begins with municipality-level regressions of aggregate income measures on the concentration of descendants and a set of controls.[31] Most municipality-level outcomes are available only in 2010. Therefore, we run regressions for this year only. The regression equation in the municipality-level analysis is:

$$y_{ms} = \beta C_m^{Desc} + X_m' \Gamma + \phi_s + \epsilon_{ms} \tag{1}$$

where y_{ms} is the outcome of interest in municipality m of state s; C_m^{Desc} is the concentration

[31] For the Spread sample, we do similar exercises that expand the set of municipality-level outcomes to include additional indicators in the income dimension, and also some indicators in the education and health dimensions.

of descendants in the municipality; X_m is a vector of municipality-level characteristics; ϕ_s are state fixed effects; and ϵ_{ms} is an idiosyncratic error term.

The vector of municipality-level controls X_m includes total yearly rainfall and the annual average temperature in the municipality (average and standard deviation in the 1981–2010 period). It also includes the total population in the municipality and its area (both in logs), distance to the state capital (log), average elevation and average Terrain Ruggedness Index (TRI), and the average distance of the municipality's economic center to historical immigrants' settlements in the states of São Paulo and Rio Grande do Sul.[32] Finally, the vector includes dummies for biomes and soil types (dummies = 1 if a given soil type or biome covers 5% or more of the municipality area).

The reasons for including the average distance to historical immigrants' settlements in our vector of controls are twofold. First, this measure is highly correlated with the distance from each municipality to Brazil's main ports for agricultural exports (e.g., the port of *Paranaguá* in the South) and financial center (the city of São Paulo in the Southeast). Second, when interacted with the TRI, the average distance to historical settlements enters our IV analyses as an excluded instrument. Therefore, having the non-interacted terms of the instrument (TRI and distance) as controls in all specifications is convenient for comparing results across OLS and IV regressions.

Regressions at the municipality-level allow us to identify associations between the concentration of descendants and a series of outcomes of interest like average income and wages, but those regressions are limited. In municipality-level regressions, we can neither identify nor control for the wage premium of descendants. Thus, we cannot check if a possible effect of the concentration of descendants on average wages in a municipality is due to simple composition effects or actual wage spillovers. Likewise, we cannot explore heterogeneity in results by ancestry and many other interesting dimensions like gender, race, education, or occupation in municipality-level regressions. Finally, though we could include shares of demographic characteristics at the municipality level to control for possible confounders, an individual-level analysis allows more detail and flexibility when controlling for characteristics of the individuals, firms, and municipalities. The individual-level data also allows the inclusion of individual fixed effects. Thus, we turn to individual-level regressions in our main analysis.

[32] Although other states in the Injection sample also had settlements, we consider only the settlements in São Paulo and Rio Grande do Sul when calculating distance measures used as controls and to constructed instrumental variables later on. These are the states for which there is evidence of long-term impacts in the economic literature, and for which there is information on the current municipality that corresponds to the location of the original settlements in these states. Including the historical settlements in the state of Paraná—the only one for which we can also obtain the corresponding current municipalities—does not affect our calculation of average distances and our main results.

5.1.2 Regressions at the individual level

The main outcome of interest in our individual-level regressions is the log of the worker's hourly wage. Wage and hours worked are observed in the same data we use to classify workers into ancestry groups and construct several control variables, making the choice of hourly wage as the outcome particularly convenient. Hourly wage is also an outcome of relevance since it serves both as a measure of the worker's welfare (income for consumption) and as a proxy for her labor productivity.

We want to know whether the concentration of descendants affects the wages of all workers—descendants and locals—in a given municipality-year. To do so, we regress a worker's log hourly wage on the concentration of descendants in the worker's municipality in a given year, a descendant dummy equal to one when that worker is a descendant (i.e., has a non-Iberian surname), and the interaction between these two variables. This way, we can control for a possible descendant wage premium in our regressions while also allowing the concentration of descendants to affect locals and descendants differently.

We also add an extensive set of individual-level controls and firm characteristics to our regressions. The coefficient on the concentration of descendants and the coefficient on the interaction of the concentration with the descendant dummy should reflect the effect of these variables only on the part of wages in a given municipality that is not explained by individual-level characteristics: the municipality wage premium. Finally, we again include several municipality-level characteristics as controls and state dummies in some regression specifications, and—when using 2004–2017 data—year dummies as well.

The basic regression equation used in our individual-level analysis is:

$$y_{imst} = \beta_1 C_{mt}^{Desc} + \beta_2 Desc_i + \beta_3 C_{mt}^{Desc} \times Desc_i + Z_{it}'\Pi + X_m'\Gamma + \phi_s + \delta_t + \varepsilon_{imst} \qquad (2)$$

where y_{imst} is the log wage of worker i in municipality m of state s in year t; C_{mt}^{Desc} is the concentration of descendants in the municipality-year; $Desc_i$ is the descendant dummy, Z_{it} is a vector of worker characteristics composed by dummies for gender and categories of race, age brackets, education levels, job tenure, firm size (number of employees), industry, and occupation; X_m is a vector of municipality-level characteristics; ϕ_s are state fixed effects; δ_t are year dummies; and ε_{imst} is an idiosyncratic error term. In all regressions, we cluster standard errors by municipality. In quadratic specifications, we add the square of the concentration of descendants and the interaction of this square with the descendant dummy.

We chose the set of controls used in our regressions carefully to mitigate concerns of omitted variable bias. We include relevant climate, geographic, and socioeconomic

characteristics that could affect the outcome of interest and bias our estimates. In particular, the inclusion of state fixed effects helps control for unobserved characteristics that vary at the state level. Ideally, we would use municipality fixed effects for this end, but their inclusion would exhaust the variation of our main explanatory variable (the concentration of descendants) in a cross-sectional analysis.

Most of our analyses use only individual-level data from RAIS in 2010. This period restriction is done both for computational convenience and because 2010 is the year for which more information on other outcomes of interest at the municipality level is available. By focusing on that year, we can better link the results from the municipality-level and individual-level regressions. When using data for 2004–2017, we take advantage of the structure in our data to construct a panel of municipalities and individuals. We use this panel to run regressions with different levels of fixed effects: state, municipality, individual, and state plus individual.

5.2 Identification strategy

The coefficient of interest β in equation (1) is identified out of the variation of the concentration of descendants between municipalities (C_m^{Desc}). The coefficient of interest β_1 in equation (2) is identified out of the variation of the concentration of descendants between municipalities and also within a municipality over time, when we use data for 2004–2017 (note the subscript t in C_{mt}^{Desc}).[33] Because most of the identifying variation comes from differences in the concentration of descendants between municipalities, we add an extensive set of controls in our regressions to reduce concerns that results are due to unobserved (or uncontrolled for) characteristics of the municipalities and not from differences in the concentration of descendants.

To address concerns of endogeneity in the concentration of descendants and attenuation bias (due to measurement error) in its coefficient, we use an instrumental variables strategy, adapting equation (2), which we implement with two-stage least squares. Controls in our regressions help to mitigate omitted variable bias, but endogeneity concerns remain. Descendants may concentrate in places where they expect to fare better and thus have a positive impact on the local economy, which could translate into higher wages. In this case, OLS results would be biased upward. On the other hand, descendants may concentrate in places with low wages to benefit from cheap local labor. In this case, OLS

[33] Specifically, in cross-sectional regressions with 2010 data, the coefficient of interest is identified out of the variation between all municipalities in the sample or between all municipalities in the same state (when state fixed effects are used). In panel regressions, however, the identifying variation comes from within each municipality or from the subset of workers who change municipalities.

results would be biased downward. Finally, we recall that our main explanatory variable is a proxy, and as such, it is subject to measurement error. OLS results may suffer from attenuation bias (assuming classical measurement error). An instrumental variables approach can help with both issues.

Our instrument uses the interaction of two terms. The first is the average distance from each municipality to all injection points of historical immigration in the states of Rio Grande do Sul and São Paulo (the settlements discussed in 2.1). We calculate a single measure, averaging the distance of each municipality to all injection points in these two states. Following the logic that distance to injection points increases migration costs, we expect this measure to be negatively correlated with the concentration of descendants in the Spread sample (the focus of this part of our analysis).

The second term in our interaction is the Terrain Ruggedness Index, a measure that serves as a proxy for how suitable a given municipality is for modern agriculture. Modern agriculture, particularly when focused on grain production, requires large plots for mechanization and gains of scale to become viable. Anecdotal evidence and historical accounts establish a link between the south-to-north internal migration of the descendants of historical non-Iberian immigrants and modern agriculture in Brazil (see Section 2.4). Flatland, the opposite of rugged terrain, is a characteristic that may have attracted descendants engaged in agriculture to particular places in the study region. Thus, we expect the average ruggedness of the terrain in the municipalities to be negatively correlated with the concentration of descendants. Figure A3 in the appendix gives a visual representation of the instrument and shows that there exists enough variation in the instrument, even within each state. This is important for our identification since most of our regressions include state fixed effects in all stages.

Less rugged terrain can benefit everyone but it would be especially attractive to potential internal migrants with a comparative advantage in modern agriculture. Because many of those who moved from the vicinity of the injection points in southern Brazil to the agricultural frontier had a comparative advantage in modern agriculture we expect that this specific group of internal migrants would benefit the most from this favorable geographic characteristic.[34]

The migration of those engaged in agriculture in the first moment may have attracted descendants working in different sectors later. Moreover, once transplanted from the southern parts of Brazil to its agricultural frontier, the descendants may have spread

[34]The comparative advantage of southern Brazilians, most of them from non-Iberian ancestry, in modern agriculture is a claim repeated by Rezende (2002), Alves (2016) and other authors who study the expansion of the agricultural frontier in Brazil. It also appears in anecdotal and historical accounts of this expansion (Wagner and Bernardi, 1995; Jepson, 2006*a*,*b*; Santos, 2008).

to the adjacent areas, ending up in municipalities of the Spread sample that are not necessarily part of the agricultural frontier. Therefore, we expect our instrument to work as a predictor of the concentration of descendants in all municipalities of the Spread sample, despite the fact that it was constructed with the migration to the frontier in mind. In fact, it does. Even when we control for state fixed effects as the results in Section 6.4 show.

The rationale behind the instrument is that both the closeness to injection points and the suitability of the potential destinations for modern agriculture (represented by low levels of terrain ruggedness) worked as pulling factors for descendants contemplating a move to the agricultural frontier in Brazil. Closeness to an injection point creates incentives to the migration of descendants to a given municipality by reducing migration costs and facilitating network formation. This pull, however, is weakened if the potential destination is rugged, making it difficult for the descendants to explore their comparative advantage in modern agriculture. Conversely, a municipality on the agricultural frontier with high availability of flat farmland attracts descendants with the expertise, disposition, and the means to engage in modern agriculture. This effect, however, is weaker, if the municipality of destination is far from a historical injection point, which means that migration costs are higher.

If the rationale discussed above is correct, both terms used in our instrument should be negatively correlated with the concentration of descendants. The interaction term, on the other hand, should be positively correlated. In fact, that is what we observe in our first-stage regressions (see Table A3 in the appendix). Both of the non-interacted terms (distance and terrain ruggedness), are negatively correlated because we are measuring the distance (opposite of closeness, or low migration costs) and the ruggedness (opposite of flatness, or suitability to modern agriculture). The interaction of the two terms is positively correlated with the concentration of descendants because the negative effect of distance is attenuated (less important to potential movers) if the ruggedness is high or, conversely, because the negative effect of ruggedness is attenuated if the potential destination is too far from the injection points making migration more costly.

If we interpret the negative correlation of the two terms used in our instrument as a first-order effect on the concentration of descendants, the positive correlation of their interaction would be the second order effect. This is central for our identification strategy because a valid instrument must not only have good predictive power but also satisfy the exclusion restriction. Both terms in our instrument, distance to injection points and terrain ruggedness, work well as predictors but are unlikely to satisfy the exclusion restriction. Distance to injection points is correlated with distance to the economic center of the country and with distance to ports in many cases. Terrain ruggedness, on the other hand, clearly influences the economic development of the municipalities via modern agriculture, which

is correlated with but not exclusively accompanied by a high concentration of descendants. The exclusion restriction for the interaction of those two terms, however, is more plausible. It only requires the second order effect to not affect the outcome of interest directly, only through the concentration of descendants. To the extent that our regressions can include the non-interacted terms and use only their interactions as the instrument, we find this non-testable restriction to be more plausible.

To instrument for the second endogenous term in our main regression—the interaction between the (endogenous) concentration of descendants and the (exogenous) descendant dummy—we simply interact our instrument with the descendant dummy.[35]

6 Results

6.1 Municipality-level

In Table 5 below, we show the results for municipality-level regressions of two different income measures on the concentration of descendants. First, in panel A, we show correlations between average income per capita and the concentration of descendants. Then, in panel B, we use average wages in the formal sector as the outcome of interest. In all specifications, we include state fixed effects. In the specifications of columns (2) and (4), we also add the set of municipality-level controls discussed in Section 5.1.1. In the appendix, in Tables A1 and A2, we show results for regressions with an extended set of municipality-level outcomes (for the Spread sample only). In addition to indicators of socioeconomic development in the income dimension, we also have indicators for the education and health dimensions.

The results reveal a consistent positive association between the concentration of descendants and average income per capita in both samples, and between the concentration and wages in the Spread sample. When comparing municipalities in 2010, we observe that one additional percentage point in the concentration of descendants is associated with 0.87% higher average income per capita in the Injection sample (panel A, column 2), and with 0.76% higher average income per capita in the Spread sample (panel A, column 4). In panel B we look at average wages in the formal sector and find close to zero coefficients in the Injection sample, but a positive results for the Spread sample. One additional percentage point in the concentration of descendants is associated with 1.23%

[35]This amounts to a triple interaction: distance times terrain ruggedness times the descendant dummy. The double interactions of the descendant dummy with distance and terrain ruggedness separately are not included.

Table 5: OLS Regressions: Income per capita and average formal sector wages at the municipality level, 2010

Sample:	Injection		Spread	
	(1)	(2)	(3)	(4)
Panel A: Log of average income per capita				
Concentration of descendants (%)	0.0060	0.0087	0.0106	0.0076
	(0.0004)***	(0.0005)***	(0.0024)***	(0.0019)***
R^2 (adjusted)	0.34	0.67	0.57	0.70
Panel B: Log average wage in the formal sector				
Concentration of descendants (%)	-0.0004	0.0008	0.0131	0.0122
	(0.0002)*	(0.0003)**	(0.0027)***	(0.0030)***
R^2 (adjusted)	0.32	0.48	0.24	0.26
N (municipalities)	2,849		2,624	
State fixed effects	Y	Y	Y	Y
Municipality-level controls		Y		Y

Notes: The dependent variable in all specifications in the top panel (A) is the log of the average income per capita in the municipality (data from Census 2010). In the bottom panel (B), the dependent variable is the log of the average wage in the formal sector in the municipality (data from RAIS 2010). The concentration of descendants (expressed in percentage points) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality-year. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. The median value of the concentration of descendants is 9.89% (avg. 16.45% and s.d. 17.68%) in the Injection sample, and it is 1.75% (avg. 2,88% and s.d. 3.82%) in the Spread sample. Municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

higher average wages in the Spread sample (panel B, column 4).[36]

Our conclusions from this exercise are twofold. First, we confirm a positive association between the concentration of descendants and average income in the municipalities in both of our samples. Second, we show that the association holds when we use a measure of income that considers all individuals in a municipality (income per capita) or one restricted to those employed in the formal sector (average formal sector wages). We also show that these associations are generally stronger in the municipalities in the Spread sample. The results in Table 5 reassure us of the assumption and motivations discussed in Section 4 and suggest that focusing on the formal sector (due to data limitations) can still be informative of results for individuals that do not participate in the formal labor market.

[36]Throughout the paper, we use the approximation $\exp(\hat{\beta}) - 1$ to interpret the coefficients from our regressions using a log-transformed dependent variable as percentage changes.

6.2 Individual-level

Table 6 shows the results of individual-level regressions of log hourly wage on the concentration of descendants, the descendant dummy, and their interaction. All regressions use data from RAIS 2010 and are presented for both the Injection and the Spread samples. In columns (1) and (4), we include only individual-level controls; in columns (2) and (5), we add state fixed effects, and in columns (3) and (6), we also have a set of municipality-level controls (our preferred specification). To make the interpretation of results for descendants and locals easier, we present marginal effects at the bottom of the table. These refer to an increase of one percentage point in the concentration of descendants over the median concentration in each sample.[37]

The results for the Injection sample tell a story of no effects for the locals and of small negative effects for descendants. In our preferred specification in column (3), with both state fixed effects and municipality-level controls, the coefficient on the concentration of descendants is -0.0003 (not statistically significant). The marginal effect of an additional percentage point in the concentration of descendants on the descendants themselves is negative (-0.11%). The result is statistically significant but quite small.

For the municipalities in the Spread sample and the individuals who work there, however, we find different results. The coefficients reveal a positive, significant, and consistent-across-specifications association between the concentration of descendants and wages. Again looking at the results for the specification with state fixed effects and municipality-level controls (column 6), we find that individuals in a municipality where the concentration of descendants is one percentage point higher have 0.80% higher wages on average. For the descendants, however, this association is reduced. The corresponding marginal effect in the bottom panel shows that one additional percentage point in the concentration of descendants is associated with 0.29% higher wages only for the group of descendants.

In both samples, we find a positive and significant wage premium for descendants, which aligns with the descriptive statics we saw in Section 3.3 and our discussion in Sections 4, and also evidence in the literature (Monasterio, 2017). Descendants earn 5.82% more than locals on average in the Injection sample (column 3) and 10.05% more in the Spread sample. The descendant wage premium appears throughout our regressions exercises with varying magnitudes, but it is always positive, significant, and sizeable. We note that these wage premia are conditional on an extensive set of worker and firm

[37] In this first set of regressions, calculating marginal effects amounts to calculating the compounded coefficient for descendants (the coefficient on the concentration plus the coefficient on the interaction) and the appropriate standard errors. In quadratic specifications, the calculation of marginal effects becomes more complex, changing with the starting point and the magnitude of the increments.

Table 6: OLS Regressions: Log hourly wages at the individual level on the concentration of descendants in the municipalities, 2010 (linear specification)

Outcome:	Log hourly wage					
Sample:		Injection			Spread	
	(1)	(2)	(3)	(4)	(5)	(6)
Concentration of descendants (%)	0.0004	-0.0020	-0.0003	0.0173	0.0093	0.0080
	(0.0005)	(0.0005)***	(0.0006)	(0.0013)***	(0.0020)***	(0.0020)***
Concentration x Descendant	-0.0031	-0.0007	-0.0009	-0.0100	-0.0066	-0.0051
	(0.0004)***	(0.0003)**	(0.0003)***	(0.0012)***	(0.0012)***	(0.0009)***
Descendant dummy	0.0929	0.0539	0.0566	0.1353	0.1086	0.0958
	(0.0075)***	(0.0061)***	(0.0053)***	(0.0077)***	(0.0085)***	(0.0066)***
R^2 (adjusted)	0.49	0.51	0.52	0.41	0.43	0.44
N (workers)		20,191,199			6,030,247	
Clusters (municipalities)		2,849			2,624	
Individual-level controls	Y	Y	Y	Y	Y	Y
State fixed effects		Y	Y		Y	Y
Municipality-level controls			Y			Y
Marginal effects of the concentration of descendants (one additional percentage point)						
for Descendants	-0.0027	-0.0028	-0.0011	0.0073	0.0026	0.0029
	(0.0004)***	(0.0004)***	(0.0005)**	(0.0009)***	(0.0012)**	(0.0017)*

Notes: The dependent variable in all specifications is the log of the worker's hourly wage. The concentration of descendants (expressed in percentage points) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality-year. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. The median value of the concentration of descendants is 9.53% (avg. 12.20% and s.d. 10.57%) in the Injection sample, and it is 2.21% (avg. 3.83% and s.d. 4.14%) in the Spread sample. Individual-level controls: dummy variables for gender and categories of age, education, race, job tenure, occupation, firm size (number of employees), and industry. Municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

characteristics used as controls in our wage regressions, including characteristics that one could expect to be correlated with ancestry (education, race, occupation, and industry). The descendant wage premium, thus, may reflect higher labor productivity of descendants or discrimination in employment and wage-setting (based on ancestry or names). To investigate the nature of the descendant wage premium, however, is beyond the scope of our study. Here we focus on the effects of the concentration of descendants on the wages of all workers, and the wages of locals in particular.

The results for the Spread sample in Table 6 suggest that elevating the median concentration of descendants by 7.32 percentage points there (so it would equal the median in the Injection sample) would correspond to 6.03% higher wages for locals and 2.15% wages for descendants. However, we do not see such large marginal effects for the Injection sample, where the median concentration of descendants is already near 10%. And we have no reason to believe any effect of the concentration on wages, if it exists,

is linear. In fact, based in our discussion in Section 4, we believe that the relationship between the concentration of descendants and wages can be non-monotonic, with effects that are large when the concentration is small and that level off when the concentration of descendants approaches the levels we see in the Injection sample. Therefore, we modify our standard regression equation to include a quadratic term for the concentration of descendants and its interaction with the descendant dummy to allow for nonlinearities.

Table 7 shows the results of individual-level regressions of log hourly wage on the concentration of descendants, the descendant dummy, and their interaction with a quadratic specification. The structure of the table and the order of the specifications is the same as in Table 6, except that the marginal effects are now calculated for both locals and descendants since both depend on the initial level over which we add one more percentage point to the concentration of descendants (the median of each sample).

The results for the Injection sample are similar to what we got with linear specifications: coefficients and marginal effects are either small and negative or zero. For the Spread sample, the results show two important differences. First, they are larger. The marginal effect of adding one percentage point to the median concentration of descendants is 2.10% higher wages for locals and 0.98% higher wages for descendants. Second, the quadratic terms are all statistically significant, showing that the relationship between wages and the concentration of descendants is non-monotonic in the municipalities of the Spread sample. If we repeat the exercise of the previous paragraph, adding 7.32 percentage points so the median in the Spread sample would be the same as in the Injection sample, the marginal effects (not shown) would be close to zero or even negative.

6.3 Panel regressions

So far, we have not used the 14 years of information in our 2004–2017 data or leveraged the panel structure of this data to include fixed effects that reduce concerns of bias caused by unobserved factors. We do this now, focusing only on the Spread sample.

Table 8 below shows results for different specifications of panel regressions. In all regressions, we use the quadratic specification and include year fixed-effects and the sets of individual-level and municipality-level controls used in previous steps. The main difference between each specification is the fixed effect added on top of year fixed effects and, consequently, the source of variation that identifies the coefficient of interest.

In column (1), we add state fixed effects, thus running a pooled OLS regression equivalent to that presented in column (6) of Table 7. The identifying variation comes from comparing the concentration of descendants in different municipalities and from variations in the concentration within a municipality over time. In column (2), we add

Table 7: OLS Regressions: Log hourly wages at the individual level on the concentration of descendants in the municipalities, 2010 (quadratic specification)

Outcome:	Log hourly wage					
Sample:		Injection			Spread	
	(1)	(2)	(3)	(4)	(5)	(6)
Concentration of descendants (%)	0.0078	-0.0022	0.0002	0.0449	0.0381	0.0247
	(0.0014)***	(0.0016)	(0.0014)	(0.0035)***	(0.0068)***	(0.0052)***
Concentration x Descendant	-0.0098	-0.0008	-0.0021	-0.0268	-0.0171	-0.0135
	(0.0016)***	(0.0010)	(0.0010)**	(0.0035)***	(0.0030)***	(0.0024)***
Concentration squared	-0.0002	0.0000	-0.0000	-0.0016	-0.0013	-0.0007
	(0.0000)***	(0.0000)	(0.0000)	(0.0002)***	(0.0002)***	(0.0002)***
Concentration squared x Descendant	0.0002	0.0000	0.0000	0.0011	0.0006	0.0005
	(0.0000)***	(0.0000)	(0.0000)	(0.0002)***	(0.0001)***	(0.0001)***
Descendant dummy	0.1305	0.0552	0.0675	0.1574	0.1269	0.1152
	(0.0148)***	(0.0101)***	(0.0099)***	(0.0126)***	(0.0114)***	(0.0094)***
R^2 (adjusted)	0.50	0.51	0.52	0.42	0.43	0.44
N (workers)		20,191,199			6,030,247	
Clusters (municipalities)		2,849			2,624	
Individual-level controls	Y	Y	Y	Y	Y	Y
State fixed effects		Y	Y		Y	Y
Municipality-level controls			Y			Y
Marginal effects of the concentration of descendants (one additional percentage point over the median)						
for Locals	0.0042	-0.0021	0.0000	0.0364	0.0313	0.0208
	(0.0009)***	(0.0011)*	(0.0010)	(0.0026)***	(0.0055)***	(0.0042)***
for Descendants	-0.0022	-0.0030	-0.0016	0.0154	0.0176	0.0098
	(0.0012)*	(0.0011)***	(0.0008)*	(0.0027)***	(0.0042)***	(0.0039)**

Notes: The dependent variable in all specifications is the log of the worker's hourly wage. The concentration of descendants (expressed in percentage points) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality-year. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. The median value of the concentration of descendants is 9.53% (avg. 12.20% and s.d. 10.57%) in the Injection sample, and it is 2.21% (avg. 3.83% and s.d. 4.14%) in the Spread sample. Individual-level controls: dummy variables for gender and categories of age, education, race, job tenure, occupation, firm size (number of employees), and industry. Municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

municipality fixed effects and identify the coefficients of interest out of the variation in the concentration of descendants within each municipality only. In column (3), we add individual fixed effects. Part of the identifying variation comes from changes in the concentration of descendants in the municipality where individuals work over time, but most identification comes from individuals who switch municipalities over the years. Finally, in column (4), we add both individual and state fixed effects to our regressions. In all specifications except the first, we adjust controls used, removing those that do not vary over time to avoid colinearities with the fixed effect used in the specification. For this reason, the last two columns, in which we use individual fixed effects, do not include the

descendant dummy (a time-invariant individual characteristic).

Table 8: Pooled OLS and Panel Regressions: Log hourly wages at the individual level on the concentration of descendants in the municipalities, 2001–2017 (quadratic specification, Spread sample)

Outcome:	Log hourly wage			
Fixed Effects:	State	Municipality	Individual	Indiv. & State
	(1)	(2)	(3)	(4)
Concentration of descendants (%)	0.0275	0.0128	0.0126	0.0106
	(0.0049)***	(0.0066)*	(0.0001)***	(0.0001)***
Concentration x Descendant	-0.0008	-0.0004	-0.0004	-0.0004
	(0.0002)***	(0.0001)***	(0.0000)***	(0.0000)***
Concentration squared	-0.0135	-0.0098	-0.0073	-0.0066
	(0.0019)***	(0.0014)***	(0.0004)***	(0.0004)***
Concentration squared x Descendant	0.0004	0.0003	0.0001	0.0001
	(0.0001)***	(0.0001)***	(0.0000)***	(0.0000)***
Descendant dummy	0.1151	0.1044		
	(0.0078)***	(0.0064)***		
R^2 (adjusted or within)	0.47	0.44	0.38	0.38
N (worker-year)	85,324,069			
Clusters (municipalities or workers)	2,680		19,217,877	
Marginal effects of the concentration of descendants (one additional p.p. over the median)				
for Locals	0.0231	0.0108	0.0101	0.0084
	(0.0040)***	(0.0059)*	(0.0001)***	(0.0001)***
for Descendants	0.0120	0.0023	0.0035	0.0024
	(0.0041)***	(0.0061)	(0.0004)***	(0.0004)***

Notes: The dependent variable in all specifications is the log of the worker's hourly wage. All specifications include year fixed effects, municipality-level controls, and individual-level controls. Time invariant controls at the level of the municipality (individual) are dropped in column 2 (columns 3 and 4). The concentration of descendants (expressed in percentage points) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality-year. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. The median value of the concentration of descendants in the sample is 2.25% (avg. 3.80% and s.d. 4.01%). Individual-level controls: dummy variables for gender and categories of age, education, race, job tenure, occupation, firm size (number of employees), and industry. Municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality (columns 1 and 2) or individual (columns 3 and 4) in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

The results in column (1) are very similar to those we obtained when using data for 2010 only, which is expected since the most significant variation in the concentration of descendants comes from comparing municipalities with different concentrations, not from the evolution of the concentration within a municipality over the years. One additional percentage point in the concentration of descendants over the median corresponds to 2.34% higher wages for locals and 1.21% higher wages for descendants. In column (2),

with the inclusion of municipality fixed effects, the results decrease by more than half. The marginal effect of increasing the concentration of descendants by one percentage point over the median in the municipalities is an increase in wages of 1.09% for locals and 0.23% for descendants (not significant). In columns (4) and (5), results are similar to the ones in column (2). One additional percentage point in the concentration of descendants over the median corresponds to 1.02% higher wages for locals (0.84% when state fixed effect are added) and 0.35% higher wages for descendants (0.24% when state fixed effect are added).

Overall, panel fixed effects suggest that accounting for unobserved characteristics of the municipalities or individuals reduces the marginal effects we have established before for the Spread sample in Table 7, column (6). However, we must keep in mind that the results from both exercises are not directly comparable. When using municipality fixed effects, only the variation within each municipality over time identifies the coefficient of interest. When using individual fixed effects, on the other hand, even though we get some identification from variation across municipalities, this variation comes from a subset of individuals who switch municipalities over the years (only a quarter of the total). If we are interested (as we are) in knowing how different concentrations of descendants in different municipalities can affect the wages of all workers living in those municipalities, then the results in Table 7, identified off of cross-sectional variation, are more informative.

6.4 Instrumental variables

The concentration of descendants in certain municipalities may be endogenous to the process that generates our results. As we discussed in Section 5.2, descendants may have concentrated in locations where they expected to affect wages positively or negatively, which would bias our results.

In this section, we present results from our instrumental variables (IV) strategy, which addresses endogeneity concerns with an instrument that relies on the interaction of two geographic characteristics of the municipalities (terrain ruggedness and distance to historical settlements). The IV strategy has the additional advantage of helping with the attenuation bias arising from using a proxy for the concentration of descendants, which inevitably brings measurement error into our analysis. The validity of the IV estimations as consistent estimates of actual effects, however, is conditional on the validity of the exclusion restriction and, in the case of attenuation bias, on the assumption of classical measurement error in our proxy. Therefore, even for the IV results, we are careful not to make general causal claims in this study. Instead, we take the collection of evidence shown across all of our results as indicative of actual impacts of the concentration of descendants on wages.

Table 9 below shows the results for IV regressions using data only from 2010. As before, we focus on the Spread sample only. The dependent variable, controls, structure of the table, and progression of specifications are similar to what we have used in the previous steps. A few differences are worth noting. First, we show results only for the linear specification. The quadratic specification has four endogenous regressors: the concentration of descendants, its square, and the interactions of both with the descendant dummy. We do not have good instruments for the two quadratic endogenous regressors, so we reduce our need for instruments by focusing on the linear specifications despite its limitations.[38] Second, the specifications in columns (1) and (2) include two municipality-level controls: the two terms used in the interactions used to create our instrument (terrain ruggedness and distance to settlements). The specification in column (3) includes these and all other municipality-level controls. Third, the regression table shows some statistics from the first stage regressions, notably the test statistics for underidentification and weak identification. Table A3 in the appendix shows the complete results for the first stage.

The results for the second stage of our IV regressions are shown in Table 9 below. Both the concentration of descendants and the interaction of the concentration with the descendant dummy are treated as endogenous variables and instrumented for in the first stage. The excluded instrument for the first endogenous variable is the interaction between the average distance from the municipalities to the injection points of historical non-Iberian immigration in the states of São Paulo and Rio Grande do Sul and the municipalities' average Terrain Ruggedness Index. For the second, we interact the first instrument and the descendant dummy.

The results show the same positive association we saw before, but the magnitude of the effects is considerably larger. The marginal effects for the locals shown in column (3), for example increases almost nine-fold, from the equivalent 0.80% we had in column (6) of Table 6 to the 7.13% higher wages we have now. We notice that in the first column, the results are smaller, although not statistically significant. Results for descendants increase even further (they are eighteen times larger). The specification used in that regressions does not include state fixed effects or municipality-level controls, so we do not place much weight on its results.

The statistics from the first stage show that we have a robust first stage in general, but that there is some loss of predictive power for the concentration of descendants when state fixed effects are included. Such a loss is expected, since the geographic variation that identifies the coefficients decreases substantially when restricted within each state.

[38] Whereas the interaction of our instrument with the descendant dummy works well to predict the concentration of descendants interacted with the descendant dummy, the same does not happen for the square of the instrument and its interactions.

Table 9: IV Regressions (2nd stage): Log hourly wages at the individual level on the concentration of descendants in the municipalities, 2010 (linear specification, Spread sample)

Outcome:	Log hourly wage		
	(1)	(2)	(3)
Concentration of descendants (%)	0.0066	0.0496	0.0689
	(0.0057)	(0.0209)**	(0.0225)***
Concentration x Descendant	-0.0081	-0.0168	-0.0167
	(0.0025)***	(0.0050)***	(0.0044)***
Descendant dummy	0.1378	0.1430	0.1362
	(0.0123)***	(0.0181)***	(0.0192)***
R^2 (centered)	0.50	0.46	0.46
N (workers):		6,030,247	
Clusters (municipalities)		2,624	
Underidentification (K-P LM Stat)	23.10	9.14	11.35
Weak identification (K-P Wald F Stat)	19.94	4.49	5.89
S-W multivariate F test of excluded instruments (weak identification)			
Concentration of descendants (%)	40.97	11.54	14.14
Concentration x Descendant	46.45	43.72	46.63
Marginal effects of the concentration of descendants (one additional p.p.)			
for Descendants	-0.0015	0.0328	0.0522
	(0.0042)	(0.0163)**	(0.0189)***
Individual-level controls	Y	Y	Y
State fixed effects		Y	Y
Municipality-level controls	Y*	Y*	Y

Notes: The dependent variable in all specifications is the log of the worker's hourly wage. The concentration of descendants (expressed in percentage points) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality-year. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. There are two endogenous regressors in all specifications: the concentration of descendants, and the interaction of this concentration with the descendant dummy. Likewise, there are two excluded instruments in the first stage of all specifications. The instrument for the first endogenous regressor (the concentration of descendants) is the interaction of the average Terrain Ruggedness Index of the municipality and the average of the distance to historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul. For the second endogenous regressor—the interaction of the concentration of descendants with the descendant—we use the interaction of the first instrument with the descendant dummy. The terms used in the interactions are included as controls in all specifications. The median value of the concentration of descendants in the sample is 2.21% (avg. 3.83% and s.d. 4.14%). Individual-level controls: dummy variables for gender and categories of age, education, race, job tenure, occupation, firm size (number of employees), and industry. Municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

There are a few possible explanations for why the results from our IV regressions are larger than the ones we obtained with OLS regressions. First, the endogenous sorting of descendants may have led them to concentrate in places where they could depress wages and take advantage of cheap local labor. Therefore, initial OLS results were biased

downward, and the IV estimation corrected for that bias and revealed the true (larger) effects. Second, our OLS estimation may have suffered from attenuation bias, which was more severe when municipality-level controls and state fixed effects were included. Third, the instrument does not predict the concentration of the descendants in general but of a specific group within them, a group that is more likely to affect wages positively. These explanations are not mutually exclusive, and some combination of them is probably the actual reason for the difference in coefficients.

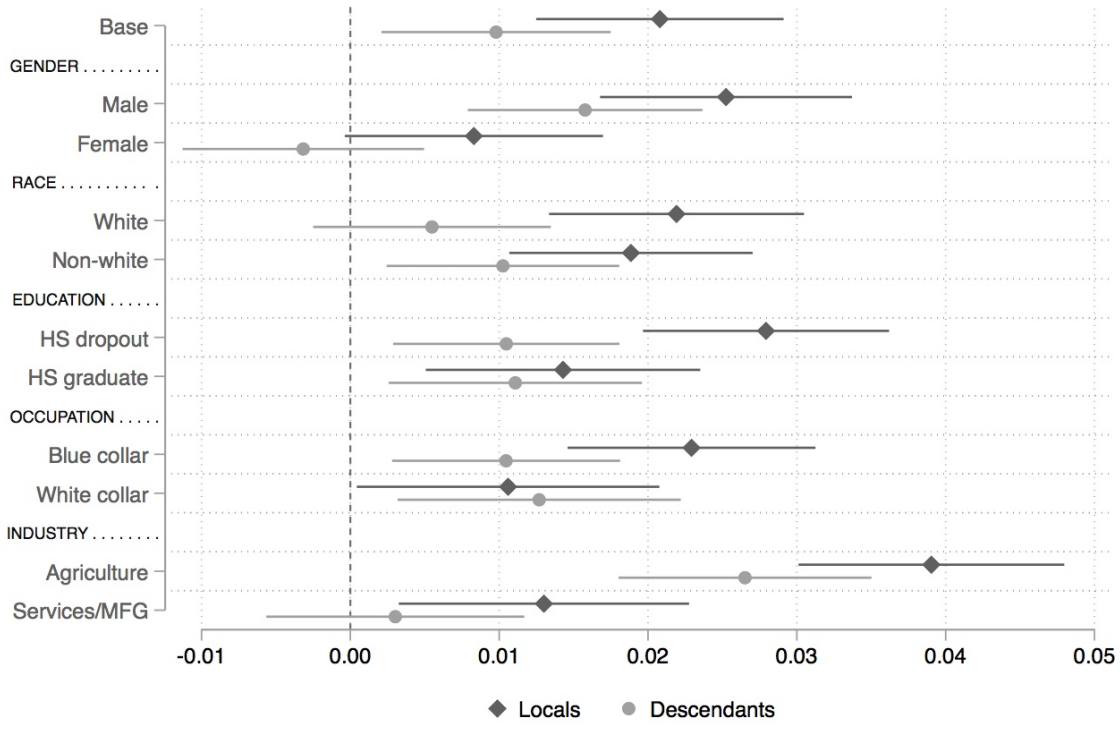
Our takeaway from the IV results is that, even though causal claims must be taken with a grain of salt due to the non-testable validity of the exclusion restriction, they point to the same positive direction for the relationship between the concentration of descendants in the municipalities and wages. Moreover, the results here align with the pattern of stronger effects for locals than for descendants, and they maintain the existences of a positive and sizeable descendant wage premium.

6.5 Heterogeneity

Figure 5 below summarizes a set of results that take advantage of the richness of information in our data to shed light on the mechanisms behind our results. This figure presents marginal effects obtained from a series of regressions that interact the concentration of descendants and its square not only with the descendant dummy but also with an additional binary indicator that splits the sample into two groups: males and females, whites and non-whites, high school dropouts and graduates, workers in blue-collar and white-collar occupations, and workers in the agricultural sector or the non-agricultural sectors (manufacturing and services). For each dimension, we present four marginal effects. For example, for gender, we show the marginal effect of one additional percentage point in the concentration of descendants over the median for local males, descendant males, local females, and descendant females. The table with the corresponding regression results is not shown. All results in the figure refer to the Spread sample in 2010. Results using a linear specification exhibit a similar pattern (see Figure A4 in the appendix).

The first two bars (base results) reproduce the marginal effects calculated in column (6) of Table 7: one additional percentage point in the concentration of descendants over the median is associated with 2.10% higher wages for local and 0.98% higher wages for descendants. In the next four bars we begin seeing the first dimension with considerable heterogeneity in the marginal effects: gender. For local males, the marginal effect corresponds to an increase of 2.56% in wages, compared to 0.83% for local females (marginally significant). Among descendants, the marginal effects are 1.59% for males and -0.32% for females (not significant). Marginal effects are larger for males, both descendants and

Figure 5: Heterogenity in the marginal effects of the concentration of descendants on wages, 2010 (quadratic specification, Spread sample)



Note: All point estimates and 95% confidence intervals shown in the figure come from regressions of the log hourly wage on the concentration of descendants, the square of the concentration, and their interactions with the descendant dummy and a binary indicator. The binary indicators split the samples into two groups according to some characteristic of the worker or her job. The regression specifications follow the format of the specification shown in column (6) of Table 7 and include individual-level and municipality-level controls and state fixed effects. Individual-level controls are adjusted according to the binary variable use to split the sample into two groups and elicit heterogeneities.

locals.

Progressing to the next group of bars, we find little heterogeneity in the racial dimension. White descendants have smaller (non-significant) marginal effects than non-white descendants, but, overall, most 95% confidence intervals of this group of bars contain the corresponding point estimate of the base estimation. A slightly different pattern emerges for the splitting over education and occupation groups. In both cases, the locals of one group (high school dropouts for education, and blue-collar workers for occupation) have higher marginal effects than everyone else (but within the confidence interval for locals in the base estimation). The point estimates for all others—local high school graduates and local white-collar workers, and descendants of any education and occupation—are all remarkably similar to the point estimate for descendants in the base estimation (a marginal effect of around 1%).

Together, these heterogeneous results suggest a pattern where low-skilled men working in agriculture benefit the most from the presence of descendants in a municipality. We note that this fact aligns with the story we tell in Section 2.4, of descendants playing a central role in the expansion of the agricultural frontier and helping the spread of advanced agriculture from southern to northern and central Brazil (Rezende, 2002; Alves, 2016). We also note that the magnitude of the results for locals working in agriculture gets closer to the IV estimates we show in Table 9. This suggests that it is, indeed, possible that our IV results are much larger than equivalent OLS estimates because our instruments are picking up mostly the concentration of descendants that is closely related to the expansion of modern agriculture in the Brazilan agricultural frontier. And are those descendants who generate the largest effects on the wages of locals.

7 Discussion of mechanisms

In section 4.5, we discuss four possible outcomes of our theoretical framework. In all of them, changes in the concentration of descendants can affect the wages of descendants and locals in a municipality. For the first three possible outcomes, we built an argument of complementarity of labor types. An increase in the concentration of descendants in a municipality can affect the wages of descendants and locals by changing their proportion in the production function of the firms. One type of labor becomes less scarce and the other less abundant. This can affect the wage differentials between the two groups. Moreover, if a higher concentration of descendants in the municipality ends up relaxing optimization constraints for the firms, we can observe wage increases for workers of both ancestries. In the fourth outcome, we built an argument of composition effects. The increase in the proportion of the labor type with higher wages drives the average wage in the municipality up.

In this section, we discuss each of these possibilities against the evidence presented in our results and bring additional evidence when we can. In the next, we list and briefly discuss other possible explanations for which we cannot directly test.

7.1 Mechanisms within the theoretical framework

***Labor complementarities**: Descendants and locals are imperfect substitutes in the production function of firms; therefore, changes in their proportion may unlock labor complementarities that benefit all workers.*

Our results agree with the hypothesis of labor complementarity in the firms' production function and market frictions that constraint firms' labor mix to the concentration of

descendants in the municipalities. When comparing workers in municipalities of the Injection sample, where the concentration is higher and better distributed, we find that descendants have a wage premium of almost 7% over the wage of locals (column 3, Table 7). We find little to no effect of their concentration on wages. If anything, an increase of one percentage point in the concentration of descendants leads to a small decrease in the wages of descendants. These results are consistent with a setting in which descendants are more productive than locals, but there is little complementarity between the two types of labor since differences in the concentration do not affect the wage of locals.

In the Spread sample, we see a different pattern of results. The wage premium of descendants is larger, over 12% in our preferred specification (column 6, Table 7). Moreover, an additional percentage point in the concentration of descendants does lead to changes in wages for both ancestry groups in this sample, with the wage of locals increasing almost twice as much as the wage of descendants. These results are consistent with our proposed setting in which firms in municipalities with low levels of the concentration of descendants can move to better technologies and unlock labor complementarities that benefit both groups when the concentration rises and relaxes optimization constraints. Also, the difference in the wage premium estimated for the Injection and Spread samples can be explained by descendants being a much more scarce type of labor in the Spread municipalities.

The fact that the wage of locals increases more than the wage of descendants may be due to a composition of two forces that go in the same direction for locals, but in opposite directions for descendants. For locals, an increase in the concentration of descendants can make local labor relatively less abundant and more valuable. The labor of locals can also become more productive if it benefits from complementarities in new technologies adopted by the firms. These two forces affect the wage of locals positively. For descendants, an increase in their concentration makes them less scarce, reducing their wage premium. However, they can also benefit from the labor complementarities unlocked with new technologies that benefit from a greater diversity of labor in the municipality. The net effect we find for descendants is still positive, but it is smaller than the effect for locals.

Finally, we note that our estimation results suggest two facts about wage spillovers. First, that they are non-monotonic, showing larger gains when the concentration is very small and decreasing in a concave fashion until being exhausted. This non-monotonic behavior shows in the results of our quadratic specifications. Second, our heterogeneity exercise in Section 6.5 suggests that labor complementarities between descendants and locals can be stronger in certain industries like agriculture, and for certain types of workers like low-skilled men.

***Composition effects**: Descendants can have higher labor productivity or better outside options; therefore, an increase in their concentration may raise the municipality average via a composition effect.*

By looking only at aggregate outcomes like income per capita or the average formal sector wage in the municipality, as we do in Table 5, we could conclude that a positive association between the concentration of descendants and the outcome of interest comes from changes in the composition of workers in a municipality. Given that descendants earn higher wages than locals, an increase in their concentration in the workforce would raise the average of the group regardless of how it affects the wages of individuals.[39] A similar argument goes for descendants having better outside options. In this case, an increase in the concentration of descendants would reflect an increase in wages that draws more descendants to be wage earners in the formal sector and elevates the average wage in a given formal labor market.

Our individual-level data, however, allow us to go beyond aggregate outcomes and to include an indicator for descendant ancestry in all our main regressions. This way, we control for the composition of workers in the municipality and can rule out the possibility that our results are due to simple composition. Instead, we verify consistent evidence of wage spillovers.

***Technology multiplier**: The concentration of descendants in the workforce reflects the number of descendants working as firm managers, who can influence the technology multiplier in the firms' production function.*

A mechanism that operates through the technology multiplier may exist if, for example, descendants that operate firms as owners or managers employ a technology that increases the productivity of labor (or a labor-augmenting technology in a version of the model with capital). Firms operated by locals could copy such technologies and increase the productivity of their workers too. Unfortunately, we cannot directly assess the technology mechanism. We lack information on the firm's technology choices and capital, and we do not have information on the ancestry of firm owners.

We can, however, identify the ancestry and the occupation level of every individual in our sample. With this information, we calculate the concentration of descendants among firm managers and directors, and among the staff separately. We then run variations of our main regression (column 6, Table 7) using the concentration of descendants

[39]Consider the following numerical examples. In one case, descendants are more productive and earn, on average, 10% more than locals. Then, an increase in their concentration from 5% to 10%, for example, all else equal, would raise the average wage by approximately 0.5% even if for each group separately, there was no change.

among all workers, among the staff of firms only, or among firm managers and directors only. The idea here is that the concentration of descendants in the management level of the firms can serve as a proxy for the presence of descendants in positions that can influence the technology decisions in the firms of a municipality. Then, if descendants bring technology or management practices that increase labor productivity of all employees in the firm via the technology multiplier, the concentration of descendants in management positions would serve as a better predictor for the wage spillovers than the concentration of descendants in general staff positions.

Table A4 in the appendix shows the results of this exercise. The concentration of descendants measured among all workers or staff members only are similar. So are the regressions coefficients in specifications that use one measure or the other. When using the concentration of descendants measured among managers and directors only, we obtain smaller results. Therefore, a mechanism that operates via descendants in management positions, although possible, is not likely to completely explain our results.

7.2 Mechanisms beyond the theoretical framework

***Preference for diversity**: High ability individuals of all ancestries prefer places with a greater diversity of ancestries.*

In all municipalities in the Spread sample, an increase in the concentration of descendants means an increase in diversity of ancestries since $C_m^{Desc} < 50\%$ for all m in that sample. Therefore, if high-productivity individuals prefer places with higher diversity (or, for whatever reason, places with a higher concentration of descendants), we could observe a composition effect similar to the one discussed in the second mechanism but unrelated to ancestry. Average wages would increase because individuals who earn more concentrate in more diverse municipalities.'

Since we control for education and other variables at the individual level (like firm tenure and age, which can capture work experience), remaining individual productivity differences would likely be part of individual-specific unobserved characteristics. We still find positive and significant effects for the concentration of descendants when controlling for individual fixed-effects, though (columns (3) and (4) in Table 8). Unobserved characteristics of the individuals that could affect their productivity regardless of ancestry do not seem to explain our results.

The magnitude of the marginal effects of an increase of one percentage point in the concentration of descendants over its median decreases when individual fixed-effects are included in our regressions. Comparing columns (1) and (4) in Table 8, we see a decrease from 2.34% to 0.84% in the marginal effect for locals and a decrease from

1.21% to 0.24% for descendants when individual fixed-effects are added to a specification that already included state fixed-effects. This reduction in marginal effects suggests that a substantial part of our main results can be explained by unobserved characteristics of individuals that sort themselves into municipalities with a higher concentration of descendants. However, we must keep in mind that the identification of the coefficients of interest in a specification with individual fixed-effects comes from a subsample of individuals who switch municipalities in 2004–2017, who are less than a quarter of the total number of individuals in our sample. We conclude that the endogenous sorting of high productivity individuals can explain part of our results, but not all.

***General internal migration**: The concentration of descendants is correlated with the concentration of (positively self-selected) internal migrants of all ancestries.*
Internal migrants, in Brazil and elsewhere, are often positively self-selected. Positive self-selection is precisely one of the explanations for why descendants have a wage premium. Nonetheless, while all internal migrants may be self-selected, not all of them are descendants. If the concentration of descendants is positively correlated with the concentration of internal migrants of any ancestry, our results may be capturing an effect of the latter on wages instead of the former.

To check this possibility, we construct several measures of internal migration. We say an individual is a "municipality migrant" if she has moved at least once to a different municipality in 2004–2017. We define "state migrants" in a similar fashion and calculate the concentration of both types in the municipalities in 2010. For this same year, we use census data to calculate the share of residents born out of the municipality and out of the state.[40] We then add these measures of internal migration as controls in versions of our main regression specification (column (6) in Table 7). We add the concentration of internal migrants and, when possible, also the internal migrant indicator, to these specifications.

The marginal effects on the wage of locals and descendants change little compared to our preferred results (results not shown). The marginal effect for locals varies between 1.78% and 2.18% across all specifications (compared to 2.10% on column (6) of Table 7). For descendants, the marginal effect in these exercises is smaller but still close to the estimate in our main specification (the range is 0.58%–0.88%). We conclude that general (contemporaneous) internal migration cannot explain our results.

***Agglomeration effects**: Descendants concentrate in municipalities with high population density, which is conducive to agglomeration effects that increase productivity and wages.*

[40]Our main data source, RAIS, does not contain information on workers' state or municipality of birth.

Our specifications with municipality-level controls include total population and municipality area as covariates, both in logs. We do so to control for agglomeration effects, a situation in which denser cities generate opportunities for mutual learning among workers and increases in productivity (Glaeser, 1999). If the concentration of descendants correlates with population density in the municipalities, our results can be driven by agglomeration effects, not complementarities in labor.

Descriptive statistics in Table 4, however, show that descendants in both samples tend to live in municipalities with fewer individual observations of workers in the formal sector, not more. A similar difference holds for the log of the total population, which is smaller for descendants on average. Therefore, we are willing to dismiss agglomeration effects as an alternative mechanism for our results. Still, we check the sensitivity of our results to different population-related controls, so that the role of agglomeration effects in our results can be better assessed.

We run different versions of our main regression (column 6, Table7). The log of the total population is either removed from our vector of municipality-level controls or replaced by a different population measure. Table A5 in the appendix shows the results for this exercise. When not controlling for any measure of population size, our coefficients drop slightly (marginal effects for locals go from 2.10% to 1.82%). This is likely due to the concentration of descendants being negatively correlated with population density. Coefficients change little when we use alternative measures like the population density in the past (1950), which is less likely to be endogenous to the concentration of descendants today. There is little change also when we use the log of the total number of individual observations. Finally, because a difference between the log of the total population and the log of the total number of observations in our data can arise from different levels of formalization in the labor market, we repeat our main regressions controlling for the degree of labor formalization in each municipality. The drop in the marginal effect for locals is larger (to 1.49%), but it remains positive and significant. We conclude that agglomeration effects or another mechanism related to the population density in the municipalities are unlikely to explain our results.

***Early investments**: The first descendants who arrived in the municipalities of the Spread sample pushed for public investments that generate positive effects today for locals and descendants alike.* The concentration of descendants likely shows hysteresis. If we observe a high concentration of descendants in a given municipality today, there is a good chance that that municipality also had a higher concentration of descendants in the near past (one or two generations ago). That would be consistent with anecdotal accounts of the expansion of the agricultural frontier in Brazil and the patterns of internal migration that populated

many municipalities in the Spread sample. Perhaps the first descendants who moved to those municipalities were better able to influence policymakers, or the descendants themselves made investments in public goods (particularly those that can affect future human capital in the municipalities, like educational infrastructure). Such early investments in public goods may be positively impacting human capital, labor productivity, and wages in the municipalities for all workers and firms today.

The lack of data informing the name of workers before 2004 makes it challenging to investigate this mechanism directly. Looking at some correlations in our data, we observe a positive association between the concentration of descendants and general measures of public goods and infrastructure like the share of families with access to electricity or connection to the sewage system. We do not find the same pattern for educational infrastructure, however, which could be more relevant to our context. A few indicators, like the number of teachers and schools per 1,000 pupils in a municipality, actually correlate negatively with the concentration of descendants (results not shown). Future work can investigate further the relationship between the concentration of descendants (past and present) and the provision of public goods. For now, however, we are willing to discard this mechanism as one of the main explanations for our results.

8 Robustness checks

In the appendix, we present a series of robustness checks divided into three tables. In each table, we present results for ten different versions of our individual-level regression. In these versions, we vary specifications, samples, and controls to assess the robustness of our results to these choices. All robustness checks are performed for the Spread sample only, in 2010. In Table A6, we show results for OLS regressions with a linear specification. In A7, we show results for OLS regressions with the quadratic specification we use in most of our analysis. Finally, in Table A8, we show results for the second stage of IV regressions with a linear specification.

Column one in all tables brings the result we obtain with our standard sample and specification. Those are comparable to column (6) in Tables 6 (OLS linear) and 7 (OLS quadratic), and to column (3) in Table 9 (IV linear). In column (2) in all tables, instead of keeping, for each worker, only their last occurrence within the year, we keep only those employed on December 31st—the date on which employers file their RAIS reports. In columns (3) and (4), we exclude individuals and municipalities at the ends of the distribution of the number of individual observations per municipality to check the sensibility of our results to municipalities with a "very small" or "very large" number

of individual observations. In column (3), we keep only individual observations within the 10th and 90th percentile of the distribution of individual observation per municipality weighted by individual. That results in dropping all individuals working in municipalities with less than 1,400 and more than 84,539 observations. In column (4), we choose a more arbitrary cutoff and drop all municipalities with less than a hundred and more than a hundred thousand observations.

In columns (5) and (6), we turn to regression controls. First, we remove individual-level controls that are potentially endogenous. Inspection of the data shows that descendants are generally more educated and more likely to work in certain industries and occupations. Therefore, we remove education, industry, and occupation dummies from the set of individual-level controls used in our regressions. Next, in column (6), we show results for regressions that add potential yields as controls. Specifically, we use the potential yields for soybean and maize under low technology and the difference in these potential yields when switching from low to high technology; the same variables are used in other studies that investigate the transformations tied to the expansion of the agricultural frontier in Brazil (Bustos et al., 2016; Bragança et al., 2021).

In columns (7) and (8) we perform the checks mentioned at the end of Section 2.3. We restrict the Spread sample first to the set of municipalities in the North and Center-West regions only, which match the definition of the "West" of Brazil in Pellegrina and Sotelo (2019). In column (8), we restrict the Spread sample to the municipalities in the regions Center-West, North, and Northeast that match the definition of "frontier municipalities" in Bustos et al. (2016).

Finally, in columns (9) and (10), we split our sample of individuals according to their ancestry. We keep only locals in column (9) and only descendants in column (10). The descendant dummy and its interaction with the concentration of descendants are, therefore, removed. With this exercise, we gain a simpler version of the IV regression with a single endogenous regressor.

Results across specifications and tables in these robustness checks are generally consistent with our main results. They show the same general pattern of a positive association between the concentration of descendants and wages, with the association being stronger for locals. The marginal effect for locals stays around 2.10%, varying up and down by no more than 35% of its original magnitude across specifications. The marginal effect for descendants fluctuates more and loses statistical significance in some cases. However, it still exhibits a similar pattern of positive results that are always below the effect for locals. The main exception to this rule is column (7) in each table, in which we restrict the sample of municipalities to the regions Center-West and North only. Results are generally smaller and less precise in this restricted sample. We note, however, that the

number of municipalities and individual observations in this restricted sample is only 34% and 42% of what we had in the original Spread sample, respectively. Therefore, losses in precision are expected. Moreover, the median concentration of descendants in this restricted sample is higher: 3.91% versus 2.21% in the original Spread sample. Smaller marginal effects for this restricted sample align with the idea that lower levels of the concentration correspond to higher returns to an increase in the presence of descendants.

The results in our robustness checks also show a positive and significant wage premium for descendants. It is generally around 10–11% in OLS regressions and approximately 14% in the IV regressions. The only exception is column (5) in each table, in which we remove individual-level controls such as education and occupation. Because descendants are more educated and have more white-collar occupations than locals, on average, their wage premium increases substantialy in a specification that does not account for these characteristics.

9 Conclusion

Our study uses a surname-based classification of ancestries to identify descendants of immigrants in Brazil and bring together two historical events that shaped the economy of the country. In the first event, massive international immigration in 1850–1960 increased the size and the diversity of the labor force in Brazil, but immigrants and their impacts concentrated in the South and Southeast regions. In the second, strong internal migration from southern to western and northern Brazil starting in the 1960s spread historical immigrants and their descendants throughout the country, and may have spread their impacts as well.

We find that internal migration spread the impacts of historical international immigration in Brazil. The concentration of descendants of historical immigrants in municipalities in northern and central Brazil is positively associated with several indicators of economic development today, in particular with higher wages. We find a wage premium of approximately 12% for descendants (column 6, Table 7). This premium suggests descendants are either more productive than locals or that they are the scarce production factor in a setting where descendant and local labor are imperfect substitutes in the production function of the firms. In accordance with a theoretical framework where there are imperfect substitutions of labor and constraints on the firm's technology choice due to the very low concentration of one labor type (descendants) in some locations, we find that an increase in the concentration of descendants generates positive wages spillovers to both locals and descendants. One additional percentage point in the concentration of

descendants in a municipality corresponds to a wage increase of 1% for descendants and 2% for non-descendants. An inspection of heterogeneity in our results shows that wage spillovers are particularly strong for low-skilled men working in the agricultural sector. This last finding matches anecdotal evidence in Brazil that asserts the role of descendants of historical immigrants in the expansion of the agricultural frontier, from the South and Southeast to the central and northern regions of the country.

Our results hold in an instrumental variables strategy that leverages geographic characteristics of the municipalities and their distance to the injection points of historical immigration in Brazil to deal with the possible endogeneity of the concentration of descendants in labor markets today. Our results are also robust to a series of different samples and specifications.

In our discussion of results, we favor labor complementarities between descendants and locals discussed in our theoretical framework as the explanation for all results. We consider other mechanisms and argue that they do not explain our results as well as labor complementarities do. However, there are other mechanisms, some of which cannot be fully investigated with the data we currently have, that are plausible and deserve further investigation. In particular, it is important to investigate the role of descendants as firm owners. Anecdotal accounts on the expansion of the agricultural frontier in Brazil mention many cases of internal migrants coming from the southern regions (the regions with a strong presence of injection points of historical non-Iberian immigration). These internal migrants would often sell their land holdings and other assets before moving to the frontier, which overlaps with a significant portion of our Spread sample. Descendants, therefore, might have brought not only their human capital but physical capital as well to the municipalities in our sample, and they may have used this capital to start and operate businesses. It is possible that, as firm owners, descendants may be increasing the stock of capital and, consequently, the labor productivity and wages of the workers in these municipalities.

It is also important to investigate further the nature of the sizeable and persistent wage premium we find for descendants. Even after controlling for an extensive set of individual-level controls—education, experience, race, occupation, industry, and others—we find that descendants earn around 12% more than locals. Differences in productivity and relative scarcity with imperfect substitution are the explanations we consider for the premium in this study. Other explanations, however, such as surname-based discrimination or relationships between ancestry and human capital quality, access to high-quality education, and connection to professional and business networks remain possible and should be explored in future research.

References

Abramitzky, R., Boustan, L. P. and Eriksson, K. (2014), 'A nation of immigrants: Assimilation and economic outcomes in the Age of Mass Migration', *Journal of Political Economy* **122**(3), 467–506.

Acemoglu, D., Johnson, S. and Robinson, J. A. (2001), 'The colonial origins of comparative development: An empirical investigation', *American Economic Review* **91**(5), 1369–1401.

Alves, E. (2016), EMBRAPA: Institutional building and technological innovations required for Cerrado agriculture, *in* A. Hosono, C. M. C. da Rocha and Y. Hongo, eds, 'Development for sustainable agriculture: The Brazilian Cerrado', Palgrave Macmillan UK, London, pp. 139–156.

Alves, V. E. L. (2005), 'A mobilidade sulista e a expansão da fronteira agrícola brasileira', *Agrária (São Paulo. Online)* **1**(2), 40–68.

Andrews, G. R. (1991), *Blacks & Whites in São Paulo, Brazil, 1888-1988*, Univ of Wisconsin Press, Madison.

Borjas, G. J. (1992), 'Ethnic capital and intergenerational mobility', *The Quarterly Journal of Economics* **107**(1), 123–150.

Borjas, G. J., Grogger, J. and Hanson, G. H. (2008), 'Imperfect substitution between immigrants and natives: A reappraisal', *NBER Working Paper No. 13887* .

Borjas, G. J. and Katz, L. F. (2007), The evolution of the Mexican-born workforce in the United States, *in* G. J. Borjas, ed., 'Mexican immigration to the United States', University of Chicago Press, pp. 13–56.

Bragança, A. (2018), 'The causes and consequences of agricultural expansion in MATOPIBA', *Revista Brasileira de Economia* **72**, 161–185.

Bragança, A., Assunção, J. and Ferraz, C. (2021), 'Human capital and technology adoption: Evidence from Brazil's green revolution', *Working Paper* .

Bustos, P., Caprettini, B. and Ponticelli, J. (2016), 'Agricultural productivity and structural transformation: Evidence from Brazil', *American Economic Review* **106**(6), 1320–65.

Bustos, P., Garber, G. and Ponticelli, J. (2017), 'Capital accumulation and structural transformation', *Working Paper* .

Card, D., Cardoso, A. R., Heining, J. and Kline, P. (2018), 'Firms and labor market inequality: Evidence and some theory', *Journal of Labor Economics* **36**(S1), S13–S70.

Carvalho, Filho, I. d. and Monasterio, L. M. (2012), 'Immigration and the origins of regional inequality: Government-sponsored European migration to southern Brazil before World War I', *Regional Science and Urban Economics* **42**(5), 794–807.

Dell, M. (2010), 'The persistent effects of Peru's mining mita', *Econometrica* **78**(6), 1863–1903.

Dix-Carneiro, R. and Kovak, B. K. (2017), 'Trade liberalization and regional dynamics', *American Economic Review* **107**(10), 2908–46.

dos Santos, S. A. (2002), 'Historical roots of the "whitening" of Brazil', *Latin American Perspectives* **29**(1), 61–82.

Droller, F. (2017), 'Migration, population composition and long run economic development: Evidence from settlements in the Pampas', *The Economic Journal* **128**(614), 2321–2352.

Ehrl, P. and Monasterio, L. M. (2017), 'Inherited cultural diversity and wages in Brazil', *Working Paper* .

Firpo, S. P. and de Pieri, R. G. (2018), 'The labor market in Brazil, 2001-2015', *IZA World of Labor* **1**(441).

Franceschetto, C. (2014), *Imigrantes: Base de dados da imigração estrangeira no Espírito Santo nos séculos XIX e XX*, Arquivo Público do Estado do Espírito Santo, Vitória.

Gerard, F., Lagos, L., Severnini, E. and Card, D. (2018), 'Assortative matching or exclusionary hiring? The impact of firm policies on racial wage differences in Brazil', *NBER Working Paper No. 25176* .

Glaeser, E. L. (1999), 'Learning in cities', *Journal of Urban Economics* **46**(2), 254–277.

Hatton, T. J. and Williamson, J. G. (1998), *The Age of Mass Migration: Causes and economic impact*, Oxford University Press.

Hosono, A. and Hongo, Y. (2012), 'Cerrado agriculture: A model of sustainable and inclusive development', *Tokyo: Japan International Cooperation Agency Research Institute* .

IBGE (2007), *Brasil: 500 anos de povoamento*, Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

Imbens, G. W. and Wooldridge, J. M. (2009), 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature* **47**(1), 5–86.

Jepson, W. (2006*a*), 'Private agricultural colonization on a Brazilian frontier, 1970–1980', *Journal of Historical Geography* **32**(4), 839–863.

Jepson, W. (2006*b*), 'Producing a modern agricultural frontier: Firms and cooperatives in eastern Mato Grosso, Brazil', *Economic Geography* **82**(3), 289–316.

Lopes, D. A. F., Silva, Filho, G. A. and Monasterio, L. M. (2017), 'Culture, institutions and school achievement in Brazil', *Working Paper* .

Monasterio, L. M. (2017), 'Surnames and ancestry in Brazil', *PloS One* **12**(5), e0176890.

Monteiro, N. d. G. (1973), *Imigração e colonizaçã em Minas 1889-1930*, Imprensa Oficial, Belo Horizonte.

Morten, M. and Oliveira, J. (2016), 'Paving the way to development: Costly migration and labor market integration', *NBER Working Paper No. 22158* .

Nishikawa, R. B. (2015), As colônias de imigrantes na província do Paraná, 1854-1889, PhD thesis, Universidade de São Paulo, São Paulo.

Nunn, N. (2009), 'The importance of history for economic development', *Annual Review of Economics* **1**(1), 65–92.

Ottaviano, G. I. P. and Peri, G. (2012), 'Rethinking the effect of immigration on wages', *Journal of the European Economic Association* **10**(1), 152–197.

Pellegrina, H. (2020), 'Trade, productivity, and the spatial organization of agriculture: Evidence from Brazil', *Working Paper* .

Pellegrina, H. S. and Sotelo, S. (2019), 'Migration, specialization, and trade: Evidence from the Brazilian march to the west', *Working Paper* .

Pérez, S. (2019), 'Southern (American) hospitality: Italians in Argentina and the US during the Age of Mass Migration', *NBER Working Paper No. 26127* .

Piazza, W. F. (1983), *Santa Catarina: sua história*, Editora da UFSC, Florianópolis.

Rezende, G. C. d. (2002), 'Ocupação agrícola e estrutura agrária no cerrado: O papel do preço da terra, dos recursos naturais e da tecnologia', *IPEA Discussion Papers* .

Riley, S. J., DeGloria, S. D. and Elliot, R. (1999), 'A terrain ruggedness index that quantifies topographic heterogeneity', *Intermountain Journal of Sciences* **5**(1-4), 23–27.

Rocha, R., Ferraz, C. and Soares, R. R. (2017), 'Human capital persistence and development', *American Economic Journal: Applied Economics* **9**(4), 105–36.

Sánchez-Alonso, B. (2007), 'The other Europeans: Immigration into Latin America and the international labour market (1870–1930)', *Revista de Historia Economica-Journal of Iberian and Latin American Economic History* **25**(3), 395–426.

Santos, R. J. (2008), *Gaúchos e Mineiros do Cerrado: Metamorfoses das diferentes temporalidades e lógicas sociais*, EDUFU, Uberlândia.

Spolaore, E. and Wacziarg, R. (2013), 'How deep are the roots of economic development?', *Journal of Economic Literature* **51**(2), 325–69.

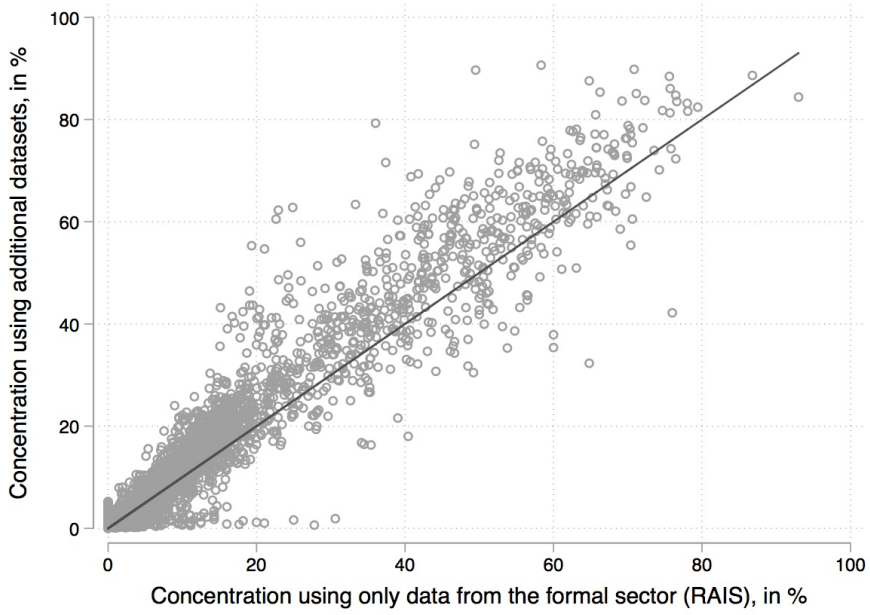
Valencia Caicedo, F. (2018), 'The Mission: Human capital transmission, economic persistence, and culture in South America', *The Quarterly Journal of Economics* **134**(1), 507–556.

Vigna, B. and Rocha, R. (2019), 'Diversity and development', *Working Paper* .

Wagner, C. and Bernardi, R. (1995), *O Brasil de Bombachas*, L&PM.

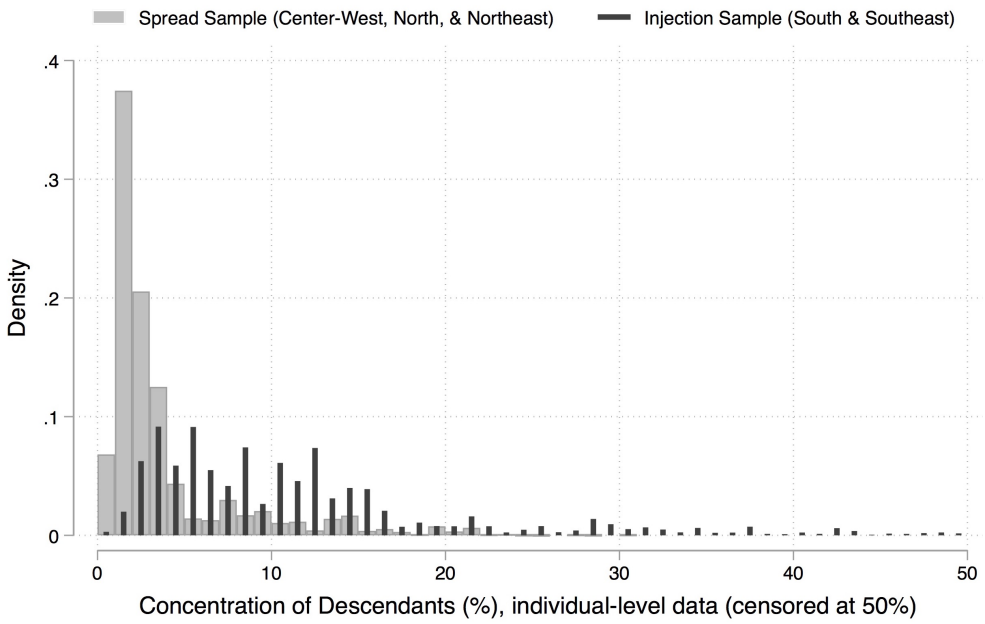
Appendix

Figure A1: Concentration of descendants in the municipalities calculated using different datasets, 2010 (National sample)



Note: The concentration of descendants in the horizontal axis uses only data from RAIS, while the one in the vertical axis includes also information from the *Cadastro Único* and the *Base Sócios* datasets. The graphs also include a 45° line.

Figure A2: Distributions of the concentration of descendants in the municipalities (averaged across individuals), 2010



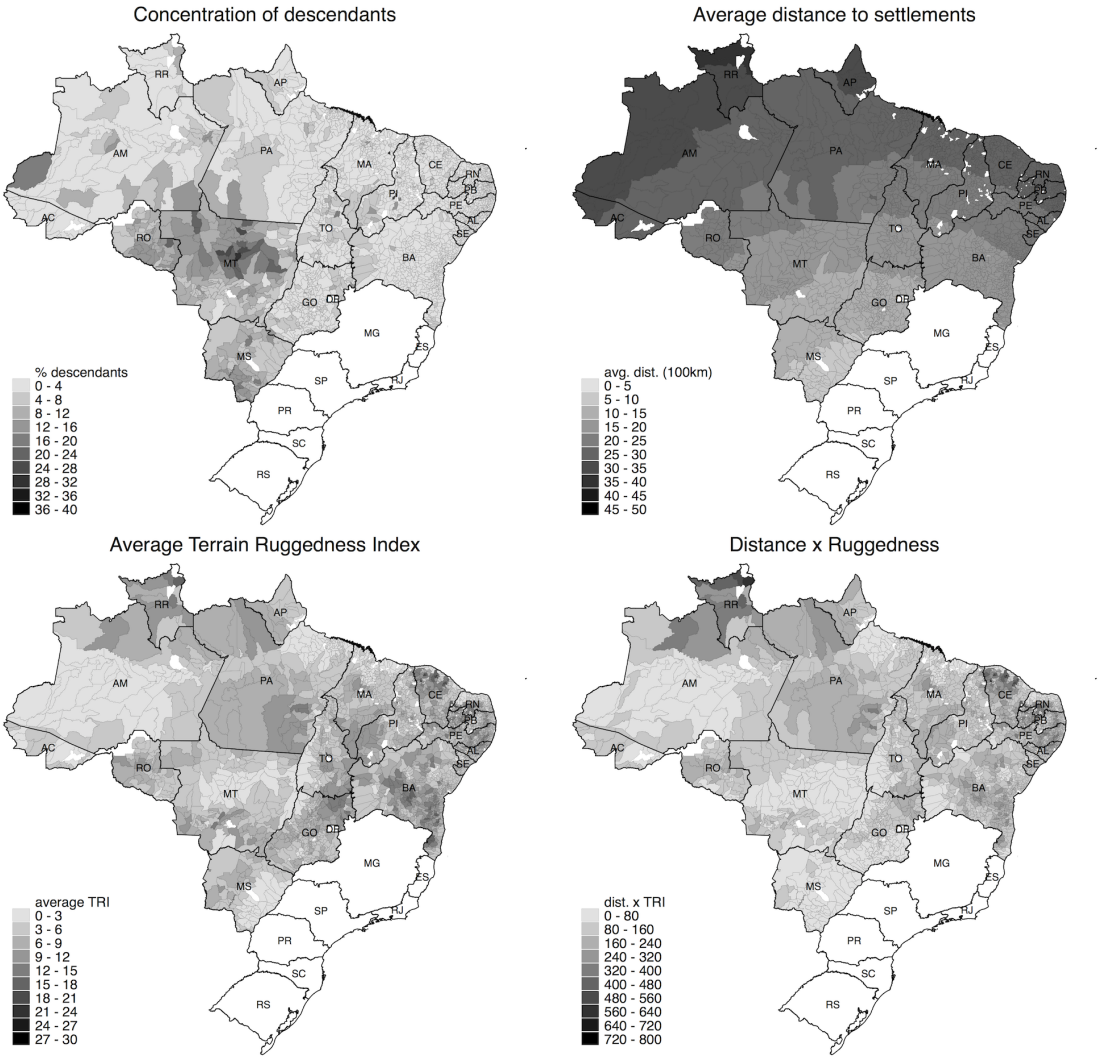
Notes: The concentration of descendants is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality in 2010. The concentration is averaged by the number of individual observations in each municipality. The Injection sample considers municipalities in all states of the regions South and Southeast. The Spread sample considers municipalities in the states of the regions Center-West, North, and Northeast. Both samples exclude state capitals and municipalities with less than five individual observations in the RAIS data in 2010 or with a missing value for the concentration of descendants. For the Injection sample, the average concentration is 11.94%, the median is 8.85%, and the standard deviation is 10.42%. For the Spread sample, the average concentration is 3.83%, the median is 2.21%, and the standard deviation is 4.14%.

Table A1: OLS Regressions: Human Development Index at the municipality-level, 2010 (Spread sample)

Outcomes:	HDI: Total (1)	HDI: Income (2)	HDI: Education (3)	HDI: Health (4)
Concentration of descendants (%)	0.0241***	0.0228***	0.0223***	0.0149***
	(0.0063)	(0.0059)	(0.0067)	(0.0054)
R^2 (adjusted)	0.54	0.60	0.42	0.54
N (municipalities)			2,624	

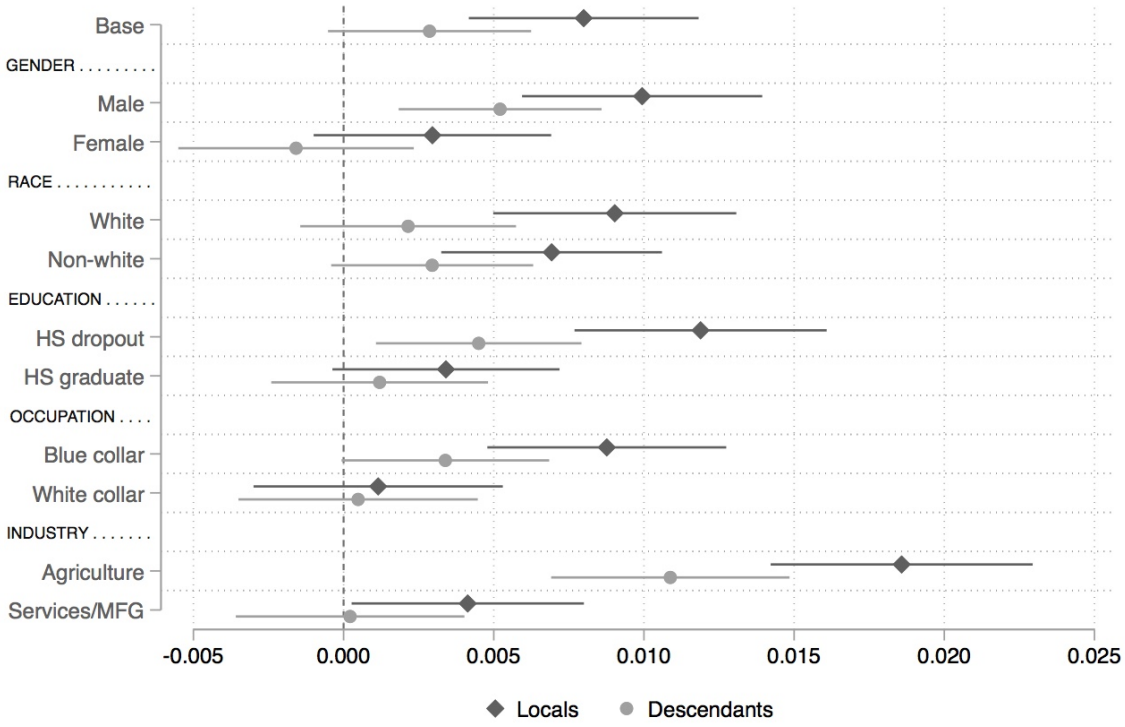
Notes: The dependent variable in the first column is the Human Development Index, which uses data from the 2010 Brazilian population census. In the remaining columns, the dependent variables are the three components that form the Human Development Index of municipalities: income, education, and health. The income component reflects income per capita, the education component combines the education level in the adult population with the educational flow of the younger population, and the health component reflects life expectancy at birth. All dependent variables were standardized to facilitate the interpretation of regression coefficients. The concentration of descendants is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality. All specifications include state fixed effects and the following municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, population density in 1950, municipality area (log), distance to the state capital (log), average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Robust standard errors in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

Figure A3: Concentration of descendants, average distance to historical immigrants' settlements, average Terrain Ruggedness Index, and their interaction (the excluded instrument)



Notes: The first map (top left) shows the concentration of descendants in the study region in 2010, the measure we instrument for in the IV regressions. The second map (top right) shows the average of the distances from each municipality to historical settlements of non-Iberian immigrants in the states of São Paulo and Rio Grande do Sul. The third map (bottom left) shows the average Terrain Ruggedness Index for the municipalities in our sample. Finally, the fourth map (bottom right) shows the actual excluded instrument, the interaction between the non-centered normalized distance and ruggedness measures. The intervals for the scale in each graph are approximately equal to one standard deviation of each variable.

Figure A4: Heterogenity in the marginal effects of the concentration of descendants on wages, 2010 (linear specification, Spread sample)



Note: All point estimates and 95% confidence intervals shown in the figure come from regressions of the log hourly wage on the concentration of descendants, the square of the concentration, and their interactions with the descendant dummy and a binary indicator. The binary indicators split the samples into two groups according to some characteristics of the worker or her job. The regression specifications follow the format of the specification shown in column (6) of table 6 and include individual-level and municipality-level controls and state fixed effects. Individual-level controls are adjusted according to the binary variable use to split the sample into two groups and elicit heterogeneities.

Table A2: OLS Regressions: Indicators of Socioeconomic Development at the municipality-level, 2010 (Spread sample)

	(1)	(2)	(3)
Panel A: Income outcomes	Log of income per capita	Unemployment rate	Gini index
Concentration of descendants (%)	0.0228***	-0.0067	0.0090
	(0.0059)	(0.0073)	(0.0063)
R^2 (adjusted)	0.60	0.26	0.35
Panel B: Education outcomes	Years of schooling	HS degree or higher	Adult literacy rate
Concentration of descendants (%)	0.0075	0.0207***	0.0352***
	(0.0069)	(0.0062)	(0.0056)
R^2 (adjusted)	0.38	0.33	0.62
Panel C: Health outcomes	Life expectancy	Infant mortality	Fertility
Concentration of descendants (%)	0.0149**	-0.0117*	-0.0076
	(0.0054)	(0.0051)	(0.0058)
R^2 (adjusted)	0.54	0.55	0.50
Panel D: Formal sector outcomes	Log earnings (Census)	Log wage (RAIS)	Munic. wage premia
Concentration of descendants (%)	0.0154*	0.0624***	0.0295***
	(0.0066)	(0.0140)	(0.0073)
R^2 (adjusted)	0.41	0.26	0.40
N (municipalities)		2,624	

Notes: The socio-economic indicators in panels A, B, and C were retrieved from the Atlas Brazil project (atlasbrasil.org.br/2013/en/). They reflect information from the 2010 Brazilian population census. The log earnings of the formally hired (first row in panel D) were retrieved from the 2010 census. The average log wage and the average local wage premia shown in the last two rows in panel D were calculated by the authors using information from RAIS in 2010. The measure of log earnings in the formal sector considers only hired workers with a formal labor contract. The municipality wage premium is the municipality fixed effect in a log-wage regression that includes the same extensive set of individual-level covariates used later in the main analyses. All dependent variables were standardized to facilitate the interpretation of regression coefficients. The concentration of descendants is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality. All specifications include state fixed effects and the following municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, population density in 1950, municipality area (log), distance to the state capital (log), average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Robust standard errors in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

Table A3: IV Regressions (1st stage): Concentration of descendants in the municipalities on the excluded instruments, 2010 (linear specification, Spread sample)

	(1)	(2)	(3)
Panel A: Concentration of descendants (%)			
Terrain ruggedness x Distance to settlements	0.0499	0.0220	0.0199
	(0.0079)***	(0.0074)***	(0.0058)***
Terrain ruggedness x Distance x Descendant	-0.0126	-0.0079	-0.0063
	(0.0039)***	(0.0019)***	(0.0014)***
Terrain Ruggedness Index	-1.4198	-0.6067	-0.5719
	(0.2204)***	(0.1964)***	(0.1521)***
Average distance to settlements (100km)	-0.5013	-0.0895	-0.0925
	(0.0559)***	(0.0852)	(0.0658)
Descendant dummy	4.4557	1.8780	1.4313
	(0.6368)***	(0.2668)***	(0.1851)***
R^2 (adjusted)	0.33	0.77	0.83
S-W multivariate F test of excluded instruments	40.97	11.54	14.14
Panel B: Concentration x Descendant			
Terrain ruggedness x Distance to settlements	0.0038	0.0024	0.0022
	(0.0007)***	(0.0011)**	(0.0008)***
Terrain ruggedness x Distance x Descendant	-0.0388	-0.0384	-0.0381
	(0.0060)***	(0.0059)***	(0.0058)***
Terrain Ruggedness Index	-0.0884	-0.0431	-0.0432
	(0.0182)***	(0.0271)	(0.0220)**
Average distance to settlements (100km)	-0.0205	0.0254	0.0119
	(0.0038)***	(0.0135)*	(0.0104)
Descendant dummy	11.8943	11.6714	11.6071
	(1.0044)***	(0.9759)***	(0.9651)***
R^2 (adjusted)	0.68	0.69	0.70
S-W multivariate F test of excluded instruments	46.45	43.72	46.63
N (workers)		6,030,247	
Clusters (municipalities)		2,624	
Individual-level controls	Y	Y	Y
State fixed effects		Y	Y
Municipality-level controls			Y

Notes: The dependent variable in panel A (first endogenous regressor) is the concentration of descendants. In panel B, the dependent variable (second endogenous regressor) is the interaction of the concentration of descendants and the individual-level descendant dummy. The instrument for the first endogenous regressor is the interaction of the average Terrain Ruggedness Index of the municipality and the average of the distance to historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul. For the second endogenous regressor, we use the interaction of the first instrument with the descendant dummy. The terms used in the interactions are included as controls in all specifications and their coefficients are reported in both panels shown here. The average distance to settlements (expressed in 100km) is an average of the distances from the economic center of a given municipality to the economic center of all municipalities with a non-Iberian historical settlement in the states of Rio Grande do Sul and São Paulo (Carvalho and Monasterio, 2012; Rocha et al., 2017). The average Terrain Ruggedness Index (TRI) was calculated using the software QGIS, the methodology by Riley, DeGloria and Elliot (1999), and topographical data from the Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) from the US Geological Survey (resolution: 15 arc-seconds). The index calculates the difference in elevation between a grid cell and its surroundings. Higher values of the index correspond to more rugged terrain. We average the index values for all grid cells in a municipality to obtain its average TRI. The individual-level and municipality-level controls are the same used in the second stage. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

Table A4: OLS Regressions: Log hourly wages at the individual level on the concentration of descendants in different occupation levels, 2010 (Spread sample)

Outcome:	Log hourly wage					
Specification:	Linear			Quadratic		
Occupaton level:	All	Staff	Management	All	Staff	Management
	(1)	(2)	(3)	(4)	(5)	(6)
Concentration of descendants (%)	0.0080	0.0076	0.0052	0.0248	0.0218	0.0109
	(0.0020)***	(0.0019)***	(0.0008)***	(0.0052)***	(0.0052)***	(0.0016)***
Concentration x Descendant	-0.0051	-0.0053	-0.0023	-0.0135	-0.0137	-0.0044
	(0.0009)***	(0.0009)***	(0.0004)***	(0.0024)***	(0.0024)***	(0.0009)***
Concentration squared				-0.0007	-0.0006	-0.0001
				(0.0002)***	(0.0002)***	(0.0000)***
Concentration squared x Descendant				0.0005	0.0005	0.0001
				(0.0001)***	(0.0001)***	(0.0000)***
Descendant dummy	0.0959	0.0952	0.0911	0.1153	0.1146	0.0996
	(0.0066)***	(0.0064)***	(0.0066)***	(0.0094)***	(0.0092)***	(0.0080)***
R^2 (adjusted)	0.44	0.44	0.44	0.44	0.44	0.44
N (workers)	6,027,379					
Clusters (municipalities)	2,502					
Marginal effects of the concentration of descendants (one additional percentage point over the median)						
for Locals	0.0080	0.0052	0.0076	0.0209	0.0093	0.0185
	(0.0020)***	(0.0019)***	(0.0008)***	(0.0043)***	(0.0042)***	(0.0013)***
for Descendants	0.0029	0.0030	0.0023	0.0099	0.0057	0.0072
	(0.0017)*	(0.0018)	(0.0008)***	(0.0039)**	(0.0041)*	(0.0013)***

Notes: The dependent variable in all specifications is the log of the worker's hourly wage. The concentration of descendants in columns (1) and (4) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality, considering workers of all occupation levels. In columns (2) and (5), only workers in the staff are considered, and in columns (3) and (6) only those working as managers and directors are considered in the computation of the concentration. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. The median value of the concentration of descendant considering workers in all occupation levels is 2.21% (avg. 3.83% and s.d. 4.14%), considering only the staff it is 2.07% (avg. 3.58% and s.d. 3.93%), and considering only managers and directors it is 5.54% (avg. 8.61% and s.d. 8.83%). All specifications include year fixed effects, municipality-level controls, and individual-level controls. Individual-level controls: dummy variables for gender and categories of age, education, race, job tenure, occupation, firm size (number of employees), and industry. Municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

Table A5: OLS Regressions: Log hourly wages at the individual level on the concentration of descendants with different population controls, 2010 (quadratic specification, Spread sample)

Outcome:	Log hourly wage				
Population control:	Log population	None	Pop. density (1950)	Log nr. obs.	% formal
	(1)	(2)	(4)	(3)	(5)
Concentration of descendants (%)	0.0247	0.0212	0.0214	0.0219	0.0174
	(0.0052)***	(0.0053)***	(0.0052)***	(0.0053)***	(0.0054)***
Concentration x Descendant	-0.0135	-0.0124	-0.0127	-0.0126	-0.0120
	(0.0024)***	(0.0024)***	(0.0024)***	(0.0024)***	(0.0024)***
Concentration squared	-0.0007	-0.0006	-0.0006	-0.0006	-0.0005
	(0.0002)***	(0.0002)***	(0.0002)***	(0.0002)***	(0.0002)**
Concentration squared x Descendant	0.0005	0.0004	0.0004	0.0004	0.0004
	(0.0001)***	(0.0001)***	(0.0001)***	(0.0001)***	(0.0001)***
Descendant dummy	0.1152	0.1126	0.1131	0.1130	0.1117
	(0.0094)***	(0.0094)***	(0.0095)***	(0.0093)***	(0.0094)***
R^2 (adjusted)	0.44	0.44	0.44	0.44	0.44
N (workers)			6,030,247		
Clusters (municipalities)			2,624		
Marginal effects of the concentration of descendants (one additional percentage point over the median)					
for Locals	0.0208	0.0180	0.0182	0.0186	0.0148
	(0.0042)***	(0.0043)***	(0.0043)***	(0.0044)***	(0.0044)***
for Descendants	0.0098	0.0078	0.0077	0.0083	0.0048
	(0.0039)**	(0.0041)*	(0.0041)*	(0.0042)**	(0.0042)

Notes: The dependent variable in all specifications is the log of the worker's hourly wage. The concentration of descendants (expressed in percentage points) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality-year. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. The median value of the concentration of descendants is 2.21% (avg. 3.83% and s.d. 4.14%). Individual-level controls: dummy variables for gender and categories of age, education, race, job tenure, occupation, firm size (number of employees), and industry. Municipality-level controls (other than the population control listed in each column): historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

Table A6: Robustness checks (OLS): Log hourly wages at the individual level on the concentration of descendants in the municipalities, 2010 (linear specification, Spread sample)

Outcome:	Log hourly wage									
Specification/check:	Basic specification	Employed in Dec 31st	Nr. of individual obs. (p10, p90)	Nr. of individual obs. (100, 100K)	Fewer controls (individual)	More controls (yields)	Spread sample wo/ Northeast	Agr. Frontier BCP (2016)	Split samples: Locals	Split samples: Descendants
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Concentration of descendants (%)	0.0080	0.0101	0.0063	0.0085	0.0095	0.0065	0.0037	0.0075	0.0077	
	(0.0020)***	(0.0019)***	(0.0016)***	(0.0022)***	(0.0026)***	(0.0020)***	(0.0016)**	(0.0025)***	(0.0020)***	
Concentration x Descendant	-0.0051	-0.0054	-0.0060	-0.0052	-0.0078	-0.0051	-0.0030	-0.0048		0.0050
	(0.0009)***	(0.0009)***	(0.0008)***	(0.0010)***	(0.0011)***	(0.0009)***	(0.0008)***	(0.0010)***		(0.0015)***
Descendant dummy	0.0958	0.1078	0.0907	0.0980	0.2172	0.0953	0.0817	0.0913		
	(0.0066)***	(0.0073)***	(0.0065)***	(0.0072)***	(0.0110)***	(0.0065)***	(0.0070)***	(0.0082)***		
R^2 (adjusted)	0.44	0.46	0.46	0.44	0.12	0.44	0.42	0.44	0.43	0.52
N (workers)	6,030,247	4,384,112	1,611,135	5,633,616	6,030,247	6,030,247	2,508,518	3,389,561	5,797,821	232,426
Clusters (municipalities)	2,624	2,604	2,110	1,898	2,624	2,624	903	1,404	2,624	2,136
Marginal effects of the concentration of descendants (one additional percentage point)										
for Descendants	0.0029	0.0047	0.0003	0.0033	0.0017	0.0014	0.0007	0.0026		0.0050
	(0.0017)*	(0.0016)***	(0.0014)	(0.0019)*	(0.0025)	(0.0017)	(0.0015)	(0.0020)		(0.0015)***

Notes: The dependent variable in all specifications is the log of the worker's hourly wage. The concentration of descendants (expressed in percentage points) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality-year. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. All specifications include year fixed effects, municipality-level controls, and individual-level controls. Individual-level controls: dummy variables for gender and categories of age, education, race, job tenure, occupation, firm size (number of employees), and industry. Municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

Table A7: Robustness checks (OLS): Log hourly wages at the individual level on the concentration of descendants in the municipalities, 2010 (quadratic specification, Spread sample)

Outcome:	Log hourly wage									
Specification/check:	Basic specification	Employed in Dec 31st	Nr. of individual obs. (p10, p90)	(100, 100K)	Fewer controls (individual)	More controls (yields)	Spread sample wo/ Northeast	Agr. Frontier BCP (2016)	Split samples Locals	Descendants
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Concentration of descendants (%)	0.0247	0.0282	0.0187	0.0277	0.0340	0.0226	0.0110	0.0198	0.0246	
	(0.0052)***	(0.0049)***	(0.0043)***	(0.0057)***	(0.0071)***	(0.0052)***	(0.0051)**	(0.0066)***	(0.0053)***	
Concentration x Descendant	-0.0135	-0.0126	-0.0105	-0.0141	-0.0175	-0.0137	-0.0078	-0.0110		0.0143
	(0.0024)***	(0.0023)***	(0.0019)***	(0.0025)***	(0.0036)***	(0.0024)***	(0.0020)***	(0.0025)***		(0.0046)***
Concentration squared	-0.0007	-0.0007	-0.0005	-0.0008	-0.0010	-0.0007	-0.0003	-0.0005	-0.0007	
	(0.0002)***	(0.0002)***	(0.0001)***	(0.0002)***	(0.0003)***	(0.0002)***	(0.0002)*	(0.0002)**	(0.0002)***	
Concentration squared x Descendant	0.0005	0.0004	0.0003	0.0005	0.0005	0.0005	0.0002	0.0003		-0.0003
	(0.0001)***	(0.0001)***	(0.0001)***	(0.0001)***	(0.0002)***	(0.0001)***	(0.0001)**	(0.0001)***		(0.0001)**
Descendant dummy	0.1152	0.1238	0.0969	0.1183	0.2376	0.1159	0.0992	0.1071		
	(0.0094)***	(0.0102)***	(0.0074)***	(0.0101)***	(0.0158)***	(0.0094)***	(0.0093)***	(0.0110)***		
R^2 (adjusted)	0.44	0.46	0.46	0.44	0.12	0.44	0.42	0.44	0.43	0.52
N (workers)	6,030,247	4,384,112	1,611,135	5,633,616	6,030,247	6,030,247	2,508,518	3,389,561	5,797,821	232,426
Clusters (municipalities)	2,624	2,604	2,110	1,898	2,624	2,624	903	1,404	2,624	2,136
Marginal effects of the concentration of descendants (one additional percentage point over the median)										
for Locals	0.0208	0.0244	0.0161	0.0233	0.0283	0.0189	0.0084	0.0170	0.0208	
	(0.0042)***	(0.0041)***	(0.0037)***	(0.0047)***	(0.0058)***	(0.0043)***	(0.0037)**	(0.0055)***	(0.0043)***	
for Descendants	0.0098	0.0138	0.0069	0.0118	0.0137	0.0076	0.0026	0.0078		0.0095
	(0.0039)**	(0.0037)***	(0.0036)*	(0.0042)***	(0.0053)***	(0.0040)*	(0.0037)	(0.0050)		(0.0026)***
Median concentration of descendants (%)	2.21	2.24	2.03	2.18	2.21	2.21	3.91	2.31	2.15	6.90

Notes: The dependent variable in all specifications is the log of the worker's hourly wage. The concentration of descendants (expressed in percentage points) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality-year. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. All specifications include year fixed effects, municipality-level controls, and individual-level controls. Individual-level controls: dummy variables for gender and categories of age, education, race, job tenure, occupation, firm size (number of employees), and industry. Municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.

Table A8: Robustness checks (IV, 2nd stage): Log hourly wages at the individual level on the concentration of descendants in the municipalities, 2010 (linear specification, Spread sample)

Outcome:	Log hourly wage									
Specification/check:	Basic specification	Employed in Dec 31st	Nr. of individual obs. (p10, p90)	(100, 100K)	Fewer controls (individual)	More controls (yields)	Spread sample wo/ Northeast	Agr. Frontier BCP (2016)	Split samples Locals	Descendants
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Concentration of descendants (%)	0.0689	0.0631	0.0477	0.0774	0.0765	0.0831	0.0915	0.0458	0.0690	
	(0.0225)***	(0.0195)***	(0.0227)**	(0.0253)***	(0.0286)***	(0.0337)**	(0.0873)	(0.0206)**	(0.0224)***	
Concentration x Descendant	-0.0167	-0.0158	-0.0159	-0.0181	-0.0174	-0.0185	-0.0187	-0.0119		0.0717
	(0.0044)***	(0.0041)***	(0.0050)***	(0.0049)***	(0.0055)***	(0.0058)***	(0.0181)	(0.0042)***		(0.0455)
Descendant dummy	0.1362	0.1438	0.1185	0.1438	0.2359	0.1407	0.1682	0.1198		
	(0.0192)***	(0.0189)***	(0.0186)***	(0.0210)***	(0.0255)***	(0.0227)***	(0.1170)	(0.0211)***		
R^2 (centered)	0.40	0.43	0.43	0.39	0.07	0.38	0.28	0.43	0.39	0.44
N (workers)	6,030,247	4,384,112	1,611,135	5,633,616	6,030,247	6,030,247	2,508,518	3,389,561	5,797,821	232,426
Clusters (municipalities)	2,624	2,604	2,110	1,898	2,624	2,624	903	1,404	2,624	2,136
Underidentification (K-P LM Stat)	11.3537	12.1623	11.8065	9.7996	11.3132	6.9722	1.0238	11.7754	11.8432	2.2796
Weak identification (K-P Wald F Stat)	5.8903	6.2452	6.2585	5.1016	5.8494	3.4773	0.5162	6.7519	12.3639	2.1608
S-W multivariate F test of excluded instruments (weak identification)										
Concentration of descendants (%)	14.14	15.49	13.24	12.09	13.89	8.03	1.06	15.39	12.36	
Concentration x Descendant dummy	46.63	48.52	45.35	49.10	46.49	44.13	12.03	25.58		2.16
Marginal effects of the concentration of descendants (one additional percentage point)										
for Descendants	0.0522	0.0474	0.0318	0.0593	0.0591	0.0646	0.0728	0.0340		0.0717
	(0.0189)***	(0.0162)***	(0.0183)*	(0.0212)***	(0.0241)**	(0.0287)**	(0.0704)	(0.0171)**		(0.0455)

Notes: The dependent variable in all specifications is the log of the worker's hourly wage. The concentration of descendants (expressed in percentage points) is given by the percentage of workers with a non-Iberian surname in the formal workforce in each municipality-year. The descendant dummy is equal to one when the worker's surname is classified as non-Iberian. There are two endogenous regressors in all specifications: the concentration of descendants, and the interaction of this concentration with the descendant dummy. Likewise, there are two excluded instruments in the first stage of all specifications. The instrument for the first endogenous regressor (the concentration of descendants) is the interaction of the average Terrain Ruggedness Index of the municipality and the average of the distance to historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul. For the second endogenous regressor—the interaction of the concentration of descendants with the descendant—we use the interaction of the first instrument with the descendant dummy. The terms used in the interactions are included as controls in all specifications. All specifications include year fixed effects, municipality-level controls, and individual-level controls. Individual-level controls: dummy variables for gender and categories of age, education, race, job tenure, occupation, firm size (number of employees), and industry. Municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), distance to the state capital (log), average elevation and average Terrain Ruggedness Index of the municipality, the average distance of the municipality economic center to historical non-Iberian settlements in the states of Rio Grande do Sul and São Paulo, and dummies for biomes and soil types (dummies = 1 if 5% or more of municipality area is covered by soil type/biome). Standard errors clustered by municipality in parentheses. Stars denote: * p<0.10; ** p<0.05; *** p<0.01.