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

Eduardo Cenci† Philipp Ehrl‡ Daniel A. F. Lopes§ Leonardo M. Monasterio¶

November 14, 2025

**Abstract**

This paper investigates how the inter-regional spread of immigrants' descendants from the 1850-1960 period affect current labor market outcomes in Brazil. We use linked employer-employee data, apply a surname-based classification to identify workers with non-Iberian ancestry (i.e. descendants) and estimate wage regressions with individual fixed effects and an instrumental variable. The data reveal pronounced differences between municipalities and ancestry groups. We find that the concentration of descendants in interior municipalities of northern and center-western Brazil is positively associated with higher wages. These wage gains, however, accrue only to the predominant population group with Iberian surnames. The same pattern, but with much less pronounced wage gains is observed in municipalities with high descendant rates, i.e. those close to immigrant settlements in the past. Our results are in accord with a simple model in which descendants and non-descendants are imperfect substitutes in the production function.

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.
The analysis draws on confidential administrative data from the Brazilian Ministry of Labor and Employment, which cannot be shared. Interested researchers may request access through the Ministry. Upon acceptance, we can provide all other dataset components assembled from public sources, as well as the complete replication code.
Philipp Ehrl thanks the Federal District Research Foundation (FAP-DF) and the National Council for Scientific and Technological Development (CNPq) for their financial support (Grant Numbers 00193-00001794/2023-77, 304782/2024-6 and 406860/2023-8, respectively). Any errors are our own. The authors declare to have no conflict of interest.

†University of Utah, corresponding author (e.cenci@utah.edu).

‡Getúlio Vargas Foundation, School of Public Policy and Government, Brasília (FGV/EPPG).

§Observatório da Indústria (CNI).

¶Instituto de Pesquisa Econômica Aplicada (IPEA) and IDP.

# 1 Introduction

Historical events have influence on present economic development (Nunn, 2009; Spolaore and Wacziarg, 2013). A common theme in this 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 et al., 2001) to the so-called Age of Mass Migration in the late 19th and early 20th centuries (Hatton and Williamson, 1998). Many studies credit positive impacts to human capital, brought to the receiving countries by historical immigrants and transmitted across generations (Borjas, 1992; Rocha et al., 2017). Yet, settlers also profoundly shaped the institutions and thus the future fortune of the receiving regions, making it a challenging task to disentangle the effect of places from people (Galor et al., 2009; Acemoglu et al., 2014).

Much of this literature focuses either on the effects on the receiving country as a whole, or on outcomes caused by regional differences in immigrant exposure. Human capital, however, moves not only across but also within countries. Although the assimilation of foreigners advances over time, cultural traits, and production know-how are remarkably persistent across generations (Abramitzky et al., 2014; Valencia, 2018; Voth, 2021). Foreign settlers attract further immigrants through networks and, as their spatial diffusion is gradual, the incidence of the original inflow unfolds unevenly across regions and cohorts. Resulting long-run effects of mass immigration on other, initially unpopulated regions, however, are not yet well understood.

In this paper, we shed light on how the inter-regional spread of historical immigrants affect current labor market outcomes in Brazil. Specifically, we use linked employer-employee data from 2008 to 2019 and a surname-based machine learning algorithm to classify workers' ancestries[1]. We will distinguish between the predominant group of workers with Iberian surnames (also referred to as *locals*) and non-Iberian immigrants' descendants (*descendants* henceforth). The research question is then how the concentration of descendants affects the wage level in municipalities. Two crucial elements in the present identification strategy are: First, the focus on regions far from the historic settlements making the labor markets under analysis free of direct influence of the historic foreigners themselves. Only decades later our so-called spread sample was heavily populated by immigrants' assimilated ancestors and Brazilians with Iberian ancestry. Second, the immigration to Brazil ceased about 1960 and ever since has been negligible. Therefore, the share

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

of current foreigners is close to zero and implies that (male) workers with a non-Iberian surname are in fact Brazilians with immigrant ancestry.

Brazil presents an intriguing setting as migrants played a central role in two epochs that shaped its economy. First, about five million immigrants arrived in Brazil between 1850 and 1960 to work as laborers in the nascent manufacturing industries of the Southeast and as farmers in the fertile lands of the South. About half of these immigrants were of Italian, German, Syrian-Lebanese, Japanese, and other non-Iberian origins (IBGE, 2007). These numerous immigrants changed the profile of the Brazilian population, historically comprising Indigenous peoples alongside individuals from the Iberian Peninsula (Portugal and Spain) and their colonies. Potentially, these non-Iberian immigrants also changed the stock and the distribution of human capital in and close to their settlement locations (Carvalho and Monasterio, 2012; Ehrl and Monasterio, 2024; Rocha et al., 2017). Second, starting around 1960, millions of internal migrants left the coastal and southern parts of Brazil towards its interior in a "March to the West" (Pellegrina and Sotelo, forthcoming). This internal migration was incentivized by the construction of roads and the new capital Brasília in the interior of the country (Bird and Straub, 2020), 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 et al., 2015). 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.

Focusing on wages as the primary outcome variable permits an analysis at the individual level with more precise identification. 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 to all or only a specific group of workers. Using individual-level outcomes like wages, therefore, allows us to separate composition and spillover

effects. 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 leverage the richness of our panel data and include fixed effects at the individual level to account for effects from spatial sorting, productivity related characteristics, discrimination, and the direct impact of individual ancestry itself. We combine the fixed effects wage regressions with an instrumental variable approach as in Combes et al. (2008); De la Roca and Puga (2017); Ehrl and Monasterio (2024) to address remaining endogeneity issues from reverse causality and measurement error. The instrumental variable is based on the location of official colonial settlements and, alternatively, on the terrain ruggedness, i.e. the suitability of land to application of large-scale modern agriculture. We argue that these variables drive the expansion of immigrants descendants to the interior of the country but are unrelated to the general wage level.

The data show that the concentration of (non-Iberian) descendants in the local workforce is positively and significantly related to the average wage level of municipalities in the interior of northern and center-western Brazil, whereas there is no relation among the regions in the South, where immigrants concentrated historically. Still, there are remarkable differences between ancestry groups. Workers in the predominant group enjoy the wage gain, whereas those with a non-Iberian ancestry experience no measurable effect. In our preferred instrumental variable specification, one additional percentage point in the concentration of descendants corresponds to a wage increase of 3% for non-descendants.

We use a simple model with standard production function, labor demand and supply, borrowed from studies that investigate imperfect substitutions between groups of workers (Moretti, 2004) to discuss the mechanisms behind our results. The observations are in line with the model, when we assume that workers with Iberian and non-Iberian ancestry are imperfect substitutes. The existence of positive spillover effect from the presence of the scarce non-Iberian ancestors, factor complementary, and diminishing marginal returns can rationalize why a higher concentration of descendants generates zero wage effects within its proper group, while wage effects for locals are highly positive.

This study contributes to three strands of literature. First, we add to the literature on the long-term consequences of the Age of Mass Migration (Hatton and Williamson, 1998). Sev-

eral studies document positive link between European immigration and development in the US and Latin America (Sánchez-Alonso, 2007; Droller, 2017; Valencia, 2018; Sequeira et al., 2020). Those studies are usually constrained to the original sites of colonization. One rare example is Von Berlepsch and Rodríguez-Pose (2021) who show that destinations of second-generation US migrants currently have a higher income per capita. By focusing on how the descendants of historical immigrants spread their human capital to other regions in a receiving country and affect its labor markets, our study broadens the current knowledge on the long-term impacts of historical events. A further advantage of our approach is the separation between immigrants' descendants human capital from immigrants' persistent influence through second-nature advantages in the settlement locations.

Second, because the internal migration flows of historical immigrants' descendants in Brazil are closely tied to the expansion of modern mechanized agriculture in the country, we contribute to the literature on the causes and consequences of the expansion of the agricultural frontier in Brazil (Bragança, 2018; Bustos et al., 2016, 2020; Pellegrina, 2022). 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. More broadly, human capital of migrants can thus be seen as another determinant for the prosperity of frontier settlements besides geography (Nunn and Puga, 2012; Bazzi et al., 2016) and the internalized frontier experience (Bazzi et al., 2020).

Third, we add to the literature that investigates imperfect substitution between domestic and immigrant labor (Borjas and Katz, 2007; Ottaviano and Peri, 2012), extending the analysis to the *descendants* of these workers. In doing so, we also connect to the literature on immigrants' assimilation (Abramitzky et al., 2014; Pérez, 2021). Non-Iberian descendants being substitutes for the predominant population group in the labor market implies that intangible assets are transmitted over generations in line with evidence surveyed in Bisin and Verdier (2011) and Voth (2021). Emergence of productive spillovers through combination of cultural backgrounds aligns with knowledge sharing, complementarity of approaches, problem solving as documented by research on the value of immigrant diversity (Ozgen, 2021).

The rest of the paper is organized as follows. Section 2 provides background information on immigration, ancestries, and surnames in Brazil. This section discusses our samples of municipalities, and presents a brief account of the expansion of the agricultural frontier in Brazil. Section 3 presents the data, explains the surname-based classification of ancestries, and shows some descriptive statistics. Section 4 contains our theoretical framework. Section 5 presents the empirical strategy, discusses identification concerns and proposed mitigation approaches. Section 6 exposes and discusses the results. We close the paper with a series of robustness checks in Section 7 and the conclusion in Section 8.

# 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 Brazilians who declare themselves as 'whites'.[2] 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 Iberian surnames today.

In the late 19th and early 20th centuries, Brazil was one of the targets of the mass immigration waves from other European and Asian countries. The largest groups of non-Iberian immigrants came from Germany, 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. Table 1 shows the exact breakdown of immigrants by country of origin for 1884–1939, the period of highest intensity of international immigration in Brazil. According to these numbers from a compilation of several Census records in IBGE (2007), the share of immigrants with non-Iberian ancestry between 1884 and 1939 is 57%.

[2]Official census surveys and administrative data in Brazil ask people to determine their skin color/race into one of five possible categories: white, mixed (*pardo*), far-eastern (*amarelo*) and native-Brazilian (*indígena*). These categories, however are often far too broad and imprecise to allow for specific applications like ours.

Table 1: Immigration to Brazil by country of origin, 1884–1939

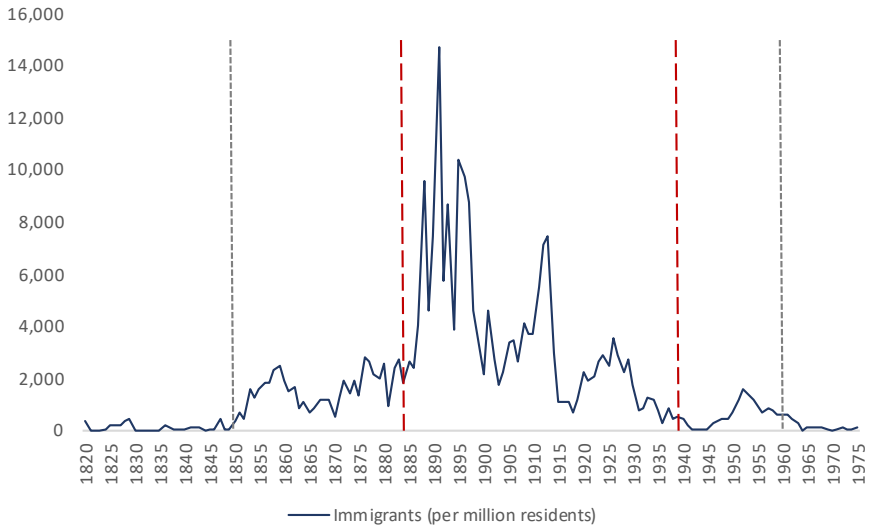
| Country of origin | Immigrants (1,000) | Share | Ancestry group | Group share |
|---|---|---|---|---|
| Italian | 1,412 | 34% | Non-Iberian | 45% |
| Japanese | 186 | 4% | | |
| German | 171 | 4% | | |
| Syrian-Lebanese | 99 | 2% | | |
| Portuguese | 1,204 | 29% | Iberian | 43% |
| Spanish | 582 |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

Notes: Data from IBGE (2007). Iberian refers to the sum of immigrants from Portugal and Spain.

International immigration was even encouraged by the Brazilian government from the 1850s through the 1930s 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. Immigration intensified from 1850 through the late 1950s (with an expected decline during World War II) and then declined sharply after 1960, see Figure 1. 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).

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, we argue that 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.

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), see Table 1.

## 2.2 Study regions – injection and spread sample

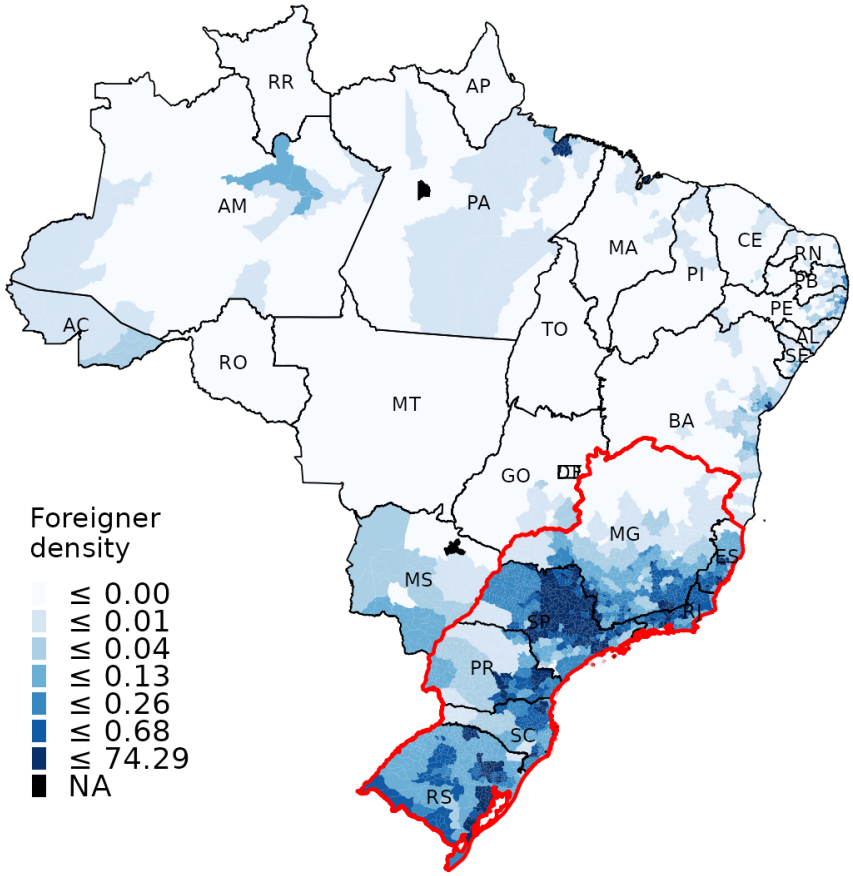
Several states in Brazil have well documented clusters (injection points) of historical non-Iberian immigration. Particularly the state-sponsored settlements (known as *colônias*) attracted a large number of non-Iberian immigrants to the South and Southeast regions of Brazil. There are historical records of non-Iberian settlements in the states of Minas Gerais (Monteiro, 1973), Espírito Santo (Franceschetto, 2014), São Paulo (Rocha et al., 2017), as well as for the three states in the South region: Paraná (Nishikawa, 2015), Santa Catarina (Piazza, 1983), and Rio Grande do Sul (Carvalho and Monasterio, 2012).[3] We refer to the main immigrant locations as 'injection points' because they served as points of attraction and destination to non-Iberian immigrants coming from different parts of the world to Brazil between 1850 and 1960. These immigrant clusters used to be unpopulated and thus developed slowly (Roche, 1969; Bublitz, 2008). Once installed, the immigrants spread only gradually to nearby regions, creating a pattern of spatial persistence (Ehrl and Monasterio, 2021).

Figure 2 shows the density of foreign immigrants in 1920 – approximately the apex of first-generation immigrants in Brazil. Note most of the current 5575 municipalities were not yet created

[3]There are occasional historical records of non-Iberian settlements in states of the other three regions in Brazil (e.g. some 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 other regions, and the settlements are smaller.

in 1920 so that the intertemporal comparison requires using Minimum Comparable Areas (AMC) provided by Ehrl (2017). The figure clearly illustrates that in 1920 foreigners were concentrated in the South and Southeast of Brazil confirming that the concentration of immigrants was persistent around the historic injection points.

Figure 2: Foreigner density in 1920 and sample definition




Notes: The map shows the number of foreign immigrants in the year 1920 per square kilometer for each of the Minimum Comparable Areas (AMC). AMCs are an aggregation of municipalities such that the defined area remains stable over the 1920–2020 period, according to Ehrl (2017). Black lines denote state boundaries. The *spread sample* encompasses all municipalities/AMC 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 *Injection Sample* is formed by 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), Rio de Janeiro (RJ), and Minas Gerais (MG) in the Southeast region—all of which have a significant and well documented presence of injection points of immigration, are at or very close to the sea, and thus present a high foreigner density. Legend categories correspond to deciles of foreigner density; the lowest are aggregated due to high frequency of near-zero values.

## 2.3 The agricultural frontier in Brazil

This section presents a brief account of the expansion of the agricultural frontier. This account is useful for understanding the migration of foreign descendants from the South and Southeast to the other regions of the country.

Following the adaptation of soybean varieties to tropical climates, migrants from the South and Southeast 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 this sector. 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 (Bragança et al., 2015; Bustos et al., 2016, 2020).

The low population density and the abundance of (mostly flat) farmland in the Brazilian *Cerrado*, the savannah-like biome that is dominant at the agricultural frontier, combined to result in low land prices that attracted farmers from the South and Southeast regions of the country. This process was further stimulated by private colonization companies, farming cooperatives, land reform initiatives, and rural development programs implemented by the national government (Jepson, 2006*a*,*b*). The frontier continued to expand and to develop in the more recent decades, stimulated by the arrival of new technologies that impact not only agricultural production, but also 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 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 (Alves, 2016). On a related note, Alves (2016, 146) argues that "compared to the local gentry, these migrants were much more advanced in terms of modern agricultural knowledge", thus corroborating the notion that migrants coming to the agricultural frontier brought a

positive shock of human capital to the region, a claim that resonates with those made by Rocha et al. (2017, 106) who claim that historical settlements in São Paulo "brought immigrants who were more educated than natives, creating a positive human capital shock".

The considerations in this section lead us to distinguish between two samples of municipalities, as indicated by the red line around states in Figure 2. The first, which we call the 'injection sample', includes all municipalities in states with documented injection points of historical immigration in 19th century Brazil. The municipalities in this sample were home to the majority of first-generation immigrants. Thus, these regions may have been impacted by the different cultural backgrounds, by accommodations made to receive the foreigners (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, but mainly in the 20th century (Pellegrina and Sotelo, forthcoming).

# 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 produce statistics on the formal labor market. This makes RAIS a high-quality annual census of all formally employed workers in Brazil. Stacked over the years, RAIS becomes a linked employer-employee panel.

The RAIS data include demographic characteristics of employees (race, gender, age, education), their remuneration, hour worker per week, and occupation. The data also include characteristics of the employers like industry, size (number of employees), and the municipality in which the establishment is located. The full name of workers, used in our surname-based classification of ancestries, is a key distinctive variable in the RAIS data which is available in the years from 2008 to 2019.

Since the RAIS registers all formal employment relations, it is common to make a few adjustments to restrict noise and capture jobs where the wage is determined by demand and supply in the market. Therefore, we exclude all foreign-born, all public servants, and the military. We keep workers between 16 and 70 years old who have a valid identifier, work more than 10 hours per month and record a positive remuneration. After adjusting remunerations for inflation and calculating hourly wages, we drop those workers with hourly wages above the 99.9th percentile. Due to the usual adoption of the husband's surname after marriage, as explained in the following subsection, we focus on male workers to have a more direct link between surname and the person's cultural heritage. Finally, we exclude municipalities with less than 30 workers and those with a zero share of descendants.

We complement this data with multiple sources for municipality-level information. Socioeconomic characteristics come from Ipeadata. 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. The geolocation 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.

### 3.2 Surname-based classification of ancestries and descendant share

This subsection describes how we classify ancestries based on workers' surnames and briefly discuss the social and legal norms surrounding surnames in Brazil.

Children in Brazil usually receive two surnames: the mother's second surname, followed by the father's second surname.[4] 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 over

[4]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.

generations. 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. In most cases, when adopted, the husband's surname becomes the second.

There are some challenges to implementing a surname classification that uses two or more surnames. In our study, we consider only the last surname of each person. We do not expect, however, a more 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 is a refined version of the algorithm described in Monasterio (2017), where more datasets are used to increase the number of distinct names and ancestries. Therefore, 16 different European-Asian, non-Iberian ancestries can be identified from the following four steps: (1) collect the second surname of all workers in the RAIS; (2) match these surnames to historical immigration records where surnames are accompanied by countries of origin; (3) link each unique surname to a country of origin or region of origin (e.g., Eastern Europe); and (4) apply a machine-learning language classification method proposed Cavnar and Trenkle (1994) to the remaining unmatched surnames.

The classification yields, for each municipality $m$ and year $t$ in our sample, the total number of workers of each ancestry. We group all ancestries into two categories: Iberians, i.e. locals (IBR) and non-Iberians, i.e. descendants (NIB). Our variable of interest in the empirical analyses is the concentration of descendants given as

$$C_{mt}^{Desc} = \frac{N_{mt}^{NIB}}{N_{mt}^{NIB} + N_{mt}^{IBR}} \tag{1}$$

Figure 3 shows the spatial distribution of the concentration of non-Iberian descendants

across Brazilian municipalities in 2008. The highest concentration of descendants is observed in the three southernmost states, where the share tends to be in the two upper quintiles of the variable's distribution. That is, in approximately half of the Southern municipalities the share of non-Iberian descendants is between 20% and 89%. Moreover, there is a high concentrations of descendants in São Paulo (SP) and Espírito Santo (ES), all places that are known for having received a significant number of historical non-Iberian immigrants. We also notice a higher concentration in the Center-West region, particularly in Mato Grosso, a state that symbolizes the expansion and modernization of the agricultural frontier in Brazil.

The Northwestern states tend to register the lowest concentration of descendants, roughly half of them not having more than 2%. The map also shows that the farther the distance from the Atlantic Ocean, the larger the area of municipalities. This pattern reflects the expansion of the agricultural frontier and the population spread, indicating that population density in northern municipalities is generally very low.

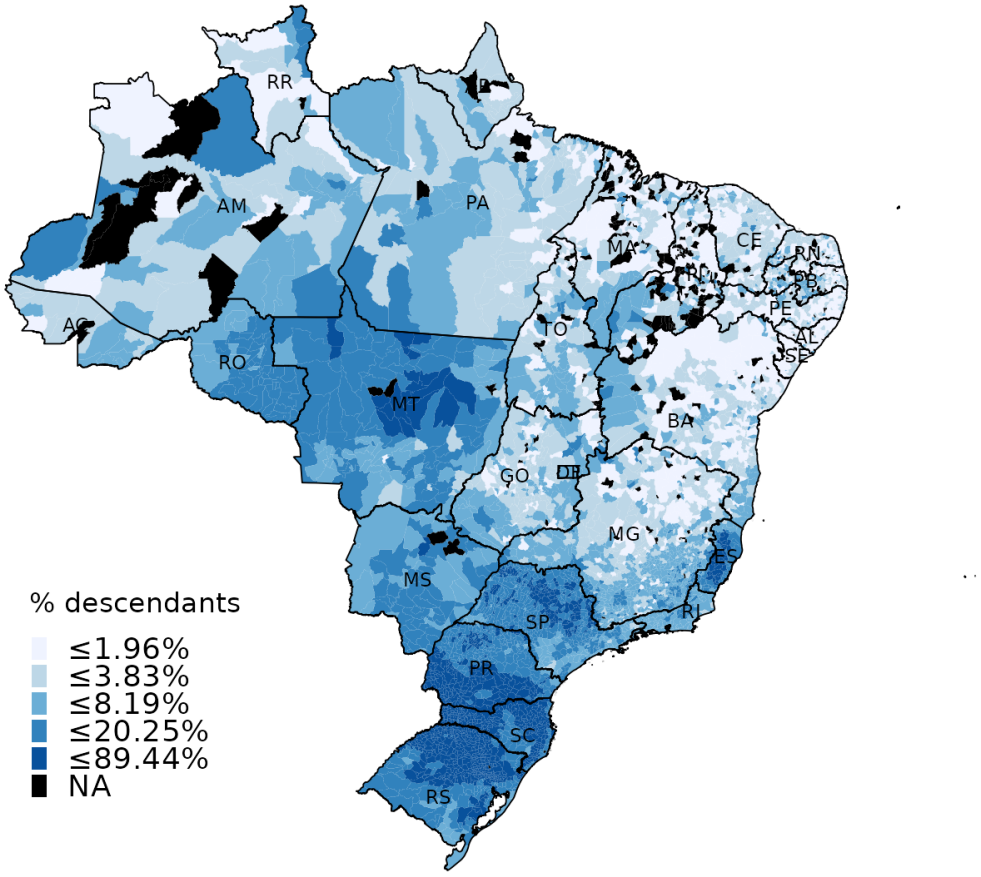
Complementary to the spatial distribution, Figure A3 in the Appendix shows the density distribution of the concentration of descendants across the municipalities in our two samples. The figure shows the distribution up to 50% only to improve visualization, since few municipalities in the injection sample and no municipality in the spread sample go beyond that level. Both distributions are skewed to the left, but beyond the share of 5%, the distribution in the injection sample is clearly the dominant one.

### 3.3 Descriptive statistics

This section by presents descriptive statistics to test for differences between our samples, and for differences between descendants and locals within each sample. Table 2 shows the mean for several variables of interest from the RAIS data and additional information from the Census in 2010. Due to the large sample sizes, 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 a quarter as meaningful.

Comparing the injection and spread sample indicates the number of workers per municipality as well as the average population size is about thrice as large in the injection sample. The share of labor formalization is also much higher in the spread sample meaning that informality,

Figure 3: Concentration of descendants in Brazilian municipalities, 2008 (%)




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 2008. The scale in the map uses quintiles to highlight variation in the concentration across space. Black lines denote state boundaries. The denomination of states and their regions are: 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; Mato Grosso do Sul (MS), Mato Grosso (MT), and Goiás (GO) in the Center-West region; Rio Grande do Sul (RS), Santa Catarina, and Paraná (PR) in the South region, and the states of São Paulo (SP), Espírito Santo (ES), Rio de Janeiro (RJ), and Minas Gerais (MG) in the Southeast region.

unemployment, and self-employment are more prevalent. In line, other indicators for economic development and socioeconomic welfare such as years of schooling, life expectancy, and average wage are also considerable higher in the injection sample. Finally, we observe that the average concentration of descendants is much higher in the injection sample. This fact is expected, since municipalities in the South and Southeast concentrated most of the arrivals and settlements of non-Iberian immigrants during the 1850–1960 period.

Table 2: Summary Statistics, Individual Level, 2010

| | injection sample | | | | spread sample | | | |
|---|---|---|---|---|---|---|---|---|
| | Locals | Descendants | Diff. (D – L) | | Locals | Descendants | Diff. (D – L) | |
| | Panel A: Information from RAIS data | | | | | | | |
| Individuals per municipality (1,000) | 477 | 332 | 145 | [0.14] | 118 | 107 | 12 | [0.06] |
| Share of descendants | 13.1% | 21.9% | -8.7% | [0.52] | 3.7% | 6.1% | -2.4% | [0.58] |
| Log of hourly wage | 2.44 | 2.66 | -0.23 | [0.21] | 2.15 | 2.45 | -0.30 | [0.33] |
| Hours worked | 42.46 | 42.94 | -0.48 | [-0.08] | 43.05 | 43.34 | -0.29 | [-0.06] |
| Age (years) | 34.61 | 35.19 | -0.58 | [0.04] | 33.81 | 34.54 | -0.73 | [0.05] |
| Education level: HS or higher | 0.52 | 0.63 | -0.12 | [0.17] | 0.50 | 0.61 | -0.12 | [0.16] |
| White collar occupation | 0.15 | 0.25 | -0.10 | [0.18] | 0.12 | 0.23 | -0.11 | [0.24] |
| Industry = Agriculture | 0.05 | 0.04 | 0.01 | [0.03] | 0.08 | 0.11 | -0.03 | [0.08] |
| Industry = MFG/Construction/Utility | 0.38 | 0.40 | -0.02 | [0.03] | 0.37 | 0.31 | 0.06 | [0.08] |
| | Panel B: Information from Census data | | | | | | | |
| Total population (1,000) | 2,575 | 1,777 | 798 | [0.15] | 862 | 773 | 89 | [0.07] |
| Expected years of schooling | 10.00 | 10.18 | -0.19 | [0.21] | 9.51 | 9.57 | -0.07 | [0.07] |
| Life expectancy at birth (years) | 76.00 | 76.17 | -0.17 | [0.11] | 73.97 | 74.39 | -0.42 | [0.16] |
| Degree of labor formalization | 70.6% | 71.6% | -1.0% | [0.09] | 56.7% | 58.3% | -1.7% | [0.10] |
| Individuals (Millions) | 12.4M | 2.1M | 14.5M | | 5.8M | 0.2M | 6M | |

Notes: The injection sample includes municipalities in the South and Southeast regions. The spread sample includes municipalities in the Center-West, North, and Northeast. Both samples exclude municipalities with fewer than 30 individual observations in our selected RAIS data or no descendants. Indicators in the top panel use individual-level data from RAIS in 2010; those in the bottom panel use municipality-level data from the Brazilian population census in 2010. Income and earnings in the Census data are in R$ of 2010. The concentration of descendants (% of workers with a non-Iberian surname in each municipality) and the movers are individuals who were observed at least once in a different municipality in the RAIS data from 2004–2015. Due to the large number of observations in all samples, all differences are statistically significant at 1%. Therefore, we show the absolute value of the normalized differences, which are unaffected by sample size, in brackets.

Regarding the two types of workers, we observe in both samples 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, descendants work in municipalities where the concentration of descendants is higher. Less obvious is the fact that the wage of descendants is significantly 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.

Table 2 further indicates that descendants and locals employed in the formal sector differ in demographic characteristics. In particular, descendants are only slightly older, but are more educated and concentrated in white collar occupations. The differences hold with similar normalized magnitudes and the same directions in both samples. Another key fact, aligned with conjectures

about the 'March to the West' is that descendants 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.

# 4 Theoretical framework

This section presents a simple formal framework to explain why wages between descendants and locals may differ and how they are related to the concentration of descendants in municipalities.

We adapt the model in Moretti (2004), who studies productivity spillovers associated with increases in the concentration of college-educated workers. Suppose each municipality is a competitive economy that produces a single output good, $Y$, using capital, $K$, and labor provided by non-descendants, $N_0$, and descendants, $N_1$. These three inputs are imperfect substitutes in the production technology of firms. For simplicity, we assume this technology has constant returns to scale and can be represented by a Cobb-Douglas production function:

$$Y = (\theta_0 N_0)^{\alpha_0} (\theta_1 N_1)^{\alpha_1} K^{1-\alpha_1-\alpha_0}, \qquad \alpha_0 > 0,\ \alpha_1 > 0,\ \alpha_0 + \alpha_1 < 1. \tag{2}$$

The terms $\theta_0$ and $\theta_1$ are the productivity shifters of each type of labor and capture productivity spillovers in the model. We assume shifters are a function of the worker's ancestry-specific human capital, $\phi$, and the spillover $f(s)$. Specifically, the model allows for human capital spillovers by letting workers' productivity depend on the share ($s$) of type 1 workers in each municipality. As in our empirical approach below, $s$, i.e. the concentration of descendants is the main variable of interest. In logs:

$$\log\left(\theta_j\right) = \phi_j + f(s), \qquad \text{for } j = 0, 1, \qquad s = \frac{N_1}{N}, \qquad N = N_1 + N_0. \tag{3}$$

We assume that wages are equal to the marginal product of each type of labor and that the productivity spillovers are external to individual firms in the municipality but internal to the municipality as a whole. That is, firms take the $\theta$'s as given and choose the amounts of labor $N_0$ and $N_1$ that maximize profits ignoring their effect on the respective productivity shifters $\theta_0$ and $\theta_1$. At the same time, spillover effects are constrained to each municipality.

We relax two assumptions in the original model. First, we let both types of workers—

descendants and locals here—have different productivities, $\theta_1 \neq \theta_0$, but we do not assume that the productivity of the first type is necessarily higher. Second, we do not assume monotonic spillovers. Instead, we let $f(s)$ take any format. In our setting, increasing the concentration of one worker type can have non-linear effects on the wages of both types.[5]

The derivatives of log wages with respect to the share of type 1 workers, $s$, show how a change in the concentration of descendants affects the wages of both worker types:

$$\frac{d\log(w_0)}{ds} = \left[\frac{1-\alpha_0}{1-s} + \frac{\alpha_1}{s}\right] + (\alpha_1 + \alpha_0) f'(s) \tag{4a}$$

and

$$\frac{d\log(w_1)}{ds} = -\left[\frac{\alpha_0}{1-s} + \frac{1-\alpha_1}{s}\right] + (\alpha_1 + \alpha_0) f'(s). \tag{4b}$$

The wage changes in both equations are composed of two different effects. The first term in square brackets is a *factor complementarity* effect in (4a) and a diminishing marginal returns effect in (4b). The second term is driven by the derivative of $f(s)$, the *spillover* effect. Although our empirical framework does not allow us to disentangle the two effects, we can still draw meaningful conclusions from the data.

Equation (4b) implies that if the relative size of one type in the production function increases, its respective marginal product and the associated wage must decrease. This results stems directly from the assumption that both types of labor are imperfect substitutes. If the observed total effect, however, is positive or indistinguishable from zero, we can conclude that a positive spillover effect is present.

The model also implies that wages of locals increase more than those of descendants when the latter share rises: $d\log(w_0)/ds > d\log(w_1)/ds$ should thus be satisfied in any case.[6]

[5] A higher $s$, thus, is not necessarily better: $f$ could be maximized at some $s^* \in (0,1)$.

[6] From equations (4a) and (4b) we have:

$$\frac{d\log(w_0)}{ds} - \frac{d\log(w_1)}{ds} = \frac{1}{s(1-s)} > 0.$$

# 5 Empirical framework

The overall aim of this paper is to investigate whether historical immigration improves contemporary local labor-market performance. In line with the historical background and the theoretical production framework exposed in the previous sections, we are specifically interested in seeing whether the share of immigrants descendants affects workers' wages. Given that the settlement of immigrants was originally concentrated in the South and South-East, we distinguish between municipalities that we termed the *injection* and *spread* sample. Moreover, we investigate whether the effects differ between immigrants' *descendants* and the remainder of the Iberian workers (*locals*).

Consider the following regression equation at the municipality level as an initial motivation for our empirical analysis.

$$y_{m(s)t} = \beta C_{mt} + Z_{mt}\Gamma + \psi_s + \phi_t + \varepsilon_{m(s)t} \tag{5}$$

where $y_{m(s)t}$ is the average log wage in municipality $m$ located in state $s$ in year $t$, $C_{mt}$ is the contemporary concentration of descendants in the municipality, $Z_{mt}$ is a vector of municipality-level characteristics, $\psi_s$ and $\phi_t$ are state and time fixed effects, respectively, and $\varepsilon_{m(s)t}$ is the error term. We run this equation separately for the injection and spread sample to test whether the correlation between wages and the descendant share captured by $\beta$ is considerably different.

Specifically, the estimated coefficient $\beta$ shows the approximate percentage increase in wages associated with one additional percentage point in the share of descendants in the municipality. In our model, a positive association between the concentration of descendants and the wage of locals in a municipality is consistent with both complementarity and spillover effects. A non-negative association with the wage of descendants, on the other hand, is consistent only with the existence of positive spillovers.

Several identification threats in equation (5) preclude a causal interpretation of the estimated coefficients. (1) The share of non-Iberian descendants is naturally imprecise because some people may have a cultural legacy from a non-Iberian country which is not (anymore) reflected in their surname. Likewise, recent Spanish and Portuguese immigrants cannot be distinguished from traditionally Brazilian residents because the linguistic origin is the same. This measurement error tends to bias $\beta$ towards zero. (2) Reverse causality may be present when either locals or descendants

systematically sort into high-wage municipalities. In fact, Ehrl and Monasterio (2024) find that non-Iberian immigrants' descendants are less mobile today, such that we can expect (further) downward bias in the estimated $\beta$ from equation (5). (3) Sorting may not only be based on the type of ancestry but also on unobservable individual characteristics as ability, mobility costs, or outside options in the labor market. Any correlation between these characteristics and the local wage level would then lead to biased estimates. (4) Attributes of municipalities like first-nature advantages and those shaped by historical immigrants affect the wage level and will bias our estimates if not properly controlled for. Political institutions and the educational system, for example, have been positively affected by the immigrants in the 19th and 20th century (Seyferth, 1994; Naritomi et al., 2012).

To deal with these threats to identification, our empirical strategy makes the following adjustments. We adopt a two-stage estimation procedure as in Combes et al. (2008), Bakens et al. (2013), De la Roca and Puga (2017), Ehrl and Monasterio (2021), among others. Exploiting the panel nature of our individual data, we run a first-stage wage regression on the exogenous time-varying worker characteristics $X_{it}$ (age and age$^2$) and fixed effects for individuals ($\mu_i$), municipalities ($\kappa_m$), and years.

$$y_{i(m)t} = X_{it}\delta' + \phi_t + \kappa_m + \mu_i + \varepsilon_{i(m)t} \tag{6}$$

The individual fixed effects account for possible bias from sorting on unobservable attributes and also nets out the influence of the worker's *own* cultural background. Consequently, the estimated municipality fixed effect $\hat{\kappa}_k$ captures remaining wage differences between the local labor markets. The second-stage regression is then given by

$$\hat{\kappa}_m = \beta C_m + Z_m\Gamma + \xi_m \tag{7}$$

That is, through the first-stage estimation we account for any dynamic effects from the composition of workers with the heterogeneous cultural backgrounds and abilities. The vector $Z_m$ has several controls that correlate with first- and second nature advantages like agricultural productivity, distance to the sea, and population density. Precisely, $Z_m$ includes measures for 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), a state capital indicator, and indicators for soil types.[7] Remaining wage differences due to the share of descendants $C_m$ thus emerge as a form of external effect from the production function.

To address remaining endogeneity issues from omitted variable bias, measurement error, and reverse causality in equation (7), we apply an instrumental variable (IV) approach. Following Ehrl and Monasterio (2024), our instrument is the average distance from each municipality to former official colonies located in the states of Rio Grande do Sul and São Paulo (the settlements discussed in Section 2).[8] The IV is supposed to be relevant since there is evidence for the spatial persistence of immigrants, see Figures2 and 3. Moreover, staying close to the historic injection points reduces migration costs and eases network formation. We thus expect the colony distance IV to be negatively correlated with the concentration of descendants in the spread sample. The instrument should also satisfy the exclusion restriction because the official colonies in the two states were chosen by government officials based on military and strategic settlement considerations (Ehrl and Monasterio, 2024).

We prefer IV models with a single instrument because exactly identified IV estimators exhibit limited bias, particularly when the first-stage coefficient's sign accords with theory (Angrist and Kolesár, 2024). Notwithstanding, to formally test the exogeneity of the colony distance, we use an alternative IV: the Terrain Ruggedness Index. The rationale behind the second IV is that the suitability of potential destinations for modern agriculture (low terrain ruggedness) worked as pulling factors for descendants contemplating a move to the agricultural frontier in Brazil since the 1960s. As we explained in section 2.3, regions with a high availability of flat farmland disproportionately attracted non-Iberian descendants with expertise and means to engage in modern agriculture coming from areas (injection sample) where agricultural production is characterized by a high productivity and capital usage (Bustos et al., 2020). Thus, we expect average terrain rugged-

[7] One may argue that estimating equations with municipality fixed effects would strengthen the control for time-constant amenities that immigrants installed. However, on the one hand, the main variable of interest ($C_m$) has very little temporal variation, making this strategy infeasible. A variance decomposition of the descendant share in the municipality level for the years 2008–2019 reveals that 98% of the variable's variation is across units and only 2% across time. On the other hand, the instrumental variable strategy is already intended to isolate the pure effect of descendant concentration on wages. Since the instrument also has no variation over time, we can only provide these estimations in a cross-section of municipalities.

[8] 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 colony distance measures. 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 (Carvalho Filho and Monasterio, 2012; Ehrl and Monasterio, 2024).

ness to negatively correlate with the concentration of descendants. Figure A2 in the Appendix gives a visual representation of the instrument.

# 6 Results

## 6.1 Municipality-level Correlations

In Table 3, we show the results for municipality-level regressions of wages on the concentration of descendants according to equation (5). In all specifications, we account for regional differences between states through fixed effects and for observable first- and second-nature advantages of municipalities. Columns (1) to (3) refer to the injection sample, whereas columns (4) to (6) are based on the spread sample of municipalities farther away from the locations of historic immigrants. For each sample, we estimate panel regressions using the years 2008-2019, as well as regressions using only the first or last year available.

Table 3: Municipality-Level Regressions of Wages on the Concentration of Descendants

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Dependent variable: log wage | | | | | |
| | injection sample | | | spread sample | | |
| Years | 2008–19 | 2008 | 2019 | 2008–19 | 2008 | 2019 |
| Descendants share | 0.031 | 0.011 | 0.031 | 1.251*** | 1.169*** | 1.479*** |
| | (0.092) | (0.101) | (0.086) | (0.314) | (0.337) | (0.282) |
| Mean of descendants | 0.15 | 0.15 | 0.14 | 0.04 | 0.04 | 0.04 |
| Adj. $R^2$ | 0.83 | 0.61 | 0.64 | 0.90 | 0.83 | 0.82 |
| Observations | 34,126 | 2,835 | 2,849 | 30,386 | 2,467 | 2,566 |
| Municipalities | 2,855 | 2,835 | 2,849 | 2,689 | 2,467 | 2,566 |
| State fixed effects | Y | Y | Y | Y | Y | Y |
| Time fixed effects | Y | | | Y | | |

Notes: The dependent variable is the log of the average formal sector wage in the municipality. Descendants (%) represents the share of workers with a non-Iberian surname among the total number of formal employments in the municipality. All regressions are weighted by the number of workers in the municipality and include the following municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), and five dummies for soil types (dummies = 1 if 5% or more of municipality area is covered by soil type). Clustered standard errors at the municipality-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The results reveal a consistent positive association between the concentration of descendants

and wages in the spread sample. In contrast, the null of no significant correlation cannot be rejected in the injection sample. The mean descendant share in each specification, reported below the coefficients in the table, confirms that municipalities in the South and South-East (0.15) have more than 3 times more non-Iberian descendants compared to the spread sample, with tiny or no variation over time. In line with this temporal stability, we find small differences in the coefficient estimates across the years. If anything, the positive correlation became stronger between 2008 and 2019. Column (6) indicates that one additional percentage point in the concentration of descendants is associated with 1.48% higher average wage in the spread sample.

Our conclusions from this exercise are twofold. First, we confirm a positive association between the concentration of descendants and average wages in municipalities. This association, however, seems to exist only for municipalities in the spread sample where descendants are much more scarce. On the other hand, the relatively flat distribution of the descendant share in the injection sample (Figure A3) already suggests there is limited information content to help explain variation in wage levels. Second, the comparison across years indicates that, due to the lack of temporal variation, panel regressions will not be more informative than cross-section specifications. In the following subsection, we exploit the variation across individuals over time to control for fixed effects but then continue with the analysis of specific years at the municipality level.

### 6.2 Two-step regressions

Table 4 shows the results from the second step according to equation (6). The first major difference with respect to the previous results in Table 3 is that regressions are performed separately for descendants and locals. Second, the first-step estimations account for worker fixed effects and thus eliminate the effect of workers' ancestry on wages as well as spatial sorting on individuals' characteristics, see equation (7). Since the previous estimates were consistent across time (and continue to be so, see Table A1 for the year 2008), we focus on the year 2019 for the rest of the paper.

The results for the injection sample tell a story of no effects for the descendants and relatively small but positive effects for the other population group, locals. In the spread sample, the relation between the concentrations of descendants and local workers' wages is also positive but 2.5 times higher. Yet, for the own population group, a higher share of descendants seems to depress wages

Table 4: Two-step Wage Regressions – Descendants vs. Locals, 2019

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable: | Estimated municipality fixed effect | | | |
| Sample: | Injection | | Spread | |
| Years: | 2019 | 2019 | 2019 | 2019 |
| Workers: | Descendants | Locals | Descendants | Locals |
| Descendants share | 0.023 | 0.266*** | -0.288*** | 0.939*** |
| | (0.029) | (0.032) | (0.093) | (0.165) |
| Mean of Descendants | 0.14 | 0.14 | 0.04 | 0.04 |
| Adj. $R^2$ | 0.60 | 0.48 | 0.45 | 0.58 |
| Observations | 2,841 | 2,845 | 2,518 | 2,562 |

Notes: The table reports coefficients from the two-step wage regressions according to equations (6) and (7). Descendants (%) represents the share of workers with a non-Iberian surname among the total number of formal employments in the municipality. All regressions are weighted by the number of workers, either locals or descendants, according to the fifth row, in the municipality and include the following municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), and five dummies for soil types (dummies = 1 if 5% or more of municipality area is covered by soil type). The controls in the first-stage worker-level wage regressions include age, age squared, as well as municipality, worker, and year fixed effects. Clustered standard errors at the municipality-level in parentheses. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

in the spread sample. Note that these observations are in line with the theoretical framework in section 4. The negative coefficient for descendants as well as the positive effect of locals' wages can thus be explained by factor complementarity and diminishing returns. The different magnitudes in the spread and injection sample may be related to different production functions and to the differences in the overall presence of non-Iberian descendants, see Table 4. This findings is at odds with other possible explanations for the observed wage externality. Since the benefits do not accrue to all groups in the population, potential benefits from non-Iberian descendants such as a higher capital stock or public goods provision do not seem to apply in the present case.

Regarding the differences between the two specifications in Tables 3 and 4, we observe that the adjustment though the first-step regression leads to a lower correlation between the share of descendants and wages. The marginal effect of an additional percentage point in the concentration of descendants is now associated with a 0.94% higher wage for locals in the spread sample. This coefficient is still highly significant, below the 1% level.[9] The reduced coefficient in the two-step

[9]Our results are not sensitive to using other type of standard errors, including bootstrap or Conley procedures. Because of adding the additional instrumentation stage in the following estimations, we prefer to report the simple, transparent clustered standard errors.

procedure indicates that without accounting for sorting based on intrinsic worker attributes, the estimates are biased upwards. That is, workers with abilities that are particularly valuable in the labor market sort into high-wage regions. The direction and relative magnitude of the implied bias is comparable to the approach in De la Roca and Puga (2017).

### 6.3 Two-step regressions with instrumentation

The concentration of descendants in municipalities may be endogenous to the process that generates our results. As we discussed previously, either locals or descendants may be systematically more mobile and attracted to high-wage regions, making the relationship mutually causal. Moreover, measurement error in the explanatory variable and unobservable municipality characteristics may also bias our results.

In this subsection, we present our preferred results from the two-step strategy with instrumental variables (IV), which addresses the remaining endogeneity concerns. Whether the IV estimations identify unbiased coefficients, 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 5 shows the results for IV regressions using data from 2019 for the spread sample where the potentially endogenous variable, the share of descendants, is instrumented by the average distance to the official, and arguably exogenous location of historical colonies in the states of Rio Grande do Sul and São Paulo in columns (1) and (4). The upper part of the table reports the second-stage coefficient, whereas the lower part displays the first-stage coefficients of the IV and related econometric statistics for the 2SLS estimation.

Colum (4) indicates that locals do benefit from a higher concentration of workers with non-Iberian ancestry. The estimated effect is highly significant, positive, and about two times higher as without instrumentation. This change indicates that measurement error and sorting of workers with Iberian ancestry resulted in considerable downward bias. Thus, an additional percentage point in the concentration of descendants is related to a 3% higher wage for locals in the spread sample. Column (1) suggests that descendants' wages are not affected by a higher concentration

Table 5: Two-step Wage Regressions with instrumentation – Descendants vs. Locals

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dependent Variable: | Estimated municipality fixed effect | | | | | |
| Sample | Spread | | | | | |
| Years | 2019 | 2019 | 2019 | 2019 | 2019 | 2019 |
| Workers | Descendants | | | Locals | | |
| Descendants share | 0.366 | 0.574 | 0.391 | 3.032*** | 6.370 | 3.321*** |
| | (0.830) | (1.082) | (0.645) | (1.042) | (4.563) | (0.994) |
| Observations | 2,518 | 2,518 | 2,518 | 2,461 | 2,562 | 2,562 |
| | IV 1.-Stage Statistics | | | | | |
| Colony Distance | -0.029*** | -0.003 | -0.003*** | -0.025*** | -0.001 | -0.002*** |
| | (0.004) | (0.002) | (0.001) | (0.004) | (0.001) | (0.001) |
| Terrain Ruggedness | | | -0.030*** | | | -0.024*** |
| | | | (0.004) | | | (0.003) |
| 1. F-Stat | 46.527 | 2.289 | 31.682 | 46.164 | 1.116 | 27.581 |
| $R^2$-partial | 0.224 | 0.025 | 0.263 | 0.291 | 0.007 | 0.289 |
| Weak IV AR | 0.207 | 0.252 | 0.799 | 6.578** | 3.476* | 12.147*** |
| Weak IV SW | 0.345 | 0.367 | 2.178 | 10.660*** | 4.542** | 33.687*** |
| Hansen J (overid) | | | 0.014 | | | 1.019 |
| Hansen J (p-value) | | | 0.907 | | | 0.313 |

Notes: The table reports coefficients from the two-step wage regressions with instrumentation according to equations (6) and (7). The dependent variable thus is the estimated municipality fixed effect obtained from the first-stage worker-level wage regressions that also include age, age squared, worker and year fixed effects as controls. Descendants (%) represents the share of workers with a non-Iberian surname among the total number of formal employments in the municipality. The descendant share is instrumented by the average of the distance to historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul. This average distance to settlements (expressed in 1,000km) 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 Sao Paulo as recorded by (Carvalho and Monasterio, 2012) and (Rocha et al., 2017), respectively. The second IV is the average Terrain Ruggedness Index (TRI), which was calculated using the methodology proposed by (Riley et al., 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 injection sample includes all municipalities in the Center-West, North, and Northeast with at least one descendant and 30 or more individual observations in our selected RAIS data. The sample excludes females. All regressions are weighted by the number of workers, either locals or descendants according to the fifth row, in the municipality. The lower part of the table reports the following first-stage IV statistics: IV coefficient and standard error, excluded IV F-statistic and partial $R^2$, weak IV Anderson–Rubin (AR) Wald and Stock–Wright LM test statistic, Hansen-Sargan overidentification test for exogeneity of IVs and its p-value. Clustered standard errors at the municipality-level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

of their own population group. Because the theoretical framework showed that the diminishing returns effect is negative, an insignificant overall effect suggests that positive spillovers compensate the former effect. The considerably high overall effect on the other population group, the workers

with Iberian ancestry, also suggests that positive spillovers are at work.

The partial R2 and F statistic in the bottom half of Table 5 show that we have a relevant IV in first stage in general. The coefficient for the concentration of descendants is negative, as expected, highly significant and similar in the descendant and local worker sample. The Anderson-Rubin and Stock-Wright weak identification test reveal an interesting difference between both samples. Both tests asses jointly whether the endogenous variable has no effect and the instrument is valid. The last part seems to be confirmed by the F-test in both samples. However, we get a rejection of the null-hypothesis only for the sample of local workers. The observed differences in columns (1) and (4) thus point out that the concentration of descendants only affects wages of local workers.

## 7 Robustness checks

**Instrument exogeneity.** The first robustness check for our IV result is integrated into Table 5. Although it is not possible to confirm the exogeneity of an instrument, we provide an approximation with the Hansen-Sargan exogeneity test in columns (3) and (6). To this end, we use an alternative IV, terrain ruggedness of municipalities.

We find that the terrain ruggedness is significantly and negatively correlated with the share of non-Iberian descendants once we also account for colony distance. The alternative IV thus also shows the expected sign. The first-stage 2SLS statistics at the bottom of Table 5 also indicate that the exogeneity of both IVs cannot be rejected according to the usual overidentification test. Per se, the IV does not seem to be particularly relevant which indicates that the distance from previous settlements still plays a major role in the expansion of the agricultural frontier. Most importantly, the estimated coefficient for the descendant share in the overidentified IV case remains close to the ones in the baseline IV regressions in columns (1) and (4).

**Focus on ancestry in the formal labor market.** 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, a potential concern for the present analysis is that the share of non-Iberian descendants in the total labor market is quite different the measure in the RAIS data. To address this shortcoming, we resort to a couple of other data sources that also inform 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 very high ($r = 0.96$). In figure A1 in the appendix we plot the concentration of descendants using only RAIS data on the same concentration using the extended data sources. The results show a concentration of observations around the 45° line, reassuring 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.

**Sample composition.** Table A3 in the Appendix presents the robustness checks regarding sample composition and institutional differences between municipalities, for the two-step wage regressions with and without instrumentation. On the one hand, we check if our results are not influenced by municipalities with the highest or lowest number of individual observations. In columns (1) and (3), we restrict the sample to within the 5th and 95th percentile of the distribution of observations per municipality weighted by the number of workers (i.e. locals). That results in dropping all individuals working in municipalities with less than 626 and more than 1,667,780 observations. On the other hand, we exclude municipalities (in the spread sample) with an elevated concentration of foreign immigrants in the year 1920. We use the 60th decile of the foreigner density as cutoff, i.e. exclude all municipalities with a foreigner density above 0.13, see Figure 2. This test serves to confirm that the municipalities with a high concentration of current descendants were not historically different from other regions. Finding divergent results from the baseline sample would also be misaligned with our interpretation that it was primarily the internal migration beginning in the 1960s that shaped the labor markets in the spread municipalities.

**Controls for institutions and human capital.** Following previous papers such as Naritomi et al. (2012), we test whether local institutional quality affects the results. The potential threat to identification is that non-Iberian descendants either chose municipalities with a high institutional quality—similar to their origin—or even shaped the public landscape over time. Similarly, another main difference that may drive our results could stem from differences in the educational system.

Therefore, the estimations in Appendix Table A3, columns (3) and (6), include proxy variables for governmental quality, access to justice, and a proxy for educational performance: test results by 9th graders in a nationwide exam.

Results across the six specifications in Table A3 are consistent with our main results. They show the same pattern of a positive association between the concentration of descendants and wages—for non-Iberian descendant workers only. The estimated marginal effect decreases slightly when we account for institutional quality in municipalities. Even when we restrict sample size, the significance level remains below the 1% mark.

**Group heterogeneity.** Finally, we test whether a specific group within the population of local workers may be particularly driving the results. To this end, we resort to the individual-level data and identify dimensions along which the effect on workers' wages may differ. Figure A4 depicts the estimated coefficient and its 95% confidence interval for the effect of the descendant share on locals and descendants from the two-step wage regressions with instrumentation distinguishing these groups by race, education, and occupation in agriculture.[10] We find little differences for the effect on local workers, except for a higher imprecision in the group of high-school dropouts and non-white workers. However, for workers with non-Iberian ancestry, a higher share of descendants seems to increase wages among high-school graduates and agricultural occupations. These differences are consistent with productivity spillovers and agricultural activity being at the center-stage of municipalities in the spread sample.

## 8 Conclusion

Our study uses a surname-based classification of workers' ancestries to identify the consequences of two historical events on current local labor markets in Brazil. First, mass immigration from 1850 to 1960 increased the size and cultural diversity of the labor force in Brazil, but immigrants concentrated in the South and Southeast. Second, internal migration to the largely unpopulated interior of Brazil starting in the 1960s spread the presence of historical immigrants and their descendants in these regions.

[10] We divided the worker sample along other dimensions than those presented in Figure A4 — such as white-collar vs. blue-collar occupations, industries, age — but results were very close to the baseline estimation and are thus omitted for the sake of space.

Our empirical strategy relies on matched employer-employee panel data from 2008 to 2019 and accounts for the effect of all intrinsic worker characteristics and spatial sorting of workers. Measurement error in the ancestry classification and reverse causality were addressed with an instrumental variable approach. We find that the concentration of descendants of historical non-Iberian immigrants in municipalities in the North and Central-West—the spread sample—is positively associated with the wage level. These benefits, however, accrue exclusively to workers with Iberian surnames, i.e., the predominant group in the population we referred to as locals.

Our observations are accordance with a theoretical framework with standard labor supply and demand where the two groups of workers are imperfect substitutions and may generate spillovers. Because we observe descendants in municipalities far from the historic injection points, our results cannot be explained by fixed factors linked to the sites of historical immigration such as land redistribution, natural endowments, or a head start in infrastructure. Therefore, we interpret our results as evidence that mobile factors were also an important driver of the persistent positive effects of historical non-Iberian immigration in Brazil, spreading the original effects from the injection points to other regions of the country. This interpretation is consistent with inter-generational transmission of task-relevant cultural traits and implies that ancestry can be economically salient, especially when heterogeneous backgrounds are combined to exploit complementarities.

This line of research could be extended at the firm-level, where it would be possible to investigate the hypothesis of labor complementarities and spillovers in greater depth. Based on the current interpretation of the results, team managers may whish to exploit the cultural capital of workers whenever possible.

## 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., Gallego, F. A. and Robinson, J. A. (2014), 'Institutions, human capital, and development', *Annual Review of Economics* **6**(1), 875–912.

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.

Angrist, J. and Kolesár, M. (2024), 'One instrument to rule them all: The bias and coverage of just-ID IV', *Journal of Econometrics* **240**(2), 105398.

Bakens, J., Mulder, P. and Nijkamp, P. (2013), 'Economic impacts of cultural diversity in the Netherlands: Productivity, utility, and sorting', *Journal of Regional Science* **53**(1), 8–36.

Bazzi, S., Fiszbein, M. and Gebresilasse, M. (2020), 'Frontier culture: The roots and persistence of "rugged individualism" in the united states', *Econometrica* **88**(6), 2329–2368.

Bazzi, S., Gaduh, A., Rothenberg, A. D. and Wong, M. (2016), 'Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia', *American Economic Review* **106**(9), 2658–98.

Bird, J. and Straub, S. (2020), 'The brasilia experiment: the heterogeneous impact of road access on spatial development in brazil', *World Development* **127**, 104739.

Bisin, A. and Verdier, T. (2011), The economics of cultural transmission and socialization, *in* 'Handbook of Social Economics', Vol. 1, Elsevier, pp. 339–416.

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

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. (2015), 'Technological change and labor selection in agriculture: Evidence from the Brazilian soybean revolution', *Working Paper* .

Bublitz, J. (2008), 'O recomeço na mata: notas para uma história ambiental da colonização alemã no rio grande do sul', *História Unisinos* **12**(3), 207–218.

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. (2020), 'Capital accumulation and structural transformation', *Quarterly Journal of Economics* **135**(2), 1037–1094.

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.

Carvalho Filho, I. 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.

Cavnar, W. B. and Trenkle, J. M. (1994), 'N-gram-based text categorization', *Ann Arbor MI* **48113**(2), 161–175.

Combes, P.-P., Duranton, G. and Gobillon, L. (2008), 'Spatial wage disparities: Sorting matters!', *Journal of Urban Economics* **63**(2), 723–742.

De la Roca, J. and Puga, D. (2017), 'Learning by working in big cities', *Review of Economic Studies* **84**(1), 106–142.

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

Ehrl, P. (2017), 'Minimum comparable areas for the period 1872-2010: An aggregation of Brazilian municipalities', *Estudos Econômicos (São Paulo)* **47**(1), 215–229.

Ehrl, P. and Monasterio, L. (2021), 'Spatial skill concentration agglomeration economies', *Journal of Regional Science* **61**(1), 140–161.

Ehrl, P. and Monasterio, L. (2024), 'Inherited cultural diversity and wages: surname-based evidence', *Journal of Economic Geography* **24**(4), 595–614.

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.

Galor, O., Moav, O. and Vollrath, D. (2009), 'Inequality in landownership, the emergence of human-capital promoting institutions, and the great divergence', *Review of Economic Studies* **76**(1), 143–179.

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.

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

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

Moretti, E. (2004), 'Estimating the social return to higher education: Evidence from longitudinal and repeated cross-sectional data', *Journal of Econometrics* **121**(1-2), 175–212.

Naritomi, J., Soares, R. R. and Assunção, J. J. (2012), 'Institutional development and colonial heritage within Brazil', *The journal of Economic History* **72**(2), 393–422.

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.

Nunn, N. and Puga, D. (2012), 'Ruggedness: The blessing of bad geography in africa', *Review of Economics and Statistics* **94**(1), 20–36.

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.

Ozgen, C. (2021), 'The economics of diversity: Innovation, productivity, and the labour market', *Journal of Economic Surveys* **35**(4), 1168–1216.

Pellegrina, H. S. (2022), 'Trade, productivity, and the spatial organization of agriculture: Evidence from brazil', *Journal of Development Economics* **156**, 102816.

Pellegrina, H. S. and Sotelo, S. (forthcoming), 'Migration, specialization, and trade: Evidence from the Brazilian march to the west', *Journal of Political Economy* .

Pérez, S. (2021), 'Southern (american) hospitality: Italians in argentina and the united states during the age of mass migration', *Economic Journal* **131**(638), 2613–2628.

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.

Roche, J. (1969), *A colonização alemã e o Rio Grande do Sul*, 1 edn, Editôra Globo, Porto Ale.

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.

Sequeira, S., Nunn, N. and Qian, N. (2020), 'Immigrants and the making of america', *The Review of Economic Studies* **87**(1), 382–419.

Seyferth, G. (1994), 'Identidade étnica, assimilação e cidadania: a imigração alemã e o estado brasileiro', *Revista Brasileira de Ciências Sociais* **9**(26), 103–122.

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

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

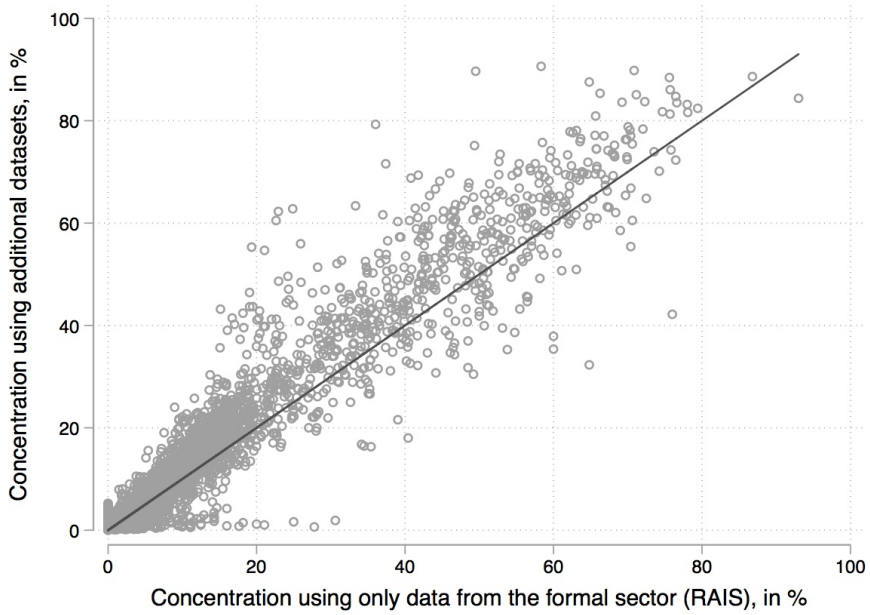
Von Berlepsch, V. and Rodríguez-Pose, A. (2021), 'The missing ingredient: distance. internal migration and its long-term economic impact in the united states', *Journal of Ethnic and Migration Studies* **47**(9), 2198–2217.

Voth, H.-J. (2021), Persistence–myth and mystery, *in* 'The Handbook of Historical Economics', Elsevier, pp. 243–267.

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

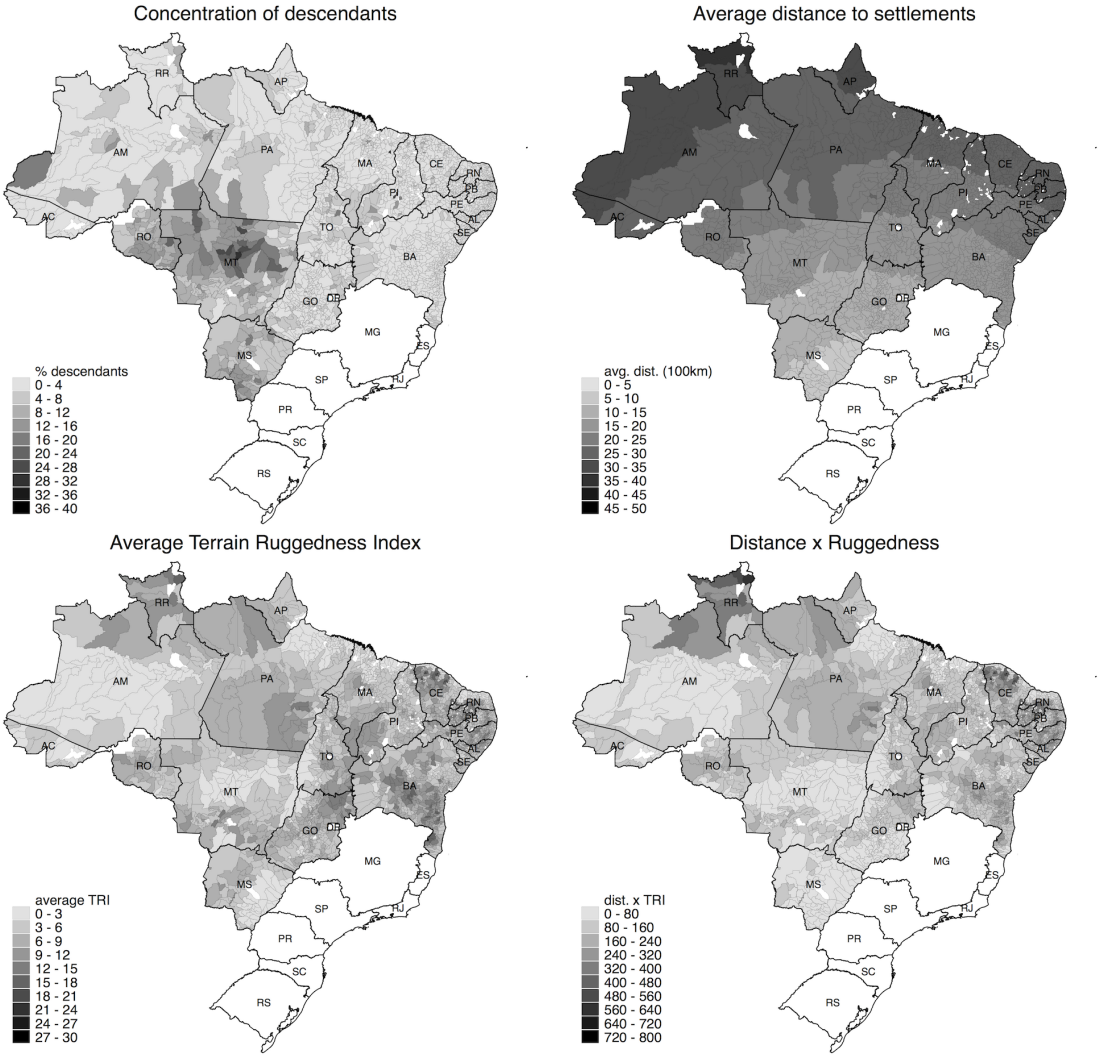
# Appendix – For Online Publication

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



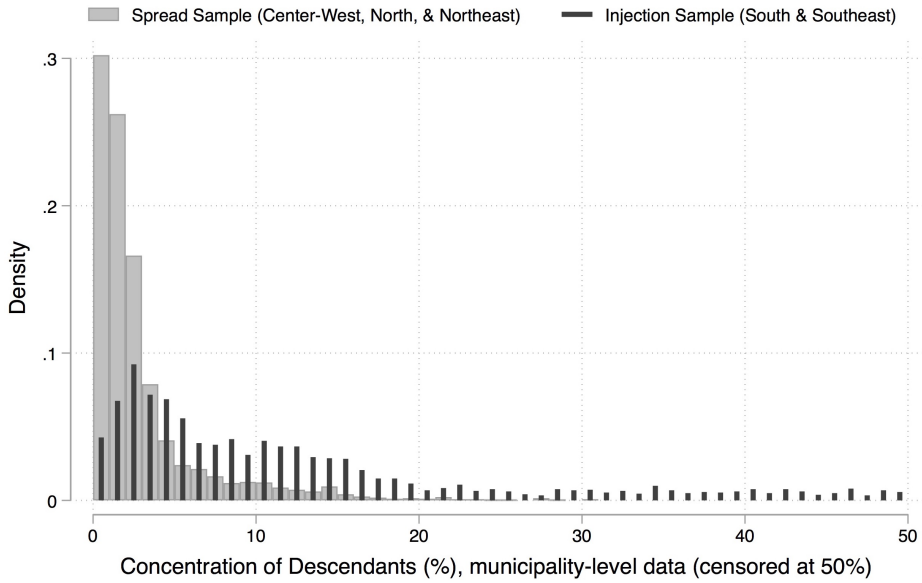

Note: The concentration of descendants in the horizontal axis uses only data from RAIS, while the one in the vertical axis includes also information from the *Cadastro Único* and the *Base Sócios* datasets. The graphs also include a 45° line.

Figure A2: The concentration of descendants, the terms used in the instrument (distance to injection points and Terrain Ruggedness Index), and their interaction



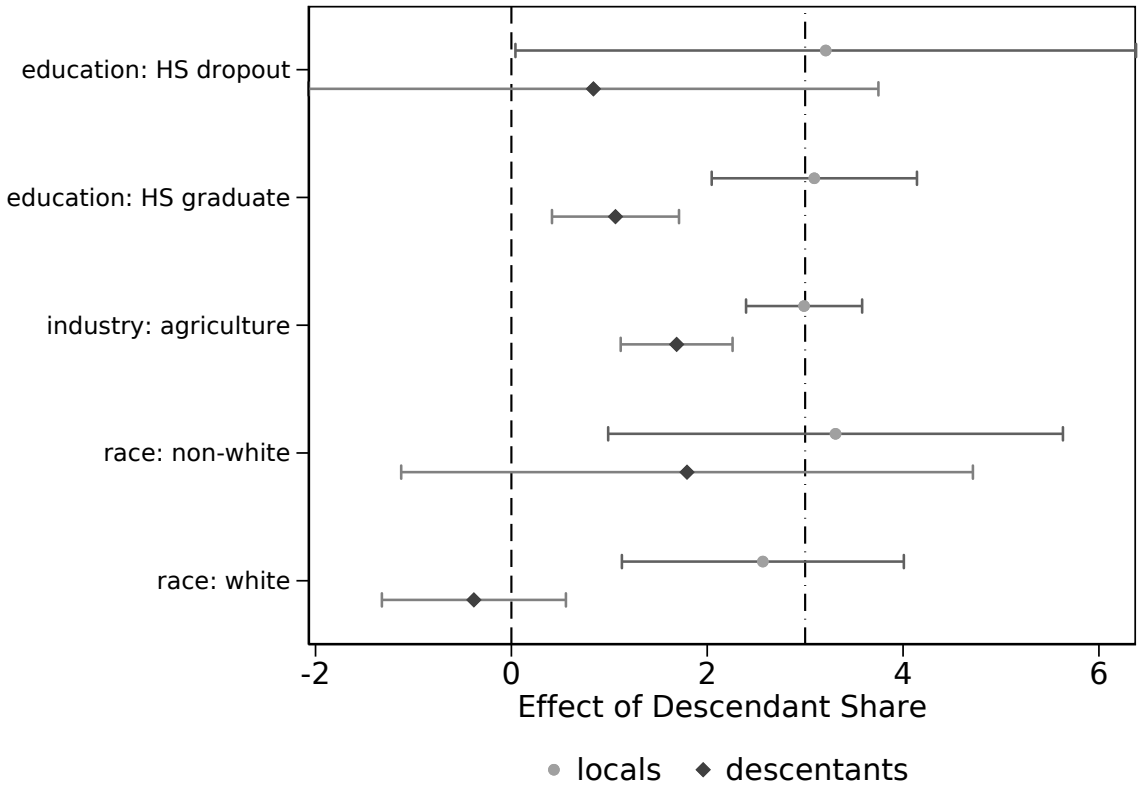

Notes: The first map (top left) shows the concentration of descendants in the study region in 2010, the measure we are instrumenting for. The second map (top right) shows the average of the distances from each municipality to the injection points of historical non-Iberian immigration in the states of São Paulo and Rio Grande do Sul. The third map (bottom left) shows the 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 A3: 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, whereas the spread sample considers municipalities in the states of the regions Center-West, North, and Northeast. Both samples exclude state capitals and municipalities with fewer 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%.

Figure A4: Heterogeneity in the marginal effects – Two-step Wage Regressions with IV




Note: All point estimates and 95% confidence intervals shown in the figure come from The second stage of the two-step wage regressions with one IV according to equation (7) for either local (grey points) or descendant workers (black diamonds). The dependent variable thus is the estimated municipality fixed effect obtained from the first-stage worker-level wage regressions that also include age, age squared, worker and year fixed effects as controls. Descendants (%) represents the share of workers with a non-Iberian surname among the total number of formal employments in the municipality. The descendant share is instrumented by the average of the distance to historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul. This average distance to settlements (expressed in 1,000km) 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 Sao Paulo as recorded by (Carvalho and Monasterio, 2012) and (Rocha et al., 2017), respectively. Estimations are based on the injection sample that includes all municipalities in the Center-West, North, and Northeast with at least one descendant and 30 or more individual observations in our selected RAIS data. The sample excludes females. All regressions are weighted by the number of workers. The binary indicators on the vertical axis split the sample into into groups according to characteristics of workers' high-school (HS) education, industry occupation and race.

Table A1: Two-step Wage Regressions – Descendants vs. Locals, 2008

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Dependent Variable: | Estimated municipality fixed effect | | | |
| Sample: | Injection | | Spread | |
| Years: | 2008 | 2008 | 2008 | 2008 |
| Workers: | Descendants | Locals | Descendants | Locals |
| Descendants share | 0.038 | 0.258*** | -0.198** | 0.847*** |
| | (0.028) | (0.032) | (0.084) | (0.143) |
| Mean of Descendants | 0.15 | 0.15 | 0.04 | 0.04 |
| Adj. $R^2$ | 0.59 | 0.46 | 0.52 | 0.61 |
| Observations | 2,823 | 2,827 | 2,376 | 2,461 |

Notes: The table reports coefficients from the two-step wage regressions according to equations (6) and (7). Descendants (%) represents the share of workers with a non-Iberian surname among the total number of formal employments in the municipality. All regressions are weighted by the number of workers, either locals or descendants, according to the fifth row, in the municipality and include the following municipality-level controls: historical average (1981–2010) and standard deviation of total yearly rainfall and average temperature, total population (log), municipality area (log), and five dummies for soil types (dummies = 1 if 5% or more of municipality area is covered by soil type). The controls in the first-stage worker-level wage regressions include age, age squared, as well as municipality, worker, and year fixed effects. Clustered standard errors at the municipality-level in parentheses. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$

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 Sao 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: Two-step Wage Regressions with and without instrumentation – Robustness

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dependent Variable: | Estimated municipality fixed effect | | | | | |
| Sample | Injection | | | | | |
| Years | 2019 | 2019 | 2019 | 2019 | 2019 | 2019 |
| Workers | Locals | | | | | |
| Instrument | | None | | Colony Distance | | |
| Check | 5-95 pct. | low 1920 foreigners | institution controls | 5-95 pct. | low 1920 foreigners | institution controls |
| Descendants share | 0.924*** | 0.661*** | 0.713*** | 2.984*** | 4.892*** | 2.567*** |
| | (0.183) | (0.172) | (0.176) | (1.073) | (1.244) | (0.739) |
| Observations | 824 | 2,428 | 2,451 | 824 | 2,428 | 2,451 |

Notes: The table reports coefficients from the two-step wage regressions with instrumentation according to equations (6) and (7). The dependent variable thus is the estimated municipality fixed effect obtained from the first-stage worker-level wage regressions that also include age, age squared, worker and year fixed effects as controls. Descendants (%) represents the share of workers with a non-Iberian surname among the total number of formal employments in the municipality. The descendant share is instrumented by the average of the distance to historical non-Iberian settlements in the states of São Paulo and Rio Grande do Sul in columns (3) to (6). The spread sample includes all municipalities in the Center-West, North, and Northeast with at least one descendant and 30 or more individual observations in our selected RAIS data. The sample excludes females. All regressions are weighted by the number of workers in the municipality. The robustness checks for the estimation with and without IV are the following: columns (1) and (4) exclude municipalities outside the 5th and 95th percentile of the distribution of observations per municipality weighted by the number of workers; Columns (2) and (5) exclude municipalities with a foreigner density in 1920 above 0.04 (the 50th decile); Columns (3) and (6) include proxy variables for governmental quality and access to justice from Naritomi et al. (2012), and the educational performance of 9th graders in the *Prova Brasil*, a nationwide exam conducted by INEP (the National Institute of Educational Studies and Research). Clustered standard errors at the municipality-level in parentheses. $^{*}$ $p < 0.10$, $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$