Migration and the epidemiological approach: time and self-selection into foreign ancestries matter

Simone Bertoli, Melchior Clerc, Jordan Loper, Èric Roca Fernández

Centre d'Études
et de Recherches
sur le Développement
International

CERDI WORKING PAPERS
2024/4

Migration and the epidemiological approach: time and self-selection into foreign ancestries matter

Simone Bertoli
Melchior Clerc
Jordan Loper
Èric Roca Fernández

To cite this working paper:

Bertoli, S., Clerc, M., Loper, J., Roca Fernández, E. "Migration and the epidemiological approach: time and self-selection into foreign ancestries matter", CERDI Working Papers, 2024/4, CERDI.

CERDI, Pôle tertiaire, 26 avenue Léon Blum, 63000 Clermont-Ferrand, France.

The authors

Simone Bertoli
Professor, Université Clermont Auvergne, CNRS, IRD, CERDI, 26 Av. Léon Blum, F63000, Clermont-Ferrand
IC Migrations, IZA
Email: simone.bertoli@uca.fr

Melchior Clerc
Ph.D candidate, Université Clermont Auvergne, CNRS, IRD, CERDI, 26 Av. Léon Blum, F63000, Clermont-Ferrand
Email: melchior.clerc@doctorant.uca.fr

Jordan Loper
Associate Professor, Université Clermont Auvergne, CNRS, IRD, CERDI, 26 Av. Léon Blum, F63000, Clermont-Ferrand
Email: jordan.loper@uca.fr

Èric Roca Fernández
Assistant Professor, Université Clermont Auvergne, CNRS, IRD, CERDI, 26 Av. Léon Blum, F63000, Clermont-Ferrand
Email: eric.roca_fernandez@uca.fr (corresponding author)

Acknowledgments

The authors gratefully acknowledge the comments provided by the guest editor Hillel Rapoport, by two anonymous referees, Travis A. Baseler, Martin Fernández, Jesús Fernández-Huertas Moraga, David Gomtsyan, Gordon Hanson, Francesca Marchetta and Gianluca Orefice, and from participants to the 16th AFD-World Bank Conference on Migration and Development (Boston U.), the 13th Conference on Immigration to OECD Countries (Paris), the 2024 Conference on Deep-Rooted Factors in Comparative Development (Brown), the 8th Conference on Understanding Voluntary and Forced Migration (Lille), the 3rd International Workshop on Migration and Family Economics (Paris), the 23rd Journées Louis-André Gérard-Varet (Marseille), the 4th International Conference on Development Economics (Marseille), and to seminar presentations at University of Naples Parthenope, IFPRI, CERDI; they also gratefully acknowledge the support received from the Agence Nationale de la Recherche of the French government through the program France 2024 (ANR-16-IDEX-0001); the usual disclaimers apply.

This work was supported by the LABEX IDGM+ (ANR-10-LABX-0014) within the program "Investissements d'Avenir" operated by the French National Research Agency (ANR).

CERDI Working Papers are available online at: https://tinyurl.com/2xwfzw8s

Director of Publication: Simone Bertoli
Editor: Catherine Araujo Bonjean
Publisher: Marie Dussol
ISSN: 2114 - 7957

Abstract

Data on individuals of immigrant origin are used in the epidemiological approach in comparative development for understanding cultural persistence, the determinants of cultural norms, and the effects of genetic traits. A widespread presumption is that this approach is exposed to attenuation bias. We describe how the increasing reliance on foreign ancestries to identify respondents' origin can invalidate this presumption. Self-selection into reporting a foreign ancestry and unobserved heterogeneity in the time elapsed since ancestral migration can overestimate the effect of interest. A simple theoretical framework describes the joint influence of these two factors on the estimates obtained from a canonical specification. We provide illustrative examples of the empirical relevance of our concerns drawing on two influential papers in the literature: Fernández and Fogli (2006) and Giuliano and Nunn (2021).

Keywords

Comparative development; migration; ancestry; culture; identity choice.

JEL Codes

F22; O12; Z10.

1 Introduction

Understanding the deep-rooted determinants of economic development is a fundamental question that has garnered considerable attention among economists. This has sparked a burgeoning literature, revealing how contemporary outcomes such as economic development, inequality, and individual behaviors are influenced by the persistent characteristics of a distant past. Furthermore, this literature has also provided abundant evidence of the persistence of cultural and genetic traits.

Isolating the long-term causal effect of such traits is a vivid challenge due to the likely emergence of numerous credible confounders. To address this, economists have extensively relied on the so-called *epidemiological approach*, an expression introduced by Raquel Fernández.[1]

This empirical strategy requires the definition of a criterion linking each individual to a single origin. The literature, which faces binding data constraints, has used four main criteria: country of birth (Antecol, 2000; Luttmer and Singhal, 2011), maternal or paternal country of birth (see, for instance, Fernández, 2007; Giuliano, 2007; Galor and Savitskiy, 2018; Galor et al., 2020; Giuliano and Nunn, 2021; Ek, 2024), foreign ancestry (see, for instance, Antecol, 2000; Guiso et al., 2006; Fernández and Fogli, 2006; Algan and Cahuc, 2010; Alesina et al., 2015; Galor and Özak, 2016; Giavazzi et al., 2019; Arbatlı et al., 2020; Giuliano and Nunn, 2021; Ek, 2021; Galor et al., 2023), and language (see, for instance, Alesina et al., 2003; Desmet et al., 2017; Giuliano and Nunn, 2021).[2] None of these criteria, except one's own country of birth, precisely define an individual's ancestry, nor do they necessarily identify a single foreign origin: parents may be born in different foreign countries, a language can be spoken in various countries, and multiple ancestries may be reported. Furthermore, the various criteria might relate an individual to different origins, e.g., an individual born in Spain might speak French at home, and declare to be of Italian ancestry. The distinctions between origins are relevant

[1] "The essence of what I call the epidemiological approach is the attempt to identify the effect of culture through the variation in economic outcomes of individuals who share the same economic and institutional environment, but whose social beliefs are potentially different." (Fernández, 2011).

[2] Additional steps are usually necessary to associate each of these four criteria to the variable(s) whose influence is tested in the econometric analysis to account for the population movements associated to the Columbian exchange; for instance, Giuliano and Nunn (2018, 2021) rely on the mapping of languages for each country from the Ethnologue combined with data on the spatial distribution of the population to aggregate data at the language (ethnicity) level for each country, while Galor et al. (2023) rely on the data from Putterman and Weil (2010).

for the econometric analysis to the extent that different origins are characterized by different norms, values of the underlying determinants of cultural traits, or genetic factors.

The multifaceted nature of one's own cultural identity, the fact that migrants from all origin countries are likely to be self-selected along some cultural traits (e.g., long-term orientation and attitude towards risk), and the fact that cultural homophily implies that migrants moving to a given destination from different countries are likely to be more similar than the populations at origin are (see, for instance, Antecol, 2000; Alesina et al., 2015, on this), all strongly suggest that the epidemiological approach is exposed to an attenuation bias. Indeed, the widely cited review of the literature by Fernández (2011) observes that:

> "It should be noted explicitly that the epidemiological approach is biased towards finding that culture does not matter. As mentioned previously, the fact that parents are only one source of cultural transmission among many and that they may have cultural attitudes that differ from the average ones in the country of ancestry implies that one is more likely to rule the cultural proxy insignificant."

This widespread presumption might not apply to the increasingly large number of papers that utilizes *self-reported* ancestry data from U.S. census or survey sources to define respondents' foreign origin. The reliance on this variable stems from concerns over using first-generation immigrants, where non-random selection based on unobservable factors specific to their origin could potentially bias the analysis.[3] Additionally, this approach is often necessitated by limitations in available data. [4] The literature has, so far, failed to consider that concerns related to non-random selection in unobservables also apply to natives of foreign ancestry, and that these can counteract the tendency towards an attenuation bias. The two main sources of these concerns are related to (i) the very large fraction of the native population that does not report a valid foreign ancestry, i.e., an ancestry that can be connected to a foreign country, and to (ii) the unobserved heterogeneity in the time elapsed since ancestral migration for natives reporting different foreign ancestries.[5]

[3]For instance, the review of the literature by Alesina and Giuliano (2015) writes that "the literature has been using mostly second-generation immigrants, who constitute a more appropriate sample than first-generation immigrants because issues of disruption and selection due to migration are more attenuated." (Alesina and Giuliano, 2015, pp. 903-4).

[4]Paternal and maternal country of birth are no longer recorded in the population censuses since 1980, and also not included in the American Community Survey.

[5]Ek (2021) expresses a concern about this dimension of unobserved heterogeneity when he writes that "census data does not identify the time of immigration or the country of birth (after 1970), of parents and

With respect to point (i), Brittingham and de la Cruz (2004) show that 19.9 percent of the total population simply does not report any ancestry in the 2000 population census. An even larger share of the population reports either a native ancestry, e.g., African American (8.8 percent), American (7.2), American Indian (2.8), a supranational ancestry,[6] e.g., Hispanic (0.9), European (0.7) or African (0.4), or (less importantly) it reports an ancestry that cannot be unambiguously connected to a current country, e.g., Basque, Czechoslovakian, Prussian, or Yugoslavian. Our analysis of the data reveals that 44.3 percent of the native population cannot be connected to a foreign country of ancestry. Natives whose ancestors came from a given foreign country might not report an ancestry in that country simply because they lack, as observed by Farley (1991), the necessary "factual knowledge" about the history of their family, or they might deliberately decide not to do so.[7] A legitimate presumption is that, if we consider a group of natives whose ancestors came to the United States from a given foreign country, those choosing to report the corresponding foreign ancestry have a greater attachment to their ancestral culture than those who chose not to report that ancestry. This, in turn, implies that the individuals entering into the estimation sample are selected in a way that likely increases the chances of finding a significant effect. The intensity of self-selection into reporting a foreign ancestry might also vary across ancestral countries, and be stronger for ancestries towards which natives have, on average, a more negative attitude because of the perception of a greater cultural distance with the United States,[8] but this is not a necessary

grandparents of the individuals included, which makes it impossible to distinguish between second and higher generations of immigrants; meaning, for example, [that] people whose ancestors arrived 20 or 200 years ago are treated identically." (p. 23).

[6]The Census Bureau refers to these supranational ancestries as "general heritages".

[7]Related to this, Duncan and Trejo (2011, 2017) and Antman et al. (2016, 2023) provide evidence of the substantial incidence of what they describe as "ethnic attrition", i.e., the propensity of individuals living in the United States who are first to third-generation immigrants from Latin American or Asian countries *not* to report that they are Hispanics or Asians when answering to questions related to race or ethnicity; ethnic attrition increases with the time elapsed since ancestral migration, it greatly varies across origin countries, and across individuals, e.g., higher-educated individuals with a Mexican ancestry are significantly less likely to identify themselves as Hispanics.

[8]As the data do not provide any objective information on the origin of one's own ancestors, e.g., the countries of birth of the grandparents or of the great-grandparents, providing evidence on the origin-specific incidence of not reporting a given ancestry is unfeasible; Fulford et al. (2020) provide an aggregate measure of the ancestral countries of origin for the population of each county in the United States; their approach cannot be deployed at the individual-level.

condition for a threat to identification.

As far as point (*ii*) is concerned, the different time profiles of past migration flows to the United States from various countries imply that the time elapsed since ancestral migration varies across groups of natives reporting different foreign ancestries.[9] As preferences are also shaped by horizontal interactions in a model *à la* Bisin and Verdier (2001), a greater elapsed time can induce a greater convergence towards the local cultural norms. For instance, Guiso et al. (2006) argue that differences in the attitudes towards redistribution of natives of different foreign ancestries could reflect differences in the time elapsed since ancestral migration,[10] and Giavazzi et al. (2019) provide empirical evidence of this for various cultural traits. The time elapsed since ancestral migration cannot be inferred from the census or from the American Community Survey, and this can confound the econometric analysis.[11] Unobserved heterogeneity in the time since ancestral migration can induce an overestimation of the extent of cultural persistence if origin countries with more distant cultural norms have a more recent history of migration to the United States. Similarly, it can confound the analysis of the effect of an underlying determinant of a cultural trait if this is systematically correlated with the time elapsed since ancestral migration. Such a correlation is likely to arise, given the gradual expansion of the countries of origin of migration flows to the United States (see, for instance, Abramitzky and Boustan, 2017, on this), and the absence of major spatial discontinuities in the geographical determinants that are analyzed in the epidemiological approach.

Our theoretical and empirical analysis is made up of four distinct but closely interrelated parts. First, a description of the canonical specification used in the epidemiological approach, and the different ways to identify the foreign origin of an individual. Second, a simple theoretical framework that describes the influence of self-selection into reporting a foreign ancestry

[9] Clearly, this second dimension of unobserved heterogeneity does not solely depend on migration history, but it is closely intertwined also with the first one, as it depends on the extent to which (say) natives who are third-generation immigrants from different origin countries still claim a foreign ancestry.

[10] "Americans with British, North European or German ancestors derive from earlier immigrants; hence, more generations were raised in the United States and forged by its culture, absorbing the belief that success is mostly determined by individual actions, which makes government intervention highly undesirable" (Guiso et al., 2006, p. 41).

[11] To the best of our knowledge, only three papers consider the time elapsed since immigration: Algan and Cahuc (2010), Alesina et al. (2015) and Giavazzi et al. (2019). These all draw on data from the General Social Survey containing information on both self-reported foreign ancestry and on the number of parents and grandparents that were born abroad to differentiate between first to (at least) fourth-generation immigrants.

and of the time elapsed since ancestral migration on the estimates obtained from the canonical specification. Third, an analysis the data from the 1980 to 2000 censuses (Ruggles et al., 2023) to document the (aggregate) incidence of not reporting a foreign ancestry, the variations over time (for a given cohort of natives) of the answers provided to the question on ancestry and, for the only ancestry for which this is possible, the incidence and the selective character of *not* reporting a Mexican ancestry. The analysis also evidences the origin-specific heterogeneity in time since ancestral migration for 109 distinct countries of ancestry. This can be, in an admittedly crude way because of binding data-constraints, measured from the data. In particular, we analyze the share of the population of a given ancestry born in the ancestral country, the spatial concentration of native of different foreign ancestries in the United States, and the incidence of the use of the ancestral language. All these three variables should decline with the average time elapsed since ancestral migration,[12] and we validate this by comparing them with a measure of the average year of immigration constructed from the 1850 to 1970 population censuses.[13] Fourth, two illustrative examples, drawing on two influential papers in the literature, namely Fernández and Fogli (2006) and Giuliano and Nunn (2021), of the empirical relevance of our concerns. We find that cultural persistence in fertility choices in Fernández and Fogli (2006) does not extend to natives that are (at least) fourth-generation immigrants, and which represent 68.9 percent of the estimation sample. With respect to Giuliano and Nunn (2021), we find that the size of the estimated effect of historical climatic variability on the propensity to speak the ancestral language in the United States, which is used as a proxy for the importance of tradition, is greatly influenced by the natives with an ancestry in a Spanish-speaking country of ancestry. These countries, which represent 21 of the 84 countries included in the estimation sample but only 11.3 percent of the observations, are characterized by a short time elapsed since ancestral migration. This favors the transmission of this ancestral language, and the incentives to speak Spanish in the United States are also clearly magnified by the large fraction of the native and foreign-born population speaking this language. Our objective here is not to challenge the findings of these two papers, which our

[12]With respect to spatial concentration, this can be due, for instance, to the fact that the importance of migrant networks (see, for instance Patel and Vella, 2013) declines over time, because the relevance of the local labor demand conditions that shaped the initial spatial distribution of the immigrants fades away, or to the fact that natives of immigrant descent move away from more expensive locations where their foreign-born ancestors tended to concentrate (Albert and Monras, 2022).

[13]Because of binding data constraints, this approach is possible only for 66 out of 109 ancestral countries.

analysis actually confirms, but rather to alert about the possible (and, in these two cases, large) confounding effect due to unobserved heterogeneity among natives of foreign ancestry.

Our paper makes two major contributions to the literature on comparative development: First, we provide a theoretical framework to think about the specific threats to identification that can arise, in the epidemiological approach, when using the answers to the questions on ethnic or ancestral origin. This is particularly important as data constraints leave little to no room for alternative approaches for native individuals. Our contribution here with respect to Giavazzi et al. (2019) is to explore the implication of the self-reported nature of foreign ancestry. Our theoretical framework neatly reveals that even an incidence of not reporting a foreign ancestry that is invariant across origins can induce a bias in the estimation, even if the confounding effect of unobserved time since ancestral migration is fully controlled for. The restriction of the estimation sample to the subset of natives who chose to report a foreign ancestry increases the chances of finding a significant effect. Second, it provides evidence about the major extent of heterogeneity across natives in the United States with distinct self-reported foreign ancestries. The variables that we build from standard data can be used as a diagnostic tool, to see whether origin countries that are clear outliers with respect to the time elapsed since ancestral migration have a large influence on the size of the estimated effect.

The rest of the paper is structured as follows: Section 2 briefly describes the canonical equation that is brought to the data, and the possible threats to identification related to three distinct criteria that can be adopted to identify the origin of each respondent; Section 3 presents a simple theoretical framework describing the threat to identification posed by self-selection into reporting a foreign ancestry and by unobserved heterogeneity in the time since ancestral migration, and derives theoretically consistent specification; Section 4 describes the data sources that we employ in the analysis; Section 5 presents the results from our descriptive empirical analysis, documenting the extent to which different foreign countries of ancestry differ with respect to the time elapsed since ancestral migration; Section 6 presents the two illustrative examples of the empirical relevance of our concerns, and Section 7 draws the main conclusions.

2 The epidemiological approach

Let us consider a set of individuals or groups of individuals indexed by i, residing in the location k within a single country d, and let o represent the origin of an individual or group, as defined later. Suppose that cross-sectional data are collected at time T. The typical regression that is brought to the data in the epidemiological approach can be written as follows, omitting the subscripts for the (only) country d and survey time T:

$$y_{iok} = \alpha w_o + \boldsymbol{\beta}'\boldsymbol{x}_o + \boldsymbol{\gamma}'\boldsymbol{x}_i[+\boldsymbol{\phi}'\boldsymbol{x}_{ok} + d_k + \lambda f(t_i)] + \epsilon_{iok} \tag{1}$$

where y_{iok} is the dependent variable, w_o is the origin-specific variable of interest, and $\boldsymbol{x}_o$ and $\boldsymbol{x}_i$ are two vectors of origin-level and individual-level variables, $\boldsymbol{x}_{ok}$ and d_k are respectively a vector of origin-location control variables and dummies for the area of residence, t_i is the time elapsed since ancestral migration for individual i, possibly transformed through a nonlinear function $f(t_i)$, and ϵ_{iok} is the error term. The notation in Eq. (1) is meant to reflect the fact that $f(t_i)$, $\boldsymbol{x}_{ok}$, and d_k are *not* systematically included.[14] Similarly, the vector $\boldsymbol{x}_o$ does not include variables measured from the population from the origin o residing in the country of destination d.

The coefficient of interest in Eq. (1) is α, and the identifying assumption is that:

$$E(w_o \times \epsilon_{iok} | \boldsymbol{x}_o, \boldsymbol{x}_i, [\boldsymbol{x}_{ok}, d_k, f(t_i)]) = 0.$$

Violations of this assumption arise, for instance, if (i) individuals from different origins descend from migrants who moved to the destination d at different points in time prior to T, λ is different from 0, and the average origin-specific time since ancestral migration is correlated with w_o, or if (ii) individuals associated with various origin countries are differently self-selected with respect to unobserved determinants of y_{iok}, when the intensity of this non-random selection is correlated with w_o. The relevance of these two potential threats to identification clearly depends on the criterion that is chosen to connect each individual (or group of individuals) in the sample to her own origin. Let us focus here on three different criteria, which implies also different sample selection criteria: country of birth, parental (paternal or maternal) country of birth, and country of foreign ancestry.

[14] Giuliano and Nunn (2021) and Arbatlı et al. (2020) are exceptions to the norm insofar as the authors include area of residence fixed-effects d_k, and the former also includes a vector $\boldsymbol{x}_{ok}$; however, individual time since ancestral migration t_i remains unaccounted for.

2.1 Country of birth

This criterion to determine the origin o restricts the sample to first-generation immigrants. Thus, it reduces the concerns related to point (i) above,[15] but, conversely, it magnifies the concerns related to point (ii) because immigrants from country o are likely to differ from the stayers along several cultural traits (see, for instance, Antecol, 2000; Alesina et al., 2015; Alesina and Giuliano, 2015, about this concern).[16] To provide just an example, migrants might be be positively self-selected with respect to their long-term orientation and negatively with respect to their risk aversion. Thus, the expected value of the error term ϵ_{iok} in Eq. (1) is likely to vary with the origin o.

2.2 Parental country of birth

This criterion for determining the origin o restricts the sample to natives with at least one foreign-born parent, i.e., second-generation immigrants. While this criterion accentuates concerns related to unobserved heterogeneity in t_i, the time elapsed since the migration to country d of the parents of individual i,[17] it can alleviate concerns related to point (ii). The extent of non-random selection in unobservables may be diminished when transitioning from the first to the second generation of migrants (see Giavazzi et al., 2019, on this), under the plausible assumption of an imperfect vertical transmission of cultural traits (Bisin and Verdier, 2001). If questions on the country of birth of the parents of the respondent are not included, then the adoption of this criterion requires a restriction of the sample to the individuals who co-reside with their parents, a potentially self-selected sample.

2.3 Country of foreign ancestry

The utilization of self-reported country of foreign ancestry narrows down the sample used for estimating Eq. (1) to native individuals who declare a foreign country as their ancestral origin. This third criterion exacerbates concerns related to unobserved heterogeneity in the

[15] These can be further mitigated by including the years since migration in the vector $\boldsymbol{x}_i$ in Eq. (1).

[16] See, for instance, Jaeger et al. (2010) and Beck Knudsen (2022) for empirical evidence on migrants' selection along cultural traits.

[17] For instance, the same survey can feature an 80 years old individual born in the United States whose parents migrated at the age of 30, and a young individual born to parents that just arrived into the country; in this example, the difference in time since parental migration may be close to a century.

time since ancestral migration t_i, as the data lacks information on the number of generations that separate individual i from ancestors who migrated to country d.

While this criterion seemingly alleviates concerns related to point (*ii*) above, as greater values of t_i can further attenuate the initial migrants' non-random selection in unobservables,[18] this dilution may *not* apply to individuals reporting foreign ancestry. Indeed, some individuals with ancestors from a specific country (e.g., Mexico) might choose not to report any ancestry or to declare an American ancestry or "general heritage" (Hispanic) ancestry, resulting in their exclusion from the estimation sample.

For natives of immigrant origin, the choice of reported ancestry reflects the evaluation of costs and benefits of each possible identity (Akerlof and Kranton, 2000) and parental efforts to transmit identity to children (Bisin and Verdier, 2001).[19] In the early stages of migration, migrants from most origin countries faced negative attitudes in the United States, as the following two examples, drawn from Fasani et al. (2019), on Irish and Chinese immigrants suggest:

> "Just outside the US borders were 'hordes of Wild Irishmen ... the turbulent and disorderly of all the world [who come to the United States in order to] distract our tranquillity.' (Massachusetts Representative Harrison Gray Otis, 1797). [...] Chinese immigrants were 'morally the most debased people on the face of the earth' (Connecticut Senator Orville Platt, 1882), who 'bring every character of vice ... [and would be] injurious in every sense of the world' (Texas Senator Samuel Bell Maxey, 1882)."

Historically, Italians also faced long-lasting negative attitudes (see, for instance, Fouka et al., 2021), as individuals originating from Latin American countries still experience in the United States (see, for instance, Chavez, 2013). Thus, natives of immigrant origin who choose to report a foreign ancestry, despite natives' potential negative attitudes, likely have a stronger attachment to this (costly) identity than their counterparts who decide not to declare this ancestry. Similarly, their intensity of non-random selection in unobservables is likely to be stronger than for the individuals reporting an ancestry that natives favorably regard.[20]

[18] Indeed, the cultural traits of natives with migrant ancestors are more and more determined by a horizontal transmission through social interactions within country d as t_i increases, and they depend less on the country of ancestral origin o.

[19] Antman and Duncan (2023) have recently developed a similar theoretical framework to the analysis of identity choice among Native Americans in the United States.

[20] Duncan and Trejo (2011) and Duncan and Trejo (2017) provide evidence of the extent to which ethnic

As the concerns related to the use of the country of foreign ancestry have not been, so far, sufficiently described or discussed in the literature, we devote the next section to a simple theoretical framework that analyzes their joint influence on the estimation of Eq. (1).

3 A simple theoretical framework

Consider two foreign countries of origin, $o = \{A, B\}$. These two countries differ with respect to the origin-specific variable of interest w_o. For the sake of simplicity, we assume that $w_o = \overline{y}_o$, with $o = \{A, B\}$, i.e., this is the cultural norm prevailing at origin. We also assume, without loss of generality, that $\overline{y}_A < \overline{y}_B$. For simplicity, we assume that the native population with ancestors in one of the these two countries is composed only by two generations j of immigrants: second (natives with foreign-born parents) and third-generation immigrants (natives with native parents and foreign-born grandparents). $s_o \in [0, 1]$, with $o = \{A, B\}$, denotes the share of the native population with ancestors in country o that is second-generation.

y^j_{io} represent the outcome for the native individual i with ancestors in country o, and $j = \{2, 3\}$ denotes her generation (second or third). We assume that $y^2_{io} = \overline{y}_o$,[21] while for third-generation immigrants there is a fraction p_o that adopts the norm of the destination country, which we denote as $\overline{y}$, i.e., $y^3_{io} = \overline{y}$ and a fraction $(1 - p_o)$ maintaining the norm of the origin country, i.e., $y^3_{io} = \overline{y}_o$.[22] The dummy variable D^3_{io} is equal to 1 if individual i adopts the norm of the destination country, and 0 otherwise.

The dummy variable F^j_{io} is equal to 1 if the native individual i who is a second or third-generation immigrant, i.e., $j = \{2, 3\}$, from country o declares country o, with $o = \{A, B\}$, as her ancestry, and 0 otherwise. We assume that $F^2_{io} = 1$, while $q_o \in [0, 1]$ represents the fraction of third-generation immigrants from country o that report an ancestry in country o.

Assume initially that $q_A = q_B = 1$, so that all natives of immigrant origin report a foreign ancestry.

If we run a regression of y^2_{io} on $\overline{y}_o$ on the sample of second-generation immigrants, we get

attrition, defined as the incidence of individuals of a given origin *not* claiming the corresponding identity, varies across origin countries.

[21] Notice that this implies that we are ruling out concerns related to a possible non-random selection in unobservables of the migrants; this allows isolating the effects due to heterogeneity in the time since ancestral migration, and of self-selection into reporting a foreign ancestry.

[22] We can include a zero-mean error term to y^2_{io} and y^3_{io} to allow for individual unobserved heterogeneity.

a value of:

$$\widehat{\alpha}^2 = \frac{\overline{y}_B^2 - \overline{y}_A^2}{\overline{y}_B - \overline{y}_A} = 1 \tag{2}$$

where $\overline{y}_o^2$ represents the average outcome for second-generation immigrants from country o, as the assumptions that we have introduced imply that second-generation immigrants adopt the norm of their country of origin, i.e., $\overline{y}_o^2 = \overline{y}_o$, with $o = \{A, B\}$.

If we run a regression of y_{io}^3 on $\overline{y}_o$ on the sample of third-generation immigrants, we get a value of:[23]

$$\widehat{\alpha}^3 = \frac{\overline{y}_B^3 - \overline{y}_A^3}{\overline{y}_B - \overline{y}_A} \tag{3}$$

where $\overline{y}_o^3 \equiv p_o\overline{y} + (1 - p_o)\overline{y}_o$, with $o = \{A, B\}$, represents the average outcome for third-generation immigrants from country o.

If the data for second and third-generation immigrants are pooled together, the estimated coefficient from a regression that does not control for the generation of a native of immigrant origin would be equal to:

$$\widehat{\alpha}^{\text{pooled}} = \frac{s_B\overline{y}_B^2 + (1 - s_B)\overline{y}_B^3 - [s_A\overline{y}_A^2 + (1 - s_A)\overline{y}_A^3]}{\overline{y}_B - \overline{y}_A}, \tag{4}$$

In the absence of self-selection into reporting a foreign ancestry, then $\widehat{\alpha}^{\text{pooled}}$ simply reflects compositional effects, related to the shares s_A and s_B of second-generation migrants from the two countries.[24] If we relax the assumption that $q_A = q_B = 1$, then selection into reporting a foreign-ancestry among third-generation immigrants would affect $\widehat{\alpha}^3$ and $\widehat{\alpha}^{\text{pooled}}$. The implications crucially depend on the correlation between D_{io}^3, the dummy that signals whether a third-generation immigrant from country o adopts the norm of the destination country, and F_{io}^3, the dummy that signals whether a third-generation immigrant from country o still reports a foreign ancestry. It is plausible to assume that $\text{corr}(D_{io}^3, F_{io}^3) < 0$: individuals who adopt the norm of the destination country are less likely to report a foreign ancestry. We denote with p_o^F the share of third-generation immigrants from country o that still claim a foreign ancestry, and which have adopted the norm of the destination country. If $\text{corr}(D_{io}^3, F_{io}^3) < 0$, then $p_o^F < p_o$.

[23] If $p_A = p_B = p$, then $\widehat{\alpha}^3 = (1 - p)\widehat{\alpha}^2 = 1 - p$.

[24] If $s_A = s_B = s$, then $\widehat{\alpha}^{\text{pooled}} = s\widehat{\alpha}^2 + (1 - s)\widehat{\alpha}^3$.

Under the assumption that the value of the social norm at destination $\bar{y} \in [\bar{y_A}, \bar{y_B}]$[25] then $\widehat{\alpha}^3$ in Eq. 3 is bounded above by $\widehat{\alpha}^2 = 1$, that is, $\widehat{\alpha}^3 < \widehat{\alpha}^2 = 1$ and, $\widehat{\alpha}^{\text{pooled}}$ in Eq. 4 is bounded between, i.e. $\widehat{\alpha}^{\text{pooled}} \in [\widehat{\alpha}^3, \widehat{\alpha}^2]$.

If this is the case, then a regression that is limited to third-generation immigrants reporting a foreign ancestry would give an estimated coefficient equal to:

$$\widehat{\alpha}^3_F = \frac{\overline{y}^3_{B,F} - \overline{y}^3_{A,F}}{\overline{y}_B - \overline{y}_A} \in [\widehat{\alpha}^3, \widehat{\alpha}^2] \tag{5}$$

where $\overline{y}^3_{o,F} \equiv p^F_o \overline{y} + (1 - p^F_o)\overline{y}_o$, with $o = \{A, B\}$, represents the average outcome for third-generation immigrants from country o reporting a foreign ancestry.

In the limit, if $\text{corr}(D^3_{io}, F^3_{io}) = -1$, then $p^F_o = 0$, i.e., all the individuals who adopt the norm of the destination country no longer report a foreign ancestry, then we have that $\widehat{\alpha}^3_F = \widehat{\alpha}^{\text{pooled}}_F = \widehat{\alpha}^2 = 1$, where $\widehat{\alpha}^{\text{pooled}}_F$ corresponds to $\widehat{\alpha}^{\text{pooled}}$ estimated *only* on the sample of individuals reporting a foreign origin. Thus, self-selection into foreign ancestry can lead to overestimate the extent of persistence among third-generation immigrants or in the pooled sample, as the individuals that converge to the norm of the destination country no longer claim a foreign ancestry.

If the norm of one of the two foreign countries coincide with the one of the destination country, e.g., say that $\overline{y}_A = \overline{y}$,[26] then we would have that $\widehat{\alpha}^3_F = \widehat{\alpha}^{\text{pooled}}_F = \widehat{\alpha}^2$ even if there is no self-selection into foreign ancestry among the individuals whose ancestors came from country A. This simply follows from the fact that (for them) the norm of their country of ancestry and of the country of destination are identical, so that $\overline{y}^3_{A,F} = \overline{y}^3_A$. Thus, self-selection into foreign ancestry is a greater source of concern for origin countries that have norms that are further away from those of the destination country.

A specification of Eq. (1) that is consistent with our theoretical framework would take the following form:[27]

$$y_{iok} = \alpha_1 w_o + \alpha_2 (w_o \times f(t_i)) + \boldsymbol{\beta}' \boldsymbol{x}_o + \boldsymbol{\gamma}' \boldsymbol{x}_i + \boldsymbol{\phi}' \boldsymbol{x}_{ok} + d_k + \lambda f(t_i) + \epsilon_{iok} \tag{6}$$

[25] This assumption appears plausible for the specific case of the USA, where the initial population consisted of a mix of British, Irish, and German migrants. Over time, the country experienced migratory inflows from other regions, which likely shifted societal norms to a point somewhere between those of the original settlers and the new arrivals.

[26] As far as many cultural traits are concerned, the norm in the United States was determined by the early migrants from Europe, so that we can expect the cultural distance between the United States and (say) the United Kingdom, Germany or Ireland to be very limited.

[27] We are grateful to an anonymous referee for suggesting this specification.

where the main difference between Eqs. (1) and (6) resides in the inclusion of a function of the time since ancestral migration $f(t_i)$, both additively and interacted with w_o. Two remarks are necessary with respect to Eq. (6): First, this equation is impossible to estimate if the data source does not provide information on t_i. This happens only with the GSS, while neither the population censuses nor the ACS provide any information on the time elapsed since ancestral migration. If t_i is not available, this could be replaced with controls for groups of ancestral countries, e.g., continents, that capture, at least partly, unobserved heterogeneity in t_i across origins, but not across individuals. Second, this specification does not address the concerns related to self-selection into reporting a foreign ancestry. In this respect, data constraints are unfortunately binding, and the fact that estimates are obtained on a selected sample should be always kept in mind when interpreting the results of the econometric analysis.

4 Data sources

Our analysis draws on the data from the three population censuses (1980, 1990 and 2000) that included the question on ancestry, as well as on the earlier population censuses (1850 to 1970) (Ruggles et al., 2023). These earlier censuses are used to validate the informational content of the three measures of the (unobserved) time since ancestral migration that we introduce below.

4.1 Population censuses in the United States

Our main data source is the 5 percent sample of the 2000 population censuses in the United States (Ruggles et al., 2023), augmented with the 5 percent samples of the 1980 and 1990 census in selected parts of the analysis.[28,29] The censuses provide a wealth of individual characteristics about all individuals residing in the United States, including undocumented immigrants, and, together with the American Community Survey, it represents a relevant data source in the literature (see, for instance, Fernández and Fogli, 2006; Giuliano and Nunn, 2021; Ek, 2021; Galor et al., 2020, 2023). The United States has historically attracted, and

[28]Other data sources commonly used in the literature focusing on the United States —the March Supplement of the Current Population Survey and the General Social Survey— allow identifying the maternal and the paternal country of birth, but have a much more limited sample size. This prevents us from building reliable measures of the origin-specific variables that we use in our analysis.

[29]The American Community Survey includes the same variables as the population census.

continues to attract, large numbers of immigrants from various origins, making it an ideal setting on which to apply the epidemiological approach both to understand the deeply-rooted causes of personal traits and to investigate how the latter persist. As mentioned before, this approach requires information on the origin of each individual, and the United States census provides four variables informative of that: place of birth, ancestry, spoken language at home (other than English) and parental country of birth.

In the remainder of this paper, we will only exploit variables about place of birth and ancestry to identify an individual's origin. Among these, only birthplace is universally available. Parental country of birth is available only for individuals co-residing with their parents, implying that the sample is limited in size and self-selected. The information about the language spoken at home is similarly not well-suited to identify one's own origin for two main reasons: first, the language spoken at home is clearly endogenous with respect to migration;[30] second, spoken languages do not, in general, uniquely identify a foreign country, e.g., someone speaking Spanish at home could originate from Mexico or El Salvador.

4.2 Variables used in the analysis

We use four main variables in the analysis, namely the birthplace of a respondent, self-reported ancestries, the maternal or paternal birthplace, and the languages other than English spoken at home by a respondent. As discussed above, the first two will be used as sample selection criteria, and to define the country of origin of each individual, while the other two variables will be used in the analysis, but not in the definition of the country of origin. We also describe how we established crosswalks between birthplaces, ancestries and languages.

4.2.1 Birthplace

The variable `bpl` describes either the state of birth in the United States, or, in general, the country of birth for foreign-born individuals. This variable is presented in two variants: the first containing country names, while the second, more comprehensive option, includes sub-regional identifiers where applicable. For instance, the latter version includes unique codes for individuals born in specific regions like Aruba or Madeira for the year 2000.

[30] Information about the mother tongue was recorded in the census only until 1970.

4.2.2 Maternal or paternal birthplace

This variable can be built combining the information on the variable `bpl` with the variables `momloc` and `poploc`, which provide the identifier (if any) of the co-resident mother and father of each respondent.

4.2.3 Self-reported ancestries

The Census Bureau first introduced the ancestry question in the 1980 census, when the questions on the maternal and paternal birthplaces, and on the mother tongue, were abandoned. This question offers respondents the opportunity to specify the "ancestry group with which [she] identifies." In 1980, the phrasing of the question was: "What is this person's ancestry?", while in 1990 and 2000 it became: "What is this person's ancestry or ethnic origin?".[31] Participation in this question is voluntary, and respondents can decide to not report any ancestry. The text of the questionnaire includes examples of ancestries, which in 2000 were the following: "Italian, Jamaican, African Am., Cambodian, Cape Verdean, Norwegian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on".[32]

Respondents can report multiple ancestries; more precisely, up to two ancestries in the 1990 and in the 2000 census (and in the ACS). The information is recorded in the variables `ancestr1` and `ancestr2`. The Census Bureau does not provide to the respondents any explicit criterion to determine the order of the ancestries that are reported. Things were different in 1980, when up to three answers were recorded. When this occurred, the ancestries were coded in alphabetical order in the variable `ancestr1` only, while the variable `ancestr2` was missing.[33] Thus, someone saying she was (in this order) of German and English descent, had German coded for `ancestr1`, and English for `ancestr2`. Conversely, someone saying she was (in this order) of German, English and French descent, would have English-French-German coded for `ancestr1`, and for `ancestr2` would be missing. Thus, the third ancestry would alter the order of the first two.

The answers that are coded can encompass both native ancestries, e.g., American, Afro-American or American Indian, and foreign ancestries. Foreign ancestries can relate directly

[31] See https://usa.ipums.org/usa-action/variables/ANCESTR1#questionnaire_text_section.
[32] The list of examples has evolved over time, as discussed below.
[33] See https://usa.ipums.org/usa-action/variables/ANCESTR1#comparability_section.

to a country (e.g. Italian), to a regional entity within a country (e.g. Sicilian), or to a supranational entity (e.g. European or Hispanic).[34] The Census Bureau treats in a specific way the American ancestry: this is considered as a valid ancestry only when no other ancestry is mentioned, and it is discarded otherwise. Thus, an individual saying, for instance, that she is of American and German ancestry would have `ancestr1` reporting German, so that one should interpret the first ancestry as actually being the first *valid* reported ancestry.

The following long quote from Farley (1991) nicely describes a few key features of the question about ancestry:

> "The ancestry question is more complicated in that it depends to a large degree on factual knowledge about the history of one's family, but also requires many people to make a decision about identification: Which one or two of several possible ethnicities does a person report? Undoubtedly, some individuals identify very strongly with a particular ancestry, and will do so regardless of clues on the census form or contemporary political events. Many others, however, may not identify strongly, so their answers may depend on ephemeral events." (Farley, 1991, p. 414).

With respect to ephemeral events, these can also be related to the census questionnaire itself. Farley (1991) observes that the ancestry question followed a question on English proficiency in 1980, which might have induced respondents without a strong attachment to an ancestry or with multiple ancestries to report an English ancestry. Similarly, the list of ancestries provided as examples has changed over time, and the changes introduced between 1980 and 1990 have exerted a substantial influence on the distribution of ancestries (Rosenwaike, 1993). Thus, one should keep in mind that the question about ancestries does not, differently from the question on the birthplace, elicit factual information, but rather a subjective judgement, that is closely intertwined with a choice concerning one's own identity.

4.2.4 Language spoken at home

The variable `language` codes the language other than English spoken at home by each respondent aged 5 and above. Individuals who only speak English at home are coded as `English`. This variable provides a limited coverage of indigenous languages of foreign countries.

We establish a crosswalk between the variables `bpl`, `ancestr1`, `ancestr2`, and `language` . More precisely, we associate to each reported ancestry the codes corresponding to birthplace,

[34] The Census Bureau refers to this latter type of ancestries as "general heritages".

e.g., we associate Belgian and Flemish ancestries to the code of the variable `bpl` corresponding to Belgium. Then, we associate to each ancestry any language that is either an official language of the associated country or that is spoken by at least 20 percent of its population, drawing on the data from Mayer and Zignago (2011). This allows us, for instance, to identify the respondents of a given foreign ancestry that are born in the ancestral country, or that speak a language associated to the ancestral country at home.

5 Empirical analysis of census data

We describe here the various steps of our empirical analysis of the data drawn from various population censuses conducted in the United States.

5.1 Identifying individuals' origin

Following the standard practice in the literature, we identify one's own origin on the basis of the first reported ancestry only.[35] As our objective is to associate the answers to a (single) foreign country, we do not consider "general heritage" ancestries that cannot be reliably linked to a country, such European, North European or Latin American, among others. Then, we aggregate distinct ancestries at the level of a country. This requires first aggregating more detailed ancestries, e.g., associate English, Northern Irish, Scottish, Scotch Irish and Welsh ancestries to Great Britain, or Acadian, Canadian and French Canadian to Canada.

We can match ancestries to 109 different foreign countries. Almost all European and American countries are separately identified. At the same time, this is not the case, with just a few exceptions, for Africa.

Overall, in the 2000 census, around 55.7 percent of the native respondents reported at least one foreign ancestry that we can exploit in the analysis.[36] Thus, we exclude from the sample 44.3 percent of the native population, with this figure corresponding to two cases that

[35] A possible alternative would be, for respondents whose first ancestry is either a native ancestry or a general heritage, to use the second reported ancestry, when the latter is non-missing and it can be connected to a foreign country; this alternative is immaterial, as most individuals with a first ancestry that we cannot exploit do not have a second ancestry that can be connected to a country.

[36] We cannot compare this figure with other papers relying on the answers to the questions on ancestry, as the share of the native population that is actually used in the analysis is not reported.

have (roughly) the same incidence: natives not reporting any foreign ancestry, and natives reporting either a native ancestry, or a foreign ancestry that we cannot connect to a country.

Binding data constraints prevent us from presenting origin-specific evidence for each of the 109 distinct foreign ancestries on the incidence of *not* reporting a foreign ancestry for natives with foreign ancestors, as we do not have information about the country of birth of one's own ancestors. However, there is a notable exception for Mexican ancestry, arising from the overlap of another question, which appears earlier in the questionnaire, with the one on ancestry. More precisely, Question 5 in the questionnaire of the 2000 census reads "Is this person Spanish/Hispanic/Latino?", proposing as a possible answer a category, "Mexican, Mexican American, Chicano". These three groups match the answers to Question 10 on ancestry that we use to identify individuals of Mexican ancestry. This overlap gives us an interesting opportunity, given that Mexico is (by far) the main non-European ancestry in the data, representing 3.9 percent of the native population, and 7.0 percent of the population reporting a valid foreign ancestry.

We can, thus, analyze the answers to the ancestry question of the 589,358 natives who, in the 5 percent sample of the 2000 census, identify themselves as "Mexican, Mexican American, Chicano" in Question 5. 60.2 percent of them report a Mexican, 12.0 percent a Mexican American, and 0.5 percent a Chicano first ancestry. Thus, only 72.7 percent of Hispanic (Mexican, Mexican American, Chicano) natives report a first ancestry that we connect to Mexico. 11.5 percent of them do not report any ancestry, 6.2 percent report Hispanic as their ancestry, 1.6 percent report a "deeper" ancestral origin (Spanish), and 1.3 percent report to be of American ancestry. Thus, if we consider that the natives who identify themselves as Hispanic (Mexican, Mexican American, Chicano) have Mexican ancestors, then we can conclude that more than one out of four of them (27.3 percent) decides not to report a Mexican ancestry. If we collapse the data at the level of each PUMA, we see that the propensity *not* to report a Mexican ancestry for natives who are Hispanic (Mexican, Mexican American, Chicano) substantially declines with the local share of the Hispanics in the local population, with the correlation between these two variables standing at -0.315. In other words, the propensity not to report a Mexican ancestry is lower when the share of the population with Mexican ancestors is larger.

The share of natives with Mexican ancestors that do not report a Mexican ancestry is large, and we can also analyze whether they are non-randomly selected with respect to a simple

measure of their attachment to the culture of the country of their ancestors: speaking Spanish at home. Among the 160,786 Hispanic (Mexican, Mexican American, Chicano) natives who do not report a Mexican ancestry, 44.3 percent speaks Spanish, while the corresponding share for the other 428,572 Hispanic (Mexican, Mexican American, Chicano) natives stands at 56.6 percent, i.e., $(56.6 - 44.3)/44.3 = 27.8$ percent higher. Thus, reporting a foreign ancestry is, at least in the case of Mexico, positively correlated with the attachment to the ancestral culture. This case, which relates to one of the main foreign ancestries in the United States, illustrates the potential non-random selection into reporting a foreign ancestry.

5.2 Main foreign ancestries among natives in 2000

We focus on the 2000 census and on the natives with a foreign ancestry that we can associate with a foreign country. The five main countries of first ancestry are Germany (12.3 percent considering all natives, 22.1 percent among people with at least one reported ancestry), Great Britain (9.7 and 17.6 percent), Ireland (7.6 and 13.6 percent), Italy (4.7 and 8.5 percent), and Mexico (3.9 and 7.0 percent), representing in total 38.4 percent of the natives (69.0 percent if we consider only individuals reporting at least one ancestry). Notice that just seven more countries represent at least 1.0 percent of the ancestries of the native population. Figure 1 plots, on a world map, the share of natives reporting a given country as their first ancestry, while Figure 2 repeats the same exercise also using the information on the second ancestry.[37,38]

Figure 5 reports the share of natives with a given first ancestry who also report a second ancestry, and Figure 6 reports the share of natives with a given ancestry (first or second) who also report another ancestry. On average, countries with a more distant migration history to the United States, such as European countries, have a higher incidence of multiple ancestries, which may capture the occurrence of mixed marriages, i.e., marriages between individuals of different ancestries, in the previous generations.

Figure 7 associates to each Public Use Microdata Area (PUMA) the most frequent ancestry among natives,[39] revealing that 23 distinct ancestries represent the main ancestry in at least

[37]Notice that this implies that, except for the American ancestry, ancestries are no longer mutually exclusive, and the sum of the shares across all countries now exceeds 100 percent.

[38]As Figures 1 and 2 do not allow visualizing differences in the share of the various ancestries when these are low, Figures 3 and 4 provide the same information using a logarithmic scale.

[39]We associate United States, American Indian, Nuevo Mexicano and Hawaiian ancestries to the United States.

one PUMA. The geographic distribution of the most frequently reported ancestries among native-born individuals at the PUMA level matches what we would expect based on historical settlement patterns and migration trends. For example, Mexican ancestry is dominant in PUMAs located near the US-Mexico border in states like Texas, California, Arizona and New Mexico. This aligns with the proximity to Mexico and history of Mexican immigration and settlement in these regions. Similarly, French ancestry is concentrated in PUMAs across Louisiana, consistent with France's colonization of that region in the 18^{th} century, while Norwegian ancestry is mostly situated in the northern parts of North Dakota and Minnesota, reflecting the large waves of Norwegian immigrants settling in the Midwest in the late 19^{th} century, especially in North and South Dakota, possibly because they searched for locations with similar climatic conditions with respect to Norway (Obolensky et al., 2024). Figures 13 and 14 report respectively the share of the native population of German and Mexican ancestries across PUMA, allowing to visualize the local incidence of these two ancestries even when they are not the main ancestry.

PUMA-level ancestry data underscores how the time elapsed since migration can vary greatly across reported ancestries, in ways that an epidemiological framework must account for. The fact that German was the most commonly reported ancestry in 1990, with roots tracing back over a century for many of those populations, provides a clear example. Unlike more recent immigrant ancestries, descendants of 19^{th} century German migrants have had longer to geographically disperse across with the United States across generations. We see evidence of this increased dispersion in the PUMA data, with high shares of German ancestry appearing more widely distributed across different regions, as compared to more concentrated patterns of ancestries with closer ancestral migration times, like Mexican nearer the southern border.

5.3 Self-reported ancestry varies over time for a given cohort

Brittingham and de la Cruz (2004) documents major changes in the share of the population that identifies with different foreign ancestries between the 1990 and the 2000 census. Notably, the share of the population reporting a German ancestry (the most common ancestry) declined from 23.3 to 15.2 percent of the population, with an absolute decline from 57.9 to 42.8 million individuals.[40] This reduction might reflect demographic events, and notably the death of older

[40]The difference with the figure that we reported above for Germany based on the 2000 census is due to

cohorts of the population who reported a German ancestry in 1990, coupled with limited incoming migration flows from Germany, and a lower propensity of new cohorts to report a foreign ancestry. Indeed, our analysis of the data reveals that the propensity to report a German ancestry greatly varies across censuses for natives born in a given year. More precisely, Figure 8 plots the share of natives born between 1940 and 1979 reporting German as a first ancestry in the 1980, 1990 and 2000 census. The restrictions on the year of birth are meant to minimize the influence of demographic events, as no individual in our analysis is aged above 60 in 2000. For each cohort, the share with a German ancestry markedly increases from 1980 to 1990, and then it abruptly declines in 2000. The changes in the treatment of multiple ancestries from 1980 to 1990 (see the discussion on this in Section 4.1) might explain the increase, but not the ensuing decline. The evolution over time in the share of natives reporting a German ancestry might also be related to the fact that the enumerators mentioned German ancestry in the fourth place of the list of examples of ancestries in 1980 and in the first place in 1990 (see Rosenwaike, 1993, for evidence on the influence of this change in wording), while German ancestry was not even mentioned in 2000. Figure 9 performs the same exercise for the Mexican ancestry, where the steady increase across birth cohorts in the share of the native population with a Mexican ancestry reflects the major increase in migration from Mexico in the 20^{th} century. We see that, for nearly all birth cohorts, the share of the native population reporting a Mexican ancestry steadily declines across the three census years. While absolute variations are clearly smaller than those in Figure 8, the relative magnitude is similar. For instance, 1.9 percent of the natives born in 1940 reported a Mexican ancestry in 1980, while this share is down to 1.5 percent in 2000, i.e., a proportional reduction exceeding 40 percent.

These numbers suggest that one should be cautious when exploiting data from the different survey waves to identify individuals' origins, and show the robustness of results when estimating epidemiological regressions using different survey waves separately.

While we have just provided two examples related to the most common European and Latin American foreign ancestry, our point is more general. The influence of the ephemeral nature of the self-reported ancestry on the econometric analysis depends on what drives the variation in the answers that a cohort of individuals gives at different points in time. Furthermore, major

our focus on the native rather than total population, and to the exclusion of individuals reporting either a native ancestry, or a general heritage as first ancestry.

swings in the incidence of a given foreign ancestry for a given cohort of natives, as we have documented for Germany and for Mexico, are consistent with reported ancestry representing a subjective decision concerning identity, which can vary over the course of one's own life.

5.4 Capturing unobserved heterogeneity across foreign ancestries

Lacking information on time since ancestral migration for all the origin countries for the individuals who self-report a foreign ancestry, we propose two variables aimed at proxying such unobserved heterogeneity, and a third variable which is a revealed measure of the attachment to one's own ancestry.

The first variable, that we denote as v_o^1, is the probability that an individual (either native or foreign-born) claiming a given ancestry was actually born in that ancestral country. This probability will be higher for more recent waves of migration, and for ancestral countries where ethnic attrition is lower. As the generations pass and migration events recede further into the past, fewer ancestry claimants will have been born abroad, even if they still identify culturally with their ancestral heritage. In essence, the variable is a proxy for the number of generations since the initial migration or migrations occurred.

The second measure, which we denote as v_o^2, exploits the strong network effects that influence where migrants initially locate within a host country. These network effects lead to clustering, with immigrants initially concentrating in just a few selected areas. Over time, as the generations pass, these network effects diminish as the migrants' descendants integrate more fully into the host society and relocate freely within its borders. To measure the geographic concentration of natives claiming each ancestry within the United States, we calculate a Herfindahl-Hirschmann index for each group. Consistent with the prior discussion, more recent migrant waves should display higher levels of geographical concentration.

The third measure, called v_o^3, is represented by the share of natives of a given foreign ancestry reporting speaking at home the language of their ancestral country. This variable, which represents a revealed measure of the importance of tradition according to Giuliano and Nunn (2021), can be meaningfully computed only for natives with an ancestry in a country in which English is not an official language (Mayer and Zignago, 2011).[41]

[41]The accuracy of this variable clearly depends on the coverage of ancestral languages in the data.

5.4.1 Share born in the ancestral country

For each foreign ancestry, we compute the share of individuals born in the ancestral country, e.g., the share of individuals of Italian ancestry that are born in Italy, that we denote as v_o^1. This variable captures the share of first-generation immigrants among all individuals of a given ancestry. Hence, it is informative of the average time elapsed since ancestral migration even when we restrict the sample to native-born individuals only. The analysis of the data clearly reveals that v_o^1 is substantially lower for European than for Latin American countries of origin, and among Latin American countries it is lower for Mexico and other Central American countries, that have a longer history of migration to the United States (see Figure 10).

To provide a few examples, this variable is equal to 2.1 percent for Germany, 0.7 percent for Ireland, 1.2 percent for Great Britain, and 3.5 percent for Italy, while it jumps to 47.2 percent for Mexico and 72.7 percent for El Salvador. The stark differences between European and Latin American countries are consistent with the different timing of migration to the United States from these regions (see notably Figure 2 in Abramitzky and Boustan, 2017), but this proxy is also informative of the differences in timing within regions, as Mexican migration preceded migration from other Latin American countries. Notice that v_o^1 is, strictly speaking and by construction, uninformative about the average time since ancestral migration among the natives of a given ancestry. However, when examining data on children co-residing with their parents, it becomes apparent that v_o^1 can serve as a reliable proxy, at least for the subsequent generation. This is evident in Figure 11, where we observe that children who identify with ancestral roots spanning a long history of migration usually have parents who were *not* born in the ancestral country. On the other hand, children who identify with more recent migratory waves typically have at least one parent who was born in the corresponding foreign country. The correlation between this variable and v_o^1 is 0.879. Thus, because different ancestries display a differentiated (average) time since migration, the latter could confound the estimates when relying on the epidemiological approach. To illustrate, suppose one aims to comprehend English proficiency and establish connections with underlying attributes of the ancestral homeland. In this scenario, the estimates would be compromised since earlier waves of migration would have had more opportunities to assimilate and acquire English proficiency.

5.4.2 Spatial concentration of natives in the United States

We use the shares s_{ok}, which corresponds to the ratio between the native population of ancestry o residing in the PUMA k and the total native population of ancestry o, to compute an origin-specific Herfindahl-Hirschman index of spatial concentration, that we denote as v_o^2. This index corresponds to the probability that two randomly drawn natives with the same ancestry reside in the same area. We compute it at the State rather than the PUMA level, to ensure a sufficient number of observations per area from each origin. The values of the index range between 3.5 percent (Great Britain) and 72.5 percent (Anguilla),[42] with an average and median value standing at 18.2 and 13.3 percent respectively. The values of v_o^2 is much lower for European countries than for Latin American countries, as can be seen from Figure 12. For instance, we have that $v_{MEX}^2 = 28.3$ percent, and the Spanish-speaking Latin American country with the lowest concentration is Bolivia, with $v_{BOL}^2 = 11.4$ percent. In contrast, a country with older migration waves like Germany displays a value $v_{DEU}^2 = 4.8$ percent. Again, the large disparities across origins speak to a differential time since migration, with higher concentration persisting for groups arriving more recently due to stronger network effects that initially influence location choice within host countries.

5.4.3 Share of natives speaking the ancestral language

The choice to speak the ancestral language at home can be regarded, following Giuliano and Nunn (2021), as a revealed measure of one's own attachment to the identity associated to a foreign ancestry. The construction of this measure, that we denote as v_o^3, is clearly meaningful only for ancestries that do not correspond to countries where English is an official language, or where at least 20 percent of the population speaks English (Mayer and Zignago, 2011) and so this can be defined only for 69 out of the 109 foreign ancestries. Figure 15 plots the value of v_o^3 for all foreign (first) ancestries corresponding to a non-English speaking country. This world map clearly reveals that no more than 10 percent of the natives of European ancestries (except for Spanish ancestry) speak their ancestral language at home, while the corresponding share of Latin American ancestries is close or above 50 percent. Figure 16 considers both first and second ancestries in the definition of v_o^3. Again, these findings highlight the importance of accounting for differential opportunities for assimilation across ancestry groups over time.

[42] Notice that the high percentage for Anguilla is mechanical, as very few natives, i.e., 95 individuals in the five percent sample from the 2000 Census, report an ancestry in this tiny Caribbean island.

5.4.4 Correlations and patterns of spatial variation

We explore here the correlation between our three origin-specific variables that we built from the 2000 census, notably the share of the population (natives and immigrants) with an ancestry in the foreign country o born in the ancestral country (v_o^1), the Herfindahl-Hirschmann Index of spatial concentration of the native population with an ancestry in the foreign country o (v_o^2), and the share of the native population with an ancestry in the foreign country o speaking the ancestral language at home (v_o^3).

The correlation between v_o^1 and v_o^2 can be computed for all the 109 foreign ancestries that we use in the analysis, while the correlations with v_o^3 are restricted to the 69 countries of foreign-ancestry where English is not an official language. The correlations are computed by weighting observations by the number of natives with an ancestry in each country o. This is the relevant measure, whenever the econometric analyses in the epidemiological approach relying on a variant of Eq. (1) are conducted with individual-level data. Furthermore, weighting observations is important, to avoid giving too much influence to ancestries reported by a limited number of natives, which mechanically have a high value of v_o^2.[43]

The correlation between v_o^1 and v_o^2 stands at 0.863, the one between v_o^1 and v_o^3 at 0.402, and the one between v_o^2 and v_o^3 at 0.500. These high correlations provide additional credence to the proxies we built to capture unobserved heterogeneities of individuals of foreign origin, that should be controlled for in the epidemiological approach.

At least eight out of the ten origin countries with the lowest values of by v_o^1, v_o^2 or v_o^3 are European countries, whose migrants typically moved to the United States in a distant past, while Latin American countries tend to be among the ones with the highest values of these three variables. Indeed, this confirms that foreign countries with a more distant history of migration to the United States are, on average, characterized by lower values of v_o^1, v_o^2, and v_o^3, i.e., these ancestries are mostly composed by natives, who are more spatially dispersed across states, and which have a lower tendency to speak their ancestral language at home.

A simple (weighted) regression of v_o^1 on dummies for the five continents produces an R^2 of 0.804, while the corresponding R^2 when the dependent variable is v_o^2 stands at 0.700. Thus, most of the variability in these two measures is across rather than within continent. However, this is not the case with v_o^3, as the R^2 stands at 0.101, i.e., this variable is characterized by

[43] However, we obtain similar results when computing unweighted correlations, which are relevant for analysis conducted at the ancestry-level.

a greater variability within rather than across continents. This, in turn, implies that the inclusion of continental dummies in Eq. (1) would absorb a substantial portion of the possible confounding effect of differences across origin countries in t_i, but not in the attachment to the foreign ancestry, as captured by the choice of the language spoken at home.

5.5 Evidence from earlier population censuses

We can draw on data from old population censuses in the United States, from 1850 to 1970, to construct a measure of the average year of immigration from each country of origin. In particular, the census has been including a question on the country of birth for foreign-born individuals since 1850.[44] This variable can be informative about the origin-specific average time elapsed since ancestral migration for natives that report different foreign ancestries in recent population censuses.

This exercise is interesting, but subject to an inherent limitation, and to some constraints related to the data. As far as the inherent limitation is concerned, the threat to the epidemiological approach stems from the unobserved heterogeneity in the average time since ancestral migration for the individuals that *chose* to report a given foreign ancestry, and *not* in the average time since ancestral migration for *all* natives that have one or more ancestors originating from a given foreign country. The two can differ, as (*i*) foreign ancestry is self-reported, and (*ii*) return migration was relatively high during the Era of Mass Migration (see Bandiera et al., 2013, on this), so that not all immigrants that one can observe in older censuses have necessarily left descendants in the United States. The analysis of old population censuses is, thus, not directly informative about the relevant dimension of unobserved heterogeneity across origins, which also depends on the differences in fertility in the United States of groups of natives with different foreign ancestors.

With respect to the data limitations, the countries of birth that are separately recorded change across censuses (e.g., Finland does not appear as a separate entry until 1870, systematic coverage of Central American countries appears only in 1920), and a given country can refer to different territories in various population censuses (e.g., Prussia). These two points imply that the set of foreign countries for which we can analyze the data is substantially smaller (just 66

[44] See https://usa.ipums.org/usa-action/variables/bpl#availability_section (last accessed on June 18, 2024).

countries) than the set of 109 foreign ancestries that we can identify in recent censuses,[45] and the comparison of the data across different censuses for a given country might be problematic.

With these caveats in mind, we have analyzed the data from 1850 to the 1970 population censuses to provide evidence of the validity of our proxies. More precisely, we pool together the data for all individuals born in the foreign country o across census years $t = 1850, 1860, ..., 1970$. Denoting with t_{io} the year in which an individual i born in o is observed in the data, we compute the origin-specific average value of t_{io}, denoted as $\bar{t}_o$, using individuals aged between 30 and 35 years at the time of the survey. The restriction to an age group smaller than the distance between two censuses is meant to ensure that each individual is counted only once,[46] while the choice of the specific age group is meant to maximize the probability that these individuals left descendants in the United States.

The year of immigration of each individual is unknown,[47] and $\bar{t}_o$ is higher than the (unobserved) average year of immigration. Furthermore, we can also observe that no individual in our dataset can have immigrated before 1815, so our measure particularly underestimates the time since ancestral migration for, say, Germany, Great Britain or Ireland. This is the reason why we also draw on the 1850 census to record the total stock of immigrants from each country in that census year, to get a sense of the size of older flows.

Table 1 presents the results obtained from this exercise. In particular, it provides, for the 25 main countries of foreign ancestry for the native population according to the 2000 census, the value of $\bar{t}_o$, and also the stock of individuals born in country o recorded in the 1850 census. For instance, we see that $\bar{t}_{DEU} = 1908$, while $\bar{t}_{MEX} = 1971$. Table 1 also reveals a substantial

[45]The set of origin countries is determined by data availability constraints. More in detail, our data source (https://sda.usa.ipums.org/sdaweb/analysis/?dataset=all_usa_samples) only provides information on the number of immigrants for a selected number of countries: Canada, Mexico, Cuba, Denmark, Finland, Iceland, Norway, Sweden, England, Scotland, Wales, United Kingdom, Ireland, Belgium, France, Liechtenstein, Luxembourg, Monaco, Netherlands, Switzerland, Albania, Andorra, Gibraltar, Greece, Italy, Malta, Portugal, San Marino, Spain, Austria, Bulgaria, Czechoslovakia, Germany, Hungary, Poland, Romania, Yugoslavia, Estonia, Latvia, Lithuania, Other USSR/Russia, China, Japan, Korea, Cambodia (Kampuchea), Indonesia, Laos, Malaysia, Philippines, Singapore, Thailand, Vietnam, Afghanistan, India, Iran, Nepal, Bahrain, Cyprus, Iraq, Iraq/Saudi Arabia, Israel/Palestine, Jordan, Kuwait, Lebanon, Qatar, Saudi Arabia, Syria, Turkey, United Arab Emirates, Yemen Arab Republic (North), Australia and New Zealand.

[46]Older censuses are full count.

[47]The year of immigration has been recorded only between 1900 and 1930, and in 1970, and it is missing in the other seven population censuses that we use in the analysis; see https://usa.ipums.org/usa-action/variables/YRIMMIG#availability_section (last accessed on June 18, 2024).

variability in the timing of past migration flows from different European countries: the average here in which Irish immigrants are observed in the census is 1887, and the stock of individuals born in Ireland in the 1850 census stood at 984,851, while Italian immigrants were, on average, observed in 1930, and they only accounted for 2,684 individuals in the 1850 census.

We can use these data to correlate $\bar{t}_o$ with v_o^1, the share of the native population with ancestry o born in the ancestral country. The correlation between these two variables stands at 0.728.

Figure 17 plots, in three separate panels, the value of $\bar{t}_o$ against each of the three proxies for the unobserved time since ancestral migration that we computed from the 2000 census.[48] Figure 17 reveals that a higher average year of arrival is systematically associated with a higher share of individuals claiming a foreign ancestry that were born in the the ancestral country, a higher value of the Herfindahl-Hirschmann index of spatial concentration of the native population of a given ancestry, and a higher percentage of native individuals that claim an ancestry speaking the ancestral language at home.

5.6 The role of the three origin-specific proxies

The three variables v_o^1, v_o^2 and v_o^3 can be used to see whether some groups of natives of foreign ancestry stand out with respect to one (or more) of the three proxies for the average origin-specific time since ancestral migration. If this is the case, then one should test whether the size and significance of the estimate obtained from bringing Eq. (1) to the data is sensitive to the exclusion of these groups of natives. If variation induced by restricting the estimation sample is in line with the effect that is expected on the basis of the average value of w_o and of the origin-specific proxies, then the estimate obtained on the original sample could be confounded by unobserved heterogeneity in the time since ancestral migration. Importantly, we do *not* suggest to augment the canonical epidemiological specification presented in Eq. (1) by replacing the unobserved $f(t_i)$ with one of these three proxies. Indeed, such an approach would not allow capturing the unobserved heterogeneity within a given ancestral country, and these variables could represent "bad controls", as they could be correlated with unobserved determinants of the outcome variables.[49] For instance, immigration policy in the United States

[48] The first two panels of Figure 17 separately identify, using different markers, English-speaking and non-English speaking countries of ancestry.

[49] We are grateful to an anonymous referee for bringing this to our attention.

in the 20^{th} century favored immigration from countries that were perceived as more culturally close, so that the three proxies are likely to be a function of the distance in cultural norms between a given origin and the United States. Similarly, natives of a foreign ancestry that have a greater attachment to their ancestral culture are more likely to be spatially concentrated, and this would influence the value of v_o^2 that we observe in the data.

6 Two illustrative examples

We provide here two examples of the empirical relevance of our concerns related to the use on foreign ancestry in the epidemiological approach, drawing on two important papers in the literature, namely Fernández and Fogli (2006) and Giuliano and Nunn (2021).

The choice of focusing on Fernández and Fogli (2006) is motivated by four main reasons: First, this is an early and seminal contribution to the literature using the epidemiological approach, totalling 588 citations on Google Scholar (information retrieved on September 4, 2024). Second, it explains an outcome (fertility) for natives of foreign ancestry in the United States using the corresponding variable for the ancestral country (more precisely, the total fertility rate in 1950), and this facilitates the exposition of the source and direction of the possible bias due to unobserved heterogeneity in the average time since ancestral migration. Third, it uses data from various waves of the General Social Survey (GSS), which allow, following Giavazzi et al. (2019), to distinguish between natives that are second, third and (at least) fourth-generation immigrants. This is particularly convenient, as it allows demonstrating the extent of unobserved heterogeneity (even at the individual-level) in the time elapsed since ancestral migration. Fourth, even though the replication data are not publicly available, we managed to reconstruct the original estimation sample, and to successfully reproduce the original results.

The choice of focusing on Giuliano and Nunn (2021) is motivated by three main reasons: First, this is a seminal contribution to the literature using the epidemiological approach, totalling 410 citations on Google Scholar (information retrieved on September 4, 2024). Second, their analysis of the influence of historical climatic variability in the ancestral country of origin on the choice to speak the ancestral language in the United States, which represents a revealed measure of the importance of tradition, allows a clear exposition of our concerns, as the countries in their estimation sample with a shorter time elapsed since ancestral migration

are clearly characterized by a lower historical climatic variability. Third, the replication data for this paper are available online.

6.1 Fernández and Fogli (2006)

We describe here, using the notation that we introduced in Eq. (1), the specification estimated by Fernández and Fogli (2006), and describe in detail their estimation sample.

Fernández and Fogli (2006) draw their data from nine waves of the GSS (1977, 1978, 1980 and 1982 to 1987). More precisely, they focus on a sample of married native women aged 29 to 50 at the time of the survey, and reporting as their main ancestry a foreign country. Their final estimation sample consists of 1,145 women from 14 countries of ancestry: Canada, Denmark, Finland, France, Germany, Ireland, Italy, Mexico, Netherlands, Norway, Russia, Spain, Sweden and the United Kingdom.[50] These data are used to estimate the following equation:

$$y_{iokt} = \alpha \overline{y}_o + [\boldsymbol{\gamma}' \boldsymbol{x}_i] + d_k + d_t + \epsilon_{iokt} \tag{7}$$

where t denotes the year in which the woman i, with an ancestry in country o and residing in region k,[51] was surveyed. The dependent variable y_{iokt} in Eq. (7) is the number of children, and $\overline{y}_o$ is the total fertility rate (TFR) in 1950 in the ancestral country o. The vector $\boldsymbol{x}_i$ includes three specific portions: (i) the number of siblings of woman i, (ii) a quadratic polynomial in age, and a dummy variable for the level of education (high school, some college and college), and (iii) the number of completed years of schooling of the father and of the mother of woman i.[52]

[50] Fernández and Fogli (2006) exclude six countries (Czechoslovakia, Hungary, Poland, Romania, Yugoslavia, and Lithuania) that become socially planned economy in the 1940s, and also eight countries (Austria, Greece, Japan, Puerto Rico, Switzerland, Portugal, and Belgium) with less than 10 observations in the data.

[51] The region k corresponds to nine categories: New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific.

[52] These two latter variables require an important clarification, as their inclusion determines a non-negligible reduction in sample size (from 1,145 to 921); consider, for instance, the education of the father (`paeduc`): the GSS records this information only if the woman i was co-residing with her father at the age of 16 (see, for instance, Q18 in the questionnaire for the GSS 1982, available at: https://gss.norc.org/documents/quex/1982%20GSS%20Quex.pdf); if the woman was not co-residing with her father, this variable is either missing, or it actually records the level of education of the (co-residing) step father, or even of another household member who is a male relative of the woman; if we use the variable `family16`, which describes the composition of the household in which a woman was living at the age 16, we see that the women (195 out of 1,145) for whom

The sample that is used to estimated Eq. (7) does *not* include any restriction with respect to the time since ancestral migration. Fernández and Fogli (2006) find that $\overline{y}_o$ has a positive and significant effect on fertility. These results are interpreted as follows:

> "[I]t is not only personal experience (as reflected in the number of siblings a woman has) that matters to a woman's fertility, but also that her culture (as embodied in the TFR in her country of ancestry) plays a role. [...] we find that even after controlling for various characteristics and family background of a woman, both her own personal experience and her culture play a role in influencing her fertility." (Fernández and Fogli, 2006, p. 561)

Fernández and Fogli (2006) write that "unfortunately we cannot distinguish second-generation Americans from those who have been in the US for longer" (p. 554), but this is feasible with the current version of the GSS data used by Fernández and Fogli (2006). The GSS provides information about whether each of the two parents is native or foreign-born, and also a variable counting the number of grand-parents that were born abroad.[53] Giavazzi et al. (2019) use these variables to define: second-generation immigrants the natives declaring a foreign ancestry and with at least a foreign-born parent, third-generation immigrants the natives declaring a foreign ancestry with two natives parents and at least two foreign-born grand-parents, and as (at least) fourth-generation immigrants the natives declaring a foreign ancestry with native parents and at least three native grand-parents.

The data that are necessary to apply the definition by Giavazzi et al. (2019) are available for 1,097 out of 1,145 native women in the estimation sample by Fernández and Fogli (2006). When we apply these definitions, 98 women (8.6 percent) are second-generation immigrants, 257 are third-generation immigrants (22.5 percent), 742 are fourth-generation or more (64.8 percent), and 48 women (4.2 percent) have missing values of the underlying variables, so their time since ancestral migration is unknown. Two key remarks about these figures: First, the number of generations since ancestral migration (measured from the number of genera-

the variable `paeduc` is missing have a higher number of children (2.48 and 2.32 respectively), and this variable is lower for the 61 out of 950 women for which `paeduc` records the level of education of the step-father or of another male relative (9.79 and 10.54 years respectively); thus, the inclusion of these two variables, whose informational content depends on `family16`, determine a non-random selection of the sample with respect to unobserved determinants of the number of children; the inclusion in the vector $\boldsymbol{x}_i$ of the number of siblings of woman i could be regarded as a sufficient statistics for the wealth and human capital effect that the level of education of the mother and of the father should control for (see Fernández and Fogli, 2006, p. 559, on this).

[53] Notice that the survey does not provide information on the country of birth of the parents and of the grand-parents.

tions separating a native respondent from the closest ancestor who is born abroad) is precisely known only for 98+257 = 355 women (31.1 percent), who are either second or third-generation migrants. We do not count here the 742 native women who are at least fourth-generation immigrants, as the GSS does not allow to identify the closest foreign ancestor for an individual whose parents and grandparents are all native-born (something that happens for 623 of these women). Second, there are major differences across countries of ancestry in the time since ancestral migration (see Table 2). Notably, for nine out of 14 ancestral countries, the highest number of women are (at least) fourth-generation immigrants, for four ancestral countries (Finland, Italy, Russia and Sweden) it is third-generation immigrants, and just for a single country (Mexico) the highest number of women is second-generation immigrants. These differences are consistent with the evidence provided in our discussion at point 1.(a) above based on the analysis of 1850 to 1970 population censuses.

Table 3 successfully replicates the original specifications in Table 3a in Fernández and Fogli (2006), while Table 4 estimates the six (out of nine) specifications of Table 3a in Fernández and Fogli (2006) which include the TFR in 1950 among the regressors for the 1,097 women for which we can apply the definition by Giavazzi et al. (2019).

Table 5 extends Eq. (7), introducing a set of dummies for the generation of each native women (with second-generation immigrant women representing the omitted category), and an interaction between these dummies and $\overline{y}_o$, i.e., the TFR in the foreign country of ancestry o.[54] In this specification, the coefficient α gives the effect of $\overline{y}_o$ on the number of children of women who are second-generation immigrants, while (for the other two generations), we need to combine α with the coefficient of the relevant interaction effect. The results in Table 5 reveal a significant effect of $\overline{y}_o$ on the number of children for women that are second or third-generation immigrants, e.g., $(98 + 257)/1,097 = 31.1$ percent of the sample in the first data column, while the effect is not significant for women that are at least fourth-generation immigrants.[55]

This pattern in the data is consistent with our theoretical framework in Section 3. However, there are legitimate competing interpretations of this pattern in the data: First, the variability in the TFR in 1950 in the ancestral country is lower among women that are at least fourth-

[54] Again, we do so only for the six data columns in Table 3a in Fernández and Fogli (2006) which include the TFR in 1950 among the regressors.

[55] We obtain similar results when estimating Eq. (7) separately for second-generation (96 observations), third-generation (257 obs.) and (at least) fourth-generation immigrant women (741 obs.).

generation immigrants. Notably, Mexico has (by far) the highest value of $\overline{y}_o$ in the estimation sample: the Mexican TFR in 1950 stood at 6.87, while the average and standard deviation in the estimation sample stand at 3.01 and 1.20 respectively (see Table 2 in Fernández and Fogli, 2006); women of Mexican ancestry represent $38/1,145 = 3.3$ percent of the original estimation sample, but just $9/742 = 1.2$ percent of the sample of women who are fourth-generation immigrants. Second, the fertility norm at origin (measured in 1950) is less accurate for women whose ancestors moved to the United States a longer time ago. Third, the composition of the estimation sample across ancestral countries varies significantly, we thus might be picking up an heterogeneity in the effect across countries rather than over the time elapsed since ancestral migration. Fourth, mixed marriages in previous generations imply that fourth-generation (or more) immigrant women are more likely to have multiple ancestries, so that the absence of an effect for them can be due to the attenuation bias that is put forward in when discussing the epidemiological approach (e.g., Fernández, 2011; Ek, 2021).

The theoretical framework presented in Section 3 implies that our estimate should be interpreted as capturing the effect of the TFR in 1950 on the subsample of natives choosing to report a foreign ancestry. The analysis of the data from the GSS reveals that there is a non-negligible share of natives that have at least one foreign-born parent or grandparent that do *not* report any foreign ancestry. If these individuals are less attached to the culture of their origin country, as the analysis of Mexican ancestry in Section 5.1 reveals, then their exclusion from the estimation sample could induce an upward bias in the estimated value of α, provided that the fertility norm at origin differs from the one in the United States.

6.2 Giuliano and Nunn (2021)

We describe here, using the notation that we introduced in Eq. (1), the specification estimated by Giuliano and Nunn (2021), and describe in detail their estimation sample.

Giuliano and Nunn (2021) draw on the data from the 2000 census to analyze whether the choice to speak a language other than English at home, which is used as a proxy for the importance of tradition, is significantly influenced by the historical climatic variability in the ancestral country. In particular, a higher climatic variability should decrease the importance of tradition, and be associated with a lower propensity to speak a foreign language (possibly the ancestral language) at home.

Giuliano and Nunn (2021) focus on a sample of 3,343,097 native individuals reporting a foreign ancestry that is connected to 84 distinct foreign countries, that do *not* have English as an official language. These observations are used to estimate the following equation:

$$y_{iok} = \alpha w_o + \boldsymbol{\beta}' \boldsymbol{x}_o + \boldsymbol{\gamma}' \boldsymbol{x}_i + \phi x_{ok} + d_k + \epsilon_{iok} \tag{8}$$

The dependent variable y_{iok} takes the value of 1 if the native individual i with an ancestry in country o residing in the metropolitan area k in the United States speaks a language other than English at home, and 0 otherwise. The origin-specific variable of interest w_o is a measure of historical climatic variability associated to the ancestral country o, the vector of origin-specific control variables $\boldsymbol{x}_o$ includes the (historical) distance from the equator, measures of historical economic development and political complexity,[56] the logarithm of GDP of country o in 2000, and the linguistic distance between the main language spoken in country o and English. The variable x_{ok} is the fraction of the population residing in the metropolitan area k who is born in the ancestral country o. The vector $\boldsymbol{x}_i$ of individual-level variables includes a quadratic in age, a sex dummy, and dummy variable for being married, and educational-attainment fixed effects, labour-force-status fixed effects, the natural log of annual income, and a rural/urban indicator variable.[57]

In the benchmark specification of Eq. (8), corresponding to the first data column of Table 5 in Giuliano and Nunn (2021), the estimated value of α stands at -0.447. A one standard-

[56]The variable w_o and the first three historical elements in the vector $\boldsymbol{x}_o$ are taken from Giuliano and Nunn (2018), and measured as follows: data from the Ethnologue and gridded population data from LandScan are used to estimate the share of the population π_{ol} in country o speaking a language l; each language l is associated to an ethnic group e in the Ethnographic Atlas; then, the variables are either directly defined for this ethnic group, e.g., political complexity, or measured from the ethnic homeland associated to each ethnic group, e.g., historical climatic variability, historical distance from the equator; the variables for each ethnic group are then averaged using π_{ol} as weights.

[57]Bertoli et al. (2024) document that the actual specification that is estimated by Giuliano and Nunn (2021) depart from the description given by the Authors in several dimensions. Notably, the estimation sample includes natives in seven countries or territories that have English has an official language, and around 25 percent of the observations in the estimation sample corresponds to instances in which the metropolitan area is "Not identifiable" or "Not in an MSA"; these individuals are treated as if they were residing in an unique metropolitan area, which is used in the definition of the fixed effects d_k and for the origin-location specific controls x_{ok}; we do not correct any of the inconsistencies documented in Bertoli et al. (2024) here, as a corrigendum published by Paola Giuliano and Nathan Nunn describes these as "imprecisions in wording and omissions in the text" (see Giuliano and Nunn, 2024).

deviation increase in historical climatic variability at origin gives rise to a $-0.447 \times 0.072 = -0.032$, i.e., 3.2 percentage points reduction in the share of the population of a given ancestry speaking a foreign language at home, a sizeable reduction, as the incidence of speaking a foreign language at home stands at 12 percent.

One out of four ancestral countries in the estimation sample (21 out of 84) is an American country, and Eq. (8) does not include a set of dummies for the continent to which country o belongs to. 388,421 natives, corresponding to 11.6 percent of the estimation sample, report an ancestry in one of these 21 countries. All these countries, except Canada,[58] have a recent history of migration to the United States, as described by our first proxy (share of the population born in the ancestral country), a factor that can contribute to the use of the ancestral language at home. Furthermore, 18 out of the 21 American countries of ancestry are Spanish-speaking countries.[59] If we focus on the 20 Spanish-speaking countries of ancestry,[60] 63.4 percent of the 378,081 observations for natives reporting an ancestry in one of these countries speak a language other than English at home, more than 10 times above the corresponding share in the rest of the sample, which stands at 5.9 percent.

This stark difference might reflect the much greater incentives to speak Spanish in the United States than any other ancestral language. These incentives go well beyond the local size of the population born in one's own ancestral country o and residing in the metropolitan area k. The variable x_{ok} is included in Eq. (8) "to account for the possibility that one's incentives to learn and speak one's ancestral language may be greater the more people there are in the same location *whose mother tongue is the ancestral language*" (Giuliano and Nunn, 2021, p. 1565-7, emphasis added). However, this variable fails to adequately capture the incentives to speak a language (Spanish) that is spoken in a large number of different ancestral countries. A telling example of this is the extent to which natives of Spanish ancestry stand out with respect to natives with other European ancestries in the estimation sample: 42.7 percent of 47,998 natives of Spanish ancestry speak a foreign language at home,[61] as opposed

[58] Canada represents one of the seven countries or territories that have English as an official language, but that are kept in the estimation sample by Giuliano and Nunn (2021).

[59] The only American countries of ancestry that are not Spanish-speaking are Brazil, Canada and Haiti.

[60] The two non-American Spanish-speaking countries of ancestry in the estimation sample are the Philippines and Spain.

[61] This might also reflect the fact that (say) a native with Mexican parents who have Spanish origin might report a Spanish rather than a Mexican ancestry; the phrasing of the question on ancestry does not rule out this possibility.

to 5.0 percent of natives with other European (non English-speaking) ancestries.

Thus, a legitimate concern is that the estimated value of α is confounded by (i) the more limited time since ancestral migration for natives with an ancestry in a Latin American Spanish-speaking country,[62] and by (ii) the inability of Eq. (8) to adequately control for the specificity of the Spanish language in the United States, with the two effects acting in the same direction. As Spanish-speaking countries are characterized by a high historical climatic variability (see Figure 4 in Giuliano and Nunn, 2021, on this), then the unobserved heterogeneity at point (i) and (ii) above could induce an overestimation of the effect of historical climatic variability.

The empirical relevance of our concern about a possible overestimation of α can be gauged by two simple modifications of the specification of Eq. (8), namely the inclusion of continent fixed effects, and the exclusion of the 378,081 observations for Spanish-speaking countries. With these two modifications, the estimated value of α moves from -0.477 (p-value 0.010) to -0.187 (p-value 0.007).[63] Thus, the effect of historical climatic variability remains statistically significant, but the magnitude of the estimated effect is reduced by 58.2 percent. A one standard-deviation increase in historical climatic variability at origin gives rise to a $-0.187 \times 0.068 = -0.013$, i.e., 1.3 percentage points reduction in the share of the population of a given ancestry speaking a foreign language at home, compared to the 3.2 percentage points reduction corresponding to the original estimate.

We do not claim that the estimated value of α from our specification represents the true effect of historical climatic variability on the propensity to speak a foreign language, nor that the estimated effect in Giuliano and Nunn (2021) is entirely spurious. We simply aim at emphasizing the relevance of a small subset of observations (11.6 percent of the original estimation sample), corresponding to Spanish-speaking countries, in shaping the size of the estimated effect in Giuliano and Nunn (2021). Spanish is a language with a unique position in the United States: it is spoken at home by 28.1 million (foreign-born and native) individuals in 2000, out of 47.0 million speaking a language other than English at home (Shin and Bruno, 2003), and reflecting the major immigration flows from many Latin American countries in the recent past. This analysis represents an illustrative example of the empirical relevance of our concerns about the possible overestimation induced by unobserved heterogeneity across

[62] Differently from the GSS, the 2000 census unfortunately does not provide any information (at the individual-level) on the time elapsed since ancestral migration.

[63] We obtain similar results when implementing just one of these two modifications.

groups of different foreign ancestries in the time since ancestral migration.

Importantly, the theoretical framework presented in Section 3 implies that our estimate should be interpreted as capturing the effect of historical climatic variability on the subsample of natives choosing to report a foreign ancestry. This could still induce an upward bias in the estimated effect if the natives deciding to report a foreign ancestry are more attached to the culture of their ancestral country, if the evidence provided in Section 5.1 for natives of Mexican ancestry also applied to other ancestries.

7 Concluding remarks

A large and growing set of influential papers have relied on the first self-reported foreign ancestry of natives in the United States to identify their country of origin. This choice reflects, to a large extent, binding data constraints, as the population census (since 1980) and the ACS offer no alternative to the use of this variable.

This paper provides theoretical arguments and empirical evidence suggesting that the reliance on foreign ancestry can give rise to specific threats to identification for the epidemiological approach. The widespread presumption that this approach is exposed to an attenuation bias might not apply when estimating a canonical specification on the subset of natives reporting a foreign ancestry. The threats to identification arise from the high share of the native population not reporting a foreign ancestry, and by unobserved heterogeneity in the time since ancestral migration.

We suggest that researchers should verify whether the magnitude of the estimated effect is sensitive to the inclusion in the estimation sample of countries of ancestry with a recent migration history to the United States, even though they might constitute a limited share of the observations. Furthermore, even if this dimension of unobserved heterogeneity is kept at bay, all the results should be interpreted on the basis of the fact that the estimation sample has been selected in a way that can magnify the chances of finding a significant effect, as the individuals choosing to report a foreign ancestry are presumably more attached to the culture of their ancestral countries. The data that are currently available do not, unfortunately, allow to overcome the inherently selected nature of the estimation sample constituted by natives of foreign ancestry.

References

Abramitzky, R. and L. Boustan (2017): "Immigration in American Economic History," *Journal of Economic Literature*, 55, 1311–45.

Akerlof, G. A. and R. E. Kranton (2000): "Economics and identity," *Quarterly Journal of Economics*, 115, 715–753.

Albert, C. and J. Monras (2022): "Immigration and Spatial Equilibrium: The Role of Expenditures in the Country of Origin," *American Economic Review*, 112, 3763–3802.

Alesina, A., Y. Algan, P. Cahuc, and P. Giuliano (2015): "Family Values and the Regulation of Labor," *Journal of the European Economic Association*, 13, 599–630.

Alesina, A., A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg (2003): "Fractionalization," *Journal of Economic Growth*, 8, 155–194.

Alesina, A. and P. Giuliano (2015): "Culture and Institutions," *Journal of Economic Literature*, 53, 898–944.

Algan, Y. and P. Cahuc (2010): "Inherited trust and growth," *American Economic Review*, 100, 2060–2092.

Antecol, H. (2000): "An examination of cross-country differences in the gender gap in labor force participation rates," *Labour Economics*, 7, 409–426.

Antman, F., B. Duncan, and S. J. Trejo (2016): "Ethnic attrition and the observed health of later-generation Mexican Americans," *American Economic Review*, 106, 467–471.

Antman, F. M. and B. Duncan (2023): "American Indian Casinos and Native American Self-Identification," *Journal of the European Economic Association*, 21, 2547–2585.

Antman, F. M., B. Duncan, and S. J. Trejo (2023): "Hispanic Americans in the Labor Market: Patterns over Time and across Generations," *Journal of Economic Perspectives*, 37, 169–198.

Arbatli, C. E., Q. H. Ashraf, O. Galor, and M. Klemp (2020): "Diversity and conflict," *Econometrica*, 88, 727–797.

Bandiera, O., I. Rasul, and M. Viarengo (2013): "The Making of Modern America: Migratory Flows in the Age of Mass Migration," *Journal of Development Economics*, 102, 23–47, migration and Development.

Beck Knudsen, A. S. (2022): "Those Who Stayed: Selection and Cultural Change in the Age of Mass Migration," Mimeo, https://annesofiebeckknudsen.com/wp-content/uploads/2022/08/thosewhostayed.pdf.

Bertoli, S., M. Clerc, J. Loper, and E. Roca Fernández (2024): "Understanding Cultural Persistence and Change: a replication exercise of Giuliano and Nunn (2021)," *Economic Inquiry*, https://doi.org/10.1111/ecin.13242.

Bisin, A. and T. Verdier (2001): "The economics of cultural transmission and the dynamics of preferences," *Journal of Economic Theory*, 97, 298–319.

Brittingham, A. and G. P. de la Cruz (2004): "Ancestry: 2000," US Census Bureau, 2000 Census Brief, available at https://www.census.gov/history/pdf/ancestry.pdf.

Chavez, L. R. (2013): *The Latino Threat: Constructing Immigrants, Citizens and the Nation*, Stanford University Press, second edition.

Desmet, K., I. Ortuño-Ortín, and R. Wacziarg (2017): "Culture, Ethnicity, and Diversity," *American Economic Review*, 107, 2479–2513.

Duncan, B. and S. J. Trejo (2011): "Tracking intergenerational progress for immigrant groups: The problem of ethnic attrition," *American Economic Review*, 101, 603–608.

——— (2017): "The complexity of immigrant generations: Implications for assessing the socioeconomic integration of Hispanics and Asians," *ILR Review*, 70, 1146–1175.

Ek, A. (2021): "Cross-country differences in preferences for leisure," *Labour Economics*, 72, 102054.

——— (2024): "Cultural Values and Productivity," *Journal of Political Economy*, 124, 295–335.

Farley, R. (1991): "The New Census Question about Ancestry: What Did It Tell Us?" *Demography*, 28, 411–429.

Fasani, F., G. Mastrobuoni, E. G. Owens, and P. Pinotti (2019): *Does immigration increase crime*, Cambridge University Press.

Fernández, R. (2011): "Does culture matter?" in *Handbook of Social Economics*, ed. by J. Benhabib, A. Bisin, and M. O. Jackson, Elsevier, vol. 1, 481–510.

Fernández, R. (2007): "Women, Work, and Culture," *Journal of the European Economic Association*, 5, 305–332.

Fernández, R. and A. Fogli (2006): "Fertility: The Role of Culture and Family Experience," *Journal of the European Economic Association*, 4, 552–561.

Fouka, V., S. Mazumder, and M. Tabellini (2021): "From Immigrants to Americans: Race and Assimilation during the Great Migration," *Review of Economic Studies*, 89, 811–842.

Fulford, S. L., I. Petkov, and F. Schiantarelli (2020): "Does it matter where you came from? Ancestry composition and economic performance of US counties, 1850–2010," *Journal of Economic Growth*, 25, 341–380.

Galor, O., M. Klemp, and D. Wainstock (2023): "Roots of Inequality," NBER Working Paper 31580.

Galor, O. and O. Özak (2016): "The Agricultural Origins of Time Preference," *American Economic Review*, 106, 3064–3103.

Galor, O. and V. Savitskiy (2018): "Climatic Roots of Loss Aversion," NBER Working Paper 25273.

Galor, O., O. Özak, and A. Sarid (2020): "Linguistic Traits and Human Capital Formation," *AEA Papers and Proceedings*, 110, 309–13.

Giavazzi, F., I. Petkov, and F. Schiantarelli (2019): "Culture: Persistence and evolution," *Journal of Economic Growth*, 24, 117–154.

Giuliano, P. (2007): "Living arrangements in Western Europe: Does cultural origin matter?" *Journal of the European Economic Association*, 5, 927–952.

Giuliano, P. and N. Nunn (2018): "Ancestral characteristics of modern populations," *Economic History of Developing Regions*, 33, 1–17.

——— (2021): "Understanding cultural persistence and change," *The Review of Economic Studies*, 88, 1541–1581.

——— (2024): "Correction to: Understanding Cultural Persistence and Change," *The Review of Economic Studies*, 91, 597.

Guiso, L., P. Sapienza, and L. Zingales (2006): "Does culture affect economic outcomes?" *Journal of Economic Perspectives*, 20, 23–48.

Jaeger, D. A., T. Dohmen, A. Falk, D. Huffman, U. Sunde, and H. Bonin (2010): "Direct evidence on risk attitudes and migration," *The Review of Economics and Statistics*, 92, 684–689.

Luttmer, E. F. P. and M. Singhal (2011): "Culture, context, and the taste for redistribution," *American Economic Journal: Economic Policy*, 3, 157–179.

Mayer, T. and S. Zignago (2011): "Notes on CEPII's distances measures: The GeoDist database," Working Papers 2011-25, CEPII.

Obolensky, M., M. Tabellini, and C. Taylor (2024): "Homeward Bound: How Migrants Seek Out Familiar Climates," Working Paper 32035, National Bureau of Economic Research.

Patel, K. and F. Vella (2013): "Immigrant networks and their implications for occupational choice and wages," *Review of Economics and Statistics*, 95, 1249–1277.

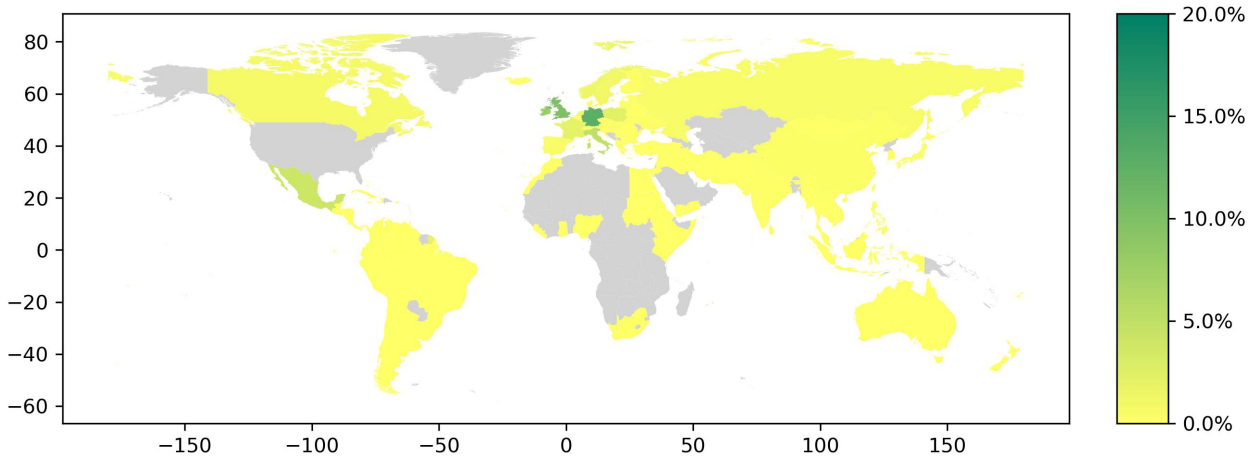
Putterman, L. and D. N. Weil (2010): "Post-1500 population flows and the long-run determinants of economic growth and inequality," *The Quarterly Journal of Economics*, 125, 1627–1682.

Rosenwaike, I. (1993): "Ancestry in the United States Census, 1980-1990," *Social Science Research*, 22, 383–390.

Ruggles, S., S. Flood, M. Sobek, D. Brockman, G. Cooper, S. Richards, and M. Schouweiler (2023): "IPUMS USA: Version 13.0 [dataset]," Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V13.0.

Shin, H. B. and R. Bruno (2003): "Language Use and English-Speaking Ability: 2000," Census 2000 Brief, US Census Bureau, available at https://www2.census.gov/library/publications/decennial/2000/briefs/c2kbr-29.pdf.

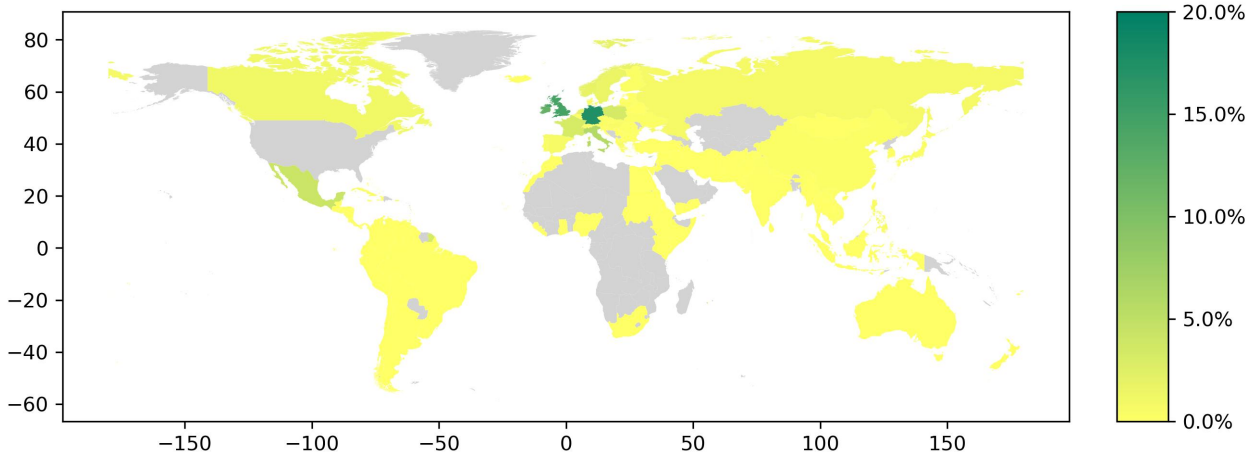
Figure 1: Proportional Distribution of Native Ancestry by Country in 2000



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: For every country, we assign a value representing the proportion of individuals who identify that country as their first ancestry.

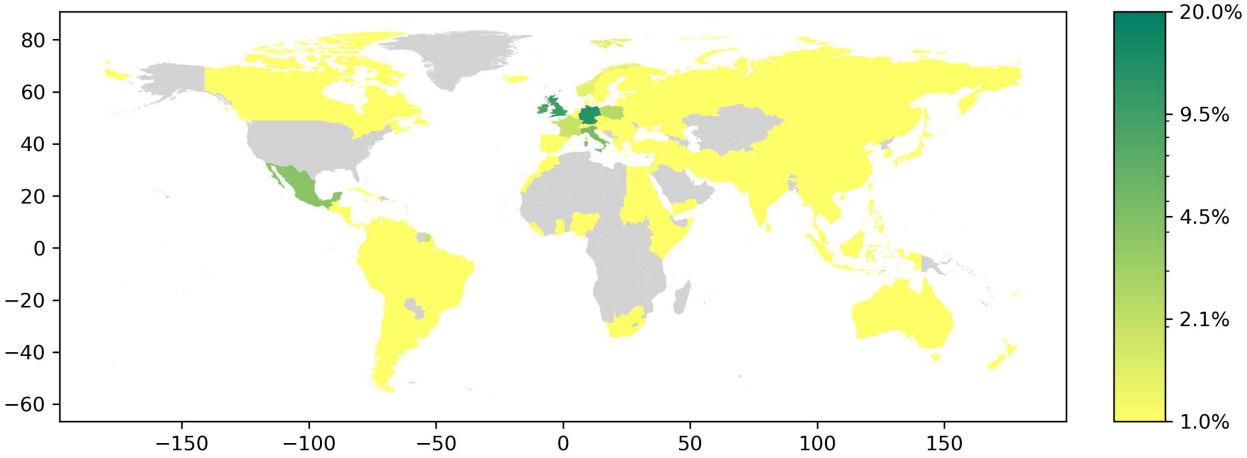
Figure 2: Proportional Distribution of Native Ancestry by Country in 2000, considering first and second ancestry



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: For every country, we assign a value representing the proportion of individuals who identify that country as their first or second ancestry. By construction, the sum of these shares exceed 100 percent.

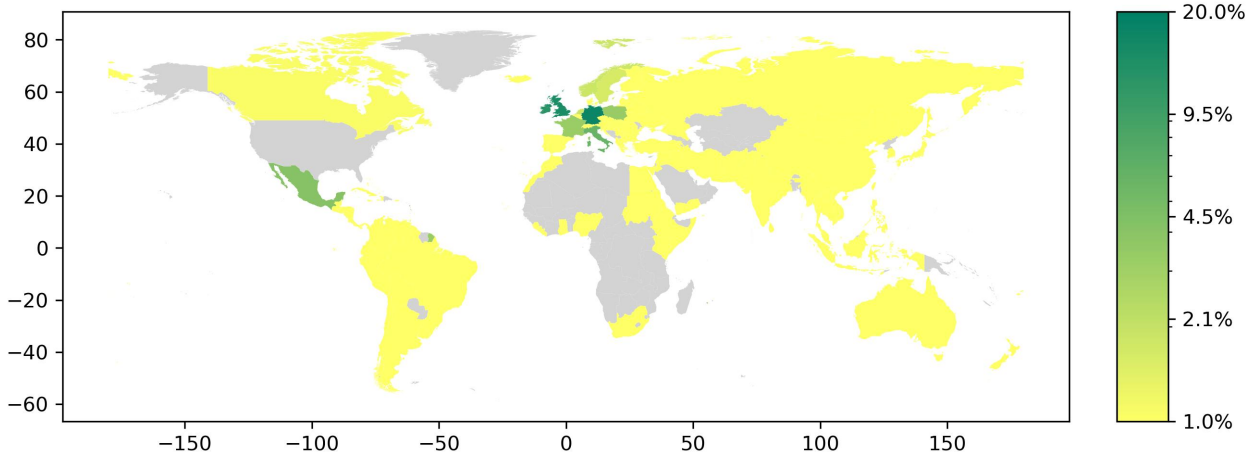
Figure 3: Proportional Distribution of Native Ancestry by Country in 2000 (logarithmic scale)



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: For every country, we assign a value representing the proportion of individuals who identify that country as their first ancestry.

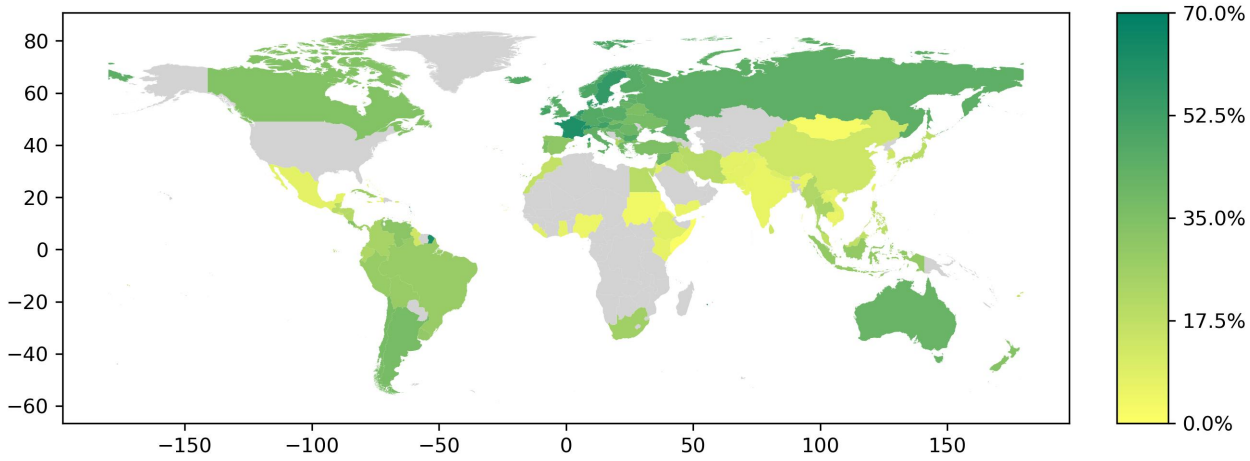
Figure 4: Proportional Distribution of Native Ancestry by Country in 2000, considering first and second ancestry (logarithmic scale)



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: For every country, we assign a value representing the proportion of individuals who identify that country as their first or second ancestry. By construction, the sum of these shares exceed 100 percent.

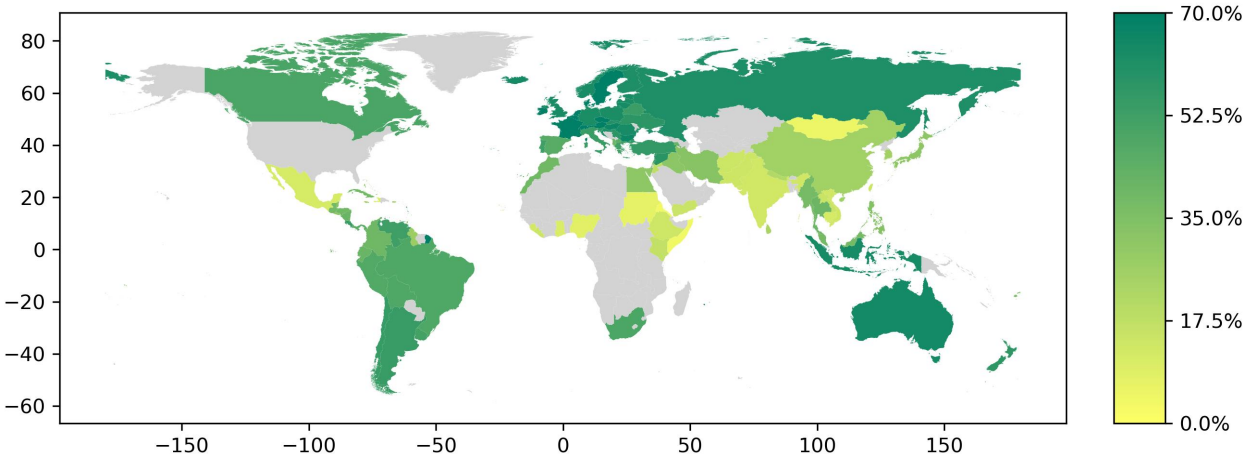
Figure 5: Incidence of multiple ancestries



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

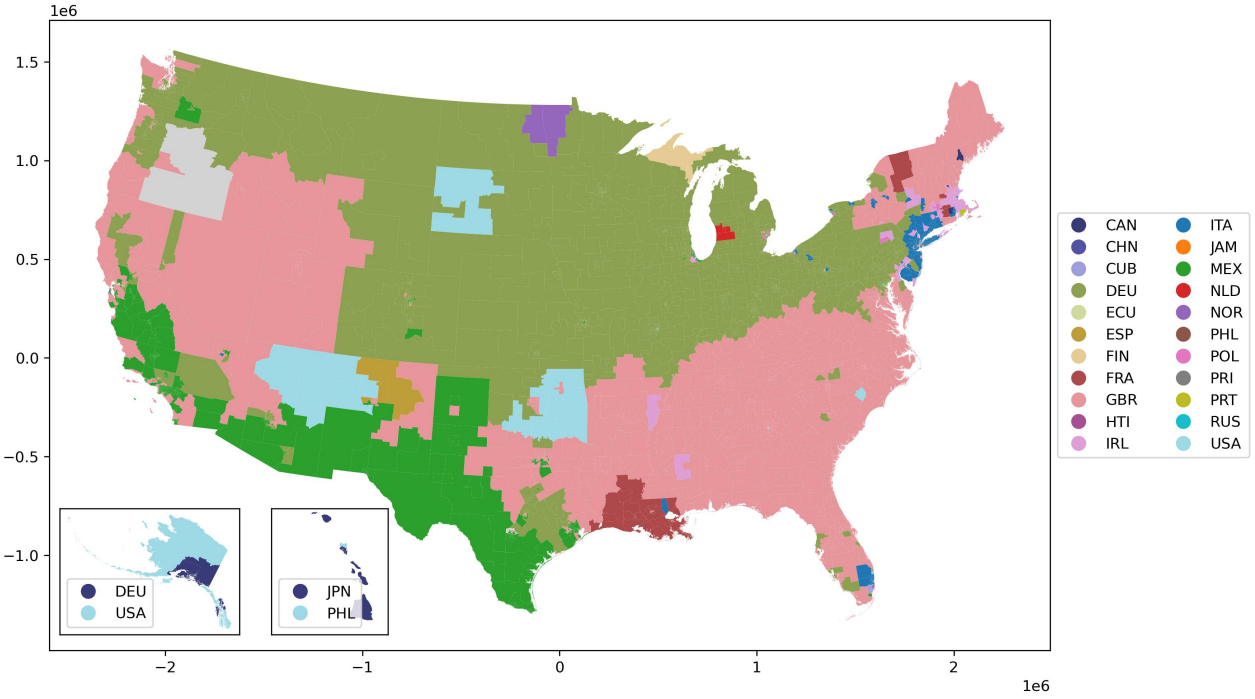
Notes: We examine the percentage of individuals with a specific ancestral background who identify as having a secondary ancestry. For example, we analyze the proportion of individuals with Italian as their primary ancestry who also report having a secondary ancestry.

Figure 6: Incidence of multiple ancestries
(first and second ancestry)



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).
Notes: for each ancestry, we compute the share of natives who also report a second ancestry, e.g., we compute the share of natives with Italian first or second ancestry that also report another ancestry.

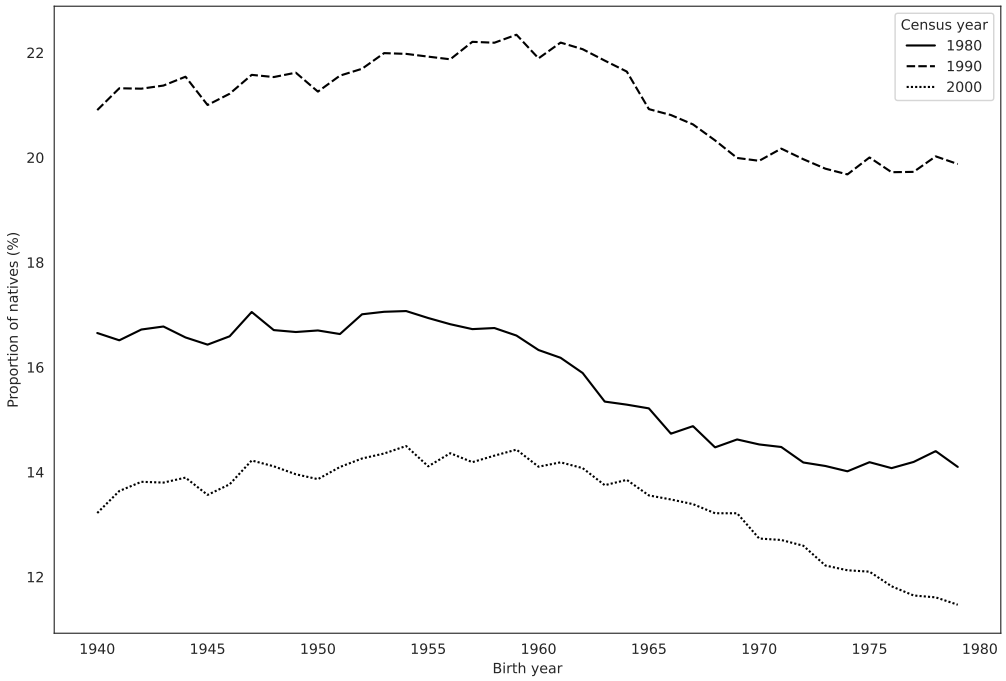
Figure 7: Most prevalent ancestry in each PUMA



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

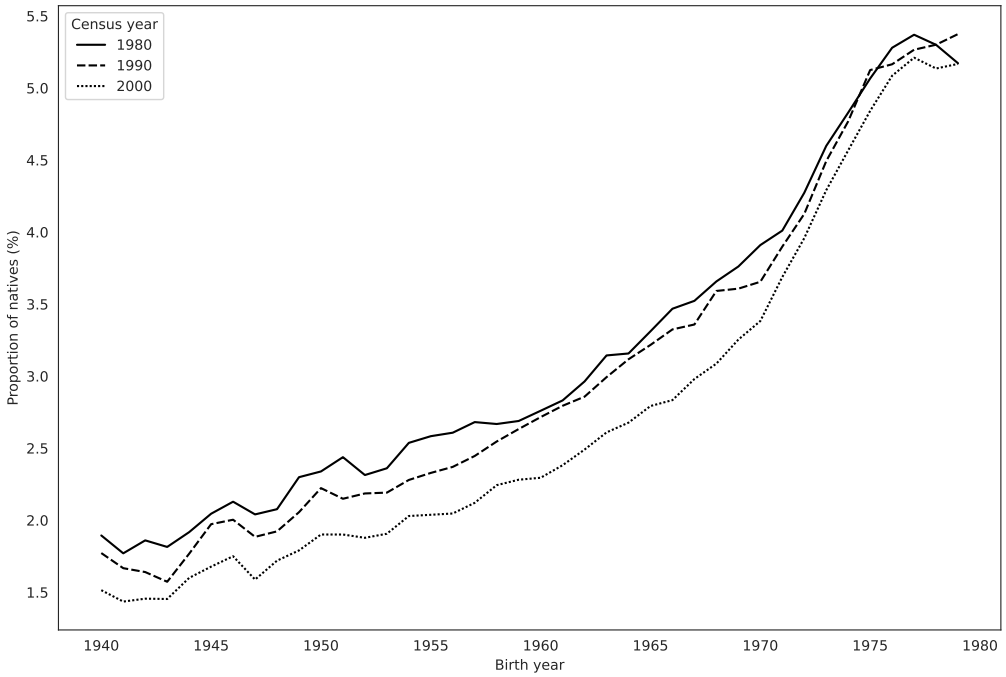
Notes: For each Public Use Microdata Area (PUMA), we identify and assign the most prevalent ancestral heritage among native residents.

Figure 8: Natives with German ancestry by birth cohort, different census years



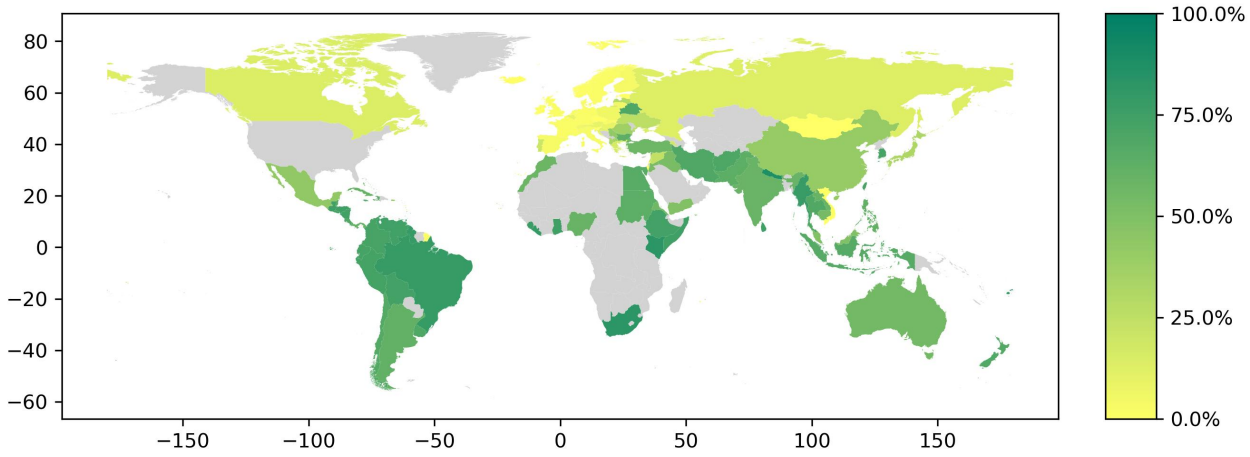
Data sources: Authors' elaboration on the 1980, 1990 and 2000 Census (Ruggles et al., 2023).

Figure 9: Natives with Mexican ancestry by birth cohort, different census years



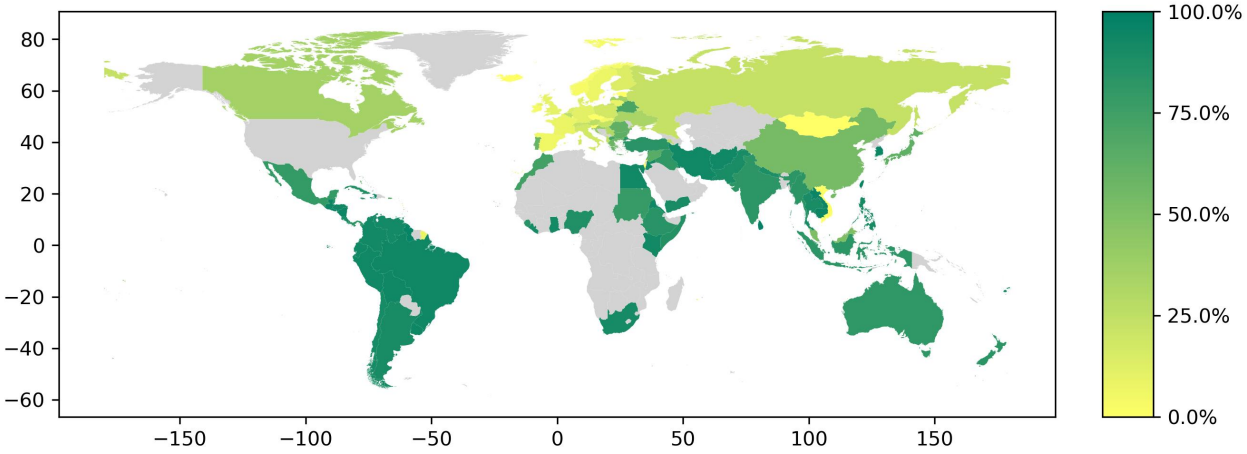
Data sources: Authors' elaboration on the 1980, 1990 and 2000 Census (Ruggles et al., 2023).

Figure 10: Share of individuals born in the ancestral country



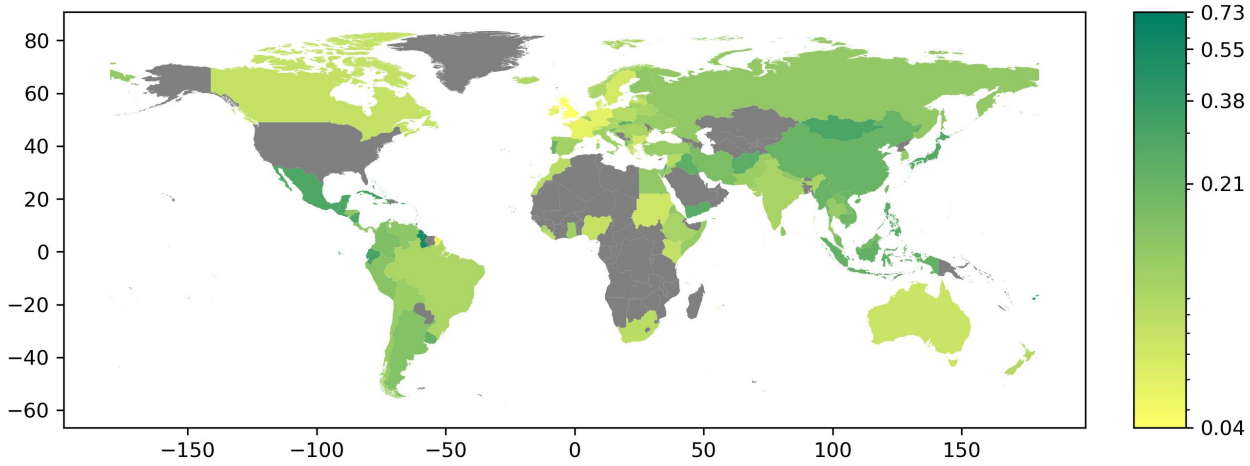
Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Figure 11: Share of natives co-residing with at least one parent whose first ancestry coincides with the maternal or paternal country of birth



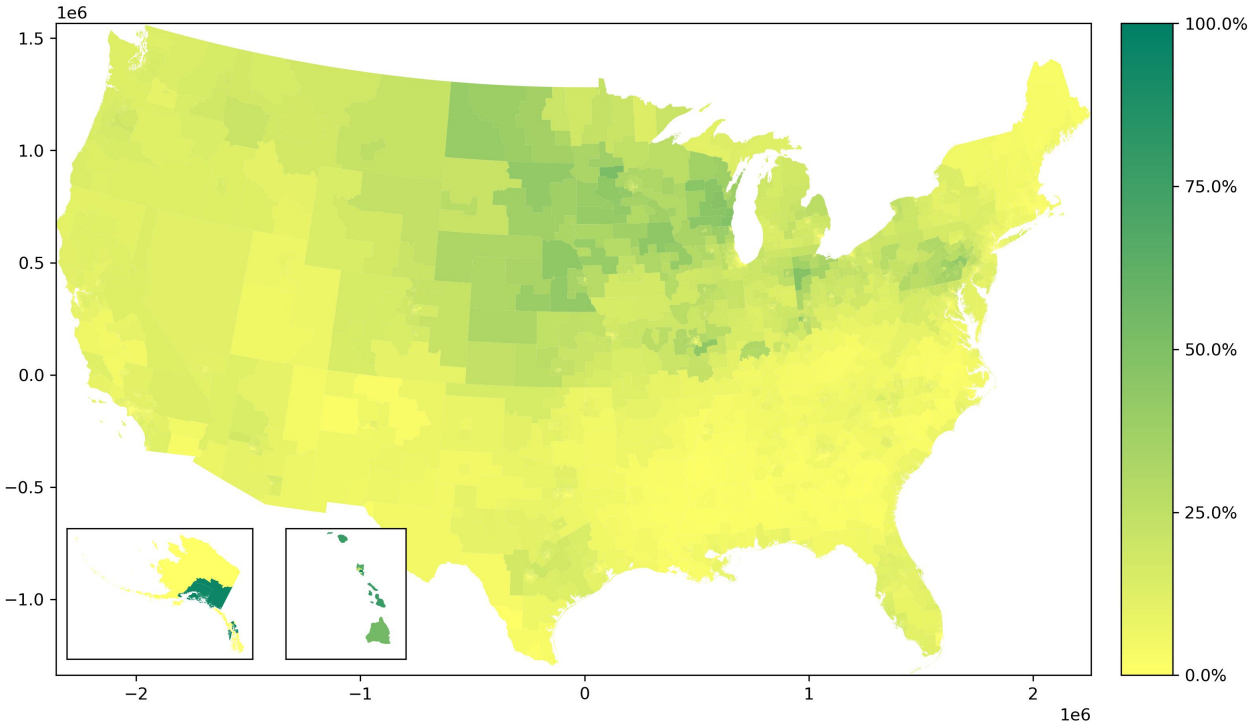
Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Figure 12: Herfindahl-Hirschman Index of spatial concentration of natives of foreign ancestry



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

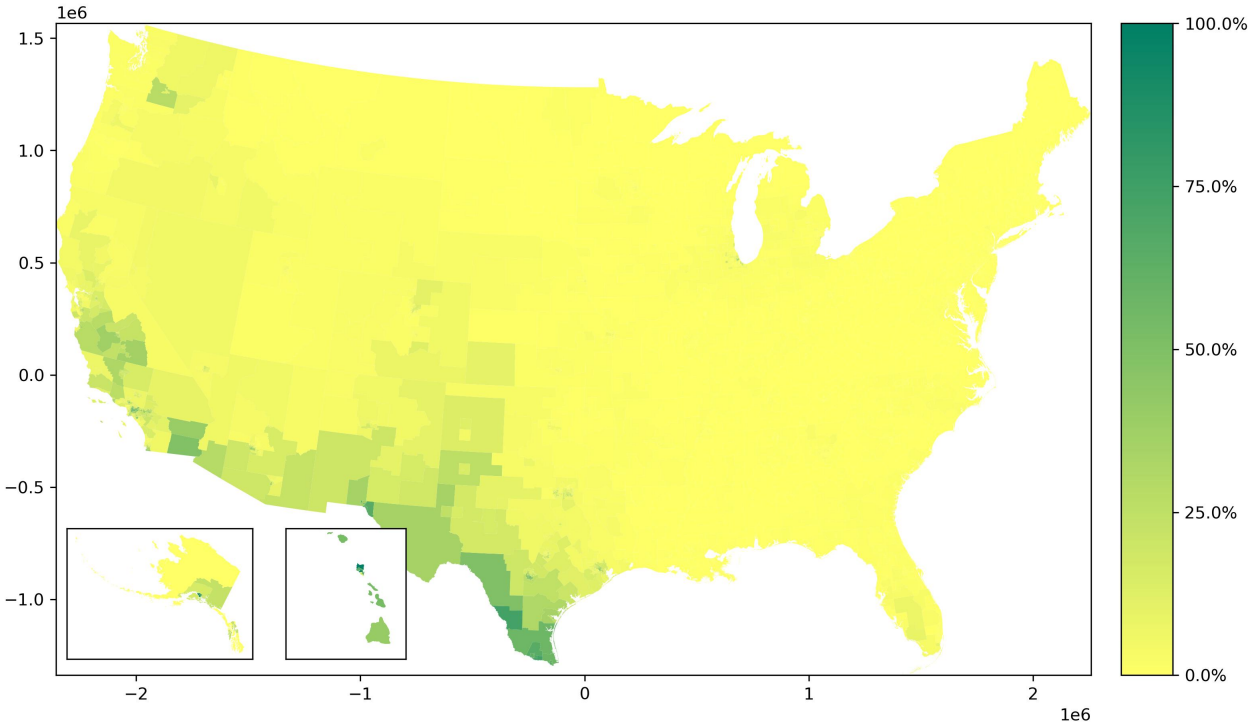
Figure 13: German ancestry among natives in each PUMA



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: This figure present the percentage of native-born population that identify with the German ancestry as first ancestry. The sample is composed of individuals born in the USA that report at least one ancestry.

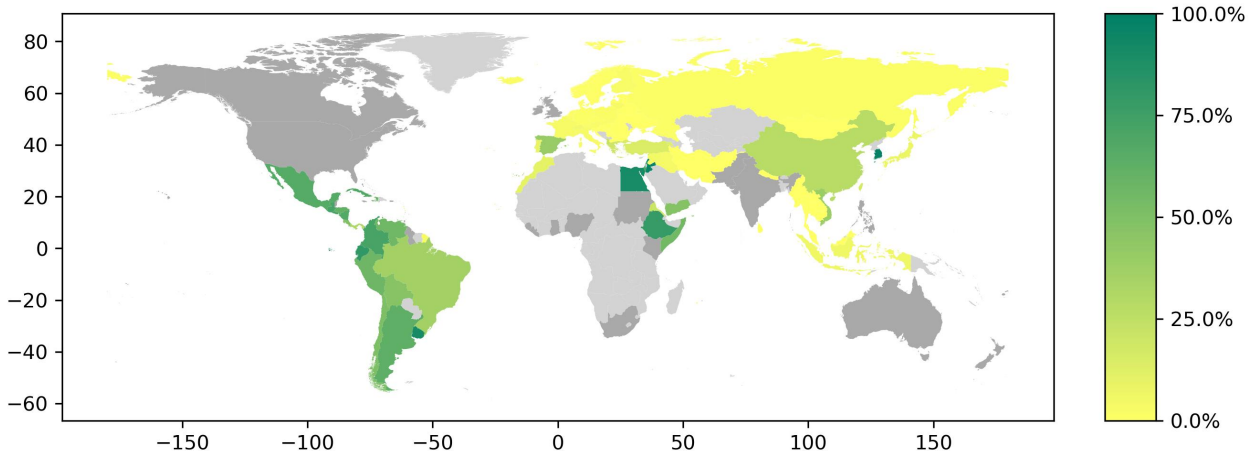
Figure 14: Mexican ancestry among natives in each PUMA



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: This figure present the percentage of native-born population that identify with the Mexican ancestry as first ancestry. The sample is composed of individuals born in the USA that report at least one ancestry.

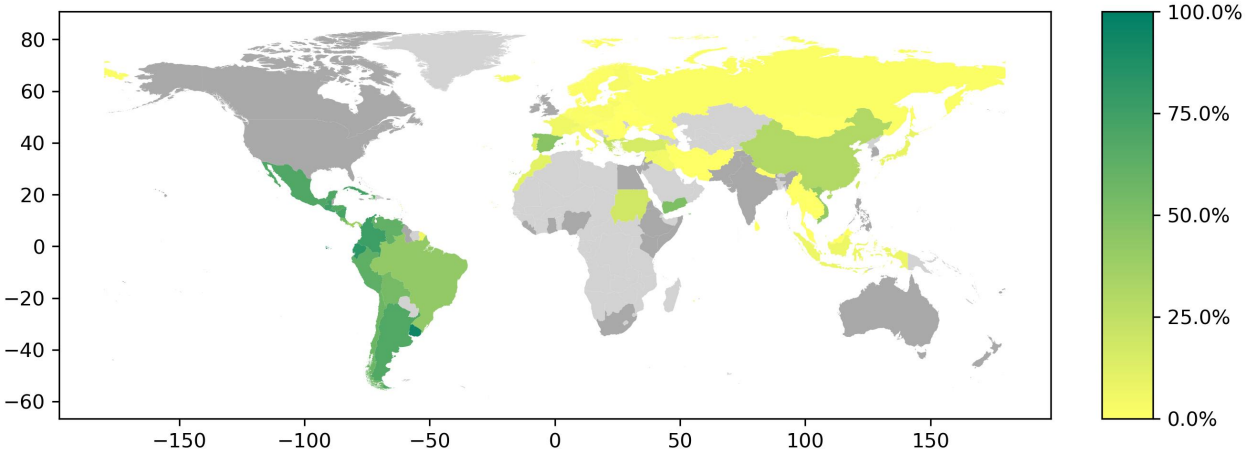
Figure 15: Share of natives speaking the ancestral language, by country of ancestry



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

Notes: This figure displays the percentage of individuals who claim ancestry in a particular country (first ancestry only) and also speak an official language of that country. The analysis includes only individuals with at least one reported ancestry, and countries where English is an official language are represented in light gray.

Figure 16: Share of natives speaking the ancestral language, by country of ancestry (first and second ancestry)



Data sources: Authors' elaboration on the 2000 Census (Ruggles et al., 2023).

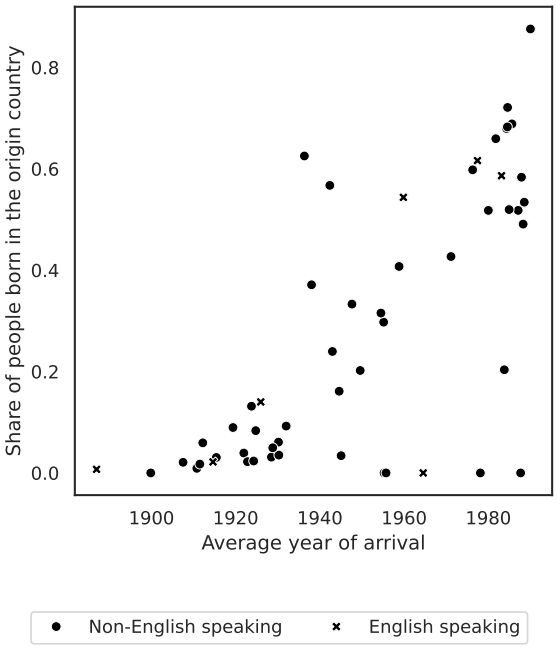
Notes: This figure displays the percentage of individuals who claim ancestry in a particular country (first or second ancestry) and also speak an official language of that country. The analysis includes only individuals with at least one reported ancestry, and countries where English is an official language are represented in light gray.

Table 1: Average arrival time in the USA for different countries

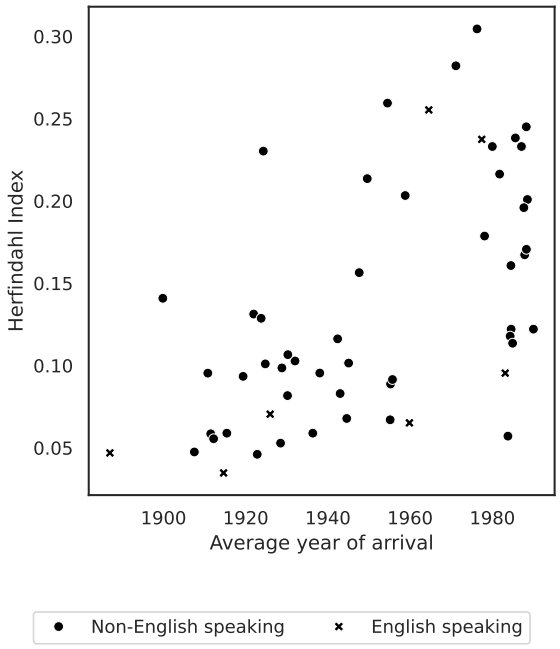
Country	Native-born	Average year of arrival	Residents in 1850
Germany	22.06	1908	596,181
UK	17.56	1915	387,710
Ireland	13.65	1887	984,851
Italy	8.50	1930	2,684
Mexico	7.03	1971	15,038
Poland	4.10	1930	3,208
France	3.43	1923	61,433
Norway	2.39	1911	12,749
Netherlands	1.81	1928	11,662
Sweden	1.74	1911	3,440
Canada	1.47	1926	135,300
Other USSR/Russia	1.05	1924	716
Spain	0.91	1945	2,634
Denmark	0.61	1915	2,080
Hungary	0.56	1925	100
Greece	0.51	1945	92
Portugal	0.49	1950	1,848
Philippines	0.49	1977	-
China	0.46	1959	671
Japan	0.45	1954	-
Switzerland	0.38	1912	14,546
Czechoslovakia	0.34	1924	97
Finland	0.33	1922	-
Lithuania	0.28	1929	-
Austria	0.27	1919	898

Notes: This table reports, for each of the 25 main foreign countries of ancestry for the native population in the 2000 census, the average year $\bar{t}_o$ in which the individuals aged 30 to 35 born in each foreign country o were observed in the data (1850 to 1970 census), and the (total) stock of individuals born in country c in the 1850 census.

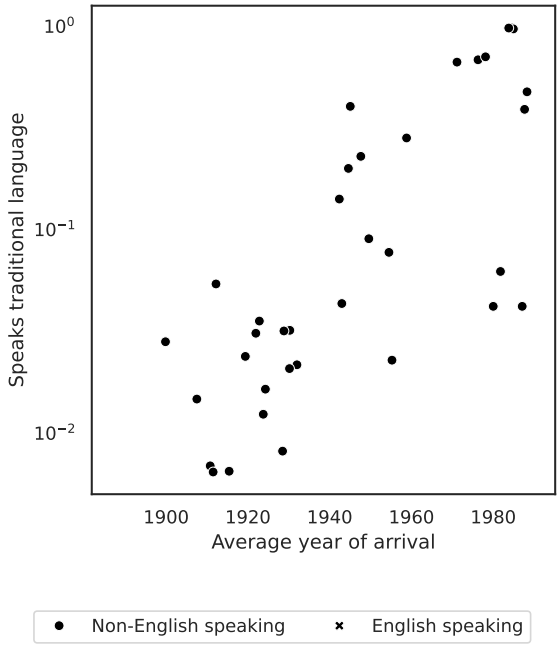
Figure 17: Average year of arrival and our proxies



(a) % of ind. born in the ancestral country



(b) Herfindahl index



(c) Speaking the ancestral language

Table 2: Countries of origin and generations of migrants in Fernández and Fogli (2006)

	Generation			
Country of origin	Second	Third	Fourth	**Total**
Germany	19	52	218	289
UK	9	28	222	259
Ireland	4	34	174	212
Italy	23	56	11	90
Canada	9	13	17	39
Mexico	14	12	9	35
France	2	6	24	32
Norway	3	10	18	31
Sweden	2	16	13	31
Netherlands	5	7	14	26
Russia	5	12	1	18
Denmark	0	4	9	13
Finland	2	6	3	11
Spain	1	1	9	11
Total	98	257	742	1,097

Notes: This table reports the distribution (number of observations) of generations of migrants by country of origin in GSS 1977, 1978, 1980 and 1982-1987, following Fernández and Fogli (2006). The sample comprises married women aged between 29 and 50 years old, born in the U.S. and reporting a foreign ancestry. Generations of migrants follow the definition of Giavazzi et al. (2019). "Second" are 2nd generation migrants (i.e. respondents born in the US and at least one of their parents was born abroad). "Third." are 3rd generation migrants (i.e. respondents born in the US, all of their parents are born in the US and at least two of their grandparents were born abroad). "Fourth" are 4th (and more) generation migrants (i.e. born in the US, all their parents are born in the US and at most one their grandparent was born abroad). This sample is used in the estimations reported in column 1 in Tables 4 and 5.

Table 3: Replication of Fernández and Fogli (2006): Fertility, culture, and siblings

	Dependant variable is Children								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
TFR 1950	0.166***		0.101***	0.117***		0.097***	0.135***		0.118**
	(0.025)		(0.026)	(0.016)		(0.020)	(0.036)		(0.042)
SIBS		0.093***	0.086***		0.044***	0.039**		0.045**	0.039*
		(0.012)	(0.014)		(0.012)	(0.014)		(0.017)	(0.019)
Age				0.281**	0.304***	0.299***	0.345***	0.364***	0.359***
				(0.098)	(0.095)	(0.095)	(0.079)	(0.083)	(0.083)
Age sq				−0.003*	−0.003**	−0.003**	−0.003***	−0.004***	−0.004***
				(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
High School				−0.814***	−0.764***	−0.738***	−0.629***	−0.574***	−0.564***
				(0.128)	(0.122)	(0.124)	(0.152)	(0.170)	(0.162)
Some College				−0.921***	−0.865***	−0.822***	−0.619***	−0.553**	−0.532**
				(0.063)	(0.072)	(0.055)	(0.181)	(0.194)	(0.178)
College				−1.362***	−1.292***	−1.250***	−1.143***	−1.071***	−1.057***
				(0.100)	(0.106)	(0.108)	(0.128)	(0.147)	(0.134)
Mother's education							−0.060**	−0.063***	−0.056**
							(0.024)	(0.019)	(0.021)
Father's education							0.027**	0.027**	0.028**
							(0.010)	(0.010)	(0.010)
Obs.	1145	1144	1144	1144	1143	1143	922	921	921
Adj. R^2	0.037	0.060	0.062	0.203	0.206	0.208	0.223	0.225	0.229
Region FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table replicates Table 3a in Fernández and Fogli (2006). It reports the results of estimating Eq. 7 on individual-level data from the GSS (1977-1987). An observation is a married women of foreign ancestry, born in the US and between 29 and 50 years of age. The dependant variable is the number of children born to a woman. TFR 1950 if the total fertility rate in the woman's country of ancestry in 1950. In-sample countries of ancestry are Canada, Denmark, UK, Finland, France, Germany, Ireland, Italy, Mexico, Netherlands, Norway, Russia, Spain, and Sweden. SIBS is the number of siblings a women has. The individual controls include: age, age squared, and a set of dummy variables to capture the level of education (below high school [omitted], high school degree ("High School"), some college ("Some College"), and at least a college degree ("college"). The parental controls include mother's education and father's education in number of years. Estimations also include region of residence fixed effects, as well as year of survey fixed effects. Standard errors are clustered at country of ancestry level. $^{*}\, p < 0.1$, $^{**}\, p < 0.05$, $^{***}\, p < 0.01$.

Table 4: Replication of Fernández and Fogli (2006), sample of individuals with information on their generation of migration

Dependant variable is Children						
	(1)	(2)	(3)	(4)	(5)	(6)
TFR 1950	0.149***	0.094***	0.115***	0.100***	0.139***	0.127**
	(0.027)	(0.031)	(0.019)	(0.024)	(0.039)	(0.044)
SIBS		0.079***		0.032**		0.033
		(0.016)		(0.014)		(0.019)
Age			0.273**	0.287***	0.343***	0.355***
			(0.092)	(0.089)	(0.090)	(0.092)
Age sq			−0.003*	−0.003**	−0.003**	−0.004**
			(0.001)	(0.001)	(0.001)	(0.001)
High School			−0.757***	−0.699***	−0.567***	−0.516***
			(0.109)	(0.110)	(0.115)	(0.124)
Some College			−0.847***	−0.769***	−0.539***	−0.469***
			(0.070)	(0.064)	(0.160)	(0.155)
College			−1.280***	−1.191***	−1.047***	−0.977***
			(0.121)	(0.125)	(0.120)	(0.119)
Mother's education					−0.065**	−0.063**
					(0.023)	(0.021)
Father's education					0.028**	0.028***
					(0.010)	(0.009)
Obs.	1097	1096	1096	1095	893	892
Adj. R^2	0.036	0.056	0.193	0.197	0.217	0.221
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table replicates Table 3a in Fernández and Fogli (2006), keeping women for who we are able to retrieve information about their generation of migration. Observations, outcome variable and control variables are defined in Table 3. Estimations also include region of residence fixed effects, as well as year of survey fixed effects. Standard errors are clustered at country of ancestry level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 5: Cultural convergence in Fernández and Fogli (2006): heterogeneity by generations of migrants

Dependant variable is Children						
	(1)	(2)	(3)	(4)	(5)	(6)
TFR 1950	0.193***	0.143***	0.215***	0.198***	0.199**	0.184**
	(0.037)	(0.043)	(0.045)	(0.046)	(0.069)	(0.072)
SIBS		0.079***		0.033**		0.031
		(0.015)		(0.014)		(0.019)
TFR 1950 × 3rd gen. migrant	0.033	0.035	−0.013	−0.010	0.017	0.017
	(0.047)	(0.043)	(0.035)	(0.036)	(0.049)	(0.049)
TFR 1950 × 4th gen. migrant	−0.143**	−0.154**	−0.210**	−0.210**	−0.125	−0.121
	(0.056)	(0.056)	(0.087)	(0.085)	(0.083)	(0.082)
TFR 1950 + TFR 1950 × 3rd gen. mig.	0.225***	0.178**	0.202***	0.188***	0.215**	0.201*
	(0.058)	(0.060)	(0.054)	(0.054)	(0.097)	(0.098)
TFR 1950 + TFR 1950 × 4th gen. mig.	0.050	-0.012	0.005	-0.012	0.074*	0.063
	(0.041)	(0.040)	(0.051)	(0.053)	(0.036)	(0.039)
Mean outcome 2nd gen. mig.	2.633	2.633	2.598	2.598	2.537	2.537
Mean outcome 3rd gen. mig.	2.265	2.265	2.265	2.265	2.265	2.265
Mean outcome 4th gen. mig.	2.307	2.306	2.307	2.306	2.296	2.295
Obs.	1097	1096	1096	1095	893	892
Adj. R^2	0.036	0.057	0.194	0.198	0.215	0.220
Individual controls	No	No	Yes	Yes	Yes	Yes
Parental controls	No	No	No	No	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table explores heterogeneities by generation of migration in Table 3a in Fernández and Fogli (2006). We work on Fernández and Fogli (2006) initial sample defined in Table 3, while keeping women for who we are able to retrieve information about their generation of migration. "3rd gen. migrant" is a dummy indicating a woman being a 3rd generation migrant (i.e.) "4th gen. migrant" is a dummy indicating a women being a 4th generation migrant. These two last definitions are based on Giavazzi et al. (2019). These two dummies are also separately included in the regressions. The dependant variable is the number of children born to a woman (we report in the second part of the table the mean dependant variable of each generation of migrants in our regression samples). TFR 1950 if the total fertility rate in the woman's country of ancestry in 1950. SIBS is the number of siblings a women has. We report in the second part of the table the estimates and standard errors of the total effect of TFR 1950 for 3rd generation migrants and 4th generation migrants. The individual controls include: age, age squared, and a set of dummy variables to capture the level of education (below high school [omitted], high school degree ("High School"), some college ("Some College"), and at least a college degree ("college"). The parental controls include mother's education and father's education in number of years. Estimations also include region of residence fixed effects, as well as year of survey fixed effects. Standard errors are clustered at country of ancestry level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.