Migrant Self-Selection: Anthropometric Evidence from the Mass Migration of Italians to the United States, 1907–1925*

Yannay Spitzer
yannay.spitzer@brown.edu

Brown University

Ariell Zimran
ariell.zimran@u.northwestern.edu

Northwestern University

August 8, 2014

Abstract

Are migrants positively or negatively self-selected from within their populations of origin? We study this fundamental and persistent question of the economics of migration using data on one of the largest flows of free migration ever—that of Italians to the United States between 1907 and 1925. We exploit never-before-used stature data in the Ellis Island arrival records—from which we transcribed the heights and other personal information of a random sample of 50,000 Italian passengers—combined with Italian province-birth cohort height distributions and our own geo-matching of millions of Italian passengers to their places of origin in order to construct a novel data set for our analysis. Relying on the well-established relationship between population average stature and living standards, we quantify migrant self-selection by comparing the heights of migrants to the height distributions of their respective birth cohorts in their provinces of origin. Our analysis reveals opposite patterns of self-selection across and within Italian provinces. Italian migrants were shorter, on average, than all Italians of the same birth cohort, suggesting negative self-selection on the national level. However, when compared only to the distribution of stature in their own provinces of origin, we find that Italian passengers were, on average, taller, indicating positive self-selection on the local level. Moreover, we find that the degree of self-selection from a province and birth cohort was decreasing in its average stature, suggesting that less-developed province-cohorts, where liquidity constraints to migration were more likely to bind, provided relatively higher quality migrants. The findings of this research demonstrate the importance of distinguishing between self-selection from a country as a whole and self-selection from within a particular sub-national region. Comparisons of migrants to their national-level origins, which are the norm in the literature on migrant self-selection, may fail to capture a significant portion of the self-selection occurring within a group of potential migrants from a particular sub-national region.

---

1 Introduction

> [A]lthough drawn from classes low in the economic scale, the new immigrants as a rule are the strongest, the most enterprising, and the best of their class . . . .
>
> (The Dillingham Commission, US Congress, 1911, p. 24)

Between 1892 and 1925, nearly four million Italians immigrated to the United States—the largest single flow during the Age of Mass Migration (Ferenczi and Wilcox, 1929, Tables 2–3, pp. 384–393). This phenomenon, part of a general contemporaneous trend of growth in migration to the United States from other southern and eastern European countries, sparked a debate over the policy of nearly total openness of the United States to immigration (Goldin, 1994). Public debate focused primarily on the "quality" of the southern and eastern European migrants.[1] Groups favoring the restriction of immigration warned of a decline in the quality of immigrants, arguing that these immigrants, unlike those arriving *en masse* from northern and western Europe in prior decades, represented the poor, incapable, uneducated, unskilled, and criminal elements of their origin countries; that is, that they were negatively selected from within their environments of origin. In the late 1910s and early 1920s, after decades of agitation, such sentiment finally prevailed with the passage of sweeping immigration restrictions, culminating in the Immigration Act of 1924, which effectively ended unfettered large-scale immigration from Italy and other countries in the European periphery (Hatton and Williamson, 1998, ch. 9).

Despite these allegations and the abundance of research, both modern and contemporary, that they precipitated (e.g., Gomellini and Ó Gráda, 2013; Hall, 1904; Stolz and Baten, 2012; US Congress, 1911), the question of whether migrants during the Age of Mass Migration were positively or negatively self-selected remains unresolved.[2] Even in the modern context, determining the nature and causes of migrant self-selection remains at the forefront of research in the economics of migration (Borjas, 1987; Chiquiar and Hanson, 2005; Fernández-Huertas Moraga, 2013; McKenzie and Rapoport, 2010), and is crucial in understanding the effects of migration on the source and host economies (c.f., Biavaschi and Elsner, 2013). If, for example, migrants are positively self-selected from within their populations of origin, then emigration, by disproportionately leading to the exit of more productive individuals from the sending economy, may harm it (Bhagwati, 1976; Di Maria and Stryszowski, 2009; Docquier and Rapoport, 2012; Mattoo, Neagu, and Özden, 2008; Todaro,

---

[1] We use the term "quality" here to refer to any traits that affect an individual's productivity. Examples include education, skill, health, wealth, and intelligence. Proponents of immigration restriction in the early 20th century had an even broader definition, arguing, for example, that these new immigrants were more likely to be involved in criminal activity, or lacked a history of self-governance that would be crucial to their assimilation in the United States.

[2] Abramitzky, Boustan, and Eriksson (2012, 2013) also study the self-selection of migrants in the Age of Mass Migration, focusing on Norwegians.

1996). Conversely, the receiving economy may benefit from the influx of these productive individuals. If migrants are negatively self-selected, the opposite may occur.

However, empirical answers regarding migrant self-selection remain elusive, primarily due to a number of data limitations that make direct comparisons between migrants and the population at risk for migration difficult, if not impossible in some contexts. In particular, a lack of representative data on the source population often confounds efforts to quantify migrant self-selection. Even when comparison data are available, they generally cannot be disaggregated to geographic levels below the country of origin, raising the possibility that the nature of self-selection from within source populations is obscured by composition effects across sub-national units. Moreover, most data on migrant quality are observed only after arrival in the host country, raising the possibility that they do not reflect the pre-migration characteristics of migrants. In most of the few cases in which these issues can be overcome, the measure typically used to compare migrants to stayers is occupation, which, although informative regarding individual skill and human capital, comprises only a rough measure of a migrant's economic capability, reflecting only limited aspects of it.

In the present research we study self-selection into migration using stature to measure migrant quality. This approach is grounded in a large body of research, which has established that the average stature of a large group is indicative of the group's average economic capability—an amalgamation of many of its facets, such as skill (Komlos, 1990), education (Case, Paxson, and Islam, 2009), income, (Deaton, 2007; Persico, Postlewaite, and Silverman, 2004), wealth (Floud, Wachter, and Gregory, 1990), health (Fogel, 1986; Steckel, 1995), childhood environment (Bailey, Hatton, and Inwood, 2014), and cognitive ability (Case and Paxson, 2008), that all determine a prospective migrant's contribution to his home economy, and his labor market outcomes in the host economy. Stature is thus, at an aggregate level, a proxy for economic capability and productivity in a broader sense than are other commonly used measures, such as occupation-implied skill. Moreover, adult stature is fixed for those in a relatively broad age range and, for such individuals, is unaffected by migration. The premise of this paper is that migrant self-selection can be quantified by comparing the average stature of migrants to that of the populations of origin. If, for example, the population of migrants is taller, on average, than the overall population in the sending economy, then it can be deduced that migrants are more economically productive than non-migrants on average.

Applying this approach to the Italian migration to the United States enables a very reliable comparison of the migrant population to the source population, as the stature distribution for the source population of Italian adult males is known.[3] Moreover, this distribution is known at a geographically disaggregated

_____

[3]In most countries, military height data are only available for a self-selected group of individuals choosing to join the military. In Italy, however, all males were required to be measured by the military. The resulting data were collected by

level, enabling us to avoid obstacles stemming from the fact that the Italian migration was composed of individuals originating in many heterogeneous provinces, and to explore the relationship between provincial characteristics and different features of migration from each province. Studying a historical episode of migration also carries many advantages that are unavailable in the study of modern migration. This approach effectively avoids difficulties created by the fact that modern migratory flows are censored by restrictive immigration policies, and are thus not representative of the latent supply of those willing and able to migrate. Studying historical migration in which such barriers did not exist, allows scholars to cleanly identify migrant self-selection at the source, to learn the mechanisms that determine the nature of migrant self-selection, and to use these insights to make inferences regarding the effects of changes in migration policy on the quantity and the quality of migrants.

In order to perform this analysis, we constructed a novel data set consisting of the stature, place of origin, and other personal information of Italian passengers from the Ellis Island arrival records database. First, we created a geolocation algorithm to assign each of the nearly five million passengers in the Ellis Island database to his province of origin based on his reported last place of residence. Next, we randomly sampled approximately 50,000 Italian passengers arriving between 1907 (when information on stature was first collected on manifests of immigrants arriving in the United States) and 1925 (when the restrictions of the Immigration Act of 1924 entered into force), and transcribed their stature and other information regarding the nature of their voyage to the United States. We then compared the heights of migrants to the distributions of Italian stature gleaned by A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) from Italian military records covering nearly all Italian males at conscription age.

This analysis reveals opposite patterns of self-selection across and within Italian provinces. Italians passing through Ellis Island were shorter, on average, than all Italians of the same birth cohort, providing evidence of negative self-selection on the national level. However, when compared only to the distribution of stature in their own provinces of origin, we find evidence that Italian passengers were, on average, taller, indicating positive self-selection on the local level. The difference between these two findings is driven by positive self-selection within southern provinces, which were the origins of a disproportionately large share of migrants, and in which the average stature was below the national average for Italy. Moreover, we find that immigrants from northern Italy tended to be negatively self-selected from within their provinces of origin— the opposite of their southern compatriots. Moreover, the degree of positive within-province self-selection of immigrants arriving in the United States after 1917 was far greater than that of immigrants arriving in the

---

A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011), and made available to us.

pre-1917 period.

We further investigate what factors determine the degree of self-selection from within provinces, thus providing a test of three major theories of migrant self-selection—relative inequality, liquidity constraints, and network connections.[4] We find that the degree of migrant self-selection was decreasing in the level of development of the province of origin (as measured by its average stature), indicating that immigrants from relatively less-developed environments were, on average, of higher quality relative to their provinces and birth cohorts of origin than those from relatively more developed environments. We also find evidence that migrants who were able to finance their own passage were more positively self-selected, on average. These results are consistent with theories that predict positive self-selection due to the need to overcome liquidity constraints to migration. We also find that individuals who migrated to join an immediate family member were, on average, shorter than those who did not. This finding is consistent with the notion that chain migration particularly helpful for lower quality migrants in overcoming liquidity constraints to migration. We do not find any robust and statistically significant evidence supporting theories that hold that the nature of migrant self-selection is determined by the relative inequality of the sending and receiving countries.

Although taken from the report of the anti-immigration Dillingham Commission, the epigraph to this paper, like our results, demonstrates the importance of distinguishing between self-selection from a country as a whole and self-selection from within a particular sub-national region, and of conceding that the two levels of self-selection may be qualitatively different. Comparisons of migrants to their national-level origins, which are the norm in the literature on migrant self-selection (e.g., Chiquiar and Hanson, 2005; Stolz and Baten, 2012), may fail to capture a significant portion of the self-selection occurring within a group of potential migrants from a particular sub-national region.

The remainder of the paper proceeds as follows. Section 2 provides the relevant historical and economic background for this study. Section 3 discusses the data construction process, and provides summary statistics for the data set used in this study. Section 4 presents the main results, which are interpreted in section 5. Section 6 evaluates various theories of migrant self-selection. Section 7 discusses possible threats to identification. Section 8 concludes.

---

[4]These theories are discussed in more detail in section 2.2.

2 Background

The issue of migrant self-selection, particularly from Italy in the early 20th century, but also in the context of modern migration, has been studied extensively by economists, modern and historical. In this section, we provide background on the mass migration of Italians in the early 20th century, as well as on the body of economic knowledge on the issue of migrant self-selection.

2.1 Historical Background

At the beginning of the 20th century, Italy lagged behind most other western European countries in terms of nearly every economic indicator. As shown in Figure 1, real wages were low, less than half their level in Britain (O'Rourke, 1997). Moreover, Italy's industrial production lagged that of its neighbors (Ciccarelli and Fenoaltea, 2013), and malaria and other diseases were endemic, particularly in the south (Foerster, 1919). As a result, living standards in Italy, measured by average stature, fell short of those of most other European countries, as depicted in Figure 2. These poor economic conditions spurred many Italians to leave their home to seek opportunity elsewhere. Such was the strength of this incentive that by the turn of the century, Italy had become the largest source of migrants to the United States, displacing such countries as Ireland, Great Britain, Sweden, and Norway. Moreover, as depicted in Figure 3, Italy led Europe in terms of relative migration, with the highest rates of emigration per capita of any European country in the period 1900–1913.[5]

In some ways, the Italian migration was typical of the Age of Mass Migration: migrants were mostly young, unskilled, and male; but in other ways, Italians were distinct from other migrants. First, they tended to distribute themselves between several destination countries, primarily the United States, Argentina, and Brazil. Between 1886 and 1895, nearly 75 percent of Italians traveling to the Americas went to Argentina or Brazil, with the remainder going to the United States. By the period 1906–1915, the United States became the lead destination for Italians, drawing more than twice the numbers of Brazil and Argentina combined. There was also considerable (mostly seasonal) migration to other European countries of a magnitude rivaling that of the flow to the United States (Hatton and Williamson, 1998, Table 6.1, p. 101). Second, roughly three-quarters of migrants were male (Hatton and Williamson, 1998, p. 102), a gender imbalance exceeding that of almost any other group. Finally, Italians were, more than any other group of migrants, likely to migrate temporarily rather than to remain abroad permanently (Gomellini and Ó Gráda, 2013). The canonical

---

[5]If we include sub-national ethnic groups, however, Russian Jews were more likely to emigrate than Italians (see Spitzer, 2013, 2014, for more details).

example is the tendency of these so-called "birds of passage" to exploit the seasonal differences between Italy and South America, traveling between the two in order to participate in both harvests (Foerster, 1919; Hatton and Williamson, 1998). Return migration also occurred from the United States; as the annual nominal wage in the United States was nearly five times that in Italy, and the cost of round trip passage would consume only 20 percent of those earnings, many Italians would travel to the United States to work, and then return or remit the money to their families in Italy (Gomellini and Ó Gráda, 2013). It has been estimated (Bandiera, Rasul, and Viarengo, 2013, Table 4, p. 37) that as many as 80 percent of Italian migrants to the United States eventually returned to Italy.

Discussing emigration from Italy as a whole obscures the considerable variation in emigration rates and the general patterns of migration across Italian provinces (Hatton and Williamson, 1998, p. 106). Whereas most Italian emigrants in the 1880s were from the north (Gomellini and Ó Gráda, 2013), the poorer south had taken the lead in terms of emigration rates by 1900 (Ferenczi and Wilcox, 1929, Table 10, pp. 432–443; Hatton and Williamson, 1998, Table 6.4, p. 107). The greater emigration rates from the south were driven primarily by the fact that the south was much poorer than the north. As a primarily rural agricultural economy characterized by lower real wages (Hatton and Williamson, 1998, pp. 115–116), southerners had a stronger push to emigrate than did northerners. In addition, the north's relative proximity to major European labor markets caused many northerners to migrate within Europe rather than overseas. Thus, the mix of Italian immigrants to the United States was primarily southern, and therefore drawn from the relatively poorer portions of the country.

Americans were aware that the bulk of Italian immigration to the United States increasingly originated in the poorer south, and many were displeased with the new growth of Italian immigration in the early 20th century. Many of those opposed to a continued openness to immigration felt that malicious forces were at work in Europe to transfer the least desirable elements of the population of Europe to the United States (Commissioner-General of Immigration, 1903; Hall, 1904), citing the immigrants' lack of skill, and (as perceived by anti-immigration activists) lack of mental and physical fortitude as evidence. Writing at the height of the migration, the Commissioner-General of Immigration (1903, p. 70) asserted that

> The great bulk of the present immigration proceeds from Italy, Austria, and Russia, and, furthermore, from some of the most undesirable sources of population of those countries. No one would object to the better classes of Italians, Austrians, and Russians coming here in large numbers; but the point is that such better element does not come.

Notably, claims of such negative selection were not seriously disputed (c.f., Douglas, 1919). Even ad-

vocates of continued openness to immigration accepted them, arguing that better measures ought to have been taken to prevent the dependent and criminal elements from entering the United States (Brandenburg, 1904), that the United States's tradition as a haven for immigrants was worth maintaining, and that the immigrants would eventually converge (even physically) to American standards through their time in the United States (Boas, 1911).

Nativist concerns led to the formation of the Dillingham Commission (US Congress, 1911), which was charged with investigating the nature and effects of the mass immigration. After collection and analysis of considerable data, the commission enumerated in great detail the negative characteristics of the immigrants, ranging from their poor living conditions to their lack of education and skill, eventually concluding that immigration restriction was necessary in order to protect the "national character." These restrictions culminated in the literacy test of 1917 and finally in the quotas of the Immigration Act of 1924, which brought an end to mass immigration to the United States from the European periphery.

2.2   The Economics of Migrant Self-Selection

Theoretical foundations for the economic analysis of migrant self-selection are laid out by Borjas (1987). According to his modification of the Roy (1951) model, the nature of self-selection into migration is determined by the relative returns to skill in the sending and receiving economies. If the returns to skill in the receiving country are higher relative to those in one source country than in another, emigrants from the latter are predicted to be more strongly positively self-selected than those from the former on the basis of skill. In most studies of migrant self-selection, the relative returns to skill are proxied by the relative inequality of the income distributions of each country.[6] When focusing on relative inequality, positive self-selection of migrants is predicted to be induced when the income distribution of the sending country is less unequal than that of the destination country. Conversely, when the sending country's income distribution is more unequal than that of the destination country, migrants are predicted to be negatively self-selected. Borjas (1987), for example, uses this framework to argue that the deteriorating performance of successive cohorts of immigrants in the latter half of the twentieth century (as measured by their earnings and integration into the American labor market) can be explained by the fact these immigrants have increasingly originated in very unequal countries, and are thus of lower quality.[7]

This so-called relative inequality model has met with mixed empirical success. Chiquiar and Hanson

---

[6]We are grateful to Timothy Hatton for pointing out this distinction.

[7]Interestingly, similar arguments were made to explain the labor market performance of the "new" immigrants from the southern and eastern European periphery as compared to that of immigrants from northern and western Europe in the early 20th century.

(2005) find that immigrants from Mexico, in which the income distribution is very unequal, are not negatively self-selected on the basis of earnings, skill, or education. Other recent empirical studies have also found evidence of positive self-selection into migration, usually with respect to skill or education, regardless of the relative inequality of income distributions (Feliciano, 2005; Gould and Moav, 2010; Grogger and Hanson, 2011). These findings are rationalized by the presence of migration costs and borrowing constraints that disproportionately inhibit migration by those in the lower tail of the income distribution (Chiquiar and Hanson, 2005; Chiswick, 1999; Mishara, 2007). Therefore, regardless of the nature of self-selection of those wishing to migrate, only those of higher quality are able to overcome liquidity constraints and actually do so, generating more positive self-selection. However, Ibarraran and Lubotsky (2007) and Fernández-Huertas Moraga (2011) dispute this finding. Several explanations have been offered to reconcile these results. Borger (2009), McKenzie and Rapoport (2010), Spitzer (2013), and Wegge (1998) argue that the direction of self-selection is indelibly tied to the strength of the potential migrant's social network. Stronger social connections in the destination country enable individuals who would otherwise be unable to overcome liquidity constraints to migration to do so, resulting in a weakening of the distortive effects of migration costs and borrowing constraints on the Roy model effects. Belot and Hatton (2012) and Fernández-Huertas Moraga (2013) attempt to reconcile these disparate theories. Belot and Hatton (2012) show that once poverty constraints are accounted for, patterns of self-selection appear to correspond to the predictions of the relative inequality model. Fernández-Huertas Moraga (2013) finds that a combination of the three explanations is required to fully account for differences in the pattern of self-selection between urban Mexico (whence migrants are negatively self-selected) and rural Mexico (whence they are positively self-selected).

A different explanation for the composition of migratory flows is given by development economists, who have recently examined the role of risk in the migration decision, which is highlighted by Harris and Todaro (1970). Bryan, Chowdhury, and Mobarak (2014) find evidence that risk aversion prevents rural-to-urban migration in developing countries. Those with greater wealth, who would be better able to bear the risk, would therefore be more likely to migrate, generating positive self-selection.

Overall, despite a vast literature studying the issue of migrant self-selection in both modern and historical contexts, a consensus on its nature and its causes and mechanisms remains elusive. Findings of different directions of self-selection in different studies make external generalizations difficult.

2.2.1 Issues of Data Availability

Empirical disagreement regarding the direction and causes of migrant self-selection can be partially attributed to data limitations that prevent or restrict comparisons between migrants and the population at risk for migration in the migrants' economy of origin. In the absence of other data, many studies have relied on aggregate data of the place of origin to study migrant self-selection (Bohlin and Eurenius, 2010; Hatton and Williamson, 1998; Lowell, 1987; Runblom and Norman, 1976). This approach is generally used when micro data are unavailable, for example, when only aggregate statistics on the volume of migration between two countries are available. In this approach, self-selection into migration is studied by comparing migration patterns across regions. If, for example, the rate of migration from higher income areas is greater than that from other areas, then the conclusion is drawn that migrants are positively self-selected on the basis of income. The approach is confounded, however, if self-selection also occurs within regions. Returning to the previous example, migrants from areas with higher average income may be poorer than non-migrants from the same area and are thus properly understood to be negatively self-selected on the basis of income.

Ideally, micro data would be used, pinning down the types of migrants and permitting the comparison of individuals within a specific (possibly sub-national) source population to one another. However, samples in which prospective migrants' quality is observed prior to migration are rare (Akee, 2010). Instead, scholars using micro data to study modern migration are often forced to rely on data collected after the migration has taken place (Chiquiar and Hanson, 2005; McKenzie and Rapoport, 2010). Except for certain indicators, measures collected after the migrants have been in the receiving economy for some time are likely to be contaminated by the experience in the destination. For example, occupations of immigrants may change depending on the labor market conditions of the receiving country.[8] Most studies in which such data are available focus only on very small migration flows, such as from Pacific island nations to the United States and New Zealand (Akee, 2010; McKenzie, Gibson, and Stillman, 2010), or from Finland to Sweden (Rooth and Saarela, 2007).

Even if pre-migration data on migrants are available, a comparison between migrants and non-migrants requires data on the distribution of productive characteristics of the population of origin. Without data on the population of origin, it cannot be determined whether migrants are positively or negatively self-selected. For example, individuals with low education in an absolute sense may in fact be highly educated relative to

---

[8]For example, Perlmann (2000) shows that the share of laborers and manufacturing workers among Jewish immigrants during the Age of Mass Migration was much higher than in the population of origin. Occupation-based self-selection is possible, but Perlmann (2000) argues that many migrants may simply have changed occupations on arrival. Ferrie (1997) also raises the issue that many immigrants tend to work in different occupations before and after migration.

their population of origin, but this cannot be determined without data on the source population. Chiquiar and Hanson (2005) show that failing to take this issue into account can lead to spurious conclusions regarding the nature of self-selection.

Fernández-Huertas Moraga's (2013) study of Mexican migrants to the United States is one of the few that is able to overcome these constraints. He compares survey data on migrants to that on non-migrants from surveys conducted before migration occurs, and is able to compare the two routs on the basis of wages, unemployment rates, and labor market participation. However, Fernández-Huertas Moraga's (2013) study, like nearly all studies of modern migration, suffer from a problem of dual selection. That is, two sample selection processes operate to determine the composition of modern migratory flows: the process that determines whether migrants find it optimal to migrate (which the literature on migrant self-selection is interested in understanding), and the selection caused by restrictive immigration policy; the latter process generally obscures the former, and comparisons between migrants and non-migrants do not reveal the nature of migrant self-selection.

2.2.2 Advantages of Historical Data

Historical data make it possible to overcome some of these hurdles. First, the problem of dual selection generally does not apply. Specifically, prior to the literacy restrictions imposed in 1917, migration to the United States from Europe was almost entirely unrestricted. Even after the literacy test was imposed, Goldin (1994) hypothesizes that migration from Europe to the United States was not significantly restricted until the quotas imposed in 1924. Thus, migrants in the Age of Mass Migration were selected only by the process about which we wish to learn: that which determines whether individuals find it optimal to migrate. It is therefore possible to identify cleanly the self-selection of migrants at the source. Second, historical data often provide access to micro data that are unavailable to researchers in modern contexts. For example, unlike most modern data, historical data are not subject to confidentiality restrictions. Moreover, much modern migration is undocumented; the lack of significant legal restrictions to migration in our study period made it unnecessary to enter the United States illicitly, ensuring that most migration was documented.

Several studies have used historical data to overcome these data limitations. Abramitzky, Boustan, and Eriksson (2013) exploit the availability of tax data in Norway to study self-selection on the basis of wealth into migration to the United States during the Age of Mass Migration. Unfortunately, Abramitzky, Boustan, and Eriksson (2013) are forced, by issues of data availability, to rely primarily on a binary indicator of whether a household owned any taxable assets, and are, in their analysis of international migration, unable

to further disaggregate wealth. The issue of coarseness also arises when occupational data are used, such as by Abramitzky, Boustan, and Eriksson (2012), Biavaschi and Elsner (2013), and Wegge (1999, 2002). Such data generally require that individuals be characterized as either skilled or unskilled, masking much useful variation in the quality of prospective migrants. It is also possible to rank occupations by median income (Abramitzky, Boustan, and Eriksson, 2012; Biavaschi and Elsner, 2013), but no variation within occupations is recovered. For instance, this approach cannot differentiate between poor and wealthy farmers, who would have had vastly different living standards. Indirect inference can also be made from the post-migration outcomes of immigrants and their children. For instance, Ferrie and Mokyr (1994) find higher rates of entrepreneurship among immigrants than natives, suggesting positive self-selection. Moreover, Abramitzky, Boustan, and Eriksson (2014) find that immigrants from some European countries in this period hold higher-paid jobs than natives on arrival, suggesting that they may also have been positively self-selected.

In most historical literature, however, even coarse data on traditional economic indicators are generally unavailable.[9] Instead, two alternative methods of measuring migrant quality relative to the population at risk for migration are common in the historical literature. Mokyr (1983) and Mokyr and Ó Gráda (1982), for instance, use age heaping, based on individual age reports, to infer the numeracy of Irish immigrants to the United States. Stolz and Baten (2012) perform a similar analysis, comparing migrants from a number of countries to census records.

Finally, even when all of these constraints can be overcome, it is generally only possible to evaluate the self-selection of migrants on a national level; that is, migrants are generally classified only by their country of origin. Analysis of self-selection on the national level, however, may obscure self-selection at the local level as a result of composition effects across sub-national entities, leading to incorrect conclusions regarding the true nature of self-selection. Fernández-Huertas Moraga's (2013) study of modern Mexican immigration to the United States is, in part, an exception to this restriction. While he does not distinguish between different geographic places of origin within Mexico in determining self-selection, he does distinguish between migrants from urban and rural areas in Mexico, finding that migrants from each sector are self-selected differently. Pooling the sample shows evidence of negative self-selection, obscuring differences in incentives and ability to migrate among different sectors of the population. As we show in the present research, Italy is a case in point, with migrants from North and South Italy exhibiting different patterns of self-selection.

---

[9]Wegge (2002) also collects data on the wealth of migrants, but systematic misreporting of this figure due to restrictions on the expatriation of cash, together with the lack of comparable data for non-migrants, makes it difficult to draw conclusions regarding self-selection with respect to wealth.

2.2.3 Stature as a Measure of Pre-Migration Living Standards

The use of stature as a measure of economic capability and productivity is grounded in a large literature. Fogel (1986, 1994), Fogel, Engerman, and Trussell (1982), Komlos and Meermann (2007), and Steckel (1995) summarize the vast literature establishing a relationship between adult stature and the standard of living experienced by a population in youth. With the genetic variation in height across individuals averaging out in comparisons of large groups to one another (Eveleth and Tanner, 1976; Frisancho, 1993; Martorell and Habicht, 1986; Silventoinen, 2003; Steckel, 1995),[10] the average stature of a population represents the difference between its gross nutrition in youth (principally, its caloric intake) and contemporaneous demands on nutrition, such as labor and disease. Thus, in addition to being correlated with traditional measures of the standard of living, such as income or GDP per capita, stature captures additional facets of welfare such as health and consumption (Steckel, 1995). The variation in stature is also informative about the degree of inequality in the population in the consumption of inputs to stature (such as food, health, and leisure) (Komlos, 1985, 1990; Komlos and Baten, 2004; Steckel, 1995)—a feature that Stolz and Baten (2012) exploit in order to measure inequality when other data are lacking.

Stature is not only correlated with inputs to individual productivity. In essence, stature is a composite measure of human economic capability—an amalgamation of all factors that ultimately determine an individual's standard of living. Thus stature reflects overall quality and economic capability in two ways. Individuals facing better conditions in childhood (e.g., more food availability, less disease, less hard work) will both become taller as adults and will also develop superior cognitive skills (Case and Paxson, 2008). These individuals tend to become more educated than their peers (Case, Paxson, and Islam, 2009), and to earn higher wages (Lundborg, Nystedt, and Rooth, 2009; Persico, Postlewaite, and Silverman, 2004) and enter into higher-skilled work (Komlos, 1990). Height might also reflect unobserved ability through an indirect channel: if the provision of a better childhood environment, which would make children taller, is correlated with parents' characteristics, such as ambition and resourcefulness, then taller children are more likely to have inherited such productive characteristics from their parents. Furthermore, for certain occupations, there are returns to strength, which is correlated with height (Bodenhorn, Guinnane, and Mroz, 2013).

Stature data can therefore be used to address several of the shortcomings of previous studies of migrant self-selection by overcoming many of the data limitations that they have faced. While other measures of pre-migration welfare, such as occupation and wealth, have their own advantages,[11] the resistance of

---

[10]The lack of a genetic difference in adult heights is particularly true when the two groups are from the same place of origin, as is the case in the present research.

[11]For example, occupational status is measured with less idiosyncratic noise, and thus can be used in cases in which only

stature to contamination by post-migration events, its correlation with unobserved ability, education, health, consumption, and pre-migration welfare, and the availability of finely measured stature data make it an attractive tool for the study of migrant selectivity. Applying this approach specifically to Italian migration also makes it possible to study self-selection from sub-national regions due to the availability of geographically disaggregated and finely measured data on the stature of the Italian population of the time. The historical coverage of the data remove the dual selection issue.

Stature has been used by several scholars to study migrant self-selection. Crimmins et al. (2005) examine the self-selection of Mexican migrants to the United States in the modern context. However, the lack of geographically disaggregated data and the confounds raised by the dual selection of migrants in modern data limit the generalizability of the conclusions. Kosack and Ward (2013) expand this approach, analyzing Mexican immigration to the United States in the early 20th century. Unfortunately, they are unable to compare the stature of migrants to that of a representative sample of Mexicans as no such sample is known to exist for their study period. Instead the average stature of migrants is compared to that of volunteer soldiers and passport applicants. As Bodenhorn, Guinnane, and Mroz (2013) point out, however, both of these comparison samples are likely to suffer from sample-select biases. Thus, it is impossible to determine whether the finding that Mexican migrants were taller than soldiers and shorter than passport applicants is an indication of the self-selection of migrants, of the comparison samples in question, or of some mixture thereof. Humphries and Leunig (2009) study the location choices of mid-nineteenth-century British seamen based on height. The scope of conclusions that can be drawn from this study are very limited in their generalizability to the self-selection of an entire population into international migration.

Our study thus improves upon previous attempts to understand self-selection into migration on a number of fronts. First, we use an easily and finely measured variable that is known to reflect living standards and other facets of quality, and whose measurement is not affected by the decision to migrate. Second, we compare migrants to data on the population of origin that are virtually free of self-selection. Third, our comparison is disaggregated to the province-birth cohort level, enabling us to study self-selection within small population bins, as well as the variation in self-selection across time and space—all while remaining cognizant of the different origin populations of the migrants. Finally, our focus on the Age of Mass Migration allows us to cleanly ascribe observed self-selection to individuals' migration decisions, rather than to a combination of these decisions and restrictive policies.

---

small samples are possible. It is also directly informative on skill and human capital. In contrast, stature requires large samples in order to eliminate idiosyncratic differences between individuals.

2.3 Self-Selection of Italian Migrants in the Age of Mass Migration

Although there have been a number of attempts to determine the nature of self-selection of Italians migrating to the United States during the Age of Mass Migration, a clear answer has eluded researchers (Gomellini and Ó Gráda, 2013). In all cases, the difficulties in studying migrant self-selection discussed above apply. Arguments of negative self-selection are advanced by Betrán and Pons (2004), who find that skill premia were falling in Italy and rising in the United States during the Age of Mass Migration. These trends indicate that unskilled laborers were disproportionately overrepresented in emigration, leading to a relative scarcity of unskilled labor in Italy. Stolz and Baten (2012) also present evidence of negative self-selection, finding that age heaping among Italian migrants was greater than among the origin population, suggesting negative self-selection on the basis of numeracy. Giffoni and Gomellini (2013), studying the relationship between migration and school dropout rates, support this view, at least partially, arguing that they find no evidence of positive self-selection of Italy. Arguments for positive selection, however, are advanced by Gomellini and Ó Gráda (2013), who point out that south Italian immigrants were more likely to be literate than their origin populations. Notably, the latter study makes a comparison of migrants to their region of origin, while the former compares immigrants to the entire country.

Anthropometric measures have also been used in this context. Danubio, Amicone, and Vargiu (2005) sample citizenship petitions filed by Italian immigrants in Massachusetts and find an average height greater than that reported by A'Hearn (2003) and Federico (2003), and computed by A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) for the population of Italy. Gomellini and Ó Gráda (2013) interpret this result as suggesting positive selection into migration. The present research builds on this strategy by disaggregating the scope of analysis to the sub-national level of Italy and by using data collected prior to migration, thus eliminating the possibility of post-migration contamination of stature through continued growth.

3   Data

The data set used in this paper is novel in two ways. First, it makes use of the stature data in the Ellis Island arrival records, which we discuss in further detail below. Second, it links Italian migrants to their places of origin with a great deal of disaggregation. In this section we discuss the collection of our data in further detail, provide summary statistics, and test whether our sample is representative of the population of migrants.

3.1 Data Sources

The primary data sources for our analysis are the province-birth cohort-level Italian stature distributions computed from military conscription records by A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011), in addition to the Ellis Island arrival records database. These two sources are discussed immediately below. We also supplement these data with population and literacy data from the Direzione Generale della Statistica e del Lavoro (1912) and the Ministero di Agricoltura, Industria e Commercio (1915, 1925).

3.1.1 Italian Stature Data

Analysis of self-selection of any kind based on stature requires a comparison sample that is known to accurately represent the population at risk for migration, or at least to represent non-migrants as a whole, without further self-selection. We are fortunate that such data exist in the Italian case. As a comparison sample for our migrants, we use height data compiled as a part of the Italian military conscription process.[12] During the period in question, all Italian males, regardless of physical condition, were required to present themselves for a medical examination, during which their heights were measured and recorded (A'Hearn, Peracchi, and Vecchi, 2009). As these data are the product of nearly the full male population of Italy (Cole, 1995), they are representative of the population as a whole. In particular, these data (as corrected by A'Hearn, Peracchi, and Vecchi, 2009; A'Hearn and Vecchi, 2011) are unlikely to suffer from issues of self-selection that are problematic in the historical heights literature.

We acquired two sets of data based on the conscription data, one from A'Hearn, Peracchi, and Vecchi (2009) and the other from A'Hearn and Vecchi (2011). The A'Hearn and Vecchi (2011) data contain the raw means and standard deviations of the height distributions of each province (except Caserta) and birth cohort from 1855–1910, as well as these values standardized to their age-20 values. We refer to the latter data as the "Unsmoothed Age 20" series. Examples of the time series of means and standard deviations of the unsmoothed age-20 distributions are presented in Figure 4. The distributions of age-20 stature may not be suitable for comparison to those of migrant stature due to the possibility of growth after age 20. Although Beard and Blaser (2002) and Frisancho (1993) show that modern populations reach terminal height by age 20, the same need not be true of our study population. Indeed, a number of studies (A'Hearn, Peracchi, and Vecchi, 2009; Fogel, Engerman, and Trussell, 1982; Frisancho, 1993) discuss the potential for malnutrition to both reduce final adult height and to delay the onset of the adolescent growth spurt (AGS), leading growth

---

[12]For a detailed description of the data and their origin and collection, see A'Hearn, Peracchi, and Vecchi (2009).

to continue into the early 20s. Similarly, A'Hearn, Peracchi, and Vecchi (2009) report that the delayed AGS may be responsible for a decline in the standard deviation of height in a population with age.

Unfortunately, the delayed AGS is a poorly quantified anthropometric phenomenon. We were not able to find any literature quantifying the effect of nutrition on the rate of growth over the lifespan, and thus we have only a limited understanding of the bias introduced by using the age-20 distributions as a basis for comparison. In particular, we do not know to what extent the bias (i.e., the continued growth after age 20) depends on average height at age 20. What we do know, however, is that shorter cohorts are likely to continue growing further into their twenties. That is, the age 20 distributions may be an image of height that is earlier in the growth process for shorter populations than for taller populations. This issue would thus create a mechanical bias toward finding stronger positive self-selection among shorter populations.[13]

We therefore take advantage of computations performed by A'Hearn, Peracchi, and Vecchi (2009). The primary computations of these authors resulted in the lines in Figure 4 labeled "Smoothed Age 20," which represent the time-smoothed age-20-corrected means and standard deviations of stature. They also adjust these distributions for continued growth to age 22, with the results represented in the series labeled "Smoothed Age 22." These distributions are based on changes in the timing of measurement over the lifespan by the Italian military,[14] but are, for the most part, out-of-sample projections performed by A'Hearn, Peracchi, and Vecchi (2009). Nonetheless, the growth that these adjusted height distributions depict relative to the age 20 distributions constitutes the most rigorous possible analysis of post-age-20 growth for the population under analysis. However, the smoothed age-22 distributions eliminate potentially valuable within-province variation over time. We therefore compute an unsmoothed age 22 distribution, labeled "Implied Age 22" in Figure 4, by adjusting the unsmoothed age-20 means by the province-birth cohort-specific difference between the smoothed age-20 and smoothed age-22 means. We perform a similar operation on the standard deviations of the distributions, which are similarly smoothed by A'Hearn, Peracchi, and Vecchi (2009) and not by A'Hearn and Vecchi (2011). By performing this correction, we produce province-birth cohort-specific height distributions normalized to age 22, at which the risk of further growth would have become negligible even in malnourished populations. We therefore consider these distributions to be the best available representations

---

[13]We have replicated the results of this paper using the age 20 distributions for comparison. The magnitude of self-selection that we find is much stronger than our main results in this paper.

[14]Specifically, there were variations in the age of measurement by the Italian military induced by the military calling up different birth cohorts at different ages. A'Hearn, Peracchi, and Vecchi (2009) report that the vast majority of birth cohorts are measured at age 20, but that for institutional reasons, some were measured as early as age 18, and others as late as age 22. Based on this variation, A'Hearn, Peracchi, and Vecchi (2009) compute the average stature at age 22 for each province and birth cohort by extrapolating from the age 20 distributions that they observe using the differences in the stature observed in cohorts measured at different ages.

of the average adult height of each birth-cohort and province.[15] The time trend in the average height of the Italian population is depicted in Figure 5.

It may be the case, however, that some smoothing of these moments is necessary. There is likely very little sampling error in the moments, as they come from nearly the entire male population; but there may be error in the reporting of ages at Ellis Island that leads us to assign passengers to the wrong birth cohort, and thus to the wrong comparison distribution. Some smoothing of the moments over time may therefore be necessary. We therefore compute for each province a kernel regression of each moment against the birth year, thus providing a smoothed version of the moments of the distributions.[16] The smoothed moments are also presented in Figure 4, and are labeled "Our Smoothed Age 22." Comparison to these distributions produces results that are not appreciably different from those driven by comparisons to unsmoothed distributions except in a small number of cases noted below.

3.1.2 Ship Manifests

Our information on the stature and other personal characteristics of migrants is taken from the Ellis Island arrival records database, which includes information on nearly all passengers who passed through the Port of New York from 1897 to 1925.[17] This database is composed of passenger manifests deposited at Ellis Island, of which Figure 6 presents an example. Some of the information on these manifests is already transcribed, while the rest is available in handwritten form on the scanned manifests. These manifests were completed upon embarkation by the steamship companies transporting the passengers to Ellis Island, and were primarily intended to fulfill two purposes. First, they were used to maintain statistics on the number of immigrants of each gender and nationality entering the United States. Second, they were part of an effort to ensure that immigrants who might become a public charge, who were ill, or were otherwise undesirable (for instance, by being anarchists or polygamists), were prevented from entering the United States (Bureau of Immigration and Naturalization, 1909). Steamship companies were therefore required to assert that they had examined all passengers, and to affirm that they did not violate any of these restrictions. Beginning in late 1906, with the passage of the Immigration Act of 1906 (US Congress, 1907), passenger manifests were required to include a physical description of the passenger, of which height was a component.

---

[15]We have also produced all of the results presented below with the age 20 distributions as the point of comparison. All results are stronger with the age 20 distributions than with the age 22 distributions.

[16]We compute our own smoother in order to provide a province-specific average over time. We do not use the smoothed means computed by A'Hearn, Peracchi, and Vecchi (2009) because they are not simply averages over time, but are instead affected by the temporal trend in other provinces.

[17]The first five years during which Ellis Island was in operation (1892–1897) are only partially covered for two reasons. First, Ellis Island at this time operated in conjunction with the older Castle Garden facility, where some immigrants were processed. Second, an 1897 fire at Ellis Island destroyed many records that were stored there.

Figure 8 presents the time series of arrivals according to both the Ellis Island data base and the official immigration statistics of the United States (Ferenczi and Wilcox, 1929, Tables 2 and 3, pp. 384–393). The former exceeds the latter for two reasons. First, it includes both immigrants and individuals entering temporarily, while the official immigration statistics only include people entering with the intention of remaining permanently. We also include in Figure 8 the Ellis Island statistics deflated by the number of individuals in our sample who report being first-time arrivals—a proxy for the number of actual immigrants. Second, the Ellis Island data include individuals who purchased passage but never embarked;[18] the official immigration statistics include only actual arrivals. The official statistics may also include individuals not included in the Ellis Island data, as the Port of New York was not the only place of entry for Italian migrants,[19] though it was the location of the bulk of arrivals.

We acquired from the Statue of Liberty-Ellis Island Foundation (SOLEIF) a subset of this database, consisting of the transcribed information of the roughly 4.8 million individuals passing through Ellis Island in this period who either reported their ethnicity as Italian, north Italian or south Italian, or whose country of origin was Italy. We restricted the sample to those arriving in 1907 or later so as to consider only those whose heights would have been recorded under the new law.[20] This restriction left approximately 2.8 million passengers in the sample. Next, we geocoded the passengers' reported last place of residence using an algorithm outlined in Appendix A.[21] As we discuss in Appendix A, a variety of tests and exercises show that this algorithm is remarkably accurate for the individuals who can be matched: the rate of false matches may be below five percent. Moreover, as shown in Figure 7, the correlation of average provincial heights of men aged 22–65 recorded at Ellis Island and average provincial heights reported by the Italian military (as adjusted by us and by A'Hearn and Vecchi, 2011) is 0.72. In section 3.3, we analyze whether there are differences between individuals who were matched, and those who were not. We formally explore the possible effects of incorrect geolocation on our results in section 7.2.

We then sampled approximately 50,000 passengers arriving after 1907, for whom we transcribed information from the original manifests that was not already transcribed by the SOLEIF.[22] The data that we

---

[18]We thank Drew Keeling for pointing this issue out to us. Our present sample includes individuals who did not embark; we will adjust the sample in future transcription.

[19]Secondary ports of arrival, such as Boston, New Orleans, and Philadelphia also received substantial migratory flows from Italy; but all of these together amounted to a small share of the total.

[20]There were also a small number of passengers who reported a place of residence in Italy but an ethnicity other than Italian, north Italian, or south Italian. We also omit these individuals from consideration.

[21]A possibly more appropriate indicator of an immigrant's origin would be the place of birth; however, unlike the last place of residence, this information is not available in digital form. If internal migration was common in Italy at the time, there would be differences between the two locations that could lead to incorrect assignment of individuals to comparison distributions. In future work, we will transcribe the place of birth of a sample of migrants and compare them to the last place of residence in order to determine the extent of possible error generated by using the last place of residence instead of the place of birth in order to select a comparison distribution for each individual.

[22]We transcribed a simple random sample of households (identified by the ordering of individuals on the manifests and by a

received in digital form (indicated by the dashed lines in Figure 6a) included the passenger's name, marital status, age, date of arrival, ethnicity, nationality, and last place of residence. We transcribed the answers to four additional questions asked regarding the migrant (indicated by dashed thick solid lines in Figure 6b): whether he had paid for his own passage, and if not, who had paid for the passage; whom he would be joining in the United States; whether he had ever been in the United States before; and his height.[23]

3.2 Summary Statistics

Figure 9 depicts the arrivals of Italian passengers in the entire 1907–1925 period, disaggregated by the province to which they were matched by our geolocation algorithm. A striking pattern is evident: first, southerners were much more likely to migrate to the United States than their northern counterparts. In particular, the rates of emigration in the regions of Sicily and Abruzzo are over 12 percent, while those in Emilia Romagna were nearer to two percent. Moreover, southerners represented a much larger proportion of all Italian passengers traveling to the United States than did northerners. Eleven provinces of southern Italy, and none in northern Italy, were the origins of more than 50,000 geolocated passengers each. Moreover, nearly all provinces from which fewer than 5,000 geolocated passengers originated are located in the north. In total, 82 percent of passengers in our geolocated sample are matched to a southern province.[24]

We restrict our sample to individuals aged 22–65 who could be matched to a province of origin by our geolocation algorithm.[25] We make this age restriction and retain it throughout the paper, as this is the range of ages over which we can be confident that terminal height has been achieved, but rapid shrinking has not begun.[26] Moreover, we see a peculiar pattern in the age distribution of male migrants, which is illustrated in Figure 10. As is typical of the Age of Mass Migration, the density of the age distribution is greatest in the early twenties. There is, however, a large dip in the distribution between ages 18 and 21, a trend that we do not observe among Italian women, and which is not present, for instance, among Russian Jewish immigrants (Spitzer, 2013, 2014). We believe that this dip, which corresponds to the age of military service, may be

---

common last name) and not of individuals. Thus, an individual traveling with one companion was twice as likely to be sampled as an individual traveling alone. Of all passengers between 1907 and 1925, nearly 75 percent traveled alone, and 94 percent traveled in groups of three or less. All further discussions are therefore corrected for this sampling technique through the use of appropriate weights.

[23]We are very grateful to Roy Mill for providing us with access to his dEntry transcription system, and for devoting considerable time and effort to making it compatible with the requirements of this project.

[24]This figure falls to 81% when individuals from Caserta, for which we do not have population stature information, are dropped.

[25]We examine whether our algorithm induced sample selection bias in section 3.3.

[26]Cline et al. (1989) show that shrinking begins essentially as soon as final height is attained, but accelerates with age. In any event, changing the end point of our sample in terms of age will not have large effects on our results, as there are relatively few older immigrants as compared to younger ones. The distribution of ages in our sample is illustrated in Figure 10. In the few cases in which our results are qualitatively affected by reducing the terminal age of our sample to 40 (which is a more conventional terminal age Silventoinen, 2003), we describe the difference.

attributable to the restrictions on legal emigration for males in this age range (Cole, 1995). We therefore suspect that migrants in this age range are self-selected differently from their fellow countrymen emigrating at a later age.[27]

We present summary statistics in Table 1. Column (1) presents summary statistics for all men and women in our geolocated sample, restricting attention to the fields for which no transcription was necessary. Consistent with official statistics (Ferenczi and Wilcox, 1929), we see that the immigrants are overwhelmingly male—more than 75 percent of our sample. Moreover, approximately 70 percent of passengers in our sample reported being married. Approximately 85 percent are matched to a province in southern Italy, as defined by the Bureau of Immigration. Columns (2) and (3) present these statistics for males and females separately. Female passengers are, on average, older than male passengers, approximately as likely to be married, and very slightly less likely to be from southern Italy.

In columns (4)–(9), we restrict attention to the sample of individuals for whom we transcribed additional information. Column (4) presents the information for all transcribed individuals, while column (5) presents it for females and column (6) presents it for males. The already-digitized information for each group is similar to that for the untranscribed sample. Based on our transcription, we classified any passenger listing any person whom they would be joining in the United States as having some connection, and any individual who reported joining an immediate family member in the United States (i.e., a sibling, parent, child, or spouse) as having an immediate family connection in the United States. Over 95 percent of male and female migrants report that they have some connection ("Any Conn."), but males were far less likely to report that this connection was an immediate family member (i.e., a parent, sibling, child, or spouse), with only 32 percent falling in this group ("Imm. Fam. Conn."= 1) as compared to nearly 74 percent of women. Similar differences are apparent in the fraction of men and women reporting having been in the United States before ("Repeater"). More than 40 percent of men reported that they had been in the United States before, compared to only 16 percent of women. A gender difference is also apparent in whether the passenger had paid for himself, with 90 percent of men paying their own passage and only 66 percent of women doing so. Moreover, an unusual relationship exists between the heights of men and women, the distributions of which are presented graphically in Figure 11. In particular, female passengers were much taller relative to male passengers than is commonly the case in modern populations (Gaulin and Boster, 1985). This relationship is discussed in more detail in Appendix B. As discussed in this Appendix, however, we find no reason to believe, based on this relationship, that there are systematic issues of accuracy in our data. However, given

---

[27]Comparisons to the age 20 distributions of province and birth cohort height indicate that those migrating at the ages of 20 and 21 are negatively self-selected.

that we do not have data on the distributions of stature for women in Italy, we exclude women from our analysis.

In column (7), we eliminate from the sample any passenger who indicated that he had been in the United States before. We make this restriction primarily for two reasons. First, the process of self-selection into return migration is not well understood (though it has recently received scholarly attention: Abramitzky, Boustan, and Eriksson, 2012, 2014; Bandiera, Rasul, and Viarengo, 2013; Crimmins et al., 2005; Ward, 2013). Distinguishing between first-time and return migration prevents our sample from being contaminated by some other form of self-selection (i.e., into return migration) and prevents us from counting the same passengers more than once. Second, these passengers may have arrived in the United States before completing their physical growth, and would thus have grown differently (Boas, 1911, 1920; Gravlee, Bernard, and Leonard, 2003; Kress, 2007; Sparks and Jantz, 2002, 2003). Therefore, our benchmark sample will be that summarized in column (7)—males aged 22–65 who reported never having been in the United States. The remaining migrants after this deletion are younger than the repeat passengers, less likely to be married or to have an immediate family connection, and very slightly shorter. All of the results discussed below are stronger when these repeat passengers are included.

Next, we summarize the geographic distribution of male heights graphically. Figure 12 presents the average male heights of each province (based on A'Hearn, Peracchi, and Vecchi, 2009 and A'Hearn and Vecchi, 2011) weighted by our passenger counts, as well as the average heights of male passengers in our sample from each province. The average military heights exhibit a strong pattern, with the tallest provinces in the north, the shortest in the south, and the middling provinces in the center. We see a similar trend in the heights of migrants, with the tallest originating in the north, and the shortest in the south. Column (7) of table 1 also shows that the average heights of male migrants in our sample was 163.80 cm.[28]

Finally, we study separately two samples that allow us to break down the analysis by time period. As is evident from Figure 8, World War I was a massive disruption to trans-Atlantic migration. Moreover, in 1917, the United States enacted the literacy test, requiring that all adults entering the United States demonstrate literacy. Both of these events fundamentally changed international migration, and there is reason to believe that post-1917 passengers might be substantially different from pre-1917 passengers, and that pre-War migrants may have differed from the post-War migrants. We therefore split the sample into pre-1917 (exclusive) and post-1917 (inclusive) subsamples, which are summarized in columns (8) and (9),

---

[28]By contrast, the average American soldier (who may have been negatively self-selected) born in the 1860s was 171 cm tall (Zehetmayer, 2011, Figure 1, p. 320).

respectively, of Table 1.[29] Most striking is the large difference in stature between the two periods: passengers in the post-1917 are more than one centimeter taller on average. They are also more than eight percentage points more likely to have an immediate family connection, six percentage points less likely to be married, and four percentage points more likely to have paid for their own passage as compared to those arriving before 1917.

3.3   Representativeness of the Geolocated Sample

Before beginning our primary analysis, we examine whether our geolocation algorithm has produced a representative sample for our analysis. First, we estimate a number of regressions of the form

$$y_i = \beta_0 + \beta_1 G_i + \varepsilon_i,$$

where y_i is some individual characteristic of interest, and G_i is an indicator equal to one if individual i is successfully matched to a province by our algorithm, and zero otherwise. The coefficient β_1 tests whether there is a difference in the mean of each characteristic between the geolocated and non-geolocated groups. Table 2 presents estimates of β_1 for a variety of individual characteristics of interest and for a variety of samples. One division of the data is based on the recorded ethnicity of passengers traveling from Italy. Beginning in 1903, the passenger manifests were required to include the ethnicity ("Race or People") of immigrants (Perlmann, 2001; Weil, 2000). North Italians and south Italians were officially considered to be two separate ethnicities. The instructions for clerks completing the passenger manifests placed the dividing line between north and south Italy at the southern extreme of the basin of the River Po.[30] Nevertheless, compliance with the official definitions of the ethnicities appears to have been lax, and some passengers were still recorded as simply Italian, without further disaggregation. We refer to these passengers as "General Italians." Figure 13 also depicts the division of Italy into North and South by the Bureau of Immigration and Naturalization. This field provides information on the probable geographic origin of individuals independent

---

[29]We could also split the sample into pre-1914 (inclusive) and post-1919 (inclusive) subsamples in order to omit the World War I years, but given the fall in the quantity of migration during this period, there is essentially no difference between this approach and ours.

[30]Specifically, the manifest defined north Italians as

> [t]he people who are native to the basin of the River Po in northern Italy (i.e., compartments of Piedmont, Lombardy, Venetia, and Emilia) and their descendants, whether residing in Italy, Switzerland, Austria-Hungary, or any other country . . . . Most of these people speak a Gallic dialect of the Italian language.

South Italians were defined as

> [t]he people who are native to that portion of Italy south of the basin of the River Po (i.e., compartments of Liguria, Tuscany, the Marches, Umbria, Rome, the Abruzzi and Molise, Campania, Apulia, Basilicata, Calabria, Sicily, and Sardinia) and their descendants . . . .

of our algorithm. We use it in Appendix A in order to test the accuracy of our algorithm, and here in order to determine if our algorithm led to imbalances between matched and unmatched individuals within geographic regions.

We begin in column (1) of Table 2 by studying all males aged 22–65 in our 1907–1925 group of passengers. Our analysis indicates that south Italians are significantly overrepresented in our geolocated sample (66.5 percent of the geolocated sample, as opposed to 58.6 percent of the non-geolocated sample) while north Italians are slightly underrepresented and general Italians are significantly underrepresented. This under-representation of general Italians is likely due to the fact that Italians traveling through non-Italian ports were less likely to be assigned a north/south ethnicity, and their locations are likely to have been recorded with less accuracy due to a lack of familiarity by clerks in other countries with Italian geography and spelling.[31] We also find that those in our geolocated sample are on average 0.08 years younger than the non-geolocated. In addition, the average birth year is 0.35 years later. There is no statistically significant difference in marriage rates between groups.

As we are interested in observing differential self-selection patterns across provinces, we must verify that the sample is balanced at the provincial level, which can be approximated (without using our geolocation algorithm) by ethnicity. In columns (2)–(4) of Table 2, we break down the the sample used in column (1) by the ethnicity of migrants. There exist statistically significant differences in the probability of being married between the geolocated and non-geolocated groups for each ethnicity, but these are small, likely reflecting the large sample sizes as much as any actual differences. Similarly, differences in age and birth year exist, but are also small. Moreover, differences in age and birth year are not particularly troubling, as all of our analyses condition on birth year by comparing the height data of migrants to the averages of their birth cohort, and height is constant with age in our range.

Next, we perform the same exercises restricting attention to the transcribed sample. Columns (5)–(8) report the difference in the means of various characteristics between transcribed males aged 22–65 who were matched to a province by our algorithm, and those who were not matched. Statistically and economically significant differences persist in matching rates across ethnicities. However, there are no statistically significant differences in age, birth year, or marital status between groups, even within ethnicities. We also compare matched and unmatched individuals on the basis of transcribed information. We find no statistically significant differences between matched and unmatched individuals on the basis of transcribed data, except among northerners with respect to the measures of social network status and whether the passengers

---

[31]For example, the modal departure port for ships characterizing all Italian passengers as simply Italian was Cherbourg, France, while the modal departure port for ships decomposing all Italian passengers by ethnicity was Naples.

paid for their own passage. In particular, matched individuals are four percentage points less likely to have any connection in the United States, nine percentage points less likely to have an immediate family connection, and four percentage points more likely to have paid for their own passage than unmatched individuals. These differences have no implications for our main results (which are based only on height). They could potentially have implications for our results regarding the mechanisms driving self-selection.

The fact that our main results are based on stature makes it particularly important to test the representativeness of our sample on this front. In particular, there are two potential dangers that we face in terms of the balancedness of our sample. First, as we compare the heights of migrants to those of the Italian population, it is important to ensure that our sample of migrants is representative of all migrants, rather than being biased upward or downward in height by our algorithm. Although not statistically significant, Table 2 shows that matched north Italians were 0.375 centimeters taller than their unmatched fellows, while south Italians were 0.009 centimeters shorter. If these difference represent a small but non-spurious bias, they would bias our estimates against our baseline results of more positive self-selection in the south. Figure 14 shows this result in greater detail: among passengers of all ethnicities, the probability of being matched is all but constant over height. Within the separate ethnicities, there is more noise among the rare heights, but essentially the same conclusion follows.

Second, since we test whether, within a province, migrants are taller or shorter than their source population, it is important to ensure that, conditional on the province of origin and its mean height, our matched sample is not taller or shorter than the unmatched group. Finally, since we test whether the trends in self-selection differ across cohorts and provinces or different mean heights, we must ensure that there are no differences in the differences between the heights of the matched and unmatched individuals across cohorts of different average stature. However, determining the province to which an individual belongs requires a successful match, which we do not have for the matched individuals. We must therefore find some way of associating unmatched individuals with provinces and birth cohorts independently of the geolocation algorithm.

To this end, we use the following procedure, which takes advantage of the fact that Italian surnames are useful indicators of geographic origins (Guglielmino and De Silvestri, 1995). First, for each surname, we determine the modal province to which individuals with that surname who could be geolocated were assigned. Then, for the purposes of this exercise only, we assign all individuals to the modal province for their surname. We then use their (known) birth year to assign them to a province and birth cohort, from which a mean height for each individual's province and birth cohort is determined. The rationale behind

this exercise is the following: family names can be used to group migrants into bins that are clustered across space. If geolocated individuals are randomly drawn from within each province and birth cohort, then we expect that the height distributions of matched and unmatched passengers would be the same within each surname bin. Mapping passengers to a province predicted by their surnames brings us as close as possible to comparing height distributions within provinces and birth cohorts and makes it possible to test whether the matched-unmatched gap changes systematically with the height of the province and birth cohort.

We first use the results of this procedure to estimate the regression equation

$$z_{ijt} = \beta_0 + \beta_1 \mu_{jt} + \beta_2 G_{ijt} + \beta_3 G_{ijt} \mu_{jt} + \varepsilon_{ijt}, \tag{1}$$

where z_{ijt} represents the standardized (by the surname-implied province-birth cohort mean and standard deviation) height of individual i, who is matched (by this procedure) to province j and birth cohort t, with mean height μ_{jt} (normalized to have mean zero), and where G_{ijt} is an indicator equal to one if individual i was matched to a province by our geolocation algorithm and to zero otherwise. We present the results of this regression in Table 3. There are two coefficients of interest. First, β_2 indicates whether there is, at the average, a systematic difference in standardized heights between the geolocated and the non-geolocated. While we find that the geolocated individuals are, on average, taller than the non-geolocated conditional on the mean height of their place of origin, this difference is statistically insignificant. This difference is somewhat concerning, as it would tend to spuriously generate our findings in section 4.3. However, it must be kept in mind that the unmatched comprise less than 15 percent of migrants, which would make any sample selection biases induced by our algorithm small. Second, β_3 indicates whether there is a systematic difference in the difference between the matched and unmatched groups between provinces and cohorts of different average heights. We find that this coefficient is positive, but that again, it is not statistically significant. Moreover, the positive sign works against our results in section 4.4. Thus, the true differential patterns are stronger than those that we measure.

We also seek to verify that these findings are not driven by the linearity assumptions of equation (1). We therefore regress non-parametrically, in Figure 15a, z_{ijt} on μ_{jt} for each of the two groups of G_{ijt}, and present an estimate of the difference between the two curves in Figure 15b.[32] Throughout the range of μ_{jt}, the confidence band includes zero. Moreover, the nature of the difference, as above, is such that it would work against our differential results, except at the upper extreme of mean heights, at which the data are

---

[32]The confidence bands in the graph are 95 percent point wise confidence intervals. We thank Anand Krishnamurthy for helpful discussions on this topic.

sparse. Thus, on the whole, these balancing tests do not show any compelling evidence of differences in heights or differential differences in heights between the matched and unmatched groups, nor compelling reason to believe that our results will be driven by imbalances in our geolocation algorithm, However, the point estimates show that we cannot rule out sample-selection bias entirely, and we will thus discuss the potential consequences of sample-selection bias below.

4   The Nature and Degree of Self-Selection

We are now equipped to begin examining the nature and degree of self-selection of Italian migrants. We first lay out a formal framework for our analysis. We then study migrants as compared to all of Italy before disaggregating the analysis to compare these migrants to their provinces of origin. We then study geographic and temporal trends in self-selection. Finally, we attempt to quantitatively translate our findings of self-selection with respect to stature into a measure of self-selection with respect to living standards.

4.1   Framework for Analysis

Before beginning our analysis, we lay out a theoretical framework to govern it. Let h_{ijt} denote the height of individual i from province j and birth cohort t. Suppose that $h_{ijt} \sim F_{jt}$, where F_{jt} is a distribution with mean μ_{jt} and variance σ_{jt}^2. Let G_t denote the mixture of F_{jt} for each t (weighted by population) and let its mean be μ_t and its variance σ_t^2. The economic history literature typically assumes that the F_{jt} and G_t are normal (c.f., A'Hearn, Peracchi, and Vecchi, 2009; Bodenhorn, Guinnane, and Mroz, 2013), though this assumption is not necessary.[33] Except in special circumstances to be specifically noted below, our analysis will not depend on an assumption of normality.[34] Instead, our analysis will be based on two statistics. The

---

[33]Moreover, this assumption need not hold. First, even if the F_{jt} are normal, their mixture, G_t, need not be. Moreover, there is evidence (e.g., A'Hearn, Peracchi, and Vecchi, 2009) that the distribution of height may not be normal.

[34]The raw data used by A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) make it possible to identify the actual distribution of Italian heights. However, we have only the means and standard deviations of these distributions, and thus limit our analysis.

first is height normalized by the all-Italy mean and standard deviation of the birth cohort of origin:[35]

$$z_{it} = \frac{h_{it} - \mu_t}{\sigma_t}.$$

The second is height normalized by the mean and standard deviation of the province and birth cohort of origin:

$$z_{ijt} = \frac{h_{ijt} - \mu_{jt}}{\sigma_{jt}}.$$

When these distributions are assumed to be normal, both statistics are distributed $N(0,1)$; but regardless of the assumptions on F_{jt} and G_t, both of these statistics have mean zero and variance of one by construction. Thus, testing whether the z_{it} and z_{ijt} in our data have mean zero is informative regarding whether or not there was self-selection of migrants, regardless of the form of F_{jt} and G_t.

4.2 Migrant Self-Selection Across Provinces

As a first examination of self-selection into migration on the basis of stature, we study the correlation between migration and average provincial heights. Figures 16a and 16b plot the number and share of migrants, respectively, from each province against the average height of that province. The average height is based on our elaborations on the A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) calculations, with means weighted across birth cohorts within provinces by the migrant counts in our data in order to make them comparable to the migrant heights, which are composed similarly. A strong negative relationship is evident in both panels. The shorter, mostly southern, provinces send both a larger absolute number of migrants and a larger share of their populations to the United States. Thus, based on provincial comparisons, one would conclude that Italian immigrants were negatively self-selected.

Figure 17a provides additional support for this conclusion. Imposing the assumption that the G_t are normal, we perform a Kolmogorov-Smirnov test of the z_{it} against a theoretical $N(0,1)$ distribution.[36] We

---

[35]We received from A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) only the moments for each province, not for the country as a whole. We therefore computed the μ_t and σ_t by weighting across distributions by 1901 population. Let N_j denote the population of province j and N denote the population of all Italy. We computed the moments as follows:

$$\mu_t = \sum_j \frac{N_j}{N} \mu_{jt}$$

and

$$\sigma_t = \left[\left(\sum_j \frac{N_j}{N} \mu_{jt}^2 \right) - \mu_t^2 \right]^{\frac{1}{2}}.$$

[36]A slight modification to the data is required in order to perform this test. The heights in the Ellis Island manifests were reported in whole inches, leading to discreteness in the distribution of heights. Comparing this discrete distribution to

29

reject the null of a $N(0,1)$ distribution, indicating some kind of self-selection if the G_t are indeed normal.

In Table 4, we remove the normality assumption and analyze migrant heights, normalized by birth cohort means and standard deviations by regression. Column (1) presents statistically significant evidence of negative self-selection for all of the samples spanning the entire study period. Specifically, the constant of the regression, which represents the mean of the z_{it}, is negative and statistically significantly different from zero, indicating negative self-selection at the national level. The sign of this self-selection is consistent with the potential bias discussed in the balancing tests; however, at 0.1 standard deviations of height (which corresponds to approximately 0.71 centimeters, or 36.7 percent of the inter-province standard deviation of average height), the self-selection is clearly too strong to be caused by our algorithm. Notably, this result corresponds to other findings with respect to Italian migration discussed in section 2.3 above, particularly those of Stolz and Baten (2012).

4.3   Self-Selection Within Provinces and Birth Cohorts

We then turn our attention to height normalized by the means and standard deviations of each individual's province and birth cohort of origin—z_{ijt}. First, we assume that the F_{jt} are normal and compare the distribution of z_{ijt} to a hypothetical $N(0,1)$ distribution.[37] The results of this comparison are depicted in Figure 17b, and the Kolmogorov-Smirnov test rejects the null of a $N(0,1)$ distribution. Next, we relax our normality assumption and analyze the z_{ijt} by regression, presenting the results in Table 5. In column (1), we regress z_{ijt} on a constant. Our results reveal that Italian migrants were on average 0.03 standard deviations of height taller than their province and birth cohort means, and that this difference is statistically significant. While smaller in magnitude than the negative national-level self-selection results discussed above, this result is non-negligible in size. In particular, 0.03 standard deviations of height corresponds to roughly 0.186 cm, or approximately 9.6 percent of the inter-province standard deviation of average height. Moreover, due to the fact that the unmatched group is small relative to the matched group, this value is approximately 6.67 times larger than what would cause concern over it being caused by sample selection bias from our algorithm.[38] Thus, changing the reference group from all of Italy to the migrants' environment of origin reverses the sign

the continuous standard normal distribution could lead to rejection of the null of a standard normal distribution even if the underlying distribution of heights is normal. Thus, for the purpose of this test, and its counterpart in section 4.3 only, we add uniformly distributed random noise with support of $[-0.5\,\text{in}, 0.5\,\text{in}]$ to the observed height, in order to account for the possibility of rounding to the nearest inch.

[37] Again, smoothing of our data through the addition of random noise is necessary to account for rounding.

[38] Let $\bar{z}_g$ denote the average standardized height of the geolocated sample (our estimate in column (1) of Table 5). Let $\bar{z}_u$ denote the average standardized height of the unmatched individuals. Moreover, the geolocated comprise approximately 85 percent of the entire sample. The average standardized height of all migrants is $\bar{z} = 0.85\bar{z}_g + 0.15\bar{z}_u$. For the true self-selection to be zero, it must be that $\bar{z}_g - \bar{z}_u = \frac{\bar{z}_g}{.15} = 6.67\bar{z}_g = 0.2$, whereas we have found (through the coefficient on the geolocation indicator in Table 3 that $\bar{z}_g - \bar{z}_u = 0.03$.

of the self-selection. This geographic decomposition, which is enabled by our geolocation algorithm, reveals results contrary to those of Stolz and Baten (2012).

4.4   Differential Self-Selection Across Origin Distributions

We next investigate the causes of the difference between the results in column (1) of Table 5 on the one hand, and column (1) of Table 4 and Figure 16 on the other, namely the finding of positive self-selection on the local level and negative self-selection on the national level. First, we study differences in passengers' stature and their self-selection across Italy. Column (2) of Table 4 shows that the negative national-level self-selection result is driven by southerners:[39] northerners are positively self-selected on average, relative to the national mean, while the more numerous southerners are negatively self-selected relative to the national mean, driving the aggregate results. Column (2) of Table 5, on the other hand, in which the dependent variable is z_{ijt}, shows precisely the opposite result, with southerners positively self-selected and northerners negatively self-selected, and migrants as a whole following the southerners. These results suggest that our findings of negative self-selection in section 4.2 are primarily a composition effect, driven by the provincial origins of our sample. Migrants were primarily from shorter southern provinces, but relatively tall within them. To be sure, the fact that the bulk of the Italian migrants came from poorer regions had some impact on the post-migration outcomes of migrants, but the fact that they are positively self-selected on the local level would also be important. We discuss this formally in section 5.1.

Figure 7 demonstrates this relationship graphically by plotting the average height of migrants from each province against the means of the population distribution for each province, weighted over birth cohorts as in Figure 16. We find that the relationship between the average stature of a province and the average stature of migrants is best linearly approximated by a line with slope statistically significantly less than one, which would be the slope if the degree of self-selection were identical across provinces. Moreover, this line is located such that migrants from shorter provinces are, on average, taller than their origins.

In order to ensure that our results are not driven by our aggregation across birth cohorts within a province and the equal weight given to each province in Figure 7, we repeat its analysis on an individual level, each individual's province-birth cohort-specific mean stature. These results are presented in column (3) of Table 5, and are quite similar to those discussed above, with the coefficient on comparison height statistically significantly less than one.

Finally, in order to explore possible non-linearities in this relationship, we fit a local linear regression to

---

[39] These are individuals matched to provinces defined as southern by the Bureau of Immigration and Naturalization, rather than individuals identifying as south Italian in the ethnicity field.

the individual height data and present the results in Figure 7b. Overall, we continue to see a trend of positive self-selection in shorter province-cohorts. When non-linearities are introduced, however, the relationship in the shortest province-cohorts is much starker, with the line of fit nearly horizontal. The trend is less evident among taller origins, where the line of fit nearly tracks the 45-degree line, which is often included in the confidence interval. In the tallest provinces and birth cohorts the line of fit is well below the 45-degree line, indicating negative self-selection.

The analysis of the relationship between the magnitude of self-selection and the average stature of the province of origin has heretofore been based on differences in actual height. Our analysis in sections 4.2 and 4.3, however, was based on standardized heights. The results of each analysis need not be the same due to heteroskedasticity across the stature distributions of different provinces and birth cohorts of origin.[40] We therefore repeat the above analysis used standardized, rather than actual height. Figure 18 is analogous to Figure 7, replacing the heights of migrants by their standardized heights, z_{ijt}. The relationship between average standardized height and provincial average height is negative. Moreover, the differential pattern of positive self-selection in shorter, primarily southern provinces and negative self-selection in taller northern provinces persists. Column (4) of Table 5 analyzes the same relationship on the individual level and again confirms that our results are not driven by aggregation of birth cohorts within provinces, or by weighting provinces equally. Finally, Figure 18b presents the analysis allowing for non-linearities. Again, we see strong positive self-selection in shorter origin groups and negative self-selection among taller groups. On the whole, the results with standardized height confirm those with unstandardized height.

4.5   Temporal Trends in Self-Selection

We also separate our analysis into two periods based on whether an individual arrived in the United States before or after the imposition of the literacy test in 1917. First, in Figure 19, we plot kernel density distributions of the z_{ijt} for the pre-1917 and post-1917 periods. We also perform Kolmogorov-Smirnov tests for whether the pre-1917 and post-1917 distributions differ from one another for the entire sample, northerners, and southerners.[41] For the southerners and the full sample, the Kolmogorov-Smirnov tests reject the null of the equality of distributions; for northerners, however, we fail to reject the null.[42] Next, in columns (3) and (4) of Table 4, we test whether there was a change in the nature of self-selection, as measured

---

[40]In particular, the standard deviation of heights was greater in the shorter province-cohorts, such that a given degree of self-selection in terms of standard deviations would translate to a greater degree of self-selection in terms of centimeters compared to provinces with a more egalitarian distribution of heights.

[41]Unlike above, these Kolmogorov-Smirnov tests do not require that we make a normality assumption or adjust the data to account for rounding.

[42]When smoothed moments are used, we can reject the null.

by z_{it}, after 1917. Column (3) shows that the self-selection is negative and statistically significant prior to 1917, and increases significantly after 1917, such that it is negative but small and statistically insignificant after 1917. In column (4) of Table 4, we examine whether the changes in the magnitude of self-selection over time differed between northerners and southerners. This is indeed what we find: the insignificant coefficient on the post-1917 variable indicates that northerners had no statistically significant increase in the degree of self-selection. The sum of the coefficients of the post-1917 variable and its interaction with southern is positive and statistically significant, indicating an increase in the magnitude of self-selection among southerners after 1917 of 0.128 standard deviations.

In columns (5) and (6) of Table 5, we repeat this analysis with z_{ijt} as the dependent variable. The results of column (5) indicate that the pre-1917 migrants from all of Italy were not, on average, statistically significantly different from their populations or origin, with the constant taking a value of only 0.004. After 1917, however, the magnitude of self-selection is very large, at 0.129 standard deviations of height, or roughly 0.8 cm or 43 percent of the inter-province standard deviation of height. However, the lack of evidence of any self-selection prior to 1917 for Italy as a whole is somewhat misleading. As shown in column (6), the constant is negative and statistically significant, indicating that pre-1917 northerners were negatively self-selected. Moreover, the sum of the estimated constant and the coefficient on southern is 0.026. While not statistically significant, the magnitude of this difference is comparable to that of the overall result in column (1). As in Table 4, we also find a statistically insignificant increase in the standardized heights of northerners (the post 1917 variable) and a statistically significant increase of 0.14 standard deviations in the standardized heights of southerners (the sum of the post-1917 and post-1917 × southern coefficients).

Next, in columns (7) and (8), we study whether there was a change in the differential self-selection after 1917. These columns repeat columns (3) and (4), respectively, but add an indicator for post 1917, as well as its interaction with average height. The coefficients on average height indicate that there was statistically significant differential self-selection prior to 1917, with the coefficient in column (7) statistically significantly less than one, and the coefficient in column (8) statistically significantly less than zero. The interaction with post-1917 in each of these columns is negative, indicating that the magnitude of differential self-selection increased after 1917, but that the difference is not statistically significant.

These findings potentially have an effect on the interpretation and generalizability of our results. One possibility is that the increase in the degree of positive self-selection after 1917 was caused by the new restrictions imposed by the Immigration Act of 1917, particularly the literacy test. If this were the case, then the post-1917 results may not be representative of the self-selection that would occur in an unrestricted

flow of migration. Instead, it would be more appropriate to focus on the pre-1917 migrants only. However, Goldin (1994) argues that the effects of the literacy test were likely small given the increase in primary schooling in southern and eastern Europe in the late 19th century.[43] In this case, the changes over time could have been caused by the effects of World War I (e.g., large scale mortality). In order to differentiate these effects, we test whether the change in the magnitude of self-selection from a province after 1917 depended on its level of literacy, which we glean from the Ministero di Agricoltura, Industria e Commercio (1915). The coefficient on the interaction of post-1917 and the male literacy rate, while not statistically significant, is large, at nearly 50 percent of the coefficient on the male literacy rate. Its negative coefficient indicates that the increase in the degree of self-selection was smaller in more literate provinces, suggesting that the change in the magnitude of self-selection after 1917 may have been caused by the implementation of the literacy test. However, a definitive conclusion cannot be drawn. The correlation between province-birth cohort average height and the literacy rate is 0.803, and we observe only one figure of literacy for each province, as opposed to one for each province and birth cohort for height. Thus, while we find evidence consistent with the change in migration being a result of the literacy test, we cannot make a definitive conclusion to this effect; in particular, we cannot rule out the potential explanation that migration patterns were altered by changes in the labor market conditions in Europe after World War I.

4.6 Lifecycle Trends in Self-Selection

Finally, we study different patterns of self-selection over the lifespan by examining the change in average standardized stature with age. In Figure 20, we plot a local linear regression of average standardized height against age for men of ages 22–45.[44] Among the complete sample, and also when northerners and southerners are analyzed separately, striking patterns appear. First, migrants under roughly age 25 are consistently negatively self-selected. Second, in the south, a rise in average standardized height with age is evident until the late 30s. Even in northern provinces, where negative self-selection appears to dominate in the 20s, an increase in standardized heights is evident in this age range.

What self-selection mechanisms could be behind these results with respect to age? We show in section 6 that our results are consistent with theories of positive self-selection induced by liquidity constraints to migration. However, if this mechanism is the only one operating to generate self-selection into migration, we would expect age to have the opposite effect; that is, we would expect the degree of self-selection to decrease

[43]Regardless of any improvement, illiteracy rates in some southern Italian provinces still exceeded 50 percent in 1911.

[44]This comparison is not contaminated by changes in average stature over time, as the heights are standardized by the province and cohort distributions. We omit ages 46–65 due to the very large confidence bands around the estimates that make the pre-age 40 trends difficult to discern.

with age, due to the ability of individuals to save over time, such that those of higher ability would be able to finance migration earlier than those of lower ability. Instead, it could be that this result is a reflection of the shorter horizon of the benefits of migration to be enjoyed by older individuals. Older migrants would have fewer years in which to enjoy the gains from migration than younger migrants, thus reducing the net benefit of migration at the time that the choice would have been made. Thus, among older individuals, only the most able, who would be most successful in the new country, would find it beneficial to migrate. Such an explanation would account both for the positive slopes in Figure 20 and the change in differential selection as shown in Table 5.

5   Interpretation of Results

The previous analysis has documented the nature of migrant self-selection on the basis of stature from within each Italian province, and from the Italian nation as a whole. In this section, we discuss the relevance and implications of these results for the understanding of migrant self-selection in general. We also interpret our results in terms of the marginal effects of stature and of migrant quality on the probability of migration.

5.1   Does Within-Province Self-Selection Matter?

Our baseline result is that migrants (mainly from south Italy) were, on average, positively self-selected from within their provinces of origin, but negatively self-selected from Italy as a whole. These results raise an issue of interpretation: conditional on knowing the degree of self-selection of a migrant from a country as a whole, does knowing his degree of local self-selection matter? If the researcher wishes to understand the effects of emigration from a particular sub-national unit on its economy, it is clear that migrants must be compared to the local (sub-national) population of origin. In so doing, it is possible to determine if, for instance, poor economic performance can be explained by a "brain drain" (Mokyr, 1983; Mokyr and Ó Gráda, 1982, 1984).[45] If the researcher is interested in predicting the labor market performance of immigrants in the receiving country and their effects on the host economy (e.g., Borjas, 1987, 1991; Chiquiar and Hanson, 2005), the answer is less clear. In the following discussion we provide a simplified theoretical setting and a condition under which the local degree of self-selection is positively correlated with future outcomes, conditional on the immigrant's observed national degree of self-selection. We suspect that overlooking self-selection within sub-national regions—as many previous authors have done (e.g., Borjas, 1987, 1991)—ignores potentially

---

[45]In Italy, however, there appears to have been a fall in the skill premium in this period, suggesting that emigrants were primarily unskilled (Betrán and Pons, 2004).

valuable information and may lead to inaccurate predictions of immigrant performance.[46]

Consider immigrant i arriving from province j and birth cohort t. Let his height be determined by a production function $h_{ijt} = h(\mu_{jt}, z_{ijt})$, where μ_{jt} is the contribution of the local environment of province j, and z_{ijt} is the contribution of i's individual quality. We abstract from noise of genetic and other sources, meaning that conditional on the effect of the local environment and individual quality, height is deterministic, but the same conclusions would follow if such noise is taken into account. In a very simple case in which $h_{ijt} = \mu_{jt} + \sigma z_{ijt}$ and z_{ijt} has mean zero and variance one (with σ being the within-province and birth cohort standard deviation of height), μ_{jt} is the average height in province j and birth cohort t, the national degree of self-selection is represented simply by the height (after subtracting the average national height) $\tilde{h}_{ijt} = h_{ijt} - \mu_t$, and the local degree of self-selection is represented by z_{ijt}. Let the immigrant's outcome w_{ijt} (standing for wage, productivity, or any other measure of value for the host economy) be determined by a deterministic production function of the same two inputs, local environment and individual quality: $w_{ijt} = w(\mu_{jt}, z_{ijt})$.

The researcher, or the policy maker, is interested in predicting an immigrant's outcome. One straightforward way to tell whether knowing the degree of local self-selection is informative above and beyond the information contained in the national degree of self-selection, is to characterize the function $\hat{w}$ that predicts the outcome w_{ijt} conditional on the two variables of self-selection: $\hat{w}_{ijt} = \hat{w}(\tilde{h}_{ijt}, z_{ijt})$. In particular, conditional on observing the immigrant's demeaned height $\tilde{h}_{ijt}$, what are the conditions under which the predicted outcome is increasing with respect to the relative height within the province z_{ijt}? The answer is that this positive relation holds under a general condition, which we argue would prevail in reasonable circumstances. In particular, denote the marginal rate of technical substitution of the height function by $\text{MRTS}^h(\mu, z) = \frac{\partial h(\mu,z)}{\partial \mu} / \frac{\partial h(\mu,z)}{\partial z}$, and similarly denote $\text{MRTS}^w(\mu, z)$. Then $\frac{\partial \hat{w}(\tilde{h}, z)}{\partial z} > 0$ if and only if

$$\text{MRTS}^h(\mu, z) > \text{MRTS}^w(\mu, z) \tag{2}$$

(see proof in Appendix C).

The intuitive interpretation of this condition is that the inputs of the local environment are relatively more important, compared to the individual quality, in affecting an output that is produced within this environment (such as height) than in affecting an output that is produced outside this environment (such as outcomes in the host country). While there may be cases to the contrary, we believe that this condition

---

[46]The opposite view, that relative ranking within a sub-group is of little significance, is echoed in a famous Mishnah proverb, attributed to Matteya ben Heresh: "...be the tail among lions rather than the head among foxes" (Avot, 4:15).

is likely to be satisfied in most standard cases. Thus, it is important to study migrant self-selection within provinces.

The empirical testing of this question is the scope of another project, in which we link migrants from Ellis Island passenger manifests to the US Census and compare the power of local and national self-selection in predicting post-immigration labor market outcomes. We are not familiar with any other studies that directly test the predictive power of local selection on top of the information provided by national selection in the context of migration. However, a case study of interest is provided by Conley and Onder (2013), who collected data on the output of graduates of PhD programs in economics over the six years after their graduation. They show that output is strongly increasing with ranking of students within the departments, to the extent that it quickly overrides the ranking of the departments themselves. For example, in five out of the ten programs ranked 21-30 in the US, the 90th percentile of graduates' output is greater than that of the 70th percentile of output in any top program (Conley and Onder, 2013, Table 1). This question is also relevant to debates in the economics of education, which compare the effectiveness of class rank and SAT scores in predicting success in college (Niu and Tienda, 2010; Pike and Saupe, 2002; Rothstein, 2004).

5.2 Marginal Effects of Height and Living Standards on Migration Probabilities

Ultimately, the goal of the results discussed above is to make inference regarding the direction and magnitude of the relationship between the probability of migration and living standards by observing the correlation between the probability of migration and stature, a correlate of standards of living. To this end, we propose the following simple model to be used in the interpretation of our results.

Suppose that the height of individual i from province j and birth cohort t is determined by a combination of standards of living and genetics. Let z_{ijt} denote the standardized height of potential migrant i from province j and birth cohort t. Suppose that

$$z_{ijt} = \alpha_{ijt} + \varepsilon_{ijt},$$

where α_{ijt} denotes the portion of height that is correlated with living standards,[47] and ε_{ijt} denotes the remaining component of height that is simply random genetic noise that is uncorrelated with living standards. We assume that α_{ijt} and ε_{ijt} are independent of on another and each have mean zero. Moreover, let ξ^2

---

[47]This may include some of the genetic variation in height to the extent that height is a signal of quality in the labor market and may lead to higher wages (Lundborg, Nystedt, and Rooth, 2009; Persico, Postlewaite, and Silverman, 2004; Schultz, 2002), or to the extent that physically larger individuals make better laborers.

denote the variance of α_{ijt} and let ψ^2 denote the variance of ε_{ijt}. Since z_{ijt} is a z-score, and due to our independence assumption, it must be the case that $\xi^2 + \psi^2 = 1$.

Furthermore, let $y_{ijt} = \mathbf{1}\{i \text{ migrates}\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function. We assume a linear probability model, according to which

$$P(y_{ijt} = 1 | \alpha_{ijt}, \varepsilon_{ijt}) = c + \delta\alpha_{ijt} + \gamma\varepsilon_{ijt} + \eta_{ijt}, \tag{3}$$

where η_{ijt} has conditional mean of 0 and c is a constant. This expression allows the probability of migration to depend on standards of living and on genetic variation of height. The coefficient on the former, δ, can be interpreted as the change in the probability of migration from a change in standards of living that is correlated with a one standard deviation increase in average height. The coefficient γ captures the extent to which taller individuals may be more or less likely to migrate simply because of their height.[48] Ultimately, the parameter of interest is δ. In principle, if we could observe α_{ijt}, estimation of equation (3) would be possible due to our assumption regarding the independence of α_{ijt} and ε_{ijt}.

Since we do not observe α_{ijt} and ε_{ijt} separately, we rewrite equation (3) using known information. In particular, we substitute $\alpha_{ijt} = z_{ijt} - \varepsilon_{ijt}$ into this equation, giving

$$P(y_{ijt} = 1 | z_{ijt}) = c + \delta(z_{ijt} - \varepsilon_{ijt}) + \gamma\varepsilon_{ijt} + \eta_{ijt}$$

$$P(y_{ijt} = 1 | z_{ijt}) = c + \delta z_{ijt} + [(\gamma - \delta)\varepsilon_{ijt} + \eta_{ijt}]. \tag{4}$$

The left-hand side of equation (4) can be calculated from our data using Bayes's theorem. In particular,

$$P(y_{ijt} = 1 | z_{ijt}) = \frac{\tau(z_{ijt} | y_{ijt} = 1) P(y_{ijt} = 1)}{\tau(z_{ijt})},$$

where the denominator of the right-hand side is taken from the normal distribution (assuming that the F_{jt} are normal), while the numerator can be learned from our data: $P(y_{ijt} = 1)$ is the fraction of the population that migrates,[49] while $\tau(z_{ijt} | y_{ijt} = 1)$ is the pdf of standardized height in our data. However, since we observe the standards of living component of height only with error (that is, we observe z_{ijt}, not α_{ijt}), there will be attenuation bias in OLS estimation of (4) due to the correlation between z_{ijt} and ε_{ijt}.[50] Using

---

[48]Differential labor market performance based on height would be included in α_{ijt}. Only differences in the propensity to migrate conditional on any correlates of living standards—for instance, if taller individuals found the cramped steerage compartments more uncomfortable than their shorter compatriots—would enter into equation (3) through ε_{ijt}.

[49]That is, $P(y_{ijt} = 1)$ is the ratio of the number of individuals passing through Ellis Island in our sample period to the total population of Italy in 1901.

[50]Note that the bias arises when we seek to learn the effects of living standards on migration probabilities. OLS estimation

standard arguments, it can be shown that

$$\text{plim}(\hat{\delta}) = (1 - \psi^2)\delta + \psi^2\gamma, \tag{5}$$

so that

$$\delta = \frac{\text{plim}(\hat{\delta}) - \gamma\psi^2}{1 - \psi^2}. \tag{6}$$

It is established in the biological height literature that genetics are responsible for approximately 80 percent of the variation in human height (Silventoinen, 2003); because we interpret ε_{ijt} as the portion of the genetic variation in height that is unrelated to living standards and because Silventoinen (2003) reports that genetics are relatively less important than environment in more deprived populations, we would expect ε_{ijt} to account for less than 80 percent of the variance in height, and we can therefore assume that $0 < \psi^2 \leq 0.8$. Under the assumption that $\gamma = 0$—that is, that individuals' propensity to migrate is affected by their height only insofar as it is correlated with their standard of living—equation (6) becomes

$$\delta = \frac{\text{plim}(\hat{\delta})}{1 - \psi^2}.$$

With our assumption on ψ^2, we then have that

$$\text{plim}(\hat{\delta}) < \delta \leq 5 \times \text{plim}(\hat{\delta}).$$

Table 6 presents the results of estimation of equation (4) for the complete sample, and for northerners and southerners separately, where $\tau(z_{ijt}|y_{ijt} = 1)$ is computed using a kernel density estimate.[51] These regressions show that a one standard deviation increase in standardized height is associated with approximately a 0.3 percentage point increase in the probability of migration (through Ellis Island) in our study period, as compared to an average probability of migration of approximately six percent in the same period throughout Italy. Under the assumption discussed above, that $\gamma = 0$, this result implies that an improvement in standards of living that would result in a one standard deviation increase in height is associated with anywhere between a 0.3 and 1.3 percentage point increase in the probability of migration for the complete sample, depending on the true value of ψ^2. For southerners, this range is 0.5 percentage points to 2.7 percentage points relative to an average emigration probability of approximately 8.5 percent. For north-

---

of equation (4) is informative about the effects of height in total on migration probability without issues of attenuation bias.

[51]Specifically, we estimate $\tau(z_{ijt}|y_{ijt} = 1)$ by computing a kernel density estimate of province-birth cohort-standardized height in our sample. The denominator is simply $\phi(z_{ijt})$, where $\phi(\cdot)$ is the standard normal density.

erners, the range is from -0.1 percentage points to -0.4 percentage points, relative to an average emigration probability of approximately two percent. Figure 21 presents the same relationship graphically using a local linear regression.

The interpretation of our estimate of δ is more complicated if we allow γ to be non-zero. As shown in equation (5), the estimate in this case is a weighted average of δ and γ. If $\gamma > 0$ and $\gamma \neq \delta$, then δ would be overestimated, with the magnitude of the bias depending on the value of ψ^2. Similarly, if $\gamma < 0$, then δ would be underestimated. In extreme cases, it is thus possible that the signs of $\text{plim}(\hat{\delta})$ and δ would differ. Therefore, in order to draw conclusions regarding the effects of living standards on migration using height data, it must be assumed that the role of the genetic component of height in the migration decision is not too large.

6   Mechanisms of Self-Selection

We also seek to evaluate the various theories of migrant self-selection presented in Section 2.2 above. The theory of Borjas (1987) and Roy (1951) hinges on the relative returns to skill in the sending and receiving economies, which are usually approximated by the relative inequality of the income distributions. Other theories predict positive self-selection into migration based on the need to overcome liquidity constraints in financing migration or due to the risk inherent in migration, though such self-selection may be moderated by network connections in the destination country (Borger, 2009; Bryan, Chowdhury, and Mobarak, 2014; Grogger and Hanson, 2011; McKenzie and Rapoport, 2010; Wegge, 1998). Belot and Hatton (2012) point out that unless this poverty constraint is taken into account, it is not possible to identify the Roy model effects on migrant self-selection, while Fernández-Huertas Moraga (2013) finds that a combination of the three is needed in order to fully explain observed patterns of migrant self-selection.

In order to perform a more formal analysis a number of additional variables are required. Testing the Borjas (1987) predictions requires some measure of inequality. Lacking any direct information on income inequality we instead use the coefficient of variation in stature as a measure of inequality. The value of this statistic in measuring inequality is evaluated favorably by Blum (2013) and it is also used by Stolz and Baten (2012) in order to evaluate the relative inequality hypothesis.[52] Specifically, we compute a separate

---

[52]Steckel (1995) also argues that the distribution of stature is informative regarding inequality.

coefficient of variation for each province j and birth cohort t according to

$$CV_{jt} = \frac{\sigma_{jt}}{\mu_{jt}} \times 100.$$

In order to account for the poverty constraint, we incorporate a number of additional variables. On the province and birth cohort level, we include the average height of the source population, which captures the level of development of the province-cohort.[53] We also include several individual-level measures of the poverty constraint, including whether the migrant paid for his own passage, and whether the migrant reported an immediate family connection in the United States; the latter is also indicative of migrants' chain migration status.

Table 7 presents a number of regressions of standardized height on combinations of these variables, always including a quadratic in age to address the selection patterns over the life cycle as discussed in section 4.6 above. In column (1), we regress the standardized heights of migrants on the coefficient of variation of stature in their respective province and birth cohort of origin. Greater inequality of the province of origin implies greater relative inequality between the province of origin and the United States, and, according to the relative inequality theory, should yield more negative self-selection. The coefficient on the measure of inequality is of the sign predicted by the relative inequality theory of Borjas (1987): greater inequality in the origin is associated with more negative self-selection; but the effect is not statistically significant and the magnitude is very small: the CV has a range of only approximately 3.5 in our sample; thus, between the lowest and highest CV in our data, the average standardized height decreases only by roughly .01 standard deviations of height. However, to the extent that liquidity constraints mask skill-based self-selection (Belot and Hatton, 2012), the results in column (1) should not necessarily be seen as refuting the relative inequality hypothesis.

In column (2), we add the demeaned average height of the province and birth cohort, finding that migrants from shorter, and thus more economically deprived, province-cohorts, are more strongly positively self-selected. This result replicates our previous results and supports the importance of liquidity constraints in determining self-selection: residents of shorter and poorer provinces are more likely to face binding constraints to financing their travel than residents of taller and wealthier provinces; these constraints would disproportionately affect lower quality migrants. In this regression, the coefficient on the CV enters insignificantly and positively, contrary to the predictions of the relative inequality model. We also account

[53]It is also possible to include literacy and industrial production from Ministero di Agricoltura, Industria e Commercio (1915) and Ciccarelli and Fenoaltea (2013), respectively, but these variables are highly correlated with population average height, and contain less variation, as they are observed at the provincial level rather than at the province-birth cohort level.

for the poverty constraint with an indicator for whether the migrant had an immediate family connection in the United States. The coefficient on this variable is not statistically significant, but is large (connected individuals are 0.027 standard deviations of height shorter, on average) and has the expected sign, indicating indicating that access to financing through chain migration can allow individuals of lower quality to migrate. We also include an indicator for whether the individual paid for his own passage. The coefficient on this indicator is large (individuals paying for their own passage were 0.099 standard deviations taller, on average) and of the expected sign (positive), but is only marginally statistically significant.

In columns (3) and (4), we divide the analysis of column (2) into pre-1917 and post-1917 periods. The pre-1917 results repeat those of the whole sample, with the coefficient on the indicator for an immediate family connection (marginally) statistically significant and more than twice its magnitude in the whole sample. The post-1917 results, however, enter with opposite signs for all variables except for average height.

On the whole, these results are consistent with theories of positive self-selection being induced by liquidity constraints and moderated by chain migration, but we find no evidence consistent with the relative inequality model.[54] However, we consider the evidence to be suggestive, not conclusive. In particular, we lack a direct measure of the returns to skill, and must therefore proxy them by inequality, of which the coefficient of variation of stature may not be a good measure. Moreover, we suspect that separately identifying the effects of different returns to skill and the effects of liquidity constraints is beyond the power of these regressions.

7   Threats to Identification

There are a number of possible causes for the results discussed in sections 4 and 6 above other than the true pattern of self-selection with respect to living standards. In the following discussion, we examine three causes of potential error. First, there may be an upward bias in the heights observed at Ellis Island due to systematic mismeasurement. Second, there may be measurement error caused by errors in our geolocation algorithm. Finally, the fact that our data covers only migrants aged 22–65 traveling to the United States may affect the generalizability of our results.

7.1   Systematic Upward Bias

Two sources of systematic upward bias in the heights observed at Ellis Island are possible: upward bias from measurement with shoes, and systematic error from self-reporting of heights. Clearly the former would lead

---

[54]When smoothed moments are used, the coefficient on the stature CV is positive and statistically significant, providing further evidence that is inconsistent with the relative inequality hypothesis.

to upward bias. With respect to the latter, it is well-established that when individuals are asked to report their own heights, their reports are systematically biased upwards (Danubio, Miranda, et al., 2008; Rowland, 1990). Thus, it is possible that our findings of positive self-selection are spurious, driven only by inaccuracies in the ship manifests that are not present in the military data, which are known to be the product of actual measurement.

Unfortunately, it is not known how the height data in the passenger manifests was gathered other than that it was entered into the manifests by the steamship companies upon embarkation in Europe. While the ships' surgeons were required to assert that they had examined each passenger—and were incentivized to do so by the requirement that shipping lines pay for the return passage of individuals found medically unfit to enter the United States (Bandiera, Rasul, and Viarengo, 2013)—it is not clear whether this examination was to include a measurement of height.[55] Furthermore, no reference is made to height in the rules of the Bureau of Immigration and Naturalization (1909), other than to stipulate that it was to be collected pursuant to the requirements of the Immigration Act of 1906. In fact, we have not been able to locate any information on the collection of these data. As a result—although the fact that the height distributions in Figure 11 do not appear pathological is reassuring—we cannot rule out self-reporting of heights and measurement in shoes, both of which would have biased heights upward, though we also have no evidence suggesting that either of these occurred.

We therefore cannot definitely conclude that our findings of positive self-selection on average are not spuriously generated by mismeasurement. Nevertheless, to the extent that measurement errors were constant across provinces, our findings regarding differential self-selection across provinces and the relationship between the degree of self-selection and the relative quality of the origin population would likely remain intact.

7.2 Bias Towards the Population Mean Through Incorrect Geolocation

As we discuss in section 3, incorrect matching to provinces by our geocoding algorithm is possible. Although we present evidence in Appendix A that our algorithm is accurate, we also study in this section the potential effects of incorrect geolocation on our results. Intuitively, such errors could generate both our positive

---

[55]With regard to the height field, the manifest asserts that it is a field that is "subject to revision by any inspection officer in the examination of aliens." No other instructions are given in any source that we were able to locate, nor does any other source discuss the collection of the height data. The shipping companies' surgeons are also made to swear that they had "made a personal examination of each of the aliens named [in the manifest], and that the foregoing Lists or Manifest Sheets . . . are, according to the best of [their] knowledge and belief, full, correct, and true in all particulars, relative to the mental and physical condition of such aliens." Whether the physical condition included height remains unclear. It is known, however, that the shipping companies had an incentive to diligently examine passengers, as they would be fined if a passenger was declared to be in violation of any restriction to immigration upon arrival in the United States.

self-selection results for the complete sample, as well as the differential self-selection results across different distributions of origin. One hypothetical danger is that the matching of most of the passengers to the south of Italy is spurious. If that were the case, then passengers would be, on average, spuriously assigned to shorter provinces, making them appear more positively self-selected than they truly were. Below, we show that our findings are likely not driven by such misassignment, even under the very conservative assumption that all misallocated passengers are uniformly distributed throughout Italy. The other, more concrete, danger is that any misallocation of passengers to the wrong province would bias the average height of migrants in each particular province toward the mean of the sample, mechanically generating an increase in the degree of self-selection with respect to the province-cohort average height.

We propose the following simple model to formalize our thought about this issue. Suppose that the true height of immigrant i from birth cohort t, who is truly from province j^*, h_{ij^*t} is determined by

$$h_{ij^*t} = \beta_0 + \mu_{j^*t} + \upsilon_{ij^*t},$$

where β_0 is the difference in means between immigrants and the whole-population distribution, μ_{j^*t} is the mean height in province j^* and birth cohort t, and υ_{ij^*t} is a determinant of individual height and has mean zero.

Importantly, this specification assumes that any self-selection that occurs is simply a mean shift of β_0, and that there is no differential self-selection by province-cohort of the type that we find above. Suppose that migrants are correctly geolocated with probability p and that incorrectly geolocated migrants are assigned randomly to a province. Then the height of individual i, who is matched to province j and birth cohort t, has mean

$$E(h_{ijt}) = \begin{cases} \beta_0 + \mu_{jt} & \text{w.p. } p \\ \beta_0 + \mu_t & \text{w.p. } 1-p \end{cases},$$

where $\beta_0 + \mu_t$ is the mean of the all-Italy distribution of migrant heights for birth cohort t. That is, when a migrant is correctly matched to a province, his height is a draw from that province's distribution of migrant heights. Conversely, if he is assigned to a province in error, he may, in reality, be from any province, and is thus drawn from the height distribution of all migrants. We can also write this model in terms of standardized height. Let z_{ij^*t} be the standardized height of individual i from birth cohort t, who is truly from province

j^*. Then the standardized height of an individual who has been matched to province j is

$$z_{ijt} = \frac{h_{ijt} - \mu_{jt}}{\sigma_{jt}}.$$

Then

$$E(z_{ijt}) = \begin{cases} \frac{\beta_0}{\sigma_{jt}} & \text{w.p. } p \\ \frac{\beta_0 + \mu_t - \mu_{jt}}{\sigma_{jt}} & \text{w.p. } 1 - p \end{cases}.$$

We first consider the effect of incorrect geolocation on our results of positive self-selection, presented in Table 4. Clearly incorrect geolocation will have no bearing on these results, as they do not depend on the province to which an individual migrant is assigned. However, incorrect geocoding may influence the results in Table 5. Suppose that $\beta_0 = 0$, so that in reality there is no self-selection of migrants, who are simply randomly drawn from the distributions of their provinces of origin. The estimate of the constant in column (1) of Table 5, which is our main estimate of the self-selection in the entire sample is

$$\bar{z} = \sum_{j,t} \frac{N_{jt}}{N} \times \frac{1}{N_{jt}} \sum_i z_{ijt},$$

where N_{jt} represents the number of migrants assigned to province j by our algorithm, and N is the total number of individuals in our sample. Based on the definitions and assumptions above, we have that

$$E(z_{ijt}) = (1 - p)\frac{\mu_t - \mu_{jt}}{\sigma_{jt}},$$

since z_{ijt} has mean of $\frac{\beta_0}{\sigma_{jt}} = 0$ with probability p. Then

$$E(\bar{z}) = (1 - p)\sum_{j,t} \frac{N_{jt}}{N}\frac{\mu_t - \mu_{jt}}{\sigma_{jt}}, \tag{7}$$

which may be positive or negative. If this expression is positive, we would erroneously conclude that there was positive self-selection when in fact there was none. Based on the value of the summation in equation (7) in our data, and our estimate in column (1) of Table 5, the value of $(1 - p)$ that would be required to produce the results in column (1) of Table 5 spuriously if the true β_0 is zero (i.e., if there is no self-selection at all) is 0.190. This degree of incorrect assignment exceeds our estimates of the rate of incorrect assignment that characterizes our algorithm. Thus, random misassignment by the geolocation algorithm is likely not behind our findings of positive self-selection.

It is important to note that this calculation is based on the conservative assumption that incorrectly matched migrants are uniformly distributed throughout Italy, resulting in the use of μ_t (the mean of the all-Italy distribution of heights) in equation (7). Instead, we might assume that our incorrectly matched individuals have the same geographic distributions as the correctly matched, and are thus primarily southern. In this cue, the μ_t in equation (7) would be replaced by a weighted average of the μ_{jt}, weighting by the number of migrants from each province. As most migrants were from the south, this weighted average would likely be less than μ_t, raising the necessary value of $(1-p)$ to spuriously generate our results.

Next, we consider the effects of incorrect geolocation on our findings of differential self-selection in column (3) of Table 5. We remove the assumption that $\beta_0 = 0$. Thus, migrants may be positively or negatively self-selected, but the magnitude of the self-selection will be the same in each province; that is, there will be no change in the degree of self-selection with average stature. Based on the framework above,

$$E(h_{ijt}) = \beta_0 + p\mu_{jt} + (1-p)\mu_t.$$

Thus, when no differential selection exists in reality, the observed differential selection will have a slope of p when migrant stature is regressed against average stature on a provincial level, as in column (3) of Table 5.[56] Therefore, under these assumptions, generating differential selection that results in a coefficient of p simply through incorrect geocoding requires that individuals be mismeasured with probability $(1-p)$. In the context of our data, generating the results in column (3) of Table 5 solely through measurement error requires that 35.6 percent of migrants be incorrectly assigned. Thus, although some mismatching likely occurs, it would have had to have been implausibly large in order to generate our differential self-selection result simply by chance.

7.3 Differences in Self-Selection Between Urban and Rural Migrants

We have not yet been able to distinguish between migrants from urban and rural areas of Italy. Our finding that southern migrants tended to be positively self-selected and northerners negatively selected may simply be due to a higher tendency for rural individuals to migrate from south Italy and for urban individuals to emigrate from north Italy.[57] In this case differences between migrant statue and population stature may simply be caused by shorter urbanites constituting the majority of migrants from one region and not anther,

---

[56]The argument could also be made in standardized height, although the derivations are somewhat more complex. In this case, the intuition is the same, as are, roughly, the implied probabilities of mismeasurement required in order to spuriously produce our results.

[57]We are grateful to Jeffrey Williamson for bringing this issue to our attention.

rather than reflective differences in individuals productive characteristics. Fortunately, our geomatching algorithm allows us to identify each individual's specific place of origin. We are currently collecting Italian census data that will allow us to classify individuals as urban or rural based on their place of origin and to determine whether the urban and rural composition of migration can be responsible for our findings with respect to self-selection on stature.

7.4 Data Coverage

Our analysis is restricted to migration to the United States. We do not observe migrants traveling to other parts of Europe or to South America—in fact, we are not aware of any micro data with coverage similar to that of the Ellis Island manifests in any country other than the United States that was the destination of a large number of Italian migrants—and thus cannot quantify self-selection among these individuals. To the extent that the degree of self-selection into migration to these destinations was different than for migration to the United States, our findings do not apply to the Italian emigration as a whole, but only to the flow between Italy and the United States. This flow in itself is very interesting to understand, as it was was the subject of an intense policy debate at the time, which resembles modern immigration policy debates. Moreover, most models of self-selection study flows between two countries; we are therefore able to test these models through the use of data on the flow from Italy to the United States.

Nonetheless, we must acknowledge that the self-selection into emigration from Italy may have differed from the self-selection into migration from Italy to the United States. For example, northerners traveling to South America may have been positively self-selected from their provinces of origin and numerous enough to outweigh the negative self selection to the United States that we observe. Our findings above would therefore be applicable to migration to the United States alone. We are currently in the process of collecting province-destination level data in order to quantify the degree to which the patterns of self-selection into migration to the United States are indicative of the larger patterns of self-selection of Italian migrants in the Age of Mass Migration.

8 Conclusions

We study the self-selection of Italians migrating to the United States during the Age of Mass Migration by comparing data on the stature of Italian migrants at their arrival in the United States to stature data on the Italian population of the time. This approach is based on the well-established relationship between stature

and a variety of measures and aspects of economic capability, such as income, skill, and intelligence. This comparison shows that the average Italian migrant was shorter than the national average, indicating that he was negatively self-selected from the country as a whole; but the average migrant was also taller than the mean of his province and birth cohort of origin, suggesting that he was positively self-selected, on average, on the local level. In addition, it is shown that migrants from poorer and less-developed areas of the country were more likely to be positively self-selected on the local level, while those from more-developed areas were more likely to be negatively self-selected on the local level. Finally, the degree of self-selection in each province is correlated with various measures of provincial development and individual financial capacity, and evidence is found consistent with theories that assign a significant role to liquidity constraints and chain migration in determining migrant self-selection, but not with the relative inequality theory of migrant self-selection.

Economists of migration have long sought to determine whether and why migrants are positively or negatively self-selected from their populations of origin, but a variety of data constraints have limited the conclusions that could be drawn. This research uses data that are free from many of these limitations to better understand migrant self-selection. Specifically, many previous studies of migrant self-selection are hampered by a lack of data. It is thus often not possible to compare migrants to the population from which they are drawn. Various indicators of productivity may be available for migrants, but their distribution in the sending population may be unknown. Alternatively, data on migrants may be collected after some time is spent in the receiving country, and may no longer represent their position in the sending population. Stature data solve both of these problems, allowing migrant self-selection to be easily characterized. Moreover, the stature data used in this project are sufficiently detailed as to allow for the first investigation of the differences between national self-selection (the type most commonly studied) and local self-selection; changing the frame of reference is shown to yield potentially different conclusions. Furthermore, a number of previous studies have relied on modern data on migration in order to determine whether the most or least productive members of a particular economy will choose to migrate. In the modern context, however, observed migrants are not simply those who wish to migrate; they must also be deemed acceptable by the receiving country. Simply comparing migrants to non-migrants thus does not directly reveal the underlying process of self-selection into migration. This research, by using data from a period in which there were very few legal restrictions to migration to the United States from Europe, effectively eliminates this issue; comparing migrants to non-migrants thus effectively identifies which portion of the population found migration to be optimal. Ultimately, this research contributes to understanding the effects of migration on the sending and receiving economies, which are widely believed to be strongly affected by the quality of migrants.

However, these results are based on a preliminary sample. As shown in Appendix D, several provinces are represented in the data by a very small number of individuals, whereas several hundred observations from each province are required to effectively analyze stature data (Komlos, 2004). Transcription of additional records is therefore necessary to ensure that our results are not driven by sampling error and that patterns in the data are not obscured by a lack of statistical power. We are therefore working to expand our data set, and to update our results accordingly.

References

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012). "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration". *The American Economic Review* 102:5, pp. 1832–1856.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2013). "Have the Poor Always Been Less Likely to Migrate? Evidence from Inheritance Practices during the Age of Mass Migration". *Journal of Development Economics* 102, pp. 2–14.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2014). "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration". *Journal of Political Economy* 122:3, pp. 467–506.

A'Hearn, Brian (2003). "Anthropometric Evidence on Living Standards in Northern Italy, 1730–1860". *The Journal of Economic History* 63, pp. 351–381.

A'Hearn, Brian, Franco Peracchi, and Giovanni Vecchi (2009). "Height and the Normal Distribution: Evidence from Italian Military Data". *Demography* 46:1, pp. 1–25.

A'Hearn, Brian and Giovanni Vecchi (2011). "Statura". In: *In Ricchezza e in Povertà: Il Benessere degli Italiani dall'Unità a Oggi*. Ed. by Giovanni Vecchi. Bologna: Il Mulino. Chap. 2, pp. 37–72.

Akee, Randall (2010). "Who Leaves? Deciphering Immigrant Self-Selection from a Developing Country". *Economic Development and Cultural Change* 58:2, pp. 323–344.

Bailey, Roy E., Timothy J. Hatton, and Kris Inwood (2014). "Health, Height and the Household at the Turn of the 20th Century". IZA Discussion Paper No. 8128.

Bandiera, Oriana, Imran Rasul, and Martina Viarengo (2013). "The Making of Modern America: Migratory Flows in the Age of Mass Migration". *Journal of Development Economics* 102, pp. 23–47.

Beard, Albertine S. and Martin J. Blaser (2002). "The Ecology of Height: The Effect of Microbial Transmission on Human Height". *Perspectives in Biology and Medicine* 45:4, pp. 475–498.

Belot, Michèle V. K. and Timothy J. Hatton (2012). "Immigrant Selection in the OECD". *Scandinavian Journal of Economics* 114:4, pp. 1105–1128.

Betrán, Concha and Maria Pons (2004). "Skilled and Unskilled Wage Differentials and Economic Integration, 1870–1930". *European Review of Economic History* 8:1, pp. 29–60.

Bhagwati, Jagdish N. (1976). "Taxing the Brain Drain". *Challenge* 19:3, pp. 34–48.

Biavaschi, Costanza and Benjamin Elsner (2013). "Let's Be Selective about Migrant Self-Selection". IZA Discussion Paper No. 7865.

Blum, Matthias (2013). "Estimating Male and Female Height Inequality". *Economics and Human Biology* In Press.

Boas, Franz (1911). *Reports of the Immigration Commission: Changes in Bodily Form of Descendants of Immigrants*. Washington: Government Printing Office.

Boas, Franz (1920). "The Influence of Environment upon Development". *Proceedings of the National Academy of Sciences of the United States of America* 6:8, pp. 489–493.

Bodenhorn, Howard, Timothy W. Guinnane, and Thomas A. Mroz (2013). "Problems of Sample-Selection Bias in the Historical Heights Literature: A Theoretical and Econometric Analysis". Mimeo., Yale University.

Bohlin, Jan and Anna-Maria Eurenius (2010). "Why They Moved—Emigration from the Swedish Countryside to the United States, 1881–1910". *Explorations in Economic History* 47, pp. 533–551.

Borger, Scott C. (2009). "Self-Selection and Liquidity Constraints in Different Migration Cost Regimes". Mimeo., UC San Diego.

Borjas, George J. (1987). "Self-Selection and the Earnings of Immigrants". *The American Economic Review* 77:4, pp. 531–553.

Borjas, George J. (1991). "Immigration and Self-Selection". In: *Immigration, Trade and the Labor Market*. Ed. by John M. Abowd and Richard B. Freeman. Chicago: University of Chicago Press. Chap. 1, pp. 29–76.

Brandenburg, Broughton (1904). *Imported Americans: The Story of the Experiences of a Disguised American and His Wife Studying the Immigration Question*. New York: Frederick A. Stokes Company.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak (2014). "Under-investment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh". Mimeo., Yale School of Management.

Bureau of Immigration and Naturalization (1909). *Immigration Laws and Regulations of July 1, 1907*. 8th ed. Washington: Government Printing Office.

Case, Anne and Christina Paxson (2008). "Stature and Status: Height, Ability, and Labor Market Outcomes". *Journal of Political Economy* 116:3, pp. 499–532.

Case, Anne, Christina Paxson, and Mahnaz Islam (2009). "Making Sense of the Labor Market Height Premium: Evidence from the British Household Panel Survey". *Economics Letters* 102, pp. 174–176.

Chiquiar, Daniel and Gordon H. Hanson (2005). "International Migration, Self-Selection and the Distribution of Wages: Evidence from Mexico and the United States". *Journal of Political Economy* 113:2, pp. 239–281.

Chiswick, Barry R. (1999). "Are Immigrants Favorably Self-Selected". *The American Economic Review, Papers and Proceedings* 89:2, pp. 181–185.

Ciccarelli, Carol and Stefano Fenoaltea (2013). "Through the Magnifying Glass: Provincial Aspects of Industrial Growth in Post-Unification Italy". *Economic History Review* 66:1, pp. 57–85.

Cline, Martha G., Keith E. Meredith, John T. Boyer, and Benjamin Burrows (1989). "Decline of Height with Age in Adults in a General Population Sample: Estimating Maximum Height and Distinguishing Birth Cohort Effects from Actual Loss of Stature with Aging". *Human Biology* 61:3, pp. 415–425.

Cole, Trafford R. (1995). *Italian Genealogical Records: How to Use Italian Civil, Ecclesiastical, & Other Records in Family History Research*. Salt Lake City: Ancestry Incorporated.

Commissioner-General of Immigration (1903). *Annual Report of the Commissioner-General of Immigration for the Fiscal Year Ended June 30, 1903*. Washington: Government Printing Office.

Conley, John P. and Ali Sina Onder (2013). "An Empirical Guide to Hiring Assistant Professors in Economics". Mimeo., Vanderbilt University.

Crimmins, E. M., B. J. Soldo, J. K. Kim, and D. E. Alley (2005). "Using Anthropometric Indicators for Mexicans in the United States and Mexico to Understand the Selection of Migrants and the 'Hispanic Paradox'". *Social Biology* 52:3–4, pp. 164–177.

Danubio, Maria Enrica, Elisa Amicone, and Rita Vargiu (2005). "Height and BMI of Italian Immigrants to the USA, 1908–1970". *Economics and Human Biology* 3, pp. 33–43.

Danubio, Maria Enrica, Gaetano Miranda, Maria Giulia Vinciguerra, Elvira Vecchi, and Fabrizio Rufo (2008). "Comparison of Self-Reported and Measured Height and Weight: Implications for Obesity Research among Young Adults". *Economics and Human Biology* 6, pp. 181–190.

Deaton, Angus (2007). "Height, Health, and Development". *Proceedings of the National Academy of Sciences* 104:33, pp. 13232–13237.

Di Maria, Corrado and Piotr Stryszowski (2009). "Migration, Human Capital Accumulation and Economic Development". *Journal of Development Economics* 90:2, pp. 306–313.

Direzione Generale della Statistica e del Lavoro (1912). *Annuario Statistico Italiano, 1911*. Vol. 1. Tipografia Nazionale di G. Bertero & C.

Docquier, Frédéric and Hillel Rapoport (2012). "Globalization, Brain Drain, and Development". *Journal of Economic Literature* 50:3, pp. 681–730.

Douglas, Paul H. (1919). "Is the New Immigration More Unskilled than the Old". *Publications of the American Statistical Association* 16:126, pp. 393–403.

Eveleth, Phyllis B. and James M. Tanner (1976). *Worldwide Variation in Human Growth*. Cambridge University Press.

Federico, Giovanni (2003). "Heights, Calories and Welfare: A New Perspective on Italian Industrialization, 1854–1913". *Economics and Human Biology* 1, pp. 289–308.

Feliciano, Cynthia (2005). "Educational Selectivity in US Immigration: How Do Immigrants Compare to Those Left Behind?" *Demography* 42:1, pp. 131–152.

Ferenczi, Imre and Walter F. Wilcox (1929). *International Migrations*. New York: National Bureau of Economic Research.

Fernández-Huertas Moraga, Jesús (2011). "New Evidence on Emigrant Selection". *The Review of Economics and Statistics* 93:1, pp. 72–96.

Fernández-Huertas Moraga, Jesús (2013). "Understanding Different Migrant Selection Patterns in Rural and Urban Mexico". *Journal of Development Economics* 103, pp. 182–201.

Ferrie, Joseph P. (1997). "The Entry into the US Labor Market of Antebellum European Immigrants, 1840–1860". *Explorations in Economic History* 34, pp. 295–330.

Ferrie, Joseph P. and Joel Mokyr (1994). "Immigration and Entrepreneurship in the Nineteenth-Century US". In: *Economic Aspects of International Migration*. Ed. by Herbert Giersch. Heidelberg: Springer-Verlag, pp. 115–138.

Floud, Roderick, Kenneth W. Wachter, and Anabel S. Gregory (1990). *Height, Health and History: Nutritional Status in the United Kingdom, 1750–1980*. Cambridge: Cambridge University Press.

Foerster, Robert F. (1919). *The Italian Emigration of Our Times*. 2nd. New York: Russell & Russell.

Fogel, Robert W. (1986). "Nutrition and the Decline in Mortality since 1700: Some Preliminary Findings". In: *Long-Term Factors in American Economic Growth*. Ed. by Stanley L. Engerman and Robert E. Gallman. Chicago: University of Chicago Press, pp. 439–556.

Fogel, Robert W. (1994). "Economic Growth, Population Theory, and Physiology: The Bearing of Long-Term Processes on the Making of Economic Policy". *The American Economic Review* 84:3, pp. 369–395.

Fogel, Robert W., Stanley L. Engerman, and James Trussell (1982). "Exploring the Uses of Data on Height: The Analysis of Long-Term Trends in Nutrition, Labor Welfare, and Labor Productivity". *Social Science History* 6:4, pp. 401–421.

Frisancho, A. Roberto (1993). *Human Adaptation and Accommodation*. Ann Arbor: The University of Michigan Press.

Gaulin, Steven and James Boster (1985). "Cross-Cultural Differences in Sexual Dimorphism: Is There Any Variance to be Explained?" *Ethology and Sociobiology* 6, pp. 219–225.

Giffoni, Francesco and Matteo Gomellini (2013). "Brain Gain in the Age of Mass Migration". Mimeo., University of Rome, La Sapienza.

Goldin, Claudia (1994). "The Political Economy of Immigration Restriction in the United States, 1890 to 1921". In: *The Regulated Economy: A Historical Approach to Political Economy.* Ed. by Claudia Goldin and Gary D. Libecap. Chicago: University of Chicago Press, pp. 223–258.

Gomellini, Matteo and Cormac Ó Gráda (2013). "Migrations". In: *The Oxford Handbook of the Italian Economy Since Unification.* Ed. by Gianni Toniolo. New York: Oxford University Press. Chap. 10, pp. 271–302.

Gould, Eric D. and Omer Moav (2010). "When is 'Too Much' Inequality Not Enough? The Selection of Israeli Emigrants". Mimeo., Hebrew University.

Gravlee, Clarence C., H. Russell Bernard, and William R. Leonard (2003). "Heredity, Environment, and Cranial Form: A Reanalysis of Boas's Immigrant Data". *American Anthropologist* 105:1, pp. 125–138.

Grogger, Jeffrey and Gordon H. Hanson (2011). "Income Maximization and the Selection and Sorting of International Migrants". *Journal of Development Economics* 95, pp. 42–57.

Guglielmino, C. R. and A. De Silvestri (1995). "Surname Sampling for the Study of the Genetic Structure of an Italian Province". *Human Biology* 67:4, pp. 613–628.

Hall, Prescott F. (1904). "Selection of Immigration". *Annals of the American Academy of Political and Social Science* 24, pp. 169–184.

Harris, John R. and Michael P. Todaro (1970). "Migration, Unemployment and Development: A Two-Sector Analysis". *The American Economic Review* 60:1, pp. 126–142.

Hatton, Timothy J. (2013). "How Have Europeans Grown So Tall?" *Oxford Economic Papers* In Press.

Hatton, Timothy J. and Bernice E. Bray (2010). "Long Run Trends in the Heights of European Men, 19th–20th Centuries". *Economics and Human Biology* 8, pp. 405–413.

Hatton, Timothy J. and Jeffrey G. Williamson (1998). *The Age of Mass Migration: Causes and Economic Impact.* New York: Oxford University Press.

Hatton, Timothy J. and Jeffrey G. Williamson (2008). *Global Migration and the World Economy: Two Centuries of Policy and Performance.* Cambridge: MIT Press.

Humphries, Jane and Timothy Leunig (2009). "Was Dick Whittington Taller than Those He Left Behind? Anthropometric Measures, Migration and the Quality of Life in Early Nineteenth Century London". *Explorations in Economic History* 46, pp. 120–131.

Ibarraran, Pablo and Darren Lubotsky (2007). "Mexican Immigration and Self-Selection: New Evidence from the 2000 Mexican Census". In: *Mexican Immigration to the United States.* Ed. by George J. Borjas. Chicago: University of Chicago Press. Chap. 5, pp. 159–192.

Komlos, John (1985). "Stature and Nutrition in the Habsburg Monarchy: The Standard of Living and Economic Development in the Eighteenth Century". *The American Historical Review* 90:5, pp. 1149–1161.

Komlos, John (1990). "Height and Social Status in Eighteenth-Century Germany". *The Journal of Interdisciplinary History* 20:4, pp. 607–621.

Komlos, John (2004). "How to (and How Not to) Analyze Deficient Height Samples". *Historical Methods* 37:4, pp. 160–173.

Komlos, John and Jörg Baten (2004). "Looking Backward and Looking Forward: Anthropometric Research and the Development of Social Science History". *Social Science History* 28:2, pp. 191–210.

Komlos, John and Lukas Meermann (2007). "The Introduction of Anthropometrics into Development and Economics". *Historical Social Research* 32:1, pp. 260–270.

Kosack, Edward and Zachary Ward (2013). "Who Crossed the Border? Self-Selection of Mexican Migrants in the Early 20th Century". Mimeo., University of Colorado at Boulder.

Kress, Margaret Rose (2007). "A Reanalysis of Boas's Hebrew Immigrant Data: Comparisons of Foreign-Born and US-Born Children Living in Early 20th Century America". MA thesis. Missoula: The University of Montana.

Lowell, B. L. (1987). *Scandinavian Exodus: Demography and Social Development of 19th-Century Rural Communities*. Boulder: Westview Press.

Lundborg, Petter, Paul Nystedt, and Dan-Olof Rooth (2009). "The Height Premium in Earnings: The Role of Physical Capacity and Cognitive and Non-Cognitive Skills". IZA Discussion Paper No. 4266.

Martorell, Reynaldo and Jean-Pierre Habicht (1986). "Growth in Early Childhod in Developing Countries". In: *Human Growth: A Comprehensive Treatise*. Ed. by Frank Falkner and James M. Tanner. Vol. 3. Plenum Press, pp. 241–262.

Mattoo, Aaditya, Ileana Cristina Neagu, and Çağlar Özden (2008). "Brain Waste? Educated Immigrants in the US Labor Market". *Journal of Development Economics* 87:2, pp. 255–269.

McKenzie, David, John Gibson, and Steven Stillman (2010). "How Important is Selection? Experimental vs. Non-Experimental Meausres of the Income Gains from Migration". *Journal of the European Economic Association* 8:4, pp. 913–945.

McKenzie, David and Hillel Rapoport (2010). "Self-Selection Patterns in Mexico-US Migration: The Role of Migration Networks". *The Review of Economics and Statistics* 92:4, pp. 811–821.

Ministero di Agricoltura, Industria e Commercio (1915). *Censimento della Popolazione del Regno d'Italia al 10 Giugno 1911*. Rome.

Ministero di Agricoltura, Industria e Commercio (1925). *Censimento della Popolazione del Regno d'Italia al 1 Dicembre 1921*. Rome.

Mishara, Prachi (2007). "Emigration and Wages in Source Countries: Evidence from Mexico". *Journal of Development Economics* 82:1, pp. 180–199.

Mokyr, Joel (1983). *Why Ireland Starved: A Quantitative and Analytical History of the Irish Economy, 1800–1850*. London: George Allen & Unwin.

Mokyr, Joel and Cormac Ó Gráda (1982). "Emigration and Poverty in Prefamine Ireland". *Explorations in Economic History* 19, pp. 360–284.

Mokyr, Joel and Cormac Ó Gráda (1984). "New Developments in Irish Population History, 1700–1850". *The Economic History Review* 37:4, pp. 473–488.

Niu, Sunny X. and Marta Tienda (2010). "Minority Student Academic Performance Under the Uniform Admission Law: Evidence from the University of Texas at Austin". *Educational Evaluation and Policy Analysis* 32:1, pp. 44–69.

O'Rourke, Kevin H. (1997). "The European Grain Invasion, 1870–1913". *The Journal of Economic History* 57:4, pp. 775–801.

Perlmann, Joel (2000). "What the Jews Brought: East European Jewish Immigration to the United States, c. 1900". In: *Immigrants, Schooling, and Social Mobility: Does Culture Make a Difference?* Ed. by Hans Vermeulen and Joel Perlmann. London: Palgrave Macmillan.

Perlmann, Joel (2001). "Race or People: Federal Race Classifications for Europeans in America, 1898–1913". Jerome Levy Economics Institute Working Paper No. 320.

Persico, Nicola, Andrew Postlewaite, and Dan Silverman (2004). "The Effect of Adolescent Experience on Labor Market Outcomes: The Case of Height". *Journal of Political Economy* 112:5, pp. 1019–1053.

Pike, Gary R. and Joseph L. Saupe (2002). "Does High School Matter? An Analysis of Three Methods of Predicting First Year Grades". *Research in Higher Education* 43:2, pp. 187–207.

Rooth, Dan-Olof and Jan Saarela (2007). "Selection in Migration and Return Migration: Evidence from Micro Data". *Economics Letters* 94:1, pp. 90–95.

Rothstein, Jesse M. (2004). "College Performance Predictions and the SAT". *Journal of Econometrics* 121, pp. 297–317.

Rowland, Michael L. (1990). "Self-Reported Weight and Height". *The American Journal of Clinical Nutrition* 52:6, pp. 1125–1133.

Roy, A. D. (1951). "Some Thoughts on the Distribution of Earnings". *Oxford Economic Papers* 3, pp. 135–146.

Runblom, H. and H. Norman (1976). *From Sweden to America: A History of the Migration.* Acta Universitatis Upsaliensis, Minneapolis Uppsala: University of Minnesota Press.

Schultz, T. Paul (2002). "Wage Gains Associated with Height as a Form of Health Human Capital". *The American Economic Review, Papers and Proceedings* 92:2, pp. 349–353.

Silventoinen, Karri (2003). "Determinants of Variation in Adult Body Height". *Journal of Biosocial Science* 35:2, pp. 263–285.

Sparks, Corey S. and Richard L. Jantz (2002). "A Reassessment of Human Cranial Plasticity: Boas Revisited". *Proceedings of the National Academy of Sciences of the United States of America* 99:23, pp. 14636–14639.

Sparks, Corey S. and Richard L. Jantz (2003). "Changing Times, Changing Faces: Franz Boas's Immigrant Study in Modern Perspective". *American Anthropologist* 105:2, pp. 333–337.

Spitzer, Yannay (2013). "Pogroms, Networks, and Migration: The Jewish Migration from the Russian Empire to the United States, 1881–1914". Mimeo., Northwestern University.

Spitzer, Yannay (2014). "The Dynamics of Mass Migration: Estimating the Effect of Income Differences on Migration in a Dynamic Model of Discrete Choice with Diffusion". Mimeo., Northwestern University.

Steckel, Richard H. (1995). "Stature and the Standard of Living". *Journal of Economic Literature* 33:4, pp. 1903–1940.

Stolz, Yvonne and Jörg Baten (2012). "Brain Drain in the Age of Mass Migration: Does Relative Inequality Explain Migrant Selectivity". *Explorations in Economic History* 49, pp. 205–220.

Todaro, Michael P. (1996). *Economic Development.* 5th. Upper Saddle River, NJ: Pearson Education.

US Congress (1907). "An Act to Establish a Bureau of Immigration and Naturalization, and to Provide for a Uniform Rule for the Naturalization of Aliens throughout the United States. Approved June 29, 1906". *The American Journal of International Law* 1:1, pp. 31–47.

US Congress (1911). *Reports of the Immigration Commission.* Washington: Government Printing Office, 61st Congress, 3rd Session, Document No. 756.

Ward, Zachary (2013). "Birds of Passage: Return Migrants, Self-Selection and Immigration Quotas". Mimeo., University of Colorado at Boulder.

Wegge, Simone A. (1998). "Chain Migration and Information Networks: Evidence from Nineteenth-Century Hesse-Cassel". *The Journal of Economic History* 58:4, pp. 957–986.

Wegge, Simone A. (1999). "To Part or Not to Part: Emigration and Inheritance Institutions in Nineteenth-Century Hesse-Cassel". *Explorations in Economic History* 36, pp. 30–55.

Wegge, Simone A. (2002). "Occupational Self-Selection of European Emigrants: Evidence from Nineteenth-Century Hesse-Cassel". *European Review of Economic History* 6:3, pp. 365–394.

Weil, Patrick (2000). "Races at the Gate: A Century of Racial Distinctions in American Immigration Policy (1865–1965)". *Georgetown Immigration Law Journal* 15, p. 627.

Zehetmayer, Matthias (2011). "The Continuation of the Antebellum Puzzle: Stature in the US, 1847–1894". *European Review of Economic History* 15, pp. 313–327.

Tables

Table 1: Summary statistics for all geolocated migrants ages 22–65.

| | All Passengers | | | Transcribed Only | | | Males | | | |
| | | | | | | | | | First-Timers Only | |
Variable	(1) All	(2) Males	(3) Females	(4) All	(5) Females	(6) All	(7) All	(8) Pre-1917	(9) Post-1917
Age	31.945 (9.027)	31.722 (8.554)	32.660 (10.373)	31.456 (8.358)	31.902 (9.415)	31.327 (8.023)	29.753 (7.669)	30.171 (7.683)	28.161 (7.404)
Married	0.696 (0.460)	0.697 (0.459)	0.693 (0.461)	0.699 (0.459)	0.700 (0.458)	0.699 (0.459)	0.612 (0.487)	0.647 (0.478)	0.482 (0.500)
Male	0.763 (0.425)			0.775 (0.418)					
Southern	0.849 (0.358)	0.850 (0.357)	0.845 (0.361)	0.846 (0.361)	0.840 (0.366)	0.848 (0.359)	0.826 (0.379)	0.830 (0.376)	0.812 (0.391)
Height (cm)				162.838 (6.816)	159.469 (6.982)	163.821 (6.445)	163.795 (6.490)	163.503 (6.675)	164.908 (5.591)
Repeater				0.377 (0.485)	0.153 (0.360)	0.442 (0.497)			
Imm. Fam. Conn.				0.413 (0.492)	0.742 (0.438)	0.318 (0.466)	0.300 (0.458)	0.284 (0.451)	0.365 (0.482)
Any Conn.				0.958 (0.201)	0.970 (0.172)	0.955 (0.208)	0.962 (0.191)	0.965 (0.184)	0.952 (0.213)
Paid for Self				0.846 (0.361)	0.659 (0.474)	0.900 (0.300)	0.915 (0.279)	0.907 (0.291)	0.947 (0.224)
Observations	1,125,373	857,477	267,896	18,561	4,837	13,724	7,496	5,905	1,591

Notes: Standard deviations in parentheses. Sample sizes are the minimum with data for all variables.

Table 2: Tests of representativeness of the geolocated sample; males ages 22–65.

Dep. Variable	All				Transcribed First-Timers			
	(1) All	(2) General	(3) Northern	(4) Southern	(5) All	(6) General	(7) Northern	(8) Southern
General Italian	-0.065^a (0.001)				-0.072^a (0.014)			
North Italian	-0.014^a (0.001)				-0.016 (0.010)			
South Italian	0.079^a (0.001)				0.087^a (0.015)			
Age	-0.083^a (0.021)	-0.156^a (0.040)	-0.452^a (0.056)	-0.026 (0.027)	0.234 (0.220)	-0.197 (0.414)	0.250 (0.529)	0.339 (0.296)
Birthyear	0.353^a (0.023)	0.520^a (0.046)	1.097^a (0.062)	0.313^a (0.030)	0.116 (0.270)	0.772 (0.527)	0.330 (0.631)	0.022 (0.356)
Married	-0.002^c (0.001)	0.006^b (0.002)	-0.012^a (0.004)	-0.013^a (0.001)	0.015 (0.015)	0.027 (0.028)	0.019 (0.039)	-0.009 (0.019)
Height (cm)					-0.103 (0.202)	-0.032 (0.385)	0.375 (0.526)	-0.091 (0.263)
Any Conn.					0.001 (0.006)	-0.005 (0.012)	-0.038^b (0.016)	0.011 (0.008)
Imm. Fam. Conn.					0.005 (0.014)	0.006 (0.026)	-0.093^b (0.036)	0.027 (0.018)
Paid for Self					0.006 (0.009)	0.009 (0.017)	0.036^c (0.020)	-0.002 (0.012)
Observations	1,055,713	253,822	119,284	682,607	8,816	2,164	1,217	5,435
Num. Geolocated	857,935	195,821	94,492	567,622	7,496	1,767	1,012	4,717

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1

Notes: The reported coefficients are from univariate regression of an individual characteristic on an indicator for being successfully geolocated by our algorithm. Robust standard errors in parentheses. In the first row of column (1), the coefficient is interpreted as follows: individuals in the geolocated sample are 6.5 percent less likely to have been general Italians than those in the non-geolocated sample. Sample sizes are the minimum number of observations with data for all variables.

58

Table 3: Differential geolocation.

Variables	(1) Std. Height
Average Height (cm, Demeaned)	-0.068^a
	(0.017)
Geolocated	0.030
	(0.034)
Average Height (cm, Demeaned) $\times$ Geolocated	0.006
	(0.018)
Constant	-0.004
	(0.031)
Observations	8657
R-squared	0.012
Fraction Geolocated	0.851

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Robust standard errors in parentheses. Dependent variable is height standardized by surname-implied province-birth cohort mean and standard deviation.

Table 4: Regressions of all-Italy birth cohort standardized height.

Variables	(1)	(2)	(3)	(4)
Southern		-0.345^a		-0.352^a
		(0.031)		(0.036)
Post-1917			0.126^a	0.085
			(0.027)	(0.063)
Post-1917 $\times$ Southern				0.043
				(0.070)
Constant	-0.116^a	0.166^a	-0.143^a	0.147^a
	(0.012)	(0.028)	(0.014)	(0.032)
Observations	7231	7231	7231	7231
R-squared	0.000	0.017	0.003	0.020
Constant + Southern		-0.179^a		-0.206^a
		(0.014)		(0.016)
Constant + Post-1917			-0.017	0.232^a
			(0.023)	(0.054)
Post-1917 + Post-1917 $\times$ Southern				0.128^a
				(0.030)

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Robust standard errors in parentheses. Dependent variable is height, standardized by all-Italy-birth cohort mean and standard deviation.

Table 5: Regressions of province and birth cohort standardized height.

Variables	(1)	(2)	(3) cm	(4)	(5)	(6)	(7) cm	(8)	(9)
Southern		0.134[a] (0.033)				0.123[a] (0.038)			
Average Height (cm, Demeaned)			0.654[a] (0.039)	−0.055[a] (0.006)			0.630[a] (0.045)	−0.058[a] (0.007)	0.220[b] (0.105)
Post-1917	0.030[b] (0.013)				0.125[a] (0.029)	0.076 (0.069)	0.947[a] (0.188)	0.164[a] (0.031)	0.164[a] (0.031)
Post-1917 × Southern						0.064 (0.076)			
Post-1917 × Average Height (cm, Demeaned)							−0.061 (0.093)	−0.015 (0.015)	
Male Literacy Rate									−0.333[a] (0.087)
Post-1917 × Male Literacy Rate									−0.155 (0.173)
Constant	0.030[b] (0.013)	−0.079[a] (0.030)	163.792[a] (0.078)	0.030[b] (0.013)	0.004 (0.015)	−0.096[a] (0.034)	163.635[a] (0.084)	0.003 (0.014)	0.196[a] (0.053)
Observations	7231	7231	7231	7231	7231	7231	7231	7231	7231
R-squared	0.000	0.002	0.038	0.010	0.002	0.005	0.040	0.014	0.006
Constant + Southern		0.055[a] (0.015)				0.026 (0.017)			
Constant + Post-1917					0.129[a] (0.025)	−0.021 (0.060)			
Post-1917 + Post-1917 × Southern						0.140[a] (0.032)			

Significance levels: [a] p<0.01, [b] p<0.05, [c] p<0.1
Notes: Robust standard errors in parentheses. In columns marked cm, dependent variable is actual height. In all other columns, dependent variable is height, standardized by province and birth cohort mean and standard deviation.

Table 6: Estimates of δ.

Variables	(1) All	(2) South	(3) North
Standardized Height	0.003	0.005	-0.001
	$[0.001, 0.004]$	$[0.003, 0.008]$	$[-0.002, 0.000]$
Constant	0.058	0.084	0.020
	$[0.058, 0.059]$	$[0.084, 0.085]$	$[0.020, 0.021]$
Observations	7046	5729	1317
Scaled $\hat{\delta}$	0.013	0.027	-0.004

Notes: Bootstrap 95% confidence intervals in square brackets. The dependent variable is migration probability, conditional on province-birth cohort-standardized height as determined by Bayes's Theorem. Estimation is weighted by the probability of migration.

Table 7: Mechanisms of self-selection.

Variables	(1) All	(2) All	(3) Pre-1917	(4) Post-1917
Stature CV	-0.004	0.004	0.052	-0.092
	(0.038)	(0.037)	(0.041)	(0.089)
Average Height (cm, Demeaned)		-0.055^a	-0.058^a	-0.064^a
		(0.006)	(0.008)	(0.012)
Imm. Fam. Conn.		-0.027	-0.061^c	0.033
		(0.028)	(0.034)	(0.051)
Paid for Self		0.099^c	0.088	-0.014
		(0.054)	(0.060)	(0.107)
Constant	0.078	-0.036	-0.212	0.521
	(0.143)	(0.150)	(0.165)	(0.350)
Observations	7231	7203	5676	1527
R-squared	0.003	0.013	0.015	0.026

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Robust standard errors in parentheses. Dependent variable is height, standardized by province and birth cohort mean and standard deviation. All regressions include a quadratic in age.

Figures
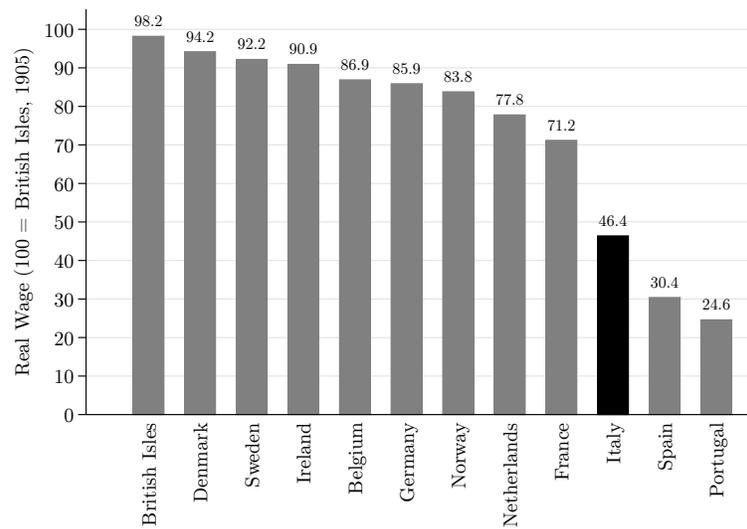
Figure 1: Real wages for several European countries. Internationally comparable PPP-adjusted averages for 1900–1913.

Source: O'Rourke (1997) as reported in Table 4.2 of Hatton and Williamson (2008).
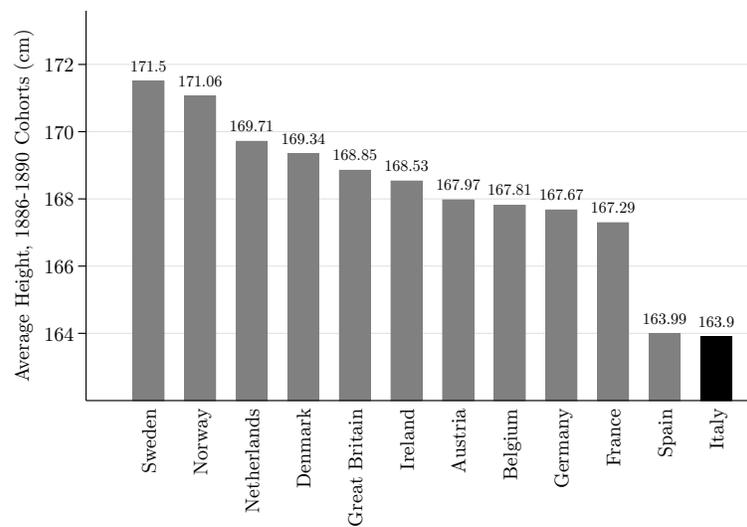
Figure 2: Average heights of males for several European countries.

Source: Hatton (2013) and Hatton and Bray (2010)

(a) 1870s

(b) 1880s

(c) 1890s

(d) 1900s
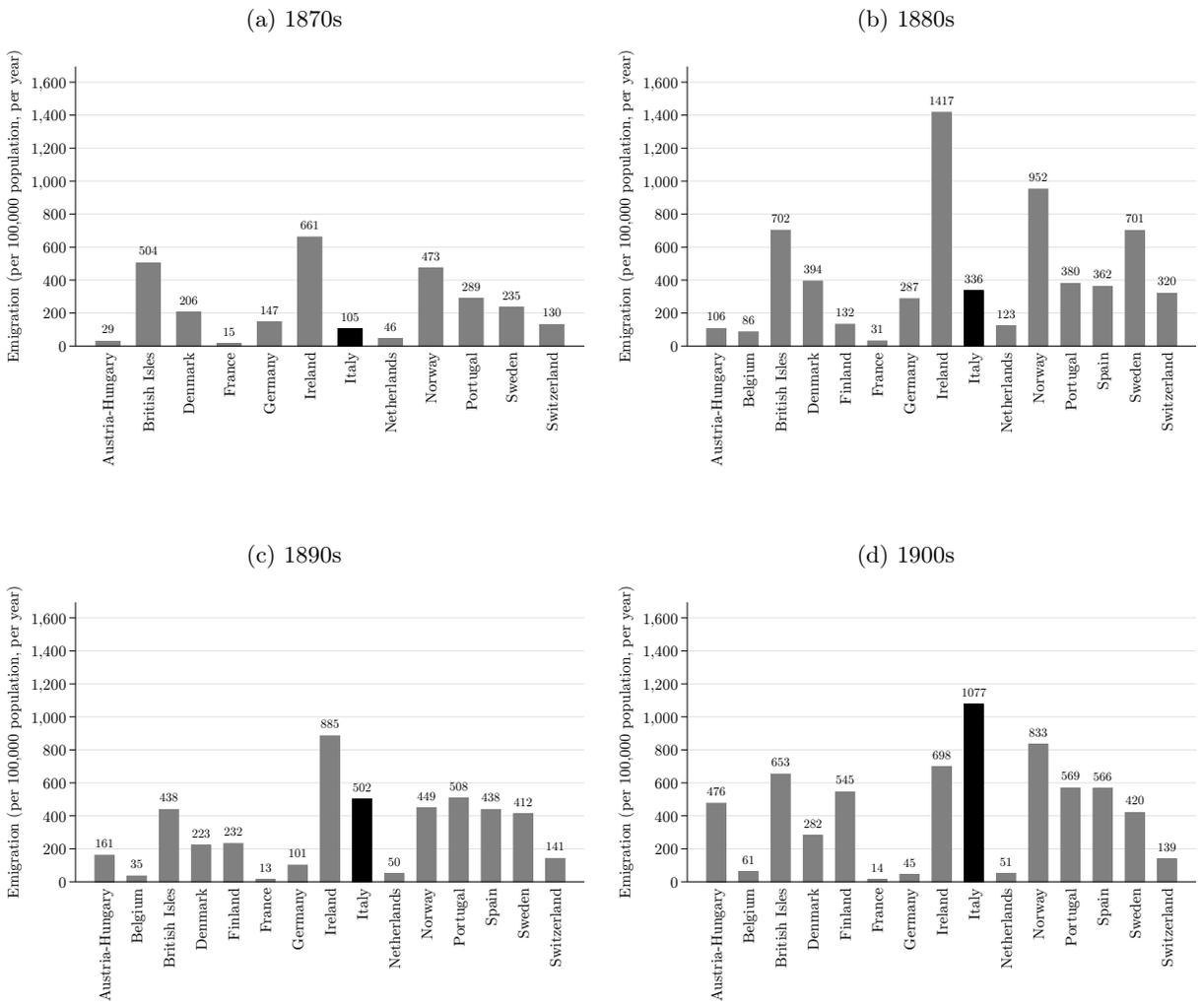
Figure 3: Emigration rates in Europe.

Source: Decadal averages taken from Table 2.1 of Hatton and Williamson (1998, p. 10), based on Text Table 9 of Ferenczi and Wilcox (1929, pp. 200–201).

(a) Means

(b) Standard Deviations
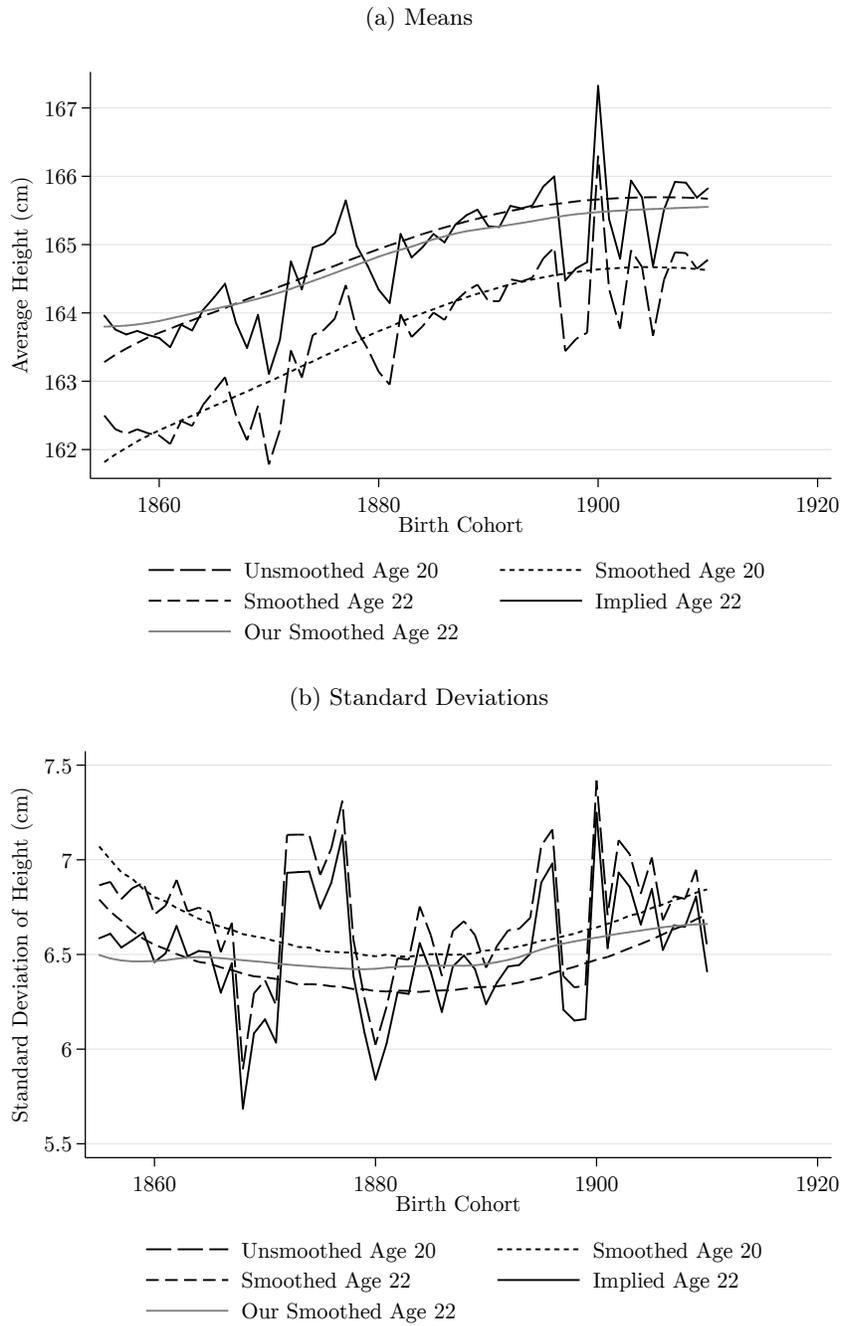
Figure 4: Example of the moments of the different height distributions received from A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011), together with moments that we derived from them, for the province of Roma.
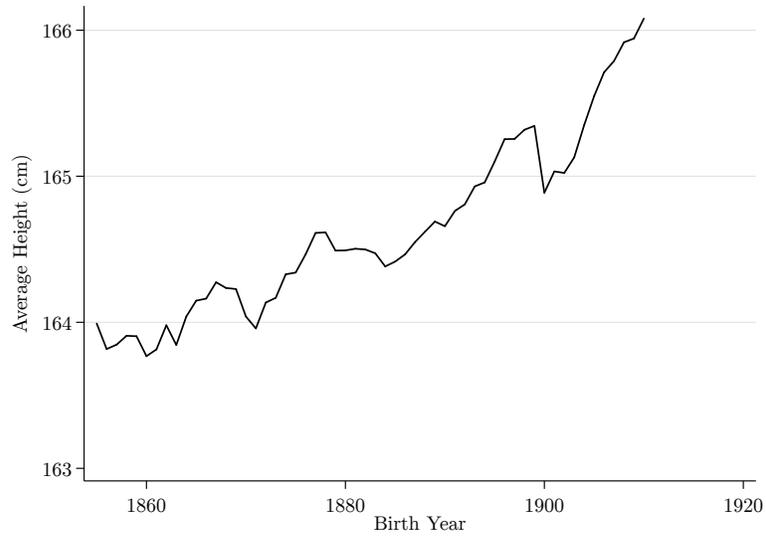
Figure 5: Trends in average height of Italian men.

Note: Mean heights are weighted within birth years across provinces by 1901 population.

Source: A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011).

(a) First page.

(b) Second page.

Figure 6: Sample manifests. Fields in dashed boxes are available in the SOLEIF files. We transcribed the fields in solid boxes.

(a) Provincial means.

Correlation: 0.72, Slope: 0.64 (SE: 0.08)

(b) Individuals.
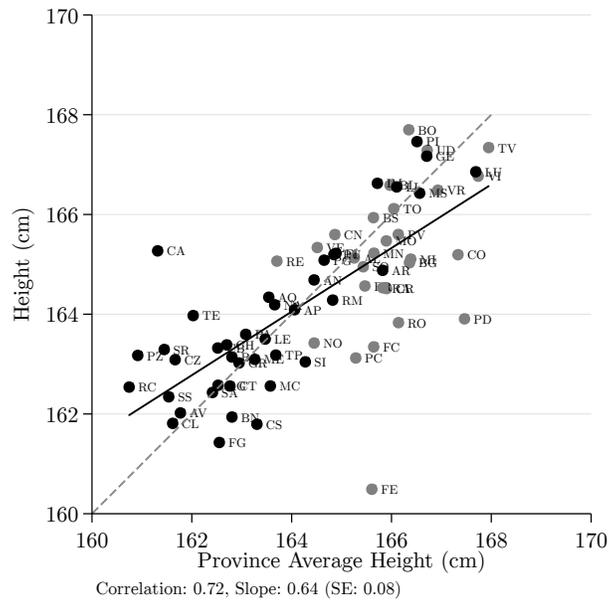
Note: Local linear regression. Epanechnikov kernel with bandwidth 1.24.
Shaded Regions are 95% confidence intervals.
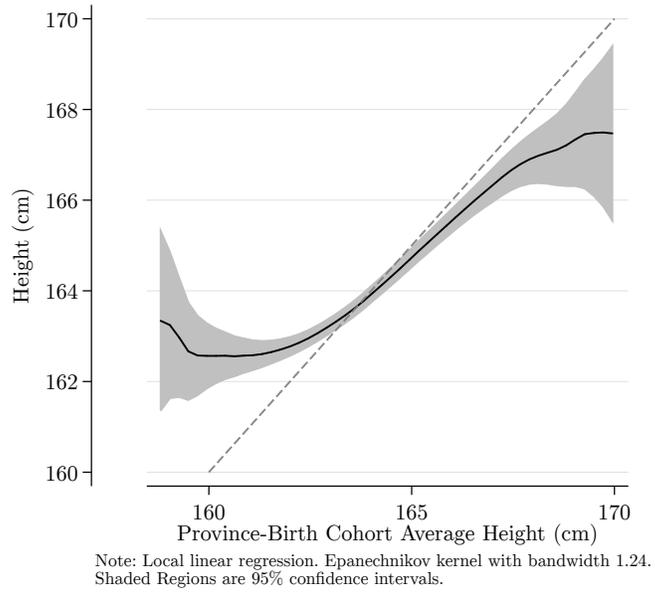
Figure 7: Heights of migrants by province-birth cohort average height.

Note: Northern provinces in gray, southern provinces in black. In Figure 7a, heights are weighted within provinces across birth cohorts by the number of migrants from each province in our sample.
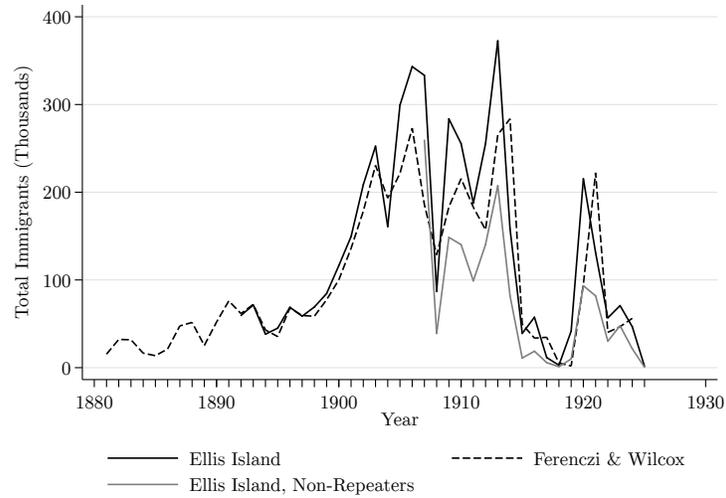
Figure 8: Italian immigration by year.

Source: Ferenczi and Wilcox (1929) and the SOLEIF data.

Total Passengers (1907-1925)

1228 - 4647
4648 - 7298
7299 - 16480
16481 - 33066
33067 - 56903
56904 - 136395

Passengers 1907-1925 to 1901 Pop.

0.005 - 0.014
0.014 - 0.023
0.023 - 0.032
0.032 - 0.074
0.074 - 0.127
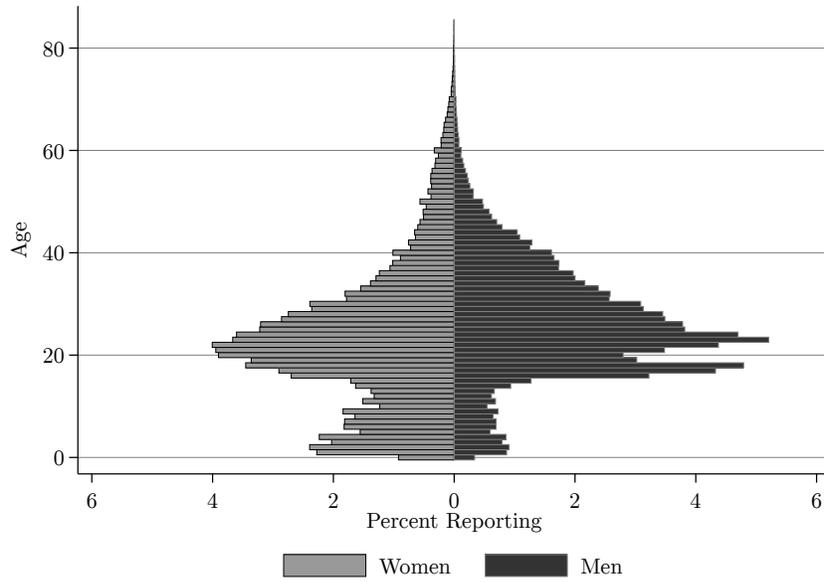0.127 - 0.193
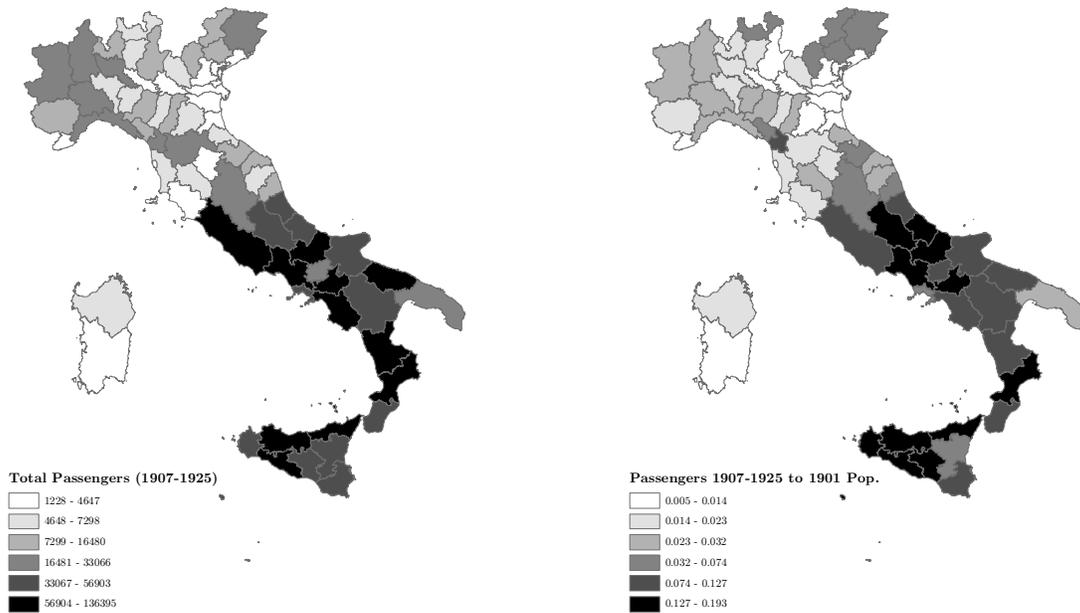
Figure 9: Origins of Italians traveling to the United States.

Figure 10: Passenger age distributions.
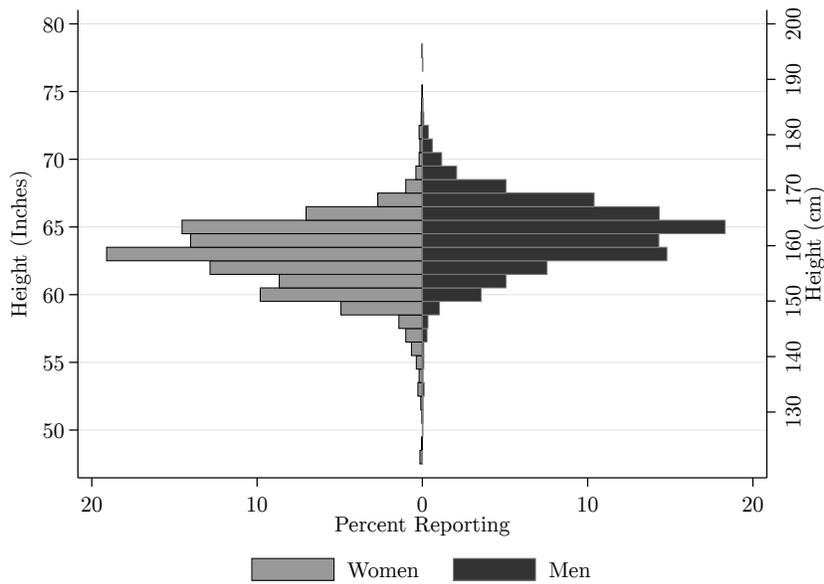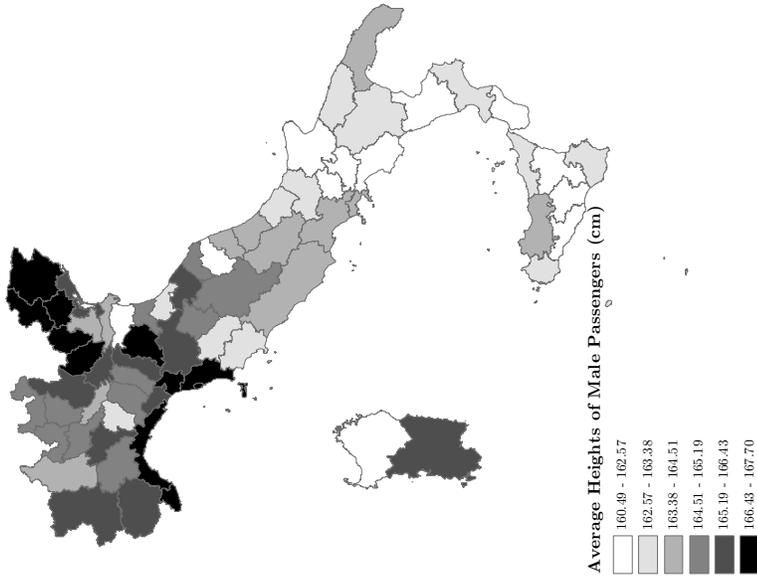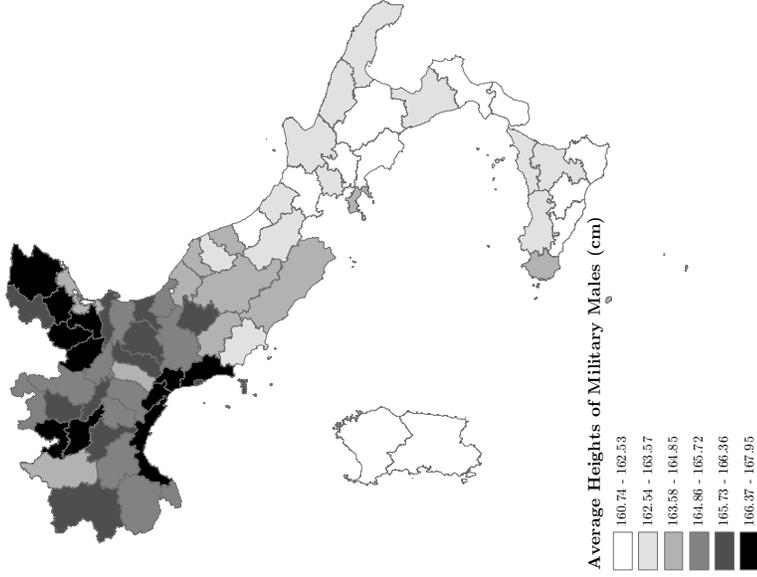

Figure 11: Passenger height distributions.

Note: Heights were almost exclusively recorded in whole inches.

Average Heights of Male Passengers (cm)

☐	160.49 - 162.57
	162.57 - 163.38
	163.38 - 164.51
	164.51 - 165.19
	165.19 - 166.43
■	166.43 - 167.70

(a) Average heights of passengers, by province.

Average Heights of Military Males (cm)

☐	160.74 - 162.53
	162.54 - 163.57
	163.58 - 164.85
	164.86 - 165.72
	165.73 - 166.36
■	166.37 - 167.95

(b) Average heights by province.

Figure 12: Average heights of passengers and populations.

Note: The province averages are weighted cross birth cohorts by the number of passengers in our data.

70

(a) Regions of Italy.

(b) Provinces of Italy.

(c) North and south Italy, according to the Bureau of Immigration and Naturalization.
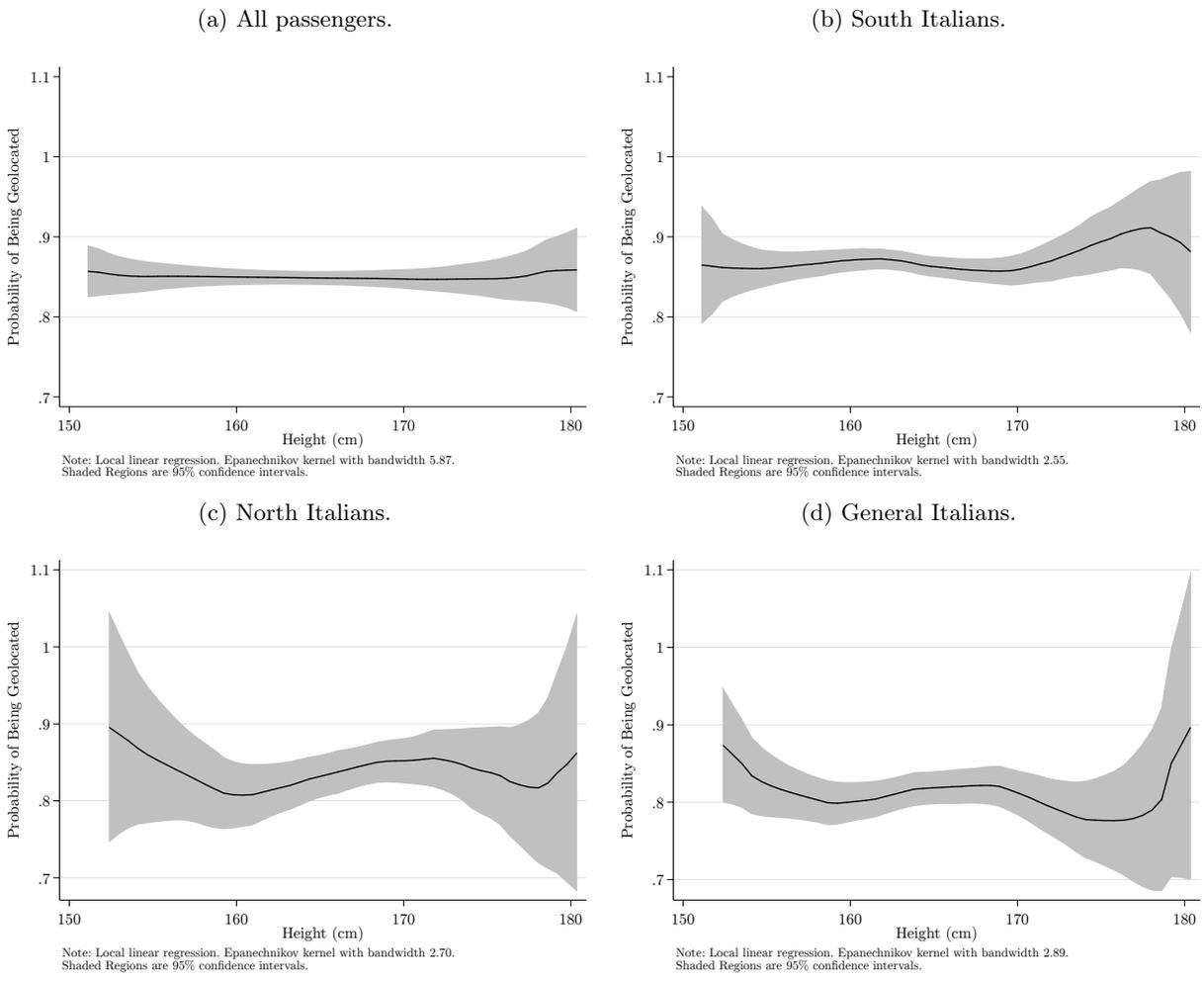
Figure 13: Geographic divisions of Italy.

(a) All passengers.

(b) South Italians.

Note: Local linear regression. Epanechnikov kernel with bandwidth 5.87.
Shaded Regions are 95% confidence intervals.

Note: Local linear regression. Epanechnikov kernel with bandwidth 2.55.
Shaded Regions are 95% confidence intervals.

(c) North Italians.

(d) General Italians.

Note: Local linear regression. Epanechnikov kernel with bandwidth 2.70.
Shaded Regions are 95% confidence intervals.

Note: Local linear regression. Epanechnikov kernel with bandwidth 2.89.
Shaded Regions are 95% confidence intervals.

Figure 14: Probability of being geolocated conditional on height.

Note: These are the results of a local linear regression of a binary variable indicating whether an individual was successfully geolocated by our algorithm against the individual's height.

(a) Actual height.
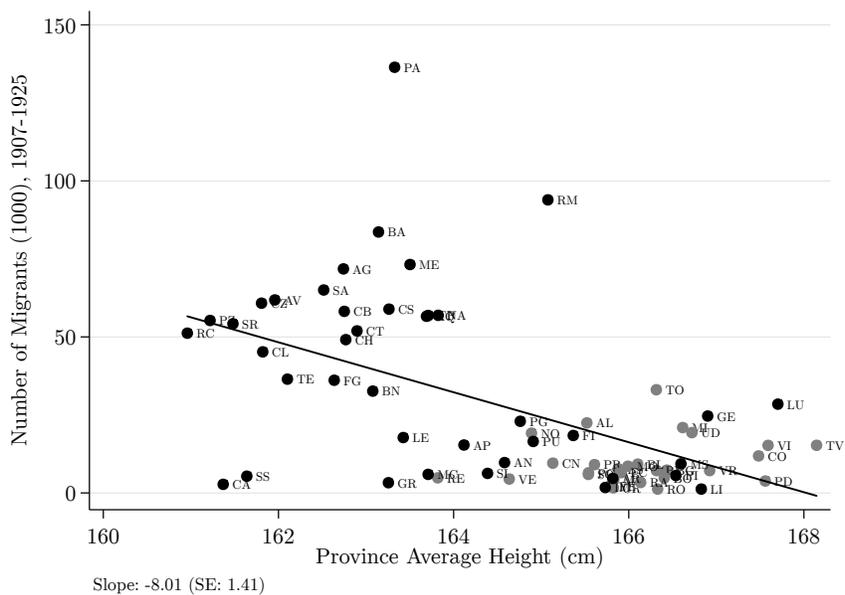

(b) Differences.
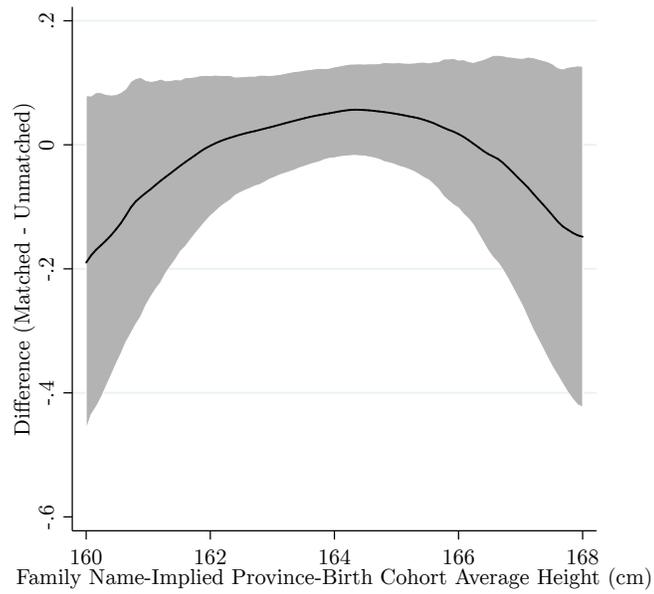
Figure 15: Tests of differential imbalance from the geomatching algorithm.

Note: Figure 15b displays the difference between the two curves in Figure 15a, together with 95 percent point wise bootstrap confidence intervals.

(a) Number of migrants vs. average heights.

Slope: -8.01 (SE: 1.41)

(b) Emigration rate vs. average heights.

Slope: -0.02 (SE: 0.00)
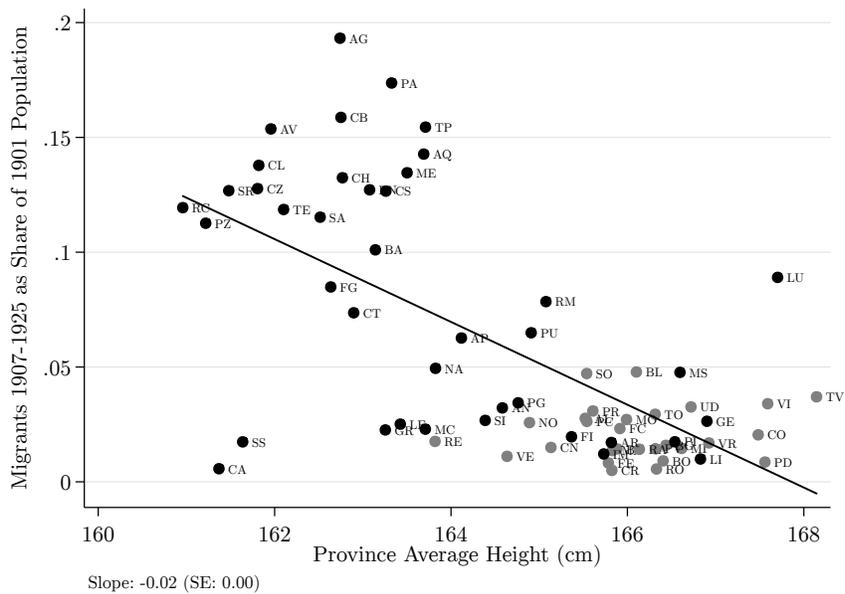
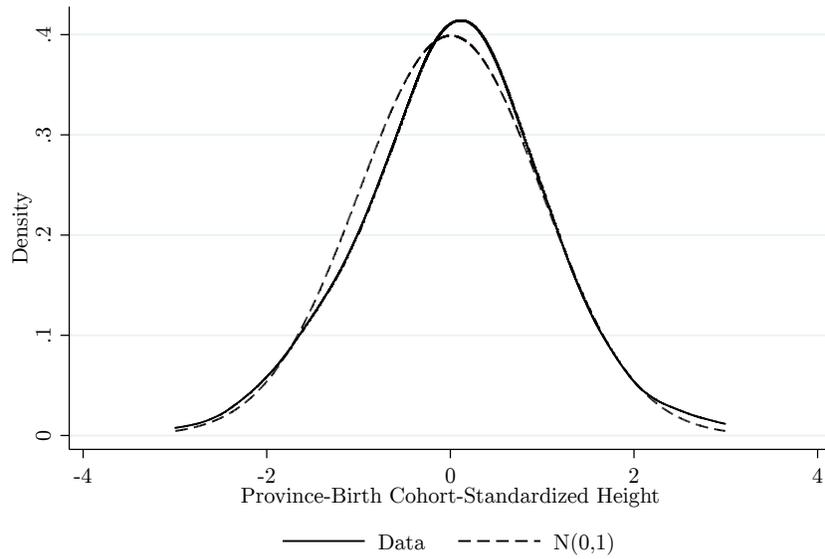Figure 16: Migration and average heights by province.

Note: Northern provinces in gray, southern provinces in black.

74

(a) All-Italy-normalized height distributions.
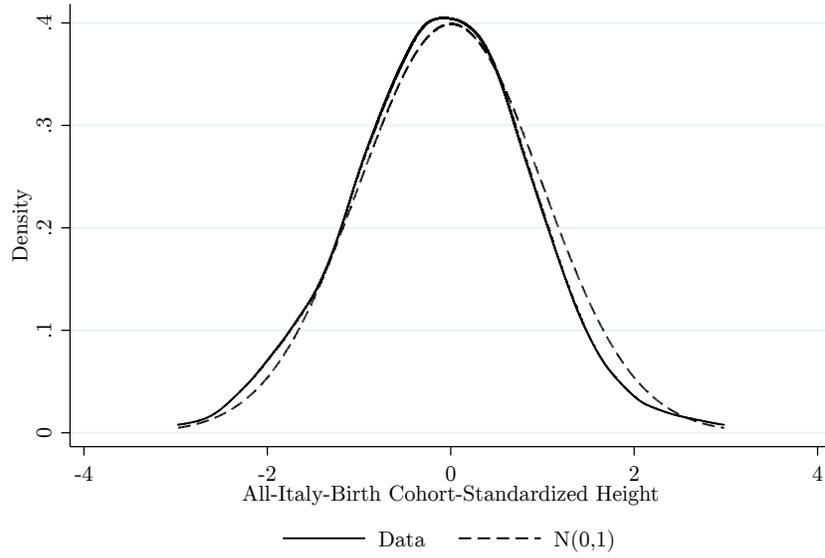
Kolmogorov-Smirnov Test: D = 0.05, p = 0.00. Kernel density estimated with bandwith 0.30.
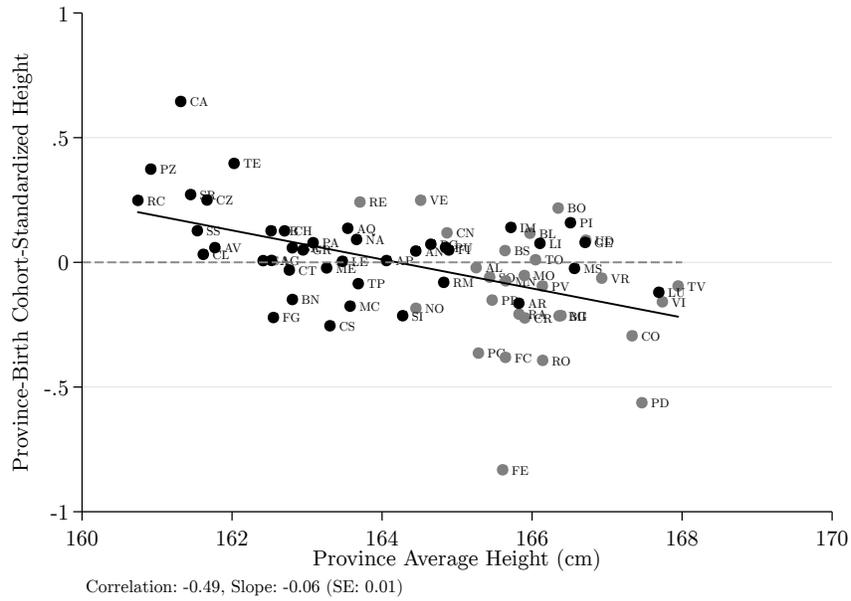
(b) Province-normalized height distributions.

Kolmogorov-Smirnov Test: D = 0.04, p = 0.00. Kernel density estimated with bandwith 0.30.

Figure 17: Distributions of migrant heights.

Note: Each figure presents a kernel density estimate of the distribution of migrant heights, standardized either by the all-Italy-birth cohort standardized height, or by the province-birth cohort standardized height. In each case, random noise that is uniformly distributed over the interval $[-0.5 \text{ in}, 0.5 \text{ in}]$ is added to the heights to correct for rounding. Kolmogorov-Smirnov tests are conducted to test whether these distributions are different from a hypothetical $N(0,1)$ distribution.

(a) Province means.

Correlation: -0.49, Slope: -0.06 (SE: 0.01)

(b) Individuals.

Note: Local linear regression. Epanechnikov kernel with bandwidth 1.26.
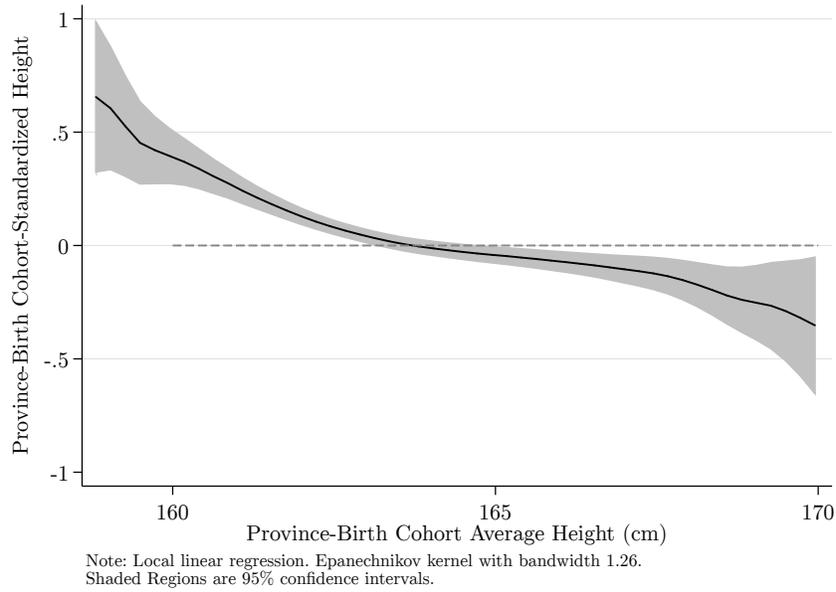Shaded Regions are 95% confidence intervals.

Figure 18: Standardized height against average province height.

Note: Northern provinces in gray, southern provinces in black. In Figure 18a, average province height is weighted within a province across birth cohorts by the number of immigrants in our sample.

(a) All Passengers

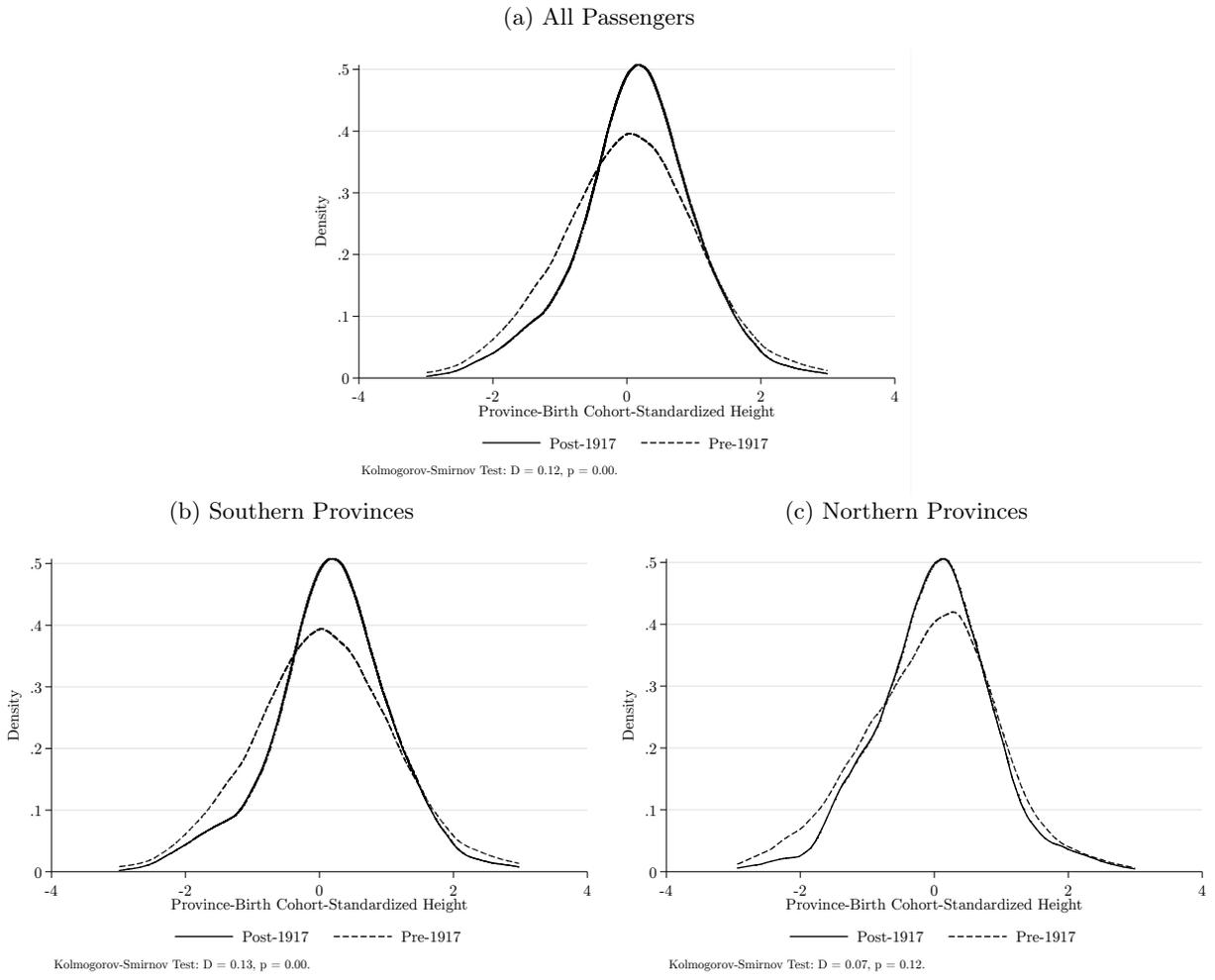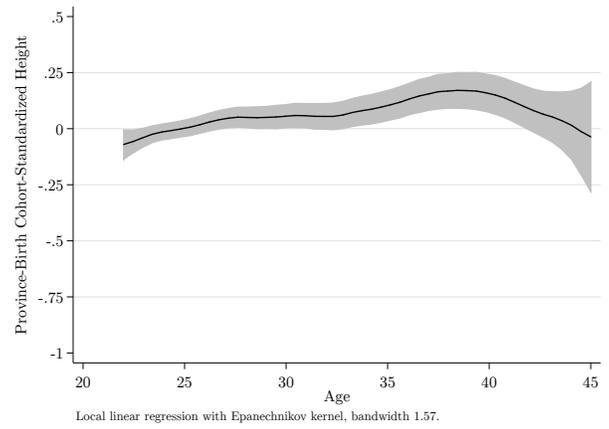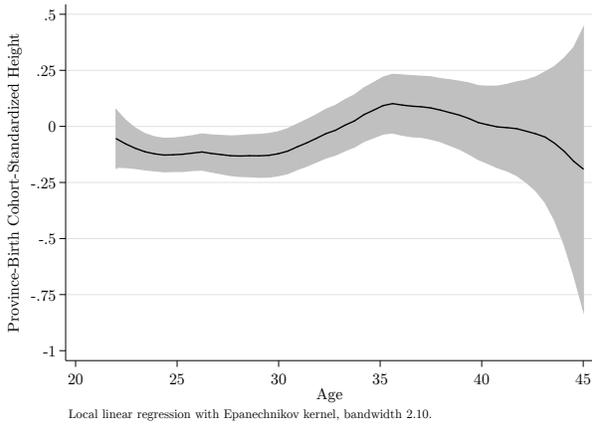(b) Southern Provinces   (c) Northern Provinces

Figure 19: Height distributions of pre-1917 (exclusive) and post-1917 (inclusive) migrants.

Note: Each graph represents a kernel density distribution of province-birth cohort standardized height for the pre-1917 and post-1917 periods. In each case, a Kolmogorov-Smirnov test is conducted to determine whether there is a statistically significant difference between the distributions. In Figures 19b and 19c, the sample is restricted to individuals matched to provinces in the legally defined south and north of Italy, respectively.

(a) All.

Local linear regression with Epanechnikov kernel, bandwidth 1.57.

(b) Northern provinces.

(c) Southern provinces.

Local linear regression with Epanechnikov kernel, bandwidth 2.10.

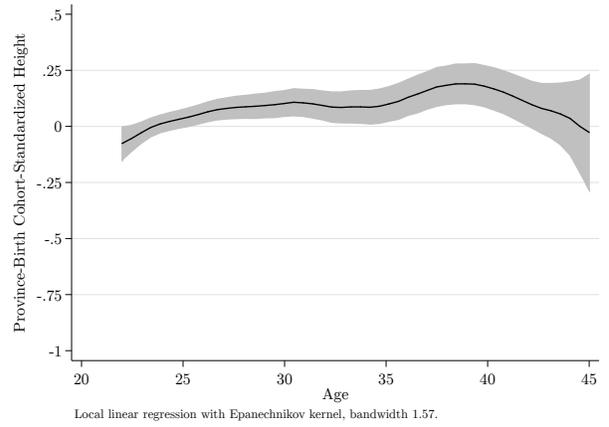Local linear regression with Epanechnikov kernel, bandwidth 1.57.

Figure 20: Self-selection by age.

Note: Each figure presents a local linear regression of province-birth cohort-standardized height on age, representing the magnitude of self-selection at each point in the life cycle. In Figures 20b and 20c, the sample is restricted to individuals matched to provinces in the legally defined south and north of Italy, respectively.

(a) All Passengers
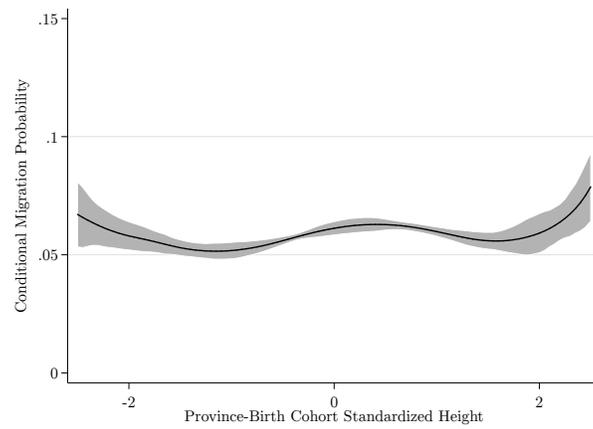
(b) Southern Provinces
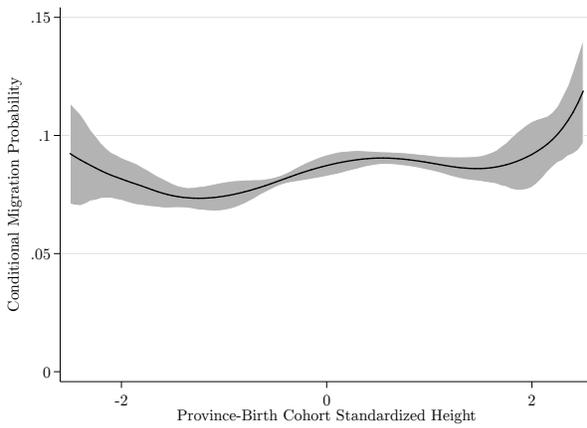
(c) Northern Provinces
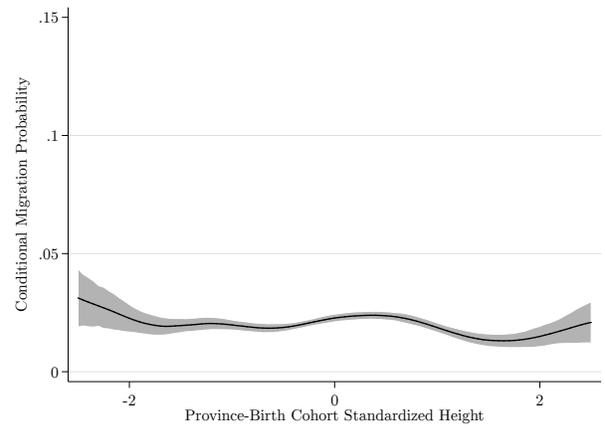
Figure 21: Migration probability conditional on province-birth cohort-standardized height.

Note: These graphs present the relationship between migration probability, conditional on province-birth cohort-standardized height and that height, determined by Bayes's Theorem. Estimation is weighted by the probability of migration. Shaded areas indicate 95 percent point wise bootstrap confidence intervals. In Figures 21b and 21c, the sample is restricted to individuals matched to provinces in the legally defined south and north of Italy, respectively.

A Implementation and Accuracy of the Geolocation Algorithm

Each migrant was assigned to a province of origin based on the last place of residence reported on the Ellis Island manifests using the following algorithm, which we implemented through the Python programming language.

1. Locations that were obviously outside of Italy (e.g., Argentina, New York) were removed from the data.

2. A search for the location listed by the migrant was conducted in Google Maps. If necessary, the search was constrained to Italy. In the case of a unique result, the coordinates of that result were recorded.

3. If there was no unique result, an attempt was made to follow Google's suggestions. If the suggestion was unique, or if there were multiple suggestions within 0.167 degrees of latitude and longitude of one another (approximately 10 miles), the centroid of the suggestions was accepted.

4. If no location could be found using the above steps, an attempt was made to make a string match between the name of the location and a list of communes of Italy. If a commune was found using this method, coordinates were recorded either if the migrant did not also list a province of origin, or if the province listed by the immigrant matches that of the commune.

5. As illustrated in Figure A.1, all immigrants who could be matched to coordinates were placed on a GIS map of historical Italy acquired from Martí-Henneberg (2005) and assigned to provinces based on the provincial borders into which they fell. These boundaries are summarized in Figure A.2.

6. Those immigrants who could not be matched to a province using the procedure above were string matched to the list of provinces of Italy where possible.

Approximately 1.9 million immigrants could be matched to a province by this algorithm. The remaining roughly 900,000 could not be matched to a province for one of two reasons: either their previous place of residence could be determined, but it was outside of the borders of 1870–1910 Italy, or the previous place of residence could not be determined from the strings available from the SOLEIF. Neither of these groups is depicted in Figure A.1.[58]

There is potential for error in our geocoding algorithm. We therefore present several pieces of evidence to suggest that our assignment is rather accurate. First, we take advantage of the distinction in ethnicity recorded at Ellis Island. In Figure A.3 we depict the fraction of the migrants assigned to each province by our geocoding algorithm who are classified as north Italian, south Italian, and general Italian. We also determine which provinces had more north Italian migrants assigned to them, and which had more south Italian migrants. Reassuringly, all southern provinces are predominantly south Italian, and most northern provinces are predominantly north Italian. There are, however, eight northern provinces to which more south Italians than north Italians are matched. In only one of these provinces, however, are a majority of matched passengers south Italian. Moreover, five of these eight provinces are in the bottom quartile in terms of absolute number of migrants (as measured by our algorithm), and seven of eight are in the bottom third. Figure A.3 also depicts the fraction of passengers from each province who are not disaggregated into north Italians and south Italians. As we expect, northern provinces consist of a higher proportion of general Italians than southern provinces.[59] Moreover, it appears that the failure to decompose Italians into north and south is primarily driven by certain ships who do not decompose Italians at all. Of the 9,905 voyages in our data, 34 percent do not decompose Italians at all. If we limit consideration to passengers classified as north Italian or south Italian, 92.5 percent are located in the correct portion of Italy.

The accuracy of our algorithm, as measured by the correct matching of north Italians to the north, and south Italians to the south, is greater in the extreme north and south of Italy than in the center, suggesting

---

[58]Tests of the representativeness of the sample of individuals who could be successfully geocoded are presented in section 3.3.

[59]In particular, 22 percent of passengers matched to a southern province are classified as general Italians, as compared to 26.8 percent of those matched to a northern province.

that much of the inaccuracy may be due to uncertainty by those completing the manifests as to whether a passenger should have been labeled as north Italian or south Italian based on his last place of residence. For example, approximately 80 percent of migrants who were assigned to Sicily, where there would have been no such uncertainty, were correctly identified as south Italian, while less than one percent were identified as north Italian. Furthermore, provinces from which relatively fewer individuals migrate to the United States are mechanically more likely to capture a greater share of inaccurately matched migrants, and therefore Sicily, which is the origin of a large number of migrants, is the most indicative of the true rate of failure of our matching algorithm. We find no reason to suspect that the matching of Sicilian migrants would be more accurate than that of other Italians.

To get a formal estimate of the failure rate of the geo-matching algorithm, we focus our attention on a group of migrants for whom two pieces of information, independent of the matching algorithm, indicate their intra-Italian ethnicity: those who in addition to being recorded by the clerks as south Italian, also departed from the port of Palermo. From among them we remove all passengers traveling in ships that did not make a complete distinction between south Italians and north Italians.[60] The remaining passengers constitute 27.9 percent of the 321,833 passengers leaving Palermo. We expect that a very large proportion of these passengers were Sicilians, which is consistent with the fact that 99.9 percent of them are recorded as south Italian. The geo-matching algorithm assigned 99.1 percent of these south Italian passengers to locations in southern Italy,[61] and 96.0 percent to Sicily specifically.[62]

Next, we perform a simple exercise that answers the following question: what is the worst rate of failure of the geo-matching algorithm that is consistent with this share of matching of Palermo passengers to Sicily? The rate of assignment to Sicily should be the sum of three elements: Sicilians who were correctly matched, non-Sicilians who were spuriously matched to Sicily, and Sicilians who were incorrectly matched, but were assigned to another place within Sicily. That is,

$$S = pS^* + s(1-p)(1-S^*) + s(1-p)S^*,$$

where S is the share of Sicilians according to the geo-matching algorithm, S^* is the true rate of Sicilians in this sample, s is the probability that a passenger would be assigned to Sicily conditional on failing to assign him to his actual last place of residence, and the object of interest is $(1-p)$, the probability that the geo-matching algorithm fails to match a migrant correctly.[63]

We assume that failed matches are proportionally distributed over the space of Italian matchable locations across Sicilian and non-Sicilian locations. That is, let $\mathbb{L}^S$ be the set of all Sicilian locations that could be matched by the algorithm, $\mathbb{L}^S = \{l_1^S, \ldots, l_{N^S}^S\}$. Similarly, let $\mathbb{L}^{\neg S} = \{l_1^{\neg S}, \ldots, l_{N^{\neg S}}^{\neg S}\}$ be the set of all such Italian locations outside of Sicily. Let l_i^* be the true last place of residence of immigrant i and l the location matched by the algorithm. The assumption is that

$$s := P(l_i \in \mathbb{L}^S | l_i \neq l_i^*) = \frac{N^S}{N^S + N^{\neg S}}.$$

Clearly, N^S and $N^{\neg S}$ are unknown, but they could be reasonably approximated in several ways. First, it is possible to count the total number of unique pairs of longitude-latitude coordinates matched by the algorithm, within and outside Sicily.[64] Of the 42,487 unique matched locations, 10.0 percent are within Sicily. Alternatively, we can use the share of current Italian communes located in Sicily; these amount to

---

[60]That is, we remove all passengers aboard ships that had at least one "Italian", without the north-south distinction.

[61]Note that we do not use ethnicity at all in the geolocation algorithm, so this is not a mechanical outcome.

[62]This rate was not driven by a tendency to blindly assign the port of departure as their last place of residence; only 14.7 percent of these passengers simply reported "Palermo" as their last place of residence.

[63]The implicit assumptions are the following: (a) the failure probability, $(1-p)$, is equal for Sicilians and non-Sicilians; and (b) the false matching rate to a location in Sicily, s, is equal for Sicilians and non-Sicilians. The former could be violated if Sicilians report their locations more clearly than other passengers, or if they are more likely to report provincial capitals or province names (which are easier to locate than small towns) than other Italians.

[64]The over-representation of Sicilians among immigrants makes this a conservative measure. Since there are more Sicilians, the range of locations we match within Sicily is probably wider than in other regions, biasing s toward more spurious matching of immigrants to Sicily.

4.8 percent of the 8,100 Italian communes. Other possible proxy measures are not far off: the area of Sicily was approximately 9.2 percent of the total area of Italy in our study period, and Sicilians were 10.9 percent of the total Italian population in 1901.

Taking as a benchmark $s = 0.100$, the rate of matchable locations in Sicily, we can write p as a function of two known variables, s and S (the rate of matches to Sicily within the sample of south Italians on completely disaggregated ships traveling from Palermo), and a single unknown variable, the true rate of Sicilians in this sample S^*:

$$p = \frac{S - s}{S^* - s}.$$

Note that this probability is decreasing in S^*, and thus a lower bound for the rate of successful matching is given when $S^* = 1$.[65] This gives

$$p \geq \frac{S - s}{1 - s} = \frac{0.960 - 0.100}{1 - 0.100} = 0.956,$$

meaning that with probability of at least 95.6 percent, the matching algorithm successfully matches a passenger to his correct last place of residence.

We can also perform a similar exercise for the port of Le Havre, which was the most common non-Italian port and mainly served north Italian passengers: 91.9 percent of passengers on fully disaggregated ships are listed as northern. Thus, we replace being southern-Italian matched outside Sicily with being matched outside of north Italy as the measure of a certainly failed match for northern Italians. Performing the exercise precisely as above provides a lower bound on p of only 0.503. However, when the measure of a failed match is redefined as being located within the *Mezzogiorno*, or traditional south of Italy (which was a subset of the definition used on the Ellis Island manifests—a fact which would likely have led to confusion among non-Italian clerks), the lower bound on p rises to 0.782. This is still a highly conservative lower bound, but it obscures the fact that the fraction of these passengers matched to a province outside of the *Mezzogiorno* is 91.2 percent. The lower bound is potentially mechanically very low, as nearly 60 percent of all possible locations are in the area north of the *Mezzogiorno*, and thus it can not be ruled out that there was a large degree of mismatch within it. Thus, repeating the exercise with the north Italians embarking from Le Havre yields a lower estimate of the lower-bound of accurate matching (although it is consistent also with correct matching rates of over 90 percent). We suspect that it is largely driven by the fact that this port covered a geographically larger emigration market compared to Palermo.

We also test the accuracy of the geolocation algorithm by applying it to the Italian migrants in the Castle Garden passenger lists. These data contain the records of over two million Italians arriving at the port of New York between 1830 and 1907, and in 1912, with the bulk of the data covering the period 1880–1907.[66] Due to the fact that this time span stretches past the actual operation of the Castle Garden facility (which was replaced by Ellis Island in the 1890s), there will be some overlap with the Ellis Island data. This overlap will not be important for this exercise, which takes advantage of the following feature of these data.

Unlike the Ellis Island records, this data source includes the passengers' last province of residence within Italy in addition to their last place of residence. Unfortunately, we have not been able to determine how the last place of residence was matched to the last province of residence by the creators of the Castle Garden data base. Unlike our algorithm, the provinces in the Castle Garden records correspond to modern Italian provinces, rather than to those that prevailed during our study period. Nonetheless, these geolocated data provide us with a setting in which the accuracy of the algorithm can be tested.

In particular, we apply our geomatching algorithm to the last place of residence listed in the Castle Garden Records, and use the coordinates generated by this process to match immigrants to the modern province with which their listed place of residence coincides. We find that in 86 percent of cases, the province to which our algorithm assigns a passenger coincides with the province to which a passenger is assigned by the Castle Garden records. This result can be interpreted in a number of ways. First, if the matching to provinces in the Castle Garden records is assumed to be perfectly accurate, this result indicates a success

---

[65]That is, when all of the passengers in this sample are Sicilian, and thus all matches to places outside Sicily are false.

[66]These data can be collected from http://www.castlegarden.org.

rate of linkage of 86 percent. Even in this very conservative case, this rate of accuracy is above the required accuracy that we have calculated would be necessary ensure that our results are not generated spuriously by incorrect linkage.

However, it is possible that disagreements between the results of our algorithm and the matching of Castle Garden may be driven by instances in which we make a correct match and the Castle Garden records are in error. In this case, the true rate of success would be greater than 86 percent. It is also possible that there are instances in which we agree with the Castle Garden records, but both we and those records have made a false match. Such a situation might arise, for example, in cases in which there was a typographic error during the completion or transcription of the manifests, and in which the incorrectly transcribed version of the last place of residence corresponds to some other location in Italy.[67] If such joint error occurred on a very large scale, it cannot be ruled out that the true rate of success is less than 86 percent. On the whole, both types of these errors likely exist. We therefore continue to search for new ways to improve our algorithm, while simultaneously asserting that a linkage rate of 86 percent, which is a conservative estimate of the accuracy of the algorithm conditional on the Castle Garden records alone (and disregarding our other arguments) is good enough to ensure that our results are not simply driven by errors in geolocation.

References

Martí-Henneberg, Jordi (2005). "The Administrative Map of Europe: Continuity and Change of the Administrative Boundaries (1850–2000)". *Geopolitics* 10, pp. 791–815.

---

[67]For example, if the commune of Lago in Salerno were mistakenly transcribed as Logo, the passenger would be matched to Logo in Como rather than to Lago.
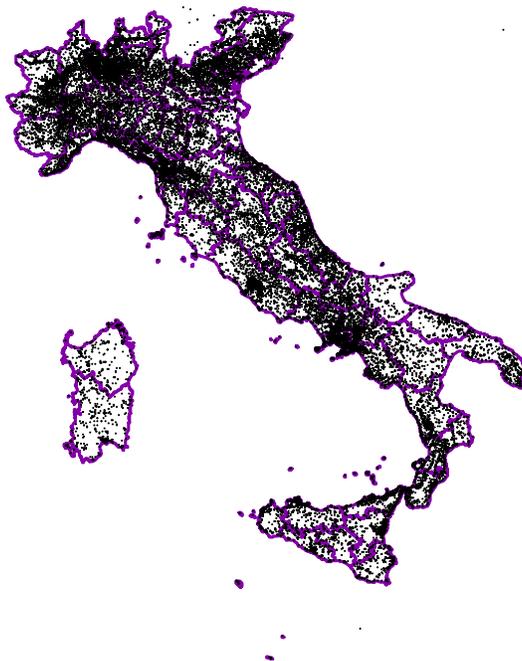
Figures

Figure A.1: Previous place of residence of all geomatched passengers, 1907–1925.

Note: The coordinates are determined through our geolocation algorithm, using the reported last place of residence of each migrant. They are placed on a map created by Martí-Henneberg (2005). The apparent sparsity of points in southern Italy is misleading; there is simply a greater heaping of individuals on certain points, such as Palermo.

(a) Regions of Italy.

(b) Provinces of Italy.

(c) North and south Italy, according to the Bureau of Immigration and Naturalization.

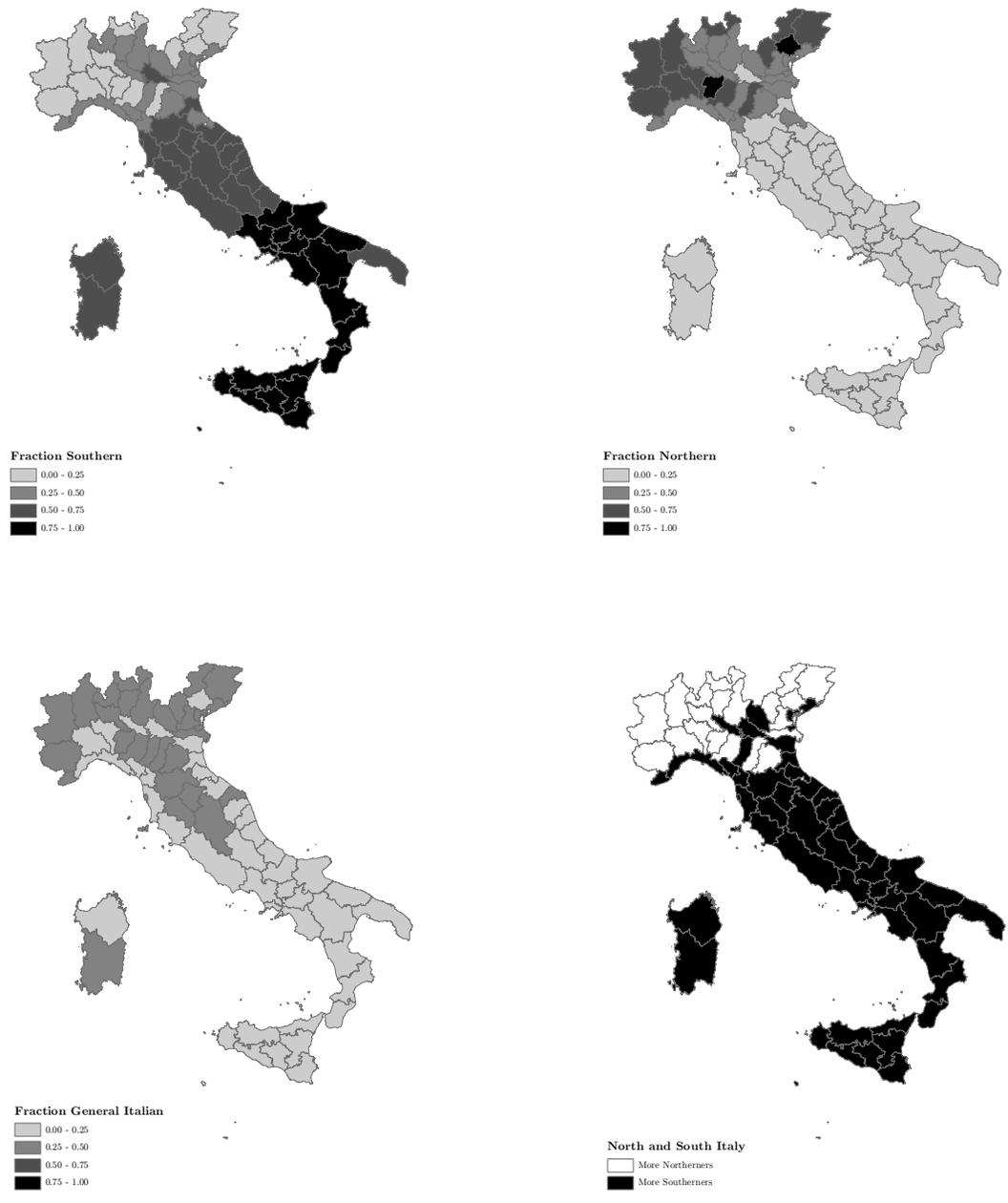Figure A.2: Geographic divisions of Italy.

Figure A.3: Ethnicity of geolocated Italian passengers.

B   The Sexual Dimorphism of Stature

The ratio of the average height of males to the average height of females—termed the sexual dimorphism of stature (SDS)—in a typical population is approximately 1.07 (Gaulin and Boster, 1985; Gustafsson and Lindenfors, 2004; Gustafsson, Werdelin, et al., 2007; Moradi, 2009).[68] In our data, however, we find that the SDS is approximately 1.03. Three interpretations of this finding are possible (as are several of these in combination). First, it is possible that it reflects some sort of error in the collection of stature data in the Ellis Island passenger manifests such that female heights are biased upwards relative to male heights. Second, it is possible that women were self-selected differently from men. However, if the ratio of average male height to average female height in Italy was in the normal range, and if the records of female height in the manifests are accurate, this would imply an astonishingly strong positive self-selection of females. As we show, however, the self-selection of males is of a reasonable magnitude, making such strong positive self-selection among women unlikely.[69] Finally, it is possible that the data are accurate reflections of the SDS in Italy (on which, to our knowledge, there are no data). In particular, Gray and Wolfe (1980) and Wolfe and Gray (1982) argue that there exists an allometric (increasing in the province average stature) relationship between the average stature of a population and the SDS; that is, that the SDS is increasing in the average height of a population.[70] As Italians of this period were rather short, a small SDS may not be entirely unusual.[71] Moreover, we do find evidence of an allometric SDS in our data, as shown in Figure B.1, which is consistent with the data being naturally generated. Since we cannot differentiate between these explanations, and since we have no direct pre-migration data for comparison, we do not include women in our analyses.[72]

References

Boas, Franz (1928). *Materials for the Study of Inheritance in Man*. New York: Columbia University Press.

Danubio, Maria Enrica, Elisa Amicone, and Rita Vargiu (2005). "Height and BMI of Italian Immigrants to the USA, 1908–1970". *Economics and Human Biology* 3, pp. 33–43.

Foerster, Robert F. (1919). *The Italian Emigration of Our Times*. 2nd. New York: Russell & Russell.

Gaulin, Steven and James Boster (1985). "Cross-Cultural Differences in Sexual Dimorphism: Is There Any Variance to be Explained?" *Ethology and Sociobiology* 6, pp. 219–225.

Gray, J. P. and L. D. Wolfe (1980). "Height and Sexual Dimorphism of Stature among Human Societies". *American Journal of Physical Anthropology* 53, pp. 441–456.

Gustafsson, Anders and Patrik Lindenfors (2004). "Human Size Evolution: No Evolutionary Allometric Relationship beteen Male and Female Stature". *Journal of Human Evolution* 47, pp. 253–266.

Gustafsson, Anders, Lars Werdelin, Birgitta S. Tullberg, and Patrik Lindenfors (2007). "Stature and Sexual Stature Dimorphism in Sweden, from the 10th to the End of the 20th Century". *American Journal of Human Biology* 19, pp. 861–870.

---

[68]This is also the case for Sicilians and Central Italians in data collected by Boas (1928).

[69]Such strong positive self-selection is not entirely impossible. Foerster (1919, pp. 380–381) reports that Italian women in the United States were more likely than those of any other immigrant group to be employed.

[70]This view appears to be controversial in the anthropometric literature. Gaulin and Boster (1985) argue that this finding is simply a vestige of small sample sizes, and that the true SDS in all populations is approximately 1.07. Gustafsson, Werdelin, et al. (2007) and Gustafsson and Lindenfors (2004) also find no evidence of an allometric relationship between stature and SDS.

[71]An SDS of 1.03 is still very small (Wolfe and Gray, 1982). Danubio, Amicone, and Vargiu (2005) find an SDS of 1.05 in their data on immigrants., but these data are known to have been subject to considerable measurement error.

[72]In principle, we could convert female heights to male equivalents (Moradi, 2009) and perform the same analysis as we do with men, but doing so requires knowing the SDS in the Italian population.

Moradi, Alexander (2009). "Towards an Objective Account of Nutrition and Health in Colonial Kenya: A Study of Stature in African Army Recruits and Civilians, 1880–1980". *The Journal of Economic History* 69:3, pp. 719–754.

Wolfe, L. D. and J. P. Gray (1982). "A Cross-Cultural Investigation into the Sexual Dimorphism of Stature". In: *Sexual Dimorphism in* Homo sapiens*: A Question of Size*. Ed. by R. L. Hall. Praeger.
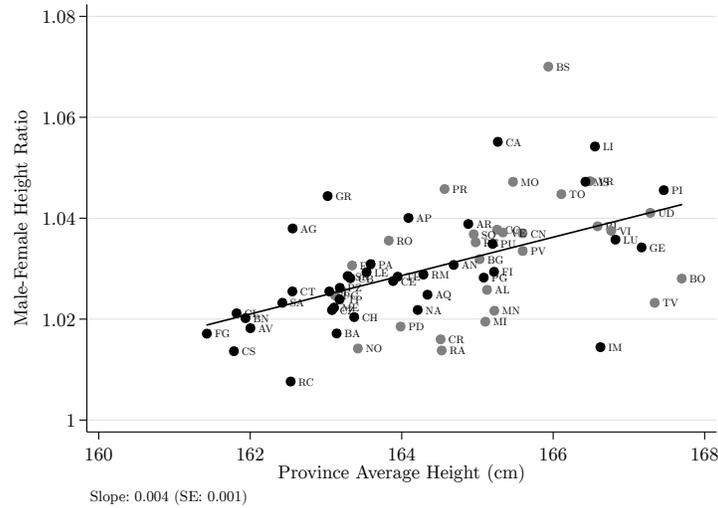
Figures

Figure B.1: Allometric relationship between the SDS and average male height.

Note: Province average heights are averaged within provinces across birth cohorts by the number of passengers in our sample.

C Proof of Expression (2)

The purpose of this appendix is to present a formal proof for the assertion put forward in section 5.1 that, conditional on the degree of national self-selection, the predicted output is increasing with respect to the degree of local self-selection.

Let the production functions of heights $h(\mu, z)$ and of host-country outcomes, that take the two arguments μ_{jt} (local inputs) and z_{ijt} (individual quality) as described in section 5.1, be continuously differentiable, strictly increasing in both arguments. Denote by $\mu_{\tilde{h}}(z)$ the inverted function mapping the local degree of self-selection z to the local inputs μ, conditional on a given national degree of self-selection $\tilde{h}$. More formally, $\mu_{\tilde{h}}(z)$ is defined as the function that for a given $\tilde{h}$ satisfies the equality

$$\tilde{h} = h(\mu_{\tilde{h}}(z), z) - \mu_t. \tag{C.1}$$

Denote by $\hat{w}(\tilde{h}, z) = w(\mu_{\tilde{h}}(z), z)$ the function predicting w conditional on a given $\tilde{h}$ and on z. Finally, denote the marginal rate of technical substitution of the height function by $\mathrm{MRTS}^h(\mu, z) = \frac{\partial h(\mu, z)}{\partial \mu} / \frac{\partial h(\mu, z)}{\partial z}$, and similarly denote $\mathrm{MRTS}^w(\mu, z)$.

Claim.

$$\frac{\partial \hat{w}(\tilde{h}, z)}{\partial z} > 0 \iff \mathrm{MRTS}^h(\mu, z) > \mathrm{MRTS}^w(\mu, z)$$

Proof. By the implicit function theorem and the definition of $\mu_{\tilde{h}}(z)$ in equation (C.1), we have that

$$\frac{\partial \mu_{\tilde{h}}(z)}{\partial z} = -\frac{\frac{\partial h(\mu, z)}{\partial z}}{\frac{\partial h(\mu, z)}{\partial \mu}}. \tag{C.2}$$

Differentiating the function $\hat{w}(\tilde{h}, z)$ with respect to z gives

$$
\begin{aligned}
\frac{\partial \hat{w}(\tilde{h}, z)}{\partial z} &= \frac{\partial w(\mu_{\tilde{h}}(z), z)}{\partial z} \\
&= \frac{\partial w(\mu, z)}{\partial \mu} \times \frac{\partial \mu_{\tilde{h}}(z)}{\partial z} + \frac{\partial w(\mu, z)}{\partial z} \\
&= -\frac{\partial w(\mu, z)}{\partial \mu} \times \frac{\frac{\partial h(\mu, z)}{\partial z}}{\frac{\partial h(\mu, z)}{\partial \mu}} + \frac{\partial w(\mu, z)}{\partial z} \\
&= \frac{\partial w(\mu, z)}{\partial \mu} \left(-\frac{1}{\mathrm{MRTS}^h(\mu, z)} + \frac{1}{\mathrm{MRTS}^w(\mu, z)} \right),
\end{aligned} \tag{C.3}
$$

where expression (C.3) follows from equation (C.2). Since by assumption $\frac{\partial w(\mu, z)}{\partial \mu} > 0$, we have that $\frac{\partial \hat{w}(\tilde{h}, z)}{\partial z} > 0$ if and only if $\mathrm{MRTS}^h(\mu, z) > \mathrm{MRTS}^w(\mu, z)$. $\qquad\square$

D   Province Abbreviations and Sample Sizes, by Region

Province	Abbreviation	N
Abruzzi		
Campobasso	CB	183
Chieti	CH	171
L'Aquila	AQ	175
Teramo	TE	157
Apulia		
Bari	BA	361
Foggia	FG	201
Lecce	LE	92
Basilicata		
Potenza	PZ	199
Calabria		
Catanzaro	CZ	211
Cosenza	CS	178
Reggio Calabria	RC	236
Campania		
Avellino	AV	212
Benevento	BN	111
Caserta	CE	344
Napoli	NA	204
Salerno	SA	206
Emilia		
Bologna	BO	26
Ferrara	FE	15
Forli-Cesena	FC	27
Modena	MO	31
Parma	PR	41
Piacenza	PC	20
Ravenna	RA	17
Reggio Emilia	RE	33
Latium		
Roma	RM	458
Liguria		
Genova	GE	103
Imperia	IM	3
Lombardy		
Bergamo	BG	49
Brescia	BS	41
Como	CO	63
Cremona	CR	5
Mantova	MN	19
Milano	MI	102
Pavia	PV	27
Sondrio	SO	23

Province	Abbreviation	N
Marches		
Ancona	AN	27
Ascoli Piceno	AP	56
Macerata	MC	29
Pesaro e Urbino	PU	84
Piedmont		
Alessandria	AL	87
Cuneo	CN	65
Novara	NO	79
Torino	TO	153
Sardinia		
Cagliari	CA	17
Sassari	SS	44
Sicily		
Agrigento	AG	278
Caltanissetta	CL	168
Catania	CT	220
Messina	ME	224
Palermo	PA	409
Siracusa	SR	238
Trapani	TP	223
Tuscany		
Arezzo	AR	22
Firenze	FI	90
Grosseto	GR	18
Livorno	LI	4
Lucca	LU	128
Massa-Carrara	MS	40
Pisa	PI	24
Siena	SI	23
Umbria		
Perugia	PG	131
Venetia		
Belluno	BL	28
Padova	PD	23
Rovigo	RO	4
Treviso	TV	99
Udine	UD	137
Venezia	VE	24
Verona	VR	35
Vicenza	VI	93