Migrant Self-Selection

Anthropometric Evidence from the Mass Migration of Italians to the United States, 1907–1925

Yannay Spitzer
yannay.spitzer@huji.ac.il

Hebrew University of Jerusalem

Ariell Zimran
ariell.zimran@u.northwestern.edu

Northwestern University

October 23, 2015

Click Here for Most Recent Version

Abstract

Are migrants positively or negatively self-selected from within their populations of origin? We study this question by comparing the heights of Italian immigrants to the United States between 1907 and 1925 to the height distributions of their respective birth cohorts in their provinces of origin. We created a novel data set based on 3.2 million Italian passengers entering Ellis Island whose last place of residence was geo-located, and on further transcription of stature and other personal information of a random subsample of 88,000 individuals. Interpreting stature as a measure of migrants' "quality," we find that the Italian migration was (a) negatively selected at the national level; but (b) positively selected at the local level; and (c) the selection varied systematically within the country, with more positive selection from shorter and poorer provinces. We show that the systematic variation in selection was not spuriously generated by differential geo-location probabilities or by non-classical measurement error caused by random errors in geo-location. We find that the patterns of selection were partly, but not entirely, driven by the post-1917 literacy requirement; they also cannot be explained by variations in rural-urban or occupational composition, though we cannot rule out that to some extent they are affected by selection across destinations. Our findings are consistent with theories of migration that highlight the importance of liquidity constraints that are partly solved by the support of networks of friends and relatives. They also highlight the significance of the distinction between selection at the local and the national levels.

Acknowledgements We are indebted to Joel Mokyr, Joseph Ferrie, Igal Hendel, and Matthew Notowidigdo for encouragement and guidance. We are grateful to Peg Zitko and the Statue of Liberty-Ellis Island Foundation for providing the Ellis Island arrival records data; to Brian A'Hearn, Franco Peracchi, and Giovanni Vecchi for sharing their computed moments of Italian stature distributions; and to Jordi Martí-Henneberg for sharing historical GIS files of Italy. We also thank Ran Abramitzky, David Figlio, Luigi Guiso, Timothy Hatton, Richard Hornbeck, Seema Jayachandran, John Komlos, Lee Lockwood, Andrea Matranga, Aviv Nevo, Paola Sapienza, Marian Smith, Richard Steckel, and Zachary Ward for helpful suggestions and insightful comments. Christine Chang and Elizabeth Nelson provided excellent research assistance. Thanks are also to Roy Mill for giving us access to the dEntry transcription system and for investing considerable time and energy adjusting it to our needs; to seminar participants at Northwestern University and the London School of Economics; and to conference participants at the 2014 Warwick Economics PhD Conference, the 2014 Cliometrics Conference, the 2014 Economic History Association Annual Meetings, the 2014 Illinois Economic Association Annual Meeting, and the 2014 Social Science History Association Conference. Financial support was provided by the Northwestern University Economics Department's Eisner Fund, the Northwestern University Center for Economic History, an Exploratory Data and Travel Grant from the Economic History Association, and an Economic History Association Dissertation Fellowship. This research was supported in part through the computational resources and staff contributions provided for the Social Sciences Computing Cluster (SSCC) at Northwestern University. Recurring funding for the SSCC is provided by the Office of the President, Weinberg College of Arts and Sciences, Kellogg School of Management, the School of Professional Studies, and Northwestern University Information Technology. This material is based upon work supported by the National Science Foundation under Grant No. SES-1425598. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. All errors are our own.

Notes A previous version of this paper was titled "Self-Selection of Immigrants on the Basis of Living Standards: Evidence from the Stature of Italian Immigrants at Ellis Island, 1907–1925."

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, vol. 1, p. 24)

Between 1892 and 1925, nearly four million Italians immigrated to the United States—the largest flow from a single country during the Age of Mass Migration. This phenomenon, part of the massive contemporaneous growth in migration to the United States from southern and eastern Europe, sparked a debate over the policy of nearly total openness of the United States to immigration. A central issue in this debate was the "quality" of the new migrants.[1] Proponents of immigration restriction argued that these immigrants represented the poor, incapable, uneducated, and unskilled elements of their countries of origin; that is, that they were negatively selected from within their populations of origin. Thus, the question of "who migrates?" took center stage in debates over immigration policy, a position that it maintains in modern contexts and in the contemporary literature on the economics of migration (Abramitzky, Boustan, and Eriksson, 2012, 2013; Borjas, 1987, 2014; Chiquiar and Hanson, 2005; Fernández-Huertas Moraga, 2013; Grogger and Hanson, 2011; Hatton and Williamson, 1998; McKenzie and Rapoport, 2010).

Interest in migrant selection is typically motivated by the view that the quality composition of migrants relative to their non-migrating peers is crucial in determining the effects of migration on the sending and receiving countries. Selection is typically measured relative to the national distribution of some quality measure in the country of origin. For example, if immigrants are on average more educated than their peers in the origin country as a whole, it is said that immigrants are positively selected on the basis of education. On the other hand, selection within local environments and the variation of such local selection across regions are rarely observed by economists or noticed by policy makers, but they are potentially important. In this paper we document a case in which immigrants were negatively selected from within their country as a whole, primarily drawn from underdeveloped provinces, and yet positively selected from within their local environments. That is, if compared to the nationwide reference group, they would have been judged to be of low quality; however, some characteristics that they possessed enabled them to excel within their populations of origin and could constitute an important part of the human capital transfer resulting from migration.

We study self-selection into migration among Italians arriving in the United States during the Age of Mass Migration (in particular between 1907 and 1925). We measure migrants' quality by their height, and

[1] We use the term "quality" here and throughout this paper to refer to any traits that affect or are positively correlated with an individual's productive capacity. Examples include education, skill, health, wealth, and intelligence.

quantify migrant selection by making comparisons to the height distributions of their cohorts of origin. This approach is grounded in a large body of research that has established that the average stature of a large group is indicative of the group's standards of living and its productive capacity (Komlos and Meermann, 2007). Height is thus a proxy for many characteristics, such as occupational 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, 1994; Jayachandran and Pande, 2015; Steckel, 1995), and cognitive ability (Case and Paxson, 2008), that all play a part in determining a prospective migrant's contribution to his home economy and his labor market outcomes in the host economy. Thus, comparisons of the stature of migrants to the stature distribution of the population at risk for migration are informative as to differences in various aspects of quality between migrants and stayers. That is, the premise of our empirical analysis is that, the taller are migrants relative to their populations of origin, the more positive is their selection into migration on the basis of characteristics that are important to policy makers and economists.

We construct a novel data set consisting of the stature, place of origin, and other personal information of Italian passengers indexed in the complete Ellis Island arrival records database. First, we geo-located the last place of residence of roughly 3.2 million Italian passengers out of a population of nearly five million. Next, we randomly sampled approximately 88,000 Italian passengers arriving at Ellis Island between 1907 (when information on migrants' stature was first recorded) and 1925, and transcribed their stature and other personal information. We then linked migrants to the population distribution of stature of their birth cohort in their province of origin, based on military records that provide nearly universal coverage of each birth cohort in each Italian province during the period (A'Hearn, Peracchi, and Vecchi, 2009; A'Hearn and Vecchi, 2011).[2] Finally, we complement these data with official Italian statistical data on emigration and other local characteristics at the level of the province, district, and township.

Our 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, indicating negative self-selection on the national level. However, this result is driven by the overrepresentation of migrants from Southern Italy, where average heights were below the Italian average, in the migratory flow to the United States. When compared only to the distributions of stature of their respective birth cohorts in their own provinces of origin, we find that Italian passengers were, on average, taller than their local birth cohort,

[2] All Italian males were required to present themselves for physical examination (including measurement of height) by the military. This data source therefore avoids the common problem that military data may not be representative of the population of interest (Bodenhorn, Guinnane, and Mroz, 2013, 2014, 2015; Mokyr and Ó Gráda, 1996; Zimran, 2015).

indicating positive self-selection on the local level.

Furthermore, Italy demonstrated a clear pattern of systematic variation of local self-selection across regions and provinces. We find that South Italians were considerably more positively self-selected on the local level than Northerners. Similarly, within regions (i.e., North and South), shorter province-cohorts tended to be the sources of more positively self-selected migrants. Thus, our three benchmark findings are as follows: negative selection at the national level; positive selection at the provincial level; and a tendency towards more positive selection from shorter province-cohorts. As a result, the most positively selected migrants, and therefore those potentially endowed with the best human capital, came from the poorest and least developed places of origin, and may thus not appear to be of particularly high quality on the absolute level, or when compared to the national average. The magnitude of the systematic variation in local self-selection that we measure is large. In particular, according to a back-of-the-envelope calculation, an improvement in quality associated with a one-standard deviation increase in average stature leads to an increase in migration probability of as much as 1.8 percent, relative to a base migration probability (at any time during our study period) of about six percent.

Despite the remarkable advantages of our data in terms of coverage, quality, representation, and disaggregation, there exist a number of potential threats to the validity of our findings. We discuss these threats and, to the extent possible, test the robustness of the patterns revealed by the data to them. First, while we find no evidence or reason to suspect that there was a systematic upward bias in heights recorded on the passenger manifests, we were unable to find information that would positively refute this possibility. Hence, we advise some caution in the interpretation of our findings regarding the national and local magnitudes of self-selection; however, the validity of our finding of systematic variations in the degree of selection throughout Italy is not threatened by this factor.

Second, the systematic variation of selection with province-cohort average height could have been spuriously generated by our geo-location algorithm through two separate mechanisms. The first is systematic variation in the probability of successful geo-location, and the second is simply through random errors in geo-location, which we show to be a possible cause of non-classical measurement error that might mechanically generate the cross-province patterns of average selection that we find. To test for differential selection into geo-location, we employ an alternative geo-location procedure that uses passengers' surnames. We find that among passengers who were geo-located through the alternative procedure, there is no evidence that the probability of geo-location in our primary algorithm varies in such a way that would spuriously generate our results. To evaluate the threat from errors in geo-location, we show that the rate of incorrect geo-location needed in order to mechanically generate our results is implausibly large, and that the our results are also present in sub-samples in which geo-location errors were particularly unlikely to have occurred.

We supplement this analysis by investigating several potential mechanisms behind our results. First, we test whether the patterns discussed above were generated solely by migration after 1917, when the new literacy requirement was put in place. We find that, as expected, this requirement, which disproportionately affected the less literate provinces, caused the degree of positive self-selection in such provinces to increase considerably. It led to greater positive self-selection after 1917 than before as well as to a widening of the local selection gap between North and South. In principle, this shift could have been entirely responsible for our findings of positive local selection and systematic variation in local selection. Indeed, the finding of all-Italian positive selection is not robust to restricting our sample to pre-1917 arrivals. However, positive local selection in the South alone, as well as the stronger positive selection from Southern and shorter provinces, were present and statistically significant prior to 1917, and were merely strengthened following the introduction of the literacy requirement. To our knowledge, this is the first systematic analysis of the effects of the 1917 restrictions on migrant self-selection. It shows that this policy that targeted one measure of quality was indeed effective in improving the selection of immigrants in terms of other measures—stature and occupation.

Next, we rule out the possibility that our results are driven by differences in the occupational composition of migrants, or by a combination of differences in their urban-rural composition and an urban height penalty. Indeed, our main results hold within each of these sectors. We also show that our results hold when taking into account the presence of multiple destinations for Italian emigrants, though we cannot rule out that the systematic variation in selection was partly driven by a possible tendency for relatively taller Northerners to migrate to non-US destinations. It is therefore possible that US-bound migration from the North may have been more poorly selected than that from the South simply due to selection across destinations, but not due to selection into emigration to any destination. Thus, we interpret our results as applying to migration from Italy to the United States, rather than to emigration from Italy in general. Finally, we evaluate evidence on various theoretical mechanisms that have been proposed to predict migrant self-selection. We find that individuals who report joining an immediate family member in the United States are more negatively selected, supporting models highlighting the combination of migration costs and migrant networks in determining the degree of migrant selection.

This paper complements a large literature that studies migrant self-selection, seeking both to quantify it and to understand its determinants. Most studies of contemporary migrant self-selection focus on the migration of Mexicans to the United States (e.g., Chiquiar and Hanson, 2005). Fernández-Huertas Moraga (2013), however, provides an important critique of this literature, arguing that the data used in such studies often suffer from incomplete coverage and sample-selection problems that often lead to qualitatively contradictory results. Recent improvements in the availability of historical data and in matching methods have enabled the circumvention of many of these limitations in studies using data from the Age of Mass Migration. Pioneering work by Abramitzky, Boustan, and Eriksson (2012, 2013) uses linkage between Norwegian and American censuses to create a data set of unprecedented quality, and measures migrant selection through comparison of occupations and wealth.

We make two primary contributions to this literature. First, we present a study of migrant selection that is supported by data of unusual clarity and completeness. The data satisfy five criteria that we identify as being essential to clean measurement of migrant selection: the two sources of height data are representative of the migrating population and the population at risk for migration, measuring quality for each group with minimal scope for selection biases; they are measured at the individual level; they capture a continuous and ordered measure of quality; they are (for those who have reached terminal height by the time of migration) unaffected by migration and immutable in preparation for migration; and the migration occurred in a period that was relatively free of restrictions on migration, making it possible to learn the supply of migrants without contamination by policy that differentially favors migrants of different levels of quality.[3] It is quite rare for all of these characteristics to be satisfied by a single data set.

Our second contribution is to highlight the importance of distinguishing between selection from a country as a whole and selection from within local, sub-national, environments. Thanks to the fine disaggregation of our data on the Italian population and our ability to link migrants to their localities of origin, we show that the two levels of selection can be qualitatively different, and that there is considerable and systematic variation in the degree of local selection within countries. As a result, comparisons of migrants to their national-level populations of origin only may fail to capture a significant portion of the selection occurring within a group of potential migrants. The epigraph to this paper highlights this distinction and helps to raise a question of interpretation. In particular, does knowledge of a migrant's degree of local selection provide information over and above his absolute quality, or his degree of national selection? While an empirical answer to this question is beyond the scope of this paper, we present a simple theoretical framework through which we argue that, under a reasonable condition, this is in fact the case, and therefore that it is important to consider local selection when attempting to predict migrants' labor market outcomes in the receiving

[3] With the exception of the 1917 literacy requirement, as explained above.

economy.

The remainder of the paper proceeds as follows. Section 2 provides the relevant historical and economic background for this paper. Section 3 discusses the construction of the data set and provides summary statistics. Section 4 presents and interprets the main results. Section 5 discusses and addresses possible threats to validity. Section 6 examines the mechanisms driving the patterns of selection that we observe. Section 7 concludes.

2 Background

2.1 A Brief Historical Background

In 1896 Italy surpassed Germany and Ireland to become the source of more immigrants to the United States per year than any other country—a position that it held (except for occasional surges by Russia and Austria-Hungary) until the closing of the Golden Door and the end of mass migration to the United States (Ferenczi and Wilcox, 1929, Table III).[4] Throughout the period 1896–1924, Italy also led all European countries in per capita emigration. This mass outflow was clearly motivated, at least in part, by a desire to escape the relatively poor standards of living faced in Italy in favor of better conditions in the United States. In particular, nominal wages in Italy were as low as one-fifth of those in the US, and Italian real wages were at only half the level of those in the industrialized economies of Western Europe (Hatton and Williamson, 2005, Table 4.2). Italy's under-development was also manifest in the height of its population, with the average Italian male markedly shorter than his peers in the United States and Western Europe (Hatton and Bray, 2010).

Though similar in motivation and (at least generally speaking) in magnitude to the migratory flows from other Southern and Eastern European countries, the Italian migration to the United States had several unique characteristics. First, Italians were more likely than migrants of any other nationality to engage in seasonal or temporary migration (Bandiera, Rasul, and Viarengo, 2013). Second, to a greater extent than the flow of any other nationality, Italian migration was divided between multiple destination countries. Although the United States became the most popular destination for Italians by the turn of the twentieth century, attracting 38 percent of migrants between 1900 and 1920 (Table A.1, other destinations, such as countries in Western Europe, South America, and the Mediterranean, remained popular. Emigrants from Northern Italy were particularly likely to travel elsewhere, with more than half of all Northern emigrants

[4] The discussion in this section is primarily based on Foerster (1919), Hatton and Williamson (1998, Ch. 6), and Gomellini and Ó Gráda (2013).

remaining on the continent. The turn towards the United States at the beginning of the twentieth century also coincided with a shift in the geographic sources of the Italian immigration to the US. In particular, the North Italian dominance of migration to the United States—which had existed previously—ended, and was surpassed by the surging migration from the poorer and less-developed Italian South.

That the number of migrants from Italy surged considerably around the turn of the century, and that they came increasingly from Italy's poorer regions, did not go unnoticed by the American public. Indeed, this growth was symptomatic of the larger process by which the sources of European emigration shifted from the relatively richer Northwest towards the poorer peripheral countries of Southern and Eastern Europe. This "new immigration" fueled nativist anxieties. A commonly expressed fear was that the United States was the recipient of Europe's undesirables—the unskilled, the uneducated, and the mentally and physically disadvantaged; in other words, that these migrants were negatively selected from their populations of origin.[5] The Commissioner-General of Immigration (1903, p. 73) expressed a viewpoint typical among many anti-immigration advocates.

> 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 [to the United States] in large numbers; but the point is that such better element does not come.

Whether or not the new immigrants from the European periphery were indeed negatively self-selected from within their populations of origin was a question of utmost importance in debates over immigration policy.

2.2 The Economics of Migrant Self-Selection: Theory and Evidence

A number of theoretical models have been proposed to explain and predict the factors determining migrant self-selection. The canonical model (henceforth, the "relative inequality model") is due to Borjas (1987), who applies Roy's (1951) model to argue that the relative returns to skill (or education) determine selection into migration. The greater are the returns to migrants' skill in the destination country relative to the sending country, the more positive will be the selection of migrants. Empirically, the fact that the wages of the skilled relative to those of the unskilled tend to be smaller in richer countries than in poorer countries leads

[5] See, for example, Hall (1904). On the political economy of the anti-immigration movement see Goldin (1994). Supporters of immigration restriction in the early 20th century often took a very broad stance on reasons for concern at the new immigration, arguing, for example, that the new immigrants were more likely to be involved in criminal activity, or that they lacked a history of self-governance that would be crucial to their assimilation in the United States.

this model to generally predict negative selection into migration between poor and rich countries in modern settings. An alternative set of models predicts positive selection into migration, either in all settings or in migration from poor to rich countries (such as Italy to the United States during the age of Mass Migration). Chiswick's (1978, 1999) human capital migration model treats migration as an investment; to the extent that part of the costs of migrating do not vary by skill (or education) levels, migration relatively is more rewarding for the higher skilled (or better educated), leading to positive self-selection. Grogger and Hanson (2011) offer a similar prediction ("generalized positive selection;" Borjas, 2014) based on a model that bases selection on absolute wage differences between the sending and receiving countries (rather than the relative differences as in the relative inequality model).[6] The tendency for the absolute difference in skilled and unskilled wages to be larger in richer countries (Hanushek and Zhang, 2009; Psacharoupolos and Patrinos, 2004) thus leads to predictions of positive self-selection into such flows.[7]

These two theories are most often tested in studies of modern migration from Mexico to the United States. A growing series of empirical studies has yet to reach a consensus on whether the Mexico-US migration is indeed positively or negatively self-selected (Chiquiar and Hanson, 2005; Feliciano, 2005; Fernández-Huertas Moraga, 2011, 2013; Ibarraran and Lubotsky, 2007; Kaestner and Malamud, 2014; McKenzie and Rapoport, 2010; Mishra, 2007; Orrenius and Zavodny, 2005). A number of other studies have been performed focusing on cross-source country variations in the degree of self-selection. Results are mixed, with Feliciano (2005) and Stolz and Baten (2012) finding evidence that supports the relative inequality model and Docquier and Marfouk (2006) and Grogger and Hanson (2011) finding evidence supporting the generalized positive selection model.

Other mechanisms affecting migrant self-selection have also been proposed. For example, the presence of migration networks may play an important role in predicting patterns of migrant self-selection. Relatively disadvantaged individuals are more likely than their more advantaged peers to face liquidity and credit constraints, preventing them from financing migration and driving selection to be more positive regardless of the underlying incentives generated by the models described above. However, this problem is ameliorated when prospective migrants are linked to friends and relatives who have already migrated and can help to reduce migration costs or to provide liquidity. Hence, thicker networks of past migrants are expected to

[6] To illustrate the differences in these two models' predictions, consider the following example, which abstracts from many of the details, including migration costs. Suppose that low skill wages in the poor country are $1 and that high skill wages are $2. Suppose that low-skill wages in the rich country are $10 and that high skill wages are $15. Then the larger absolute gap in high-skill wages leads Grogger and Hanson's (2011) model to predict positive selection into migration from the poor country to the rich country, while the smaller relative gap leads Borjas's (1987) model to predict negative selection. It should be kept in mind that these models predict the composition of a migrant flow rather than its size.

[7] The key difference between Grogger and Hanson's (2011) model and Borjas's (1987) model is technical, but strong enough to drive these theoretical differences: the former assumes linear utility of income, whereas the latter assumes logarithmic utility.

shift the distribution of migrants' quality to the left.[8] Several studies have found evidence confirming the importance of this mechanism (Beine, Docquier, and Özden, 2011; Belot and Hatton, 2012; Fernández-Huertas Moraga, 2013; McKenzie and Rapoport, 2007, 2010), particularly in the age of mass migration (Spitzer, 2015a,b; Wegge, 1998). Bryan, Chowdhury, and Mobarak (2014) cite risk aversion as an additional determinant of migrant self-selection (see also Harris and Todaro, 1970). Higher wealth households in their model have lower risk aversion and are better able to bear the risks inherent to migration, leading to more positive self-selection. Gould and Moav (2015) provide another prediction in a model stressing the different effects of general and local skills on migration probability. Overall, consensus on the mechanisms determining migrant selection remains elusive.

2.3 Data Requirements

To properly measure the degree of migrant self-selection, the data to be analyzed should ideally satisfy five conditions that jointly enable accurate and precise location of migrants in the quality distribution of their place of origin. First, the samples of individuals at risk for migration and of actual migrants must be representative of their respective populations and must provide the same measure of quality for both groups.[9] Second, the measurement of the index of quality should not be affected by the migration itself. This requirement takes two forms. In the first, the measure must not change after or as a result of migration.[10] In the second, the measure must not be manipulated in anticipation of migration. Third, the measure of quality should preferably not be too coarse,[11] and its distribution should be observed and compared within bins that are as fine as possible. This is particularly important for two reasons. In a trivial sense, migrants are positively selected in almost all cases simply because they disproportionately tend to be young adults, who are more productive and often more literate than older and non-labor force-participating prospective migrants. The non-trivial task is to quantify selection from within birth cohorts. Moreover, small geographic bins are important in order to differentiate between variation in the distribution of quality across regions and

[8] Assortative matching in forming networks has the potential to drive networks to make selection more positive. That is, if there is positive selection into network formation, the effect of networks in aiding migration may ease the migration of positively selected individuals. We will find evidence below consistent with the view of networks supporting more negative selection, however; the concern over assortative matching can thus be viewed as biasing our results towards zero.

[9] This condition is violated, for example, by the commonly used data from the Mexican Migration Project, which is not a nationally representative sample of Mexican households. Modern US censuses face similar problems in that they undercount undocumented migrants. Fernández-Huertas Moraga (2013) and Kaestner and Malamud (2014) discuss these issues in detail.

[10] For example, some studies use measures of education taken after several years in the country of destination (e.g., Chiquiar and Hanson, 2005; McKenzie and Rapoport, 2010). These studies restrict the age at migration to adulthood, which partly but not entirely removes the risk of measuring education that was affected by the migration itself.

[11] Typical coarse measures are binary indicators such as literacy, or a small ordered set of categories, such as educational attainment or broad occupational status. A common problem is that occupations may vary very little among migrants. For example, Kosack and Ward (2014) find that over 90 percent of Mexican migrants to the United States in their sample reported their occupations as either laborers or miners.

selection from within a given economic environment. Fourth, the data should ideally provide information at the individual level.[12]

A fifth condition is particularly important and rarely satisfied. In order to uncover the features of the underlying supply of migrants, data must be generated by voluntary and legally unfettered migration. If significant legal restrictions apply to migration, then what is observed is not the migration flow generated by economic conditions and the underlying model, but a masked and distorted rendering of this flow.[13] Moreover, restricting migration leads to avoidance of restrictions, resulting in undocumented—and therefore difficult to study—migration. The prevalence of legal barriers to migration in modern contexts implies that unrestricted flows exist only for migratory flows that are not particularly representative of the main migration movements of interest to economists.[14] In contrast, in the Age of Mass Migration, immigration to the US and to some other destination countries was largely unfettered and well documented, providing an excellent setting for studying migrant self-selection.[15]

2.4 Evidence from the Age of Mass Migration

A number of recent studies have benefitted from the advantageous historical context of the Age of Mass Migration and the data it offers, as well as from the increasing availability of individual-level data at a large enough scale to make possible wide coverage and representation of entire populations. Abramitzky, Boustan, and Eriksson (2013), in a study of internal and international migration among Norwegians, link census and tax roll data from individuals' childhood households to census records in adulthood for both movers and

[12] Though we focus on height as our preferred measure of quality in this paper, we also observe occupation in both the passenger records and at a fairly disaggregated level in census records. However, we do not observe the occupational distribution by birth cohort as we do with heights, and thus cannot compare migrants to the appropriate distribution of quality from which they are drawn. Moreover, a very large fraction (over 70 percent) of migrants report an occupation of either "farmer," "peasant," "farm laborer," or "laborer," providing very little useful variation. Finally, we observe no variation within these classes. This is particularly a problem within the farmer category, which in reality bundles a variety of skills and social statuses. The Italian censuses do provide distributions of farmer sub-classes (e.g., sharecropper, owner-occupier, etc.), but unfortunately these distinctions are not made in the Ellis Island data.

[13] To see this, suppose that immigration restrictions were such that only highly educated individuals were admitted. Then regardless of the underlying supply of immigrants, all observed migration would be positively selected on the basis of education. In this sense, restrictions distort the supply of migrants, the structure of which is what economists and policy makers are fundamentally interested in.

[14] Cases of selection studies on contemporary unrestricted migration dealt with the interesting but esoteric flows of Micronesians to Guam and Hawaii (Akee, 2010), Tongans to New Zealand (McKenzie, Gibson, and Stillman, 2010), and Finns to Sweden (Rooth and Saarela, 2007). Borjas, Kauppinen, and Poutvaara (2015) analyze modern Danish emigration using administrative data, thus satisfying all of the conditions identified above. This migration of a very small fraction of the population of a very wealthy country is potentially quite different from the case of mass migration from relatively poor countries that more commonly comes to the attention of economists and policy makers.

[15] Italy did have restrictions on the emigration of men of military age (Cole, 1995); however, we study only men aged 22 or older, who would not have been bound by this requirement. To be sure, there would have been delayed or foregone migration among those who might have chosen to migrate during the age of military service if they had been permitted to do so, but we suspect that the distortions generated by these effects would have been minor.

stayers.[16] They find evidence strongly suggesting that the availability of household wealth substitutes for rather than complements migration. That is, contrary to the widely held view, the effect of access to wealth or credit that would finance the costs of migration seems to have been dominated by the better conditions gained from wealth in the home country, leading to negative selection on the basis of wealth. In the same context (but with an indirect measure), Abramitzky, Boustan, and Eriksson (2012) find evidence of negative self-selection into migration from urban areas on the basis of occupation. Connor (2015) applies a similar census linkage approach to Irish migration to the United States in the early twentieth century. Measuring selection by occupational status, he finds evidence of positive selection. Using an approach similar to ours, Kosack and Ward (2014) compare the heights of Mexican immigrants to the United States in the 1920s to height distributions of two selected Mexican populations—volunteer soldiers and passport applicants. Migrants were much taller than soldiers, but only slightly shorter than passport applicants, which is argued to imply positive selection into migration.

In the case of Italy, the human capital and generalized positive selection models predict that widespread poverty (and the resulting liquidity constraints) would cause positive selection. Furthermore, according to Betrán and Pons (2004), skill premia were in fact higher in the US relative to Italy, which would also imply more positive selection according to the relative inequality model.[17] Contemporary analyses, however, did not see the Italian migration in this light. Foerster (1919), in a seminal account of the Italian migration, described Sicilian migrants as more intensely drawn from the poorer rural classes,[18] although he also praised the superior qualities of certain groups of Italian emigrants.[19] Empirical evidence is scant, coarse, and mixed. Stolz and Baten (2012) find that age heaping among Italian migrants was greater than among the origin population of the country as a whole, suggesting negative selection. Similarly (but unlike immigrants from some other western European countries), Italian immigrants were found to be less literate than the Italian population on average (Hatton and Williamson, 2005, Table 5.3). There is little doubt, however, that such evidence for countrywide negative selection merely reflects the over-representation of Southerners (Hatton and Williamson, 2005, p. 93; see discussion in Section 4.2). On the other hand, anthropometric evidence

[16] See also Wegge (1999, 2002), who studies a mid-nineteenth century sample of emigrants from German villages. Biavaschi and Elsner (2013) also study the case of the Norwegian migration.

[17] Interestingly, Italy was more unequal than the United States in terms of the height distribution (Blum and Baten, 2011). However, the predictions of the relative inequality model are based on the difference in skill premia, for which relative inequality is simply a convenient and widely available proxy.

[18] "...the movement in Sicily has selected especially from the day laborers, next those associated with the mezzadria [share-cropping] and rent contracts, lastly the small proprietors" (Foerster, 1919, p. 104).

[19] "Those who have studied the exodus from country to city have generally averred that it is the more active spirits that participate. In a large view the affirmation must apply as well to a general emigration. To venture one's all—even if that be little—amid incalculable perils and the friendlessness of differing peoples requires a certain staunchness of soul that many men lack. In energy and prowess the emigrant must run well ahead of his sessile neighbor" (Foerster, 1919, p. 419).

by Danubio, Amicone, and Vargiu (2005) found that a sample of Italian applicants for US citizenship in Massachusetts was on average taller than the average Italian. To the best of our knowledge, general evidence on the local degree of selection from Italy does not exist yet, and moreover, this is the first study to examine variations in the degree of local self-selection into migration within a country.

2.5 Stature as a Measure of Pre-Migration Quality

We use stature as a proxy measure for migrants' quality, an approach intuitively similar to a first principal component that is correlated with a variety of quality measures of interest for economists and policy makers. It is well established that adult stature reflects standards of living during childhood and adolescence (Eveleth and Tanner, 1976; Floud, 1985; Fogel, 1986; Fogel, Engerman, and Trussell, 1982; Komlos and Meermann, 2007; Steckel, 1995, 2008).[20] While individual height is overwhelmingly determined by idiosyncratic genetic factors, environmental conditions in the first 18 to 22 years of life are also important determinants (Eveleth and Tanner, 1976; Frisancho, 1993; Martorell and Habicht, 1986; Silventoinen, 2003; Steckel, 1995, 2008). In large samples, however, the genetic variation is averaged away (particularly in relatively more homogenous populations such as Italians), making average stature a measure of the balance between a cohort's gross nutritional intake and the toll of labor and disease during the growing period (Steckel, 1995). Average stature is thus a measure of the "biological standard of living"—capturing health and physical well-being, and with other measures of welfare, such as GDP per capita (Floud, 1985; Steckel, 1995).[21] Moreover, stature has been shown to be correlated with desirable labor market attributes on the individual level, including cognitive ability (Case and Paxson, 2008), education (Case, Paxson, and Islam, 2009), occupational skill (Komlos, 1990), and wages (Case and Paxson, 2008; Lundborg, Nystedt, and Rooth, 2009; Persico, Postlewaite, and Silverman, 2004).[22] Stature thus provides a measure of welfare and of productive capacity, broadly defined.[23] For this reason, height data have been used extensively by economic historians and development economists seeking to document living standards (e.g., Costa and Steckel, 1997; Deaton, 2007; Fogel et al., 1983; Jayachandran and Pande, 2015; Komlos, 1987), providing both a substitute for conventional measures

[20] Komlos and Meermann (2007, p. 200) specifically describe height as a measure of "productive capacity."

[21] Height, and more generally the biological standard of living, are generally seen as complementary to other conventional indicators (e.g, GDP per capita) in measuring living standards. These indicators provide a measure of the "economic standard of living"—essentially a measure of the quantity of goods and services available. It has been suggested that height may even be a better measure of welfare than these conventional indicators because it provides "a direct measure of welfare, much closer to what we think of as welfare or the standard of living than artificial constructs such as national income per capita or the real wage" (Floud, 1985, p. 33).

[22] If there are direct returns to height, such as through increased physical strength, then height is in itself a component of human capital.

[23] In addition to effects through health and nutrition, a variety of channels could theoretically cause the positive relation between height and desirable labor market characteristics, among them a reflection of general investment of parents in the child's upbringing, or simply physical strength.

when they are not available or are difficult to measure, and a supplement to capture many aspects of welfare (such as health) that economists find important in measuring living standards.

Relative to other measures that have been used to estimate selection into migration, the main disadvantage of stature is that it measures quality with a great degree of error, in the sense that the genetic variation across individuals within groups is uninformative. This disadvantage, however, averages away in large samples. The advantages, on the other hand, are that stature is pre-determined (for old enough migrants), it is easily measured, it cannot be manipulated in anticipation of, or in response to, migration, and it is continuous and monotone.[24] A few previous studies have used stature as a measure for estimating self-selection of migrants. Kosack and Ward's (2014) study, discussed above, complements Crimmins et al.'s (2005) use of stature to study contemporary Mexican immigration, showing that Mexican immigrants older than 50 are taller than stayers of the same age groups. Humphries and Leunig (2009b) study selection into rural migration to London in mid-nineteenth century Britain, as reflected by a group of individuals who would later become seamen. Relative to these studies, our Italian data have the advantage of being representative of both populations of interest—migrants and the population at risk for migration—and of enabling the observation of selection at the local level under the advantageous setting of the Age of Mass Migration.

3 Data

3.1 Data Sources

3.1.1 Ship Manifests

Our information on the stature and other personal characteristics of migrants is taken from the Ellis Island arrival records data base. This source contains the records of nearly all passengers who passed through the Port of New York from 1897 to 1924 (and January 1925),[25] comprising the overwhelming majority of Italian passengers entering the United States during the Age of Mass Migration.[26] This data base was compiled from passenger manifests deposited at Ellis Island, of which Figure B.2 presents an example. These manifests were completed upon embarkation (rather than upon arrival at Ellis Island, as is commonly believed) by

[24] For example, occupation could be changed in expectation for migration, and its ranking may be ambiguous—in particular, farm workers may come from a wide range of backgrounds, from penniless daily laborers to well-off owner-cultivators. Literacy is a binary measure that mutes the potentially important variations within the groups of literates and illiterates.

[25] For two reasons, data from the first five years during which Ellis Island was in operation (1892–1897) provide partial coverage of migration during this period. 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 of the records that were stored there.

[26] As can be seen in Figure B.3, the time series of the Italian immigration in our sample fairly closely tracks that of the Italian immigration from the official statistics.

the steamship companies transporting passengers to Ellis Island, and were primarily intended to fulfill three purposes. First, they were used to maintain statistics on immigrants entering the United States. Second, they were part of an effort to prevent the entry of potential immigrants who might become a public charge, who were ill, or who were considered undesirable (anarchists and polygamists). Passengers who were found to be unfit and were not admitted were deported at the expense of the shipping company. Third, the manifests were used as proof of the date of entry when immigrants began the naturalization process. Beginning in late 1906, with the passage of the Immigration Act of 1906, passenger manifests were required to include a physical description of the passenger, including his height. Heights were recorded in full inches, although in some cases we do see heights recorded in fractions of inches or centimeters. All heights are converted to centimeters for analysis. The height records were meant, at least in part, in order to facilitate identification at the time of the naturalization application.[27] We discuss potential biases in the heights data in section 5.1.

We acquired from the Statue of Liberty-Ellis Island Foundation (SOLEIF) a subset of the arrivals data base, consisting of the information transcribed, by volunteers from the LDS Church, of the roughly 4.8 million passengers entering the United States in this period who either reported their ethnicity as Italian, north Italian or south Italian, or whose country of origin was Italy.[28] Next, we geocoded the passengers' reported last place of residence using an algorithm outlined in Appendix C. As we discuss in Appendix C, a variety of tests and exercises show that this algorithm is very likely to be remarkably accurate for the individuals who can be matched, with a rate of false matches that may be about eight percent or less. In section 5.2, we test whether geolocated passengers were different from those who could not be geolocated in ways that could affect our results. We formally explore the possible effects of incorrect geolocation on our results in section 5.3.

Since the heights, as well as a number of other useful fields, were not transcribed by the LDS volunteers, we sampled 88,000 passengers arriving in 1907 or later, for whom we transcribed some of the additional fields.[29] The data that we received in digital form (indicated by the dashed lines in Figure B.2a) included

[27] When foreign citizens sought to become naturalized, they were required to show that they had entered the United States legally, and had been resident therein for a sufficient period of time. To this end, immigration officials would consult the passenger manifests and issue a Certificate of Arrival. The inclusion of the height and other physical attributes would help in verifying that the applicant was indeed the same person listed in the manifest.

[28] The US Bureau of Immigration and Naturalization considered North and South Italians to be of different ethnicities, and attempted to ensure that the distinction was accurately recorded on the manifests (Perlmann, 2001; Weil, 2000). The division between North and South was placed at the southern edge of the River Po basin. There were a number of cases, however, in which migrants were categorized as Italian without further disaggregation. These migrants were much more likely to travel from non-Italian ports, consistent with the possibility that foreign clerks would have been less likely than Italian clerks to be aware of the correct categorization for any given location of origin. 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 omit these individuals from consideration.

[29] We transcribed a simple random sample of households (identified by the ordering of individuals on the manifests and by a common last name) and not of individuals. Thus, an individual traveling with one companion was twice as likely to be sampled

the passenger's name, marital status, age, date of arrival, ethnicity, nationality, and last place of residence (as a text field). For all passengers within our sampled households, we transcribed the answers to four additional questions asked regarding the migrant (indicated by dashed thick solid lines in Figure B.2b): 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;[30] whether he had ever been in the United States before; and his height. For a smaller random sample of these passengers (about 38,000), we transcribed occupation and literacy.

3.1.2 Italian Stature Data

In order to quantify migrant self-selection, we must be able to compare migrants' heights to those of the Italian population at risk for migration. We acquired height data compiled as a part of the Italian military conscription process.[31] Following Italian unification in 1861, all Italian males of conscription age (usually but not always age 20) were required to present themselves for a medical examination, even if they had obvious grounds for exemption, such as a physical condition that would preclude service. During this examination, their heights were measured and recorded. A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) collected the one-centimeter frequencies that were compiled from the conscription records for each of the birth cohorts 1855–1910 in each province.[32] This near-complete enumeration of more than 21 million men amounts to a data set that, for this period, is probably unsurpassed in its population coverage and resolution. It nevertheless has two main shortcomings that are relatively minor compared to other sources, but which do require attention. First, although examination was mandatory, absenteeism was still on average 12 percent. According to A'Hearn, Peracchi, and Vecchi (2009, p. 6), this was largely the result of emigration during childhood and adolescence; overt draft dodging was rare. Thus, although the distributions do not encompass the entire cohort (including those who migrated during childhood to the US and to other destinations), they do reflect the population at risk for migration in adulthood, from which our migrants are drawn. The height distributions received from A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) also include a correction for the rate of absenteeism.

The second difficulty is that the age of measurement varied, and there is reason to suspect that some part

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.

[30] The manifests typically included the full name and address of the contact person, and the nature of the relationship. We only transcribed the nature of the relationship (e.g., husband, friend, brother in law, etc.).

[31] For details, see A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) and Cole (1995, pp. 145–146).

[32] Henceforth, we use the term province-cohort to denote a given birth-year cohort in a given province. A merger of the provinces of Napoli and Caserta in 1927 led A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) to treat the two provinces as one. We follow this approach throughout this paper, referring to the combined entity as Napoli.

of the population had not yet reached terminal height at the time of measurement. Candidates for military service were legally required to present themselves for measurement in the year in which they turned 20, but the actual average ages of measurement varied somewhat around that age (A'Hearn, Peracchi, and Vecchi, 2009, pp. 3–5). Although Beard and Blaser (2002) and Frisancho (1993) show that modern populations reach terminal height by age 20, the same may not have been true of Italians in our study period. Indeed, a number of studies (A'Hearn, Peracchi, and Vecchi, 2009; Fogel, Engerman, and Trussell, 1982; Frisancho, 1993; Horrell, Meredith, and Oxley, 2009; Horrell and Oxley, 2015; Steckel, 1986, 2009) discuss the potential for early-life nutritional stress to both reduce final adult height and to extend the growth period into the early twenties. Earlier measurement thus implies a potential downward bias of the mean (as well as increased variance) of the observed height distributions relative to the terminal distribution of heights. Unfortunately, the relationship between nutritional deprivation and the end of growth is poorly quantified.[33] The consensus, however, is that the same conditions that lead to shorter height are also likely to extend the growing period (Steckel, 2009). This implies that the raw means are more downward biased (and that their variance is more upward biased) among shorter populations. Failing to account for this phenomenon would lead us to spuriously find stronger positive self-selection among the shorter cohorts when studying migrants who had achieved terminal height.

We therefore rely on A'Hearn, Peracchi, and Vecchi's (2009) extrapolations of the means and standard deviations to age 22, beyond which any further growth would have been negligible even among malnourished populations. These moments were intended to address the continued growth problem while accounting for the average age at measurement of each province-cohort and we use them to generate what we consider to be the closest possible measure of the true terminal height moments for each province-cohort.[34] While the corrections from age 20 to age 22 are quite large in some cases,[35] we find little reason to suspect that

[33] While the phenomenon is known to occur, we were unable to find literature quantifying the effect of early-life nutritional stress on the rate of growth after age 20, conditional on the rate of growth prior to age 20.

[34] The age-22 moments that we received from A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) were smoothed across birth cohorts within provinces by the procedure employed to correct for measurement age, absenteeism, and other data issues; we generated unsmoothed age-22 moments from the unsmoothed age-20 moments based on the difference between the age-20 and age-22 smoothed moments. Based on variation in the age of measurement, 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. These 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 B.1, 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. We consider the unsmoothed age-22 moments to be the closest possible measure to the true terminal height moments.

[35] Using instead the age-20 measures would shift upward the estimates of the degrees of self-selection for all cohorts.

they meaningfully bias terminal height moments, or that they do so differentially across province-cohorts. However, since we suspect that there may have been some error in the recording of ages on the ship manifests, we verify that our main results still hold when we use moments that are smoothed across cohorts within provinces.[36]

3.1.3 Italian Official Statistics

We collected data on characteristics of provinces, districts (*circondari*), and townships (*comuni*) from a variety of official statistics published by the Italian Government. From the Italian Censuses of 1901 and 1911, we collected data on *comune*-level population in 1901 (Volume I, Table I), which we use to create indicators for urban residence; province-level property ownership (Volume IV, Table VIII); and province-level literacy rates from the 1911 census (Volume III, Table V). We also collect emigration data from the *Statistica della Emigrazione Italiana per l'Estero* for 1900–1920.[37] From these sources, we gathered information on the number of emigrants from each province to each destination country for each year in the range (Table V of this publication), which we will use to construct a measure of exposure to non-US-bound emigration.

3.2 Summary Statistics

We impose five main restrictions on the complete passenger data that lead to our baseline sample of choice. First, we focus on arrivals in 1907 or later, the period for which passengers' stature data are available. Second, we focus on males, as there were no stature frequencies recorded for females in Italy to which we can compare female migrants. Third, to be included in the analysis, passengers had to belong to a household that was randomized into our transcribed sample. Fourth, we omit passengers younger than 22 and older than 65 years old. The younger passengers are removed because they may not yet have achieved terminal height, and because we find clear evidence that emigration of Italian males aged 18–21 was significantly curtailed, probably due to Italian laws restricting emigration during the age of military service (Cole, 1995).[38] The older cohorts are removed to avoid the problem of shrinkage during old age.[39] Finally, to avoid assigning

[36] Rather than using the moments smoothed by A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011), we compute our own smoothed moments using province-specific kernel regressions of the moments against birth year. do so because those of A'Hearn, Peracchi, and Vecchi (2009) and A'Hearn and Vecchi (2011) are not simply averages over time, but are instead affected by the temporal trend in other provinces. We point out below any instances in which our choice of unsmoothed moments over smoothed moments has a qualitative effect on our results.

[37] After 1920, the publication was superseded by the *Annuario Statistico della Emigrazione Italiana dal 1876 al 1925*, which contains much less detailed information, usually no finer than the regional level.

[38] As can be seen in Figure B.4, there was a sharp dip in the age distribution of Italian male passengers between ages 18 and 21, a trend that was not shared by Italian females or, for instance, by Russian-Jewish males. These passengers were indeed shorter, which could be either due to the AGS problem or to distorted selection caused by the legal restrictions.

[39] Cline et al. (1989) show that shrinking begins 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

more weight to passengers who crossed the Atlantic multiple times, and to exclude cases in which some of the growth may have taken place while in the US, we restrict the sample to those making a first arrival in the United States—those who reported not having been in the US before.[40] We test whether our baseline results are meaningfully affected by the restrictions on age and report when differences are found. Table A.2 presents the sample size meeting these requirements and providing useful height data, divided by region (*compartimenti*) and province.

3.2.1 Passengers

The main descriptive statistics of our transcribed sample are presented in Table 1.[41] Column (1) reports the means for geolocated and transcribed males, aged 22–65 at arrival, including those who were making a second or subsequent entry to the US (a group comprising 44 percent of arrivals after 1907). Column (2) describes our sample of choice, which excludes the repeat entries. Most strikingly, this column reflects the well known feature of the Italian migration to the United States—that it was predominantly southern. Southerners outnumber northerners by more than four-to-one in our sample; they were also more likely to migrate to the United States, as evidenced by the fact that Southerners comprised just under two-thirds of Italy's 1901 population.[42] Average height in our sample was 163.9 cm (see also height distributions in Figure 1). Only a small fraction of passengers reported no connection in the US, but only three of ten passengers were joining an immediate family member (i.e., sibling, parent, child, or spouse)—the rest reported relying on friends, more distant relatives, or some sort of professional relation. More than ninety percent of passengers in our benchmark sample financed their own passage passage. Columns (3) and (4) compare the southern and the northern passengers (based on results of the geolocation algorithm), showing three main differences: northerners were on average two centimeters taller;[43] they were far less likely to come from an urban locality;[44] and they were much less likely to have departed from an Italian port (which was the case for over 90 percent of southerners). Northerners were also somewhat younger and less likely to be married.

Although females are not included in our benchmark sample, it is important to note a number of dif-

passengers.

[40] Boas's (1911, 1920) study of convergence of immigrant children's heights, a seminal work in the field of physical anthropology, dealt with this question with precisely our population of interest—a large part of his sample consisted of Sicilian and south Italian children in New York schools. See also Gravlee, Bernard, and Leonard (2003), Kress (2007), and Sparks and Jantz (2002, 2003).

[41] See Table A.3 for descriptive statistics for the full geolocated sample.

[42] For the emigration rates of particular provinces, see Figure B.5.

[43] The average heights of each province according to the conscription records and passenger manifests are presented on the maps on Figure B.6.

[44] The definition used for urban is a township with more than 10,000 residents in 1901.

ferences across genders (see statistics for females in Table A.3). Females were much less likely to have paid for their own passage, as many of them reported their fare paid by their husbands. Males were much more likely to have been repeaters (that is, to report having been in the United States before), reflecting the predominance of males among the "birds of passage"—passengers crossing the ocean multiple times. Finally, females were 4.5 centimeters shorter than males—a gap that is unusually small relative to those observed among modern populations (Gaulin and Boster, 1985).[45]

3.2.2 Provinces

We use a number of province-level variables, supplementing the height data from the conscription records with other Italian statistical sources. These variables provide a preliminary insight into the north-south divide (see descriptive statistics in Table A.4), and will be used in Section 6 to evaluate the mechanisms and the channels through which the patterns of self-selection are generated. The difference between the two regions is apparent both in a height gap of roughly 2.5 centimeters and a literacy gap of nearly 30 percentage points, each in favor of the north. However, a rough measure of inequality—the coefficient of variation in heights—was roughly equal in the North and South, as was the fraction owning property. Northerners were far less likely to live in an urban area (a *comune* of more than 10,000 inhabitants). Rates of total yearly emigration were very similar across regions, yet while as many as 87 percent of northern emigrants went to countries other than the United States (primarily other European countries, but also South America), the majority of southerners traveled to the United States. In Section 6 we examine whether variations in inequality, urbanization, and the degree of exposure to non-United States-bound migration help to explain the variation in self-selection across provinces.

Finally, as can be seen in Figure B.7, the Italian population was rather uniformly and almost monotonically growing taller over time, indicating a steady improvement in living standards, albeit from a very poor starting point. The South, however trailed considerably behind the North, with Southerners of the 1910

[45] The ratio of the average height of males to the average height of females—termed the sexual dimorphism of stature (SDS)—in a typical modern population is approximately 1.07 (Gaulin and Boster, 1985; Gustafsson and Lindenfors, 2004; Gustafsson et al., 2007; Moradi, 2009). 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. 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. As Italians of this period were rather short, a small SDS may not be entirely unusual. Moreover, we do find evidence of an allometric SDS in our data, 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.

cohort still roughly 0.5 centimeters shorter than northerners of the 1855 cohort. Put differently, the average rate of growth in the population was approximately 0.38 centimeters per decade, leaving Southerners about 68 years behind Northerners as measured by the differences in average heights between the regions.

4 Main Results

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, and let the heights be distributed $h_{ijt} \sim F_{jt}$, where F_{jt} is a distribution with mean μ_{jt} and variance σ^2_{jt}. Let F_t denote the stature distribution at the national level for birth cohort t and let its mean and variance be μ_t and σ^2_t.[46] Our analysis will be based on two statistics. The first is the national z-score, which is height normalized by the all-Italy mean and standard deviation of the birth cohort, $z_{it} = \frac{h_{it}-\mu_t}{\sigma_t}$. The second is the local z-score, which is height normalized by the mean and standard deviation of the birth cohort in the province of origin, $z_{ijt} = \frac{h_{ijt}-\mu_{jt}}{\sigma_{jt}}$. We will test for national- and province-level self-selection by estimating these z-scores and testing whether they are different from zero.

4.1 Countrywide Self-Selection

How did the first-time migrants compare with their peers in the whole of Italy? Figure 3a shows the distribution of their z-scores (relative to the height distribution of all Italy for their birth cohort), showing a small leftward shift relative to a normal distribution centered at zero. Imposing the assumption that the F_t are normal (for this exercise only), we perform a Kolmogorov-Smirnov test of z_{it} against a $N(0,1)$ distribution.[47] We reject the null of a $N(0,1)$ distribution, indicating that some sort of self-selection did take place if the F_t were indeed normal. In Table 2, we formally measure the degree of self-selection at the national level. Column (1) presents a regression of the countrywide z-scores on a constant. The constant, which represents the mean of the z_{it}, is negative and strongly statistically significant, indicating that the average Italian immigrant was more than 0.1 standard deviations shorter than the mean of his all-Italian cohort of origin. This difference corresponds to approximately 0.70 centimeters, or nearly 40 percent of the

[46] 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 μ_t and σ_t by weighting across provincial distributions by 1901 population. Let N_j denote the population of province j in 1901 and N denote the population of all Italy in 1901. We computed the moments as $\mu_t = \sum_j (N_j/N)\mu_{jt}$ and $\sigma^2_t = \sum_j (N_j/N)(\mu^2_{jt} + \sigma^2_{jt} - \mu^2_t)$ (Früwirth-Schnatter, 2006, pp. 10–11).

[47] 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 the continuous standard normal distribution could lead to rejection of the null of a standard normal distribution even if the underlying distribution of z-scores is precisely the standard normal. Thus, for the purpose of this test, and its counterpart in section 4.2 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.

inter-provincial standard-deviation of mean heights. It is thus quite large.[48]

In column (2), we separate the population across the north-south divide, as defined by the US Bureau of Immigration (i.e., Central Italy is included in the South). We find that the countrywide negative self-selection was in fact a result of a mixture between positively selected (with respect to the countrywide mean) northerners and negatively selected (again, on the country level) southerners, with the former being 0.164 standard deviations taller than their respective all-Italy cohort, whereas the latter were 0.169 standard deviations shorter than average. The north-south gap is not a result of different cohorts or years of arrival, as can be seen in column (3), where birth-year and arrival-year indicators are added as controls and the North-South difference remains of roughly the same magnitude.

A straightforward explanation exists for this pattern: Southerners were on average roughly 2.5 centimeters shorter than their northern counterparts (see Table A.4). Unless either group of immigrants were extremely different from their population of origin, one would expect to find that relative to the all-Italian population, southerners were negatively selected and northerners were positively selected. As can be seen in Figure 2, the shorter Southern provinces sent many more immigrants to the United States. In fact, more than four-fifths of the first-time migrants came from the South (see Table 1), and thus the overall selection of immigrants was dominated by shorter cohorts.

In sum, when the reference point is the entire population of Italy, the heights of Italian immigrants tell a story of negative self-selection; but it may well be that this was driven purely by differences across provinces. What this does not tell us is whether, within a given local environment, immigrants were negatively or positively self-selected.

4.2 Local Self-Selection

What do the heights of immigrants tell us about their selection into migration from within their local environments? We approach this question using the province-birth cohort z-score. A graphical summary of these data can be seen in Figure 3b, in which the distribution of migrant height normalized by the moments of the province-birth cohort appears to be shifted to the right relative to a standard normal distribution. A Kolmogorov-Smirnov test rejects the null of a $N(0,1)$ distribution.

The other main results of this paper are presented in Table 3. In column (1), we regress the province-birth cohort z-score on a constant. We find that Italian migrants were on average 0.037 standard deviations of

[48] We compare this magnitude to the inter-provincial standard deviation of mean heights (about 1.8 centimeters) rather than the national standard deviation of heights (about 6.5 centimeters) because the latter includes genetic variation. The former, on the other hand, represents only differences in living standards across provinces because genetics average out in the computation of province-cohort average heights.

height taller than their province and birth cohort means, and that this difference is statistically significant. This result is non-negligible: it corresponds to roughly 0.231 cm, or approximately 12.5 percent of the inter-province standard deviation of average height. Thus, changing the reference group from all of Italy to the migrants' provinces of origin reverses the sign of the self-selection. Although the migrants originated overwhelmingly in the shorter South, they were positively selected from within their local places of origin.

As with the national selection, there is also considerable variation in the degree of local self-selection across Italy. Column (2) of Table 3 divides the degree of local self-selection by North and South, showing that the positive self-selection seen in column (1) was even stronger in the South (0.065 standard deviations) and was negative in the North (-0.081 standard deviations). As with the national self-selection result, the all-Italian average positive local self-selection reflects the South's dominance in migration to the United States, masking differential patterns of selection across Italy. To further decompose these differential patterns, we regress the province-birth cohort z-score on the mean height of the province-birth cohort. The result, shown in column (3), tells a similar story to that above: the shorter the province is on average, the greater is the degree of local self-selection. The coefficient implies that immigrants from a one-centimeter taller province-birth cohort were an additional 0.064 standard deviations (or on average 0.40 centimeters) shorter than their province-birth cohort means. Thus, shorter (and less developed) provinces tended to supply the most positively locally selected migrants.

The North-South divide and the negative relationship between mean height and local selection are clearly shown in Figure 4. Collapsing the average province-birth cohort z-scores within province, with each point representing a single province,[49] the southern provinces are generally in the short and positively selected upper left, whereas the northern provinces are in the taller and negatively selected lower right. The correlation coefficient between mean province height and the mean province-birth cohort z-score is -0.52. The curve represents an individual-level non-parametric regression, and its tight confidence band suggests that the downward slope is uniformly robust across the horizontal range. The near linearity of this curve also provides evidence that the downward trend of self-selection with respect to average height does not merely reflect different patterns within the two regions. Indeed, the downward slope is evident throughout the range of province-birth cohort average heights.

Column (4) of Table 3, illustrates the robustness of this pattern across regions through a regression. In particular, we interact the province-birth cohort average height with a South indicator. The results show

[49] The computation of average province z-scores and average province heights requires averaging away birth-cohort differences. We do this by simply averaging the values for all observations in our sample, which is essentially equivalent to weighting the means over birth cohorts based on the number of passengers in each birth cohort for each province.

that the negative slope is present in both regions. It is approximately one-third weaker in the South than in the North, though the difference between the trends in the two regions is not statistically significant. Within the South, the downward trend is in fact slightly stronger than the estimated all-Italian slope. Thus the systematic variation above is also present within regions; both within the South and within the North, shorter cohorts were increasingly represented in the US by relatively taller and more positively selected immigrants.

4.3 Relevance and Interpretation

4.3.1 Local vs. National Selection

Our 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 provide any information pertinent to understanding the effects of migration? Should a policy maker in the country of destination seeking high quality migrants assign value to the fact that these shorter passengers exceed their peers within their local environments? Or should he value only the absolute measure of their quality and ignore local selection altogether? That is, does quality relative to one's peers provide information beyond absolute measures of quality? These questions have not been addressed by other studies of migration, and providing a definitive answer is an empirical task that is beyond the scope of this paper.[50] However, we suspect that this question is important in the case of selection of migrants, as well as in any case that involves human selection from a variety of groups. We therefore provide, in Appendix D, a simplified theoretical setting that illustrates the problem. In this setting, a proxy measure for quality (such as absolute height) is observed, as well as the local ranking of this measure within subgroups (e.g., the local z-score). This measure is a function of two types of unobserved inputs—individual ability and an environmental contribution—and it proxies for quality in the sense that the same two inputs also determine individual's human capital, which is unobserved at arrival.[51] We derive a condition under which the local ranking is positively correlated with outcomes in the receiving country, conditional on the observed proxy measure; that is, a condition under which an individual who is positively locally selected can be expected to outperform another individual of the same absolute level of the proxy measure who is less positively locally selected.

The intuition behind the condition is the following. This positive correlation occurs when the environ-

[50] There is evidence in other contexts that relative measures of quality are important in predicting outcomes, conditional on absolute measures (Conley and Önder, 2014; Niu and Tienda, 2010; Pike and Saupe, 2002; Rothstein, 2004).

[51] A key assumption is that the distribution of individual ability is the same in each local environment. The lack of meaningful internal migration in Italy prior to World War II (Bonifazi and Heins, 2000) suggests that sorting across locations on the basis of ability is likely not problematic.

mental input is relatively more effective in contributing to height (the observed proxy), compared with the input of the individual ability, than in contributing to the expected wage in the receiving country. Intuitively, if an individual was able to excel in his depressed environment he is likely to have high individual ability. This condition implies that the retarding effects of his environment would be lessened after having emigrated from it, with his advantageous personal ability making a greater impact on his performance in the new environment. As a result, we suspect that overlooking selection within sub-national regions and focusing only on the degree of national selection ignores potentially valuable information on immigrants' productive capacity, particularly when migrants originate in countries with large variation in economic conditions across regions.[52] Focusing on absolute measures of quality without making comparisons to the appropriate local groups of reference may overemphasize the role of different environmental conditions across different points of origin at the expense of individual quality.[53]

4.3.2 Marginal Effects on Migration Probabilities

A common approach to measuring migrant self-selection is to determine the marginal effect of some measure of quality on the probability of migration, often through estimation of a binary choice model with the measure of quality as a regressor (e.g., Abramitzky, Boustan, and Eriksson, 2013; Connor, 2015). Under the assumption that the province-cohort distribution of heights reflects an underlying distribution of quality, it is possible to transform our estimates of the degree of migrant self-selection from sections 4.1 and 4.2 into rough estimates of the marginal effects of quality on the probability of migration. These will provide a sense of the economic significance of the relationship between quality and migration.

This exercise is discussed in detail in Appendix E. In brief, it proceeds as follows. Using Bayes's theorem, the distribution of z-scores conditional on migration (learned from our data) can be transformed to yield the probability of migration conditional on z-score. However, the z-score measures quality with (primarily genetic) measurement error. If this measurement error is assumed to be of the classical form, then the standard attenuation bias results hold. If the variance of the measurement error is known, then it is possible to correct for the measurement error in order to learn the effects of standards of living on migration probabilities.

[52] This is true for some of the most important cases of contemporary immigration to the United States, such as Mexico, China, and India.

[53] Another example helps to illustrate this point. Consider the use of education as the measure of migrant quality. Consider two migrants, each with the same level of education, but suppose that one comes from a place (country or sub-national area) where education is widely available, while the other comes from a place where it is not. Then the latter migrant is likely higher in the distribution of ability (since he became educated where it was more difficult to do so) than the former, if the distributions of ability are the same across origins. In this case, the migrant from the area where education provision is poor is likely to perform better in the receiving country than the other migrant because he has higher ability and the same education. A similar conclusion would hold if the lower ability migrant had higher education, as long as the gains from the education itself did not fully offset the gains to higher ability. The condition in Appendix D formalizes this trade-off.

From the anthropometrics literature, we take as a benchmark that genetics explain about 80 percent of the variation in heights in modern environments, and less than 80 percent in poorer settings (Silventoinen, 2003). This upper bound on the genetic variation in heights enables us to place an upper bound on the effects of changes in quality on migration, and thus to approximate the results of a regression with quality as the regressor.

Table 4 presents the results of this exercise. The estimate of the effect of a unit increase z-score (or a one-standard deviation increase in height) on the probability of migration is given in the first row of this table. This also represents the lower bound of the of the estimate of the effect of an increase in quality associated with a one standard deviation increase in height on migration probability. The upper bound is given in the last row of the table. In all of Italy, an increase in quality associated with a one-standard deviation increase in height is associated with an increase in migration probability of 0.4 to 1.8 percentage points, compared to a base migration probability of about six percent over the period 1907–1925. In the south, the range is 0.8 to 4.2 percentage points, compared to a base migration probability of migration of about 8.5 percent. In the north, the range is -0.1 to -0.6 percentage points, compared to a base emigration probability of 2.7 percent. The effects of variations in quality on migration probability are thus potentially very large.

5 Robustness

There exist several possible threats to validity of the results of section 4. In this section, we discuss these possibilities and address the extent to which this may affect our findings.

5.1 Systematic Upward Bias

The measurement of heights in the Italian conscription records is well documented, and we find no reason to suspect that it contains meaningful biases after the corrections discussed in Section 3.1.1. However, there are two main possible sources of systematic upward bias in the heights reported on the passenger manifests—measurement with shoes, and self-reporting.[54] Unfortunately, we were unable to find documentation of how the height data in the passenger manifests were gathered, other than that the information 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

[54] It is well-established that when individuals are asked to report their own heights, their reports are systematically biased upwards (Danubio et al., 2008; Rowland, 1990).

requirement that shipping lines pay for the return passage of individuals found medically unfit to enter the United States—it is not clear whether this examination included 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.[56] As a result, we cannot rule out that in some cases heights were self-reported or measured with shoes and thus biased upwards.

Although the height distributions do not appear pathological (see Figure 1) and we have no concrete evidence that either of the above-mentioned potential sources of bias were prevalent, we cannot definitely conclude that our findings of positive local self-selection are not partly generated by such mismeasurement. While we believe that the size of the selection that we find is sufficiently large so as not to be fully driven by such biases, we must conclude that it should, nevertheless, be taken with some doubt. However, to the extent that measurement errors did not vary systematically across provinces, our findings regarding the relationship between the degree of self-selection and the relative quality of the origin population are robust to systematic measurement bias of this type.

5.2 Representativeness

As explained in section 3.1.1, our benchmark sample is limited to passengers who were successfully geolocated by our algorithm (henceforth, "matched" passengers). Geolocation may not have occurred at random, however. Indeed, one might worry that systematic variations in the probability to be geolocated caused the benchmark sample to be selected in ways that would bias the results. We believe that there is no particular reason for this to be the case for migrants embarking from Italian ports. For these individuals, the last place of residence was almost universally written with remarkable textual precision.[57] When the localities are written with errors in the Ellis Island files, it is typically due to handwriting that is difficult to decipher. We attribute the remarkable clarity of the place of origin to the fact that there was a legal requirement in Italy to obtain a passport prior to embarkation (Foerster, 1919, pp. 10–11), and we believe that these

[55] The manifest asserts that the height field 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 were 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.

[56] In personal correspondence with Marian Smith, the chief historian of the US Citizenship and Immigration Services in the Department of Homeland Security, we were told that ours was the first inquiry into these data since she assumed her position in 1988.

[57] For example, we rarely see phonetic ambiguities or potentially difficult cases leading to spelling errors, such as double consonants mistakenly written as a single consonant.

official documents—which would have had the commune of origin correctly written on them—were used in completing the manifests. Thus, we do not suspect, for example, that less literate passengers were less likely to be geolocated when embarking from Italy.

However, approximately 13 percent of Italian passengers in our sample embarked from ports outside of Italy, primarily from the French ports of Le Havre and Cherbourg, as well as from Trieste (part of Austria-Hungary until after World War I). There, the passport requirements were not in effect and the clerks were not Italians, increasing the chance of errors in recording Italian localities. Embarkation from non-Italian ports was certainly non-random. In particular, Northern Italians were much more likely to do so,[58] in part due to their closer proximity to French ports. Moreover, it is likely that some embarkations from outside of Italy were really cases of step migration—Italians who, for example, had already spent time in France and later moved overseas—and clearly these were not selected at random. Thus, there is reason to suspect that successful geolocation may have varied across provinces. There also remains a suspicion that it varied systematically across individuals within provinces (despite the discussion above). We investigate this problem below. However, it must be kept in mind that we were able to match 85 percent of passengers;[59] thus, any systematic differences between the matched and unmatched would have to be quite large for non-random matching to pose a serious threat to the validity of our results.

To determine whether non-random selection into geolocation might have affected our results, we first run a battery of balancing-test regressions of individual characteristics on a geolocation indicator, presenting results in Table A.5; column (5) in particular covers our benchmark sample. To briefly summarize the results of these regressions, there was indeed a clear pattern of under-representation of northerners, but characteristics other than ethnicity and geography (and a marginally statistically significant and small difference in whether the migrants had any connection in the United States) are not statistically significantly different between matched and unmatched passengers, nor are any of the differences large in magnitude.[60] This pattern threatens the validity of our result regarding national self-selection, which requires that selection into matching be unconditionally independent of height (or independent conditional on year of arrival and

[58] The share of northerners out of Italians embarking from foreign ports was approximately 55 percent according to our algorithm, and approximately 51 percent according to the ethnic classification on the manifests, compared to 9 percent and 5 percent, respectively, out of embarkations from Italian ports. Approximately 41 percent of passengers departing from non-Italian ports had no specific ethnicity recorded, whereas only about 20 percent of those departing from Italian ports had no specified intra-Italian ethnicity.

[59] This is the share among transcribed males with usable height data, aged 22–65, making a first arrival. The share of geolocated among males aged 22–65 is 79 percent for the period 1892–1925 and 81 percent for the period 1907–1925. In all cases, these figures exclude individuals whose locations were not searched for because none was listed or were clearly outside of Italy.

[60] See also Figure B.9, which reports the results of non-parametric regressions of geolocation probability on individual height by ethnic group. For the most part, no patterns are visible.

birth year); essentially, the fact that we are more likely to match southerners means that we are generally more likely to match shorter migrants. On the other hand, our key findings in section 4—pertaining to the local degree of self selection—are not threatened by non-random selection into geolocation across provinces and cohorts, provided that, within province-cohort, passengers were geolocated at random.[61] If, however, within province-cohorts, the matched were taller than the unmatched, this would cast doubt on our finding of positive average local selection. In regards to our finding that migrants from shorter province-cohorts were more positively selected, we need to verify another assumption—that differences in height between the matched and the unmatched do not vary systematically with the height of the province-cohort.[62] Unfortunately, the fact that, by definition, unmatched passengers were not matched to a place of origin precludes straightforward tests of the latter two assumptions of within province-cohort representativeness. We must therefore find a way to associate passengers with local environments without relying on the success of the geolocation algorithm.

To this end we take advantage of the fact that Italian surnames are useful indicators of geographic origins (Guglielmino and De Silvestri, 1995) and create an alternative surname-based matching of passengers to provinces. The details of this procedure and its accuracy are discussed in Appendix F. For the purposes of this exercise only, we assign all individuals, both matched and unmatched, to their surname-implied province and compare them to the height distribution of their birth cohort in that province. To the extent that surname-implied provinces represent actual provinces of origin, we are then able to test the identifying assumptions described above by comparing the matched and unmatched passengers in each surname-implied province.

In Table 5 we report the results of this exercise. Column (1) tests the identifying assumption underlying the national self-selection result (which does not require the use of the surname-implied provinces). Successfully matched individuals are found to be 0.029 standard deviations shorter on average than the unmatched conditional on birth year and arrival year fixed effects, likely due to disproportionate success in matching southerners. A negative sign on this coefficient might raise the possibility that our negative national self-selection result is spurious. However, the difference is not statistically significant, nor is the magnitude large enough to generate our result.[63] Column (2) performs the same exercise, but the dependent

[61] Formally, the identifying assumption is that $E(z_{ijt}|g_{ijt} = 1) = E(z_{ijt})$, where g_{ijt} is an indicator for successful geolocation and z_{ijt} is the z-score for passenger i from province j and birth cohort t.

[62] That is, let $\Delta z_{jt} = E(z_{ijt}|g_{ijt} = 1) - E(z_{ijt})$ be the difference between the mean z-score of the matched passengers and that of all passengers (both matched and unmatched) in province j and cohort t. The identifying assumption is $\Delta z_{jt} \perp \mu_{jt}$. The threat is that what we perceive to be a greater gap between migrants and stayers among shorter province-cohorts may in fact be a greater gap between the matched and the full population of passengers.

[63] Recall that the average degree of negative self-selection was over 3.5 times greater than this, and that we successfully matched 85 percent of passengers.

variable is the local z-score rather than the national z-score (which was the dependent variable in the first column). The coefficient on the geolocation indicator provides a test of the identifying assumption for our local self-selection result. This specification shows that matched passengers were slightly (0.002 standard deviations) and statistically insignificantly shorter than the unmatched from the same imputed province-cohort and year of arrival. The direction of this bias (which goes against our result in section 4.2), together with the small share of the unmatched and the fact that the magnitude of this difference is much smaller than the result in Table 3, again leads us to conclude that the result of average positive local selection is very unlikely to be driven by the absence of the unmatched in our benchmark sample.

Finally, to verify the assumption underlying the result of selection varying with a province's average height, we test whether any difference in heights that exists between the matched and unmatched varies systematically with the average height of a province-cohort. Figure 5 shows the non-parametric regression of the z-score with respect to the height distribution of the surname-implied province-cohort of the matched, the unmatched, and of the two groups pooled together, on the surname-implied province-cohort average height. The trend of the unmatched is steeper than that of the matched, and the trend of the pooled group is virtually indistinguishable from that of the matched.[64] In column (3) of Table 5 we perform a linear version of this test. In particular, we regress the surname-implied z-score on the surname-implied province-cohort average height, an indicator for successful linkage, and an interaction of the two. This replicates our baseline result from column (3) of Table 3 separately for the matched and unmatched. The coefficient of interest is that on the interaction, which captures a difference in the linear trend between the matched and the unmatched relative to mean height.[65] The estimated difference in trends is an order of magnitude less than the estimated trend among the unmatched, and it is statistically insignificant and positive. In order to spuriously generate our differential results, this coefficient would need to be negative and larger than it is. We therefore conclude that this result is likely not driven by differential success in geolocation. Columns (4) and (5) perform the same exercise for the North and South separately (according to surname-implied province). These results tell the same story as for the complete sample in column (3). The difference in trend between the geolocated and non-geolocated is statistically insignificant and small relative to the trend among the geolocated.

[64] Confidence intervals are omitted for clarity. They do not enable us to reject the null hypothesis that all three curves are uniformly identical across the entire range.

[65] Note that the coefficient on the surname-implied province-cohort average height is larger (more negative) than the equivalent coefficient in column (3) of Table 3. This is mechanically driven by the artificial increase in geolocation errors, as discussed below in section 5.3.

5.3 Errors in Geolocation

Another potential source of bias in our benchmark results is errors in geolocation. Although we find strong evidence that our geolocation algorithm yields few incorrect matches (see Appendix C), we need to evaluate the extent to which some of the results might be driven by such errors. The negative national selection result cannot be driven by such errors because it does not make use of the matches. We also do not suspect that errors in geolocation may have spuriously generated the average positive local selection result. Unless errors are extraordinarily biased toward locating passengers to shorter provinces (we see no reason to believe that this would be the case), then due to the fact that most migrants came from shorter provinces, random geolocation errors would on average match passengers to taller province-cohorts than their true cohort. This would bias the results toward finding more negative local selection.

The result under threat is to the finding of differential selection across province-cohorts. Errors in geolocation would imply that the distribution of passengers in each province is a mixture of passengers that truly originated from that province and passengers who originated elsewhere in Italy. The height of the latter group would presumably be drawn from the distribution of heights of all Italian passengers of their birth cohort. This implies that the average height of passengers from each province-cohort is biased towards the mean of the sample. If in reality there were no systematic variation in selection with province-cohort average height (that is, the coefficient on average height in column (3) of Table 3 is truly zero) and if the probability of a geolocation error does not depend on the true province of origin, then the average individual who is spuriously matched to a short province would be relatively tall, and vice versa. In other words, the ensuing measurement errors in the z-scores are negatively correlated with the measure of province-cohort mean height. This would create a spurious negative trend in selection relative to province-cohort mean height, similar to our results in columns (3) and (4) of Table 3. We explain this problem formally in Appendix G, where we conclude that under certain reasonable assumptions, the rate of error in geolocation would have to be nearly 40 percent in order to spuriously produce the systematic variation that we find in Table 3.[66] As explained in Appendix C, we believe that our true rate of error in geolocation is far smaller.

In Table 6 we supplement the exercises in Appendices C and G, providing another test of the risk of a spurious trend generated by errors in geolocation. We restrict the sample to a subset of passengers for whom we have additional information to support our geolocation. In columns (1)–(4), we repeat the regressions of the main results of Table 3 on a sample of passengers whose North-South ethnic identity, as noted in the

[66] That is, 40 percent of the matched would have to be matched in error.

manifests, does not disagree with the province to which the algorithm assigned them.[67] In column (3), the systematic variation is 0.052, slightly weaker than the one found in Table 3, but still strongly statistically significant. In fact, in column (4), the systematic variation within both the North and the South appear even stronger than in Table 3. In columns (5)–(8) we repeat the same exercise over a sample based on a much stricter criterion. In particular, we keep only passengers whose geolocated province is the same as their surname-implied province (see section 5.2 and Appendix F), which means that their surnames strongly confirm that their geolocation was correct. The resulting subsample is one fifth of the size of the original sample, and so the standard errors are larger. The all-Italian negative systematic variation of selection with province-cohort average height is smaller (-0.045), while the one within the South is larger (-0.075), compared to the main results, and both are statistically significant. The systematic variation within the North is also similar to that above, and remains statistically significant. Overall, we cannot rule out that the negative systematic variation is somewhat biased and made stronger due to errors in geolocation, but we find little evidence that this bias is so strong so as to have spuriously generated the qualitative finding of systematic variation of selection with respect to average height. Indeed, only the negative local selection in the North is not reaffirmed by the restrictive samples in Table 6. This difference may be due either to weaker precision caused by the smaller sample, or to the fact that relatively few migrants originated in the North, making a false match of a southerner to a northern province particularly severe in generating observed negative selection.

6 Mechanisms

6.1 The 1917 Literacy Requirement

After two and a half decades of attempts to pass such legislation, the 1917 Immigration Act imposed, for the first time, significant restrictions on European immigration to the United States (Daniels, 2004, ch. 2; Goldin, 1994; Hing, 2004, ch. 3; Zolberg, 2006, ch. 7), spelling the beginning of the end of the Age of Mass Migration. The law banned the entry of passengers over the age of 16 who, despite being physically able to read, were unable to prove basic literacy in English or in another language.[68] Shortly thereafter, the 1921 Emergency Quota Act severely limited Italian immigration by setting the national yearly quota for Italy at

[67] Individuals listed as Italian without any further detail are retained.

[68] Elderly parents, wives, and children of literate passengers (or passengers who had already entered) were allowed to enter regardless of literacy.

just over 42,000 immigrants, less than one-fifth of the total Italian immigration in Fiscal Year 1921.[69] The 1921 law did not add any explicit selection criterion over the literacy requirement,[70] although in practice it is possible that it indirectly generated the positive selection sought by the 1917 literacy test.[71]

The effects of the literacy requirement on migrant self-selection have not (to our knowledge) been studied, though Goldin (1994) suggests that the test was far less restrictive than it was initially intended to be due to rising literacy in Europe in the early twentieth century. Our sample, because it covers the period 1917–1925, enables the study of the changes in the quality o0f migrants that took place following the enactment. The 1917 literacy requirement may have significantly curtailed the migration of the less educated Italians, and in particular, it may have strongly affected the composition of migration from the South, where illiteracy was still endemic.[72] Though almost four-fifths of the passengers in our sample arrived prior to 1917, it may be the case that our findings were primarily driven by this later selection criterion. If literacy and height were positively correlated, then even if those wishing to enter the United States were drawn at random from the Italian population (i.e., no selection into migration), the literacy requirement would have mechanically created a more positive selection among the admitted passengers from the shorter and less literate, potentially generating a spurious negative correlation between selection and province-cohort average height.

The descriptive statistics support the suspicion that the literacy test had such effects. First, it appears that the literacy requirement was a binding constraint. The rate of age 22–65 illiteracy in our transcribed sub-sample (again limited to first arrivals) decreased from 46.2 percent before 1917 to only 2.2 percent after enactment (Table A.3; see also the yearly trend in Figure B.10). Second, prior to 1917, the North-South literacy gap among passengers closely reflected the very large gap across the two populations at home,[73] meaning that the test would have been much more restrictive of Southern immigration. Had the literacy requirement applied earlier, a much larger proportion of Southern migrants would not have qualified, in all probability significantly increasing the degree of selection from the South to a much greater extent than in the North. Furthermore, literacy was indeed correlated with height among passengers. Table A.7 shows that

[69] The bill passed in May 1921 and came into effect in June, near the end of FY 1921, such that FY 1922 was the first year in which the quota was effective. The measure was extended annually until the 1924 Immigration Act, which was even more restrictive, was passed.

[70] The implementation of the annual quotas, which were divided in practice into monthly quotas, was effectively entrusted to the steamship companies. Until the adjustment of the quota system by the Immigration Act of 1924, shipping companies had to limit their volumes, and wound up racing each other's ships to disembark their passengers as close as possible to, but not before, the beginning of the month (Zolberg, 2006, p. 254; Robertson, 2010, pp. 204–205).

[71] In a study of passengers entering Alaska and Washington from Canada in the period 1918–1924, Massey (2012) finds that passengers from countries subject to the quotas became more skilled relative to passengers that were not subject to the quotas.

[72] Adult male literacy was only 55.8 percent in the average Southern province in 1911, as opposed to 83.2 in the North; see Table A.4.

[73] The rates of illiteracy among pre-1917 arrivals (limited to first arrivals) were 51.8 percent among Southerners and 21.2 percent among Northerners.

literate migrants were, on average, 0.125 standard deviations taller than illiterate passengers (controlling for birth-year and arrival-year fixed effects); similar patterns are evident within provinces, as well as within regions.[74] Third, the changes in the occupational composition of passengers before and after 1917 point to changing selection as shown in Table A.8. Prior to 1917, less-skilled passengers were much more illiterate, with illiteracy as high as 58.6 percent among farmers and agricultural workers. After the enactment, the rates of illiteracy were nearly zero in every occupational class. This change was accompanied by a drastic shift in the occupational composition of passengers—the shares of professionals and of skilled workers more than doubled after 1917, whereas the share of agricultural workers fell to nearly half its previous level.

It is clear then that the post-1917 literacy requirement would have generated a selection pattern that was very similar to the one that we find for the period 1907–1925 as a whole, with increasingly positive selection among shorter province-cohorts. This threatens the interpretation of our results from section 4.2, as it is possible that they are simply driven by the effects of the literacy requirement and do not represent the actual structure of the supply of immigrants. Table 7 evaluates this threat. In column (1), we regress the province-cohort z-score on a post-1917 indicator. The coefficient is large and significant—the average local z-score increased by 0.147 standard deviations after 1917, whereas the pre-1917 z-score was effectively zero. This implies that our aggregate positive local selection result is driven by post-1917 arrivals. However, in column (2), we interact the post-1917 indicator with a South indicator to decompose the change over time by region. Consistent with the fact that the literacy constraint was primarily binding among passengers from the South, the post-1917 upward shift in selection appears to have almost entirely occurred there; the coefficient of the interaction term is 0.123 standard deviations, whereas the main effect of the post-1917 period, representing the change in selection among Northerners, was much smaller and not statistically significantly different from zero. Importantly, within the South there was already statistically significant positive self-selection prior to the literacy requirement (0.029 standard deviations). During the post-1917 period, this increased to a much higher level (0.207).[75] The North-South selection gap also existed in the earlier period, with a difference of 0.123 standard deviations in favor of the South before 1917, before doubling to 0.246. Thus, though the aggregate result is largely driven by post-1917 arrivals, it is still the case that pre-1917 Southerners were positively self-selected on the local level, and that they were more strongly positively self-selected than were Northerners. In column (3), we estimate the relation between the z-score and the cohort-province mean

[74] After 1917, the relationship between height and literacy became negative; however, due to the very small number of illiterate migrants after 1917, the standard errors are large and the coefficient is not statistically significant (Table A.7, column (6)).

[75] Note that due to the relatively small number of post-1917 passengers, the Southern selection estimated in Table 3 for both periods together is still very close to the pre-1917 estimate despite the sevenfold increase following 1917.

height before and after 1917. Consistent with the findings in column (2), the negative relationship between the z-score and the cohort-province mean height had already existed before the literacy requirement; it strengthened after 1917, although this change was neither statistically significant nor of a large magnitude. Thus, with the exception of the general local self-selection result (but including its separate southern and northern components) our results were not driven by post-1917 migration, which was subject to the literacy test.

One concern is that the literacy requirement took effect during World War I, when transatlantic migration was halted, analysis of the effects of the literacy requirement essentially entails a comparison of pre-World War I migration to post-World War I migration. Thus, it is ultimately not possible to distinguish between the effects of changes in the supply of migration induced by World War I and the causal effects of the literacy test. To affirm that the shift was indeed driven by the differential degree to which the literacy requirement bound, we test in column (4) whether the increase in selection was greater in provinces that had a larger share of illiterate males by interacting the post-1917 indicator with the 1911 rate of literacy (the latest estimate before the imposition of the literacy requirement). Controlling for province fixed-effects, the difference-in-differences coefficient is negative, as expected, and statistically significant. A naïve interpretation of the magnitude (-0.331) would imply that the North-South literacy gap of 27.4 percent (according to the census) can explain 61 percent of the widening in the z-score gap in favor of the South.[76] That the shift in selection was closely correlatd with province literacy is strong evidence that hte post-1917 pattern was driven by the literacy requirement rathern than by other shocks, such as World War I.

In summary, the 1917 literacy requirement, potentially compounded by the 1921 quotas, did not entirely produce the main patterns that we find. Instead, its imposition strengthened an extant pattern of more positive selection among the Southern and shorter province-cohorts by shifting upward the selection of migrants in both North and South Italy, with a disproportionately greater increase in the South.

6.2 Sectors

6.2.1 The Urban Height Penalty

Due to high population density, poor sanitation, and relatively costly nutrition, health conditions and diets in large European cities in the nineteenth century were notoriously disadvantageous relative to those in the countryside, causing higher rates of morbidity and mortality, and in some cases, shorter average stature. This phenomenon is known as an "urban penalty." If Italy had an urban height penalty—a condition in which

[76] This is $0.274 \times -0.331/0.149 = 0.609$.

urbanites were shorter relative to their rural peers within the same province—then it is possible that the rural-urban composition of migrants may have been responsible for the selection patterns that we observe.[77] For example, if urbanites were relatively over-represented among Northern passengers as compared to Southern ones, and if there was in fact an urban height penalty, then this would have produced more positive selection among the shorter (and Southern) province-cohorts.

Whether an urban height penalty existed in post-unification Italy is an open debate. Such a penalty was found, for example, in the US among recruits to the Union Army (Haines, Craig, and Weiss, 2003; Zimran, 2015) and during World War I (Haines and Steckel, 2000), among Bavarian prisoners (Baten and Murray, 2000), and among pre-unification North Italians (A'Hearn, 2003). Other studies, however, have found inconsistent differences, or even an urban premium.[78] In their study of British seamen in the 1840s, Humphries and Leunig (2009a) found that Londoners exhibited significant stunting, but that sailors from smaller cities and towns were only slightly shorter than their rural peers. This raises the possibility that an urban height penalty was present (to a meaningful extent) only in the highly dense or industrialized urban centers.[79] Moreover, while Southern Italy did experience some epidemics in urban areas, such as the 1884 Cholera outbreak in Naples (Snowden, 1995), this region primarily suffered from malaria (Snowden, 2006).[80] Endemic in the Southern countryside, this disease left the rural communities equally, if not more, vulnerable to potentially stunting insults to health and height.[81] There is also no clear evidence that rural households were less likely to be undernourished relative to urban ones.[82]

To test the role of urban over-representation in generating the selection patterns discussed above, we used the individual urban indicator, based on the 1901 population of the communes to which the passengers were matched. Based on this indicator, we can construct measures of the degree of urban over-representation within each group.[83] As shown in Figure B.11, the raw correlations are consistent with the urban over-representation interpretation of our results. Province-level urban over-representation is positively correlated

[77] Fernández-Huertas Moraga (2011, 2013) shows that rural and urban Mexican communities have different selection patterns. Abramitzky, Boustan, and Eriksson (2012) show similar evidence for Norway in the Age of Mass Migration.

[78] The studies on other southern European countries are cases in point; see Martínez-Carrión and Moreno-Lázaro (2007) on Castile-Leon and Southeast Spain, and Reis (2009) on Lisbon. See also Twarog (1997) on Württemburg.

[79] This is consistent with A'Hearn's (2003) finding that North Italian cities that were not provincial capitals did not have a height penalty.

[80] Another disease, particularly common in North Italy, was pellagra, ordinarily a result of a maize-based diet (though reduced consumption of meat may also cause niacin deficiency, the cause of pellagra; Prinzo, 2000). See A'Hearn (2003) for a discussion of the effects of malaria and pellagra on North Italian heights.

[81] In fact, there is a sense in which urban clusters provided a refuge from Malaria. This disease was held by some as one of the main causes for the prevalence of agro-towns—typically urban settlements situated uphill, above malaria-infested plains that were ill-suited for overnight stays—in the South. Most of the population was employed in agriculture and commuted daily to the fields, sometimes over great distances. See a discussion of the origins of the agro-town in Curtis (2013).

[82] According to Vecchi and Coppola (2006, Table 4), in 1881 the rates of undernourishment among agricultural and non-agricultural households were similar, and by 1901 a gap developed in favor of non-agricultural households.

[83] The degree of urban over-representation within a province is defined as the share of urban passengers in our sample, divided by the share of urban population within the province as recorded by the census.

with a province's average height, and the province-cohort z-score is negatively correlated with urban over-representation. In other words, urban centers are more strongly represented from taller provinces, and provinces with greater urban over-representation have weaker local selection (either less positive or more negative). However, the correlations are not strong, with coefficients of correlation of 0.078 and -0.135, and the relationships are not statistically significant. In Table 8, we test this explanation more formally. In column (1), we regress the province-cohort z-score on the province-level degree of urban over-representation. Consistent with the raw correlations, a ten percentage point increase in urban over-representation from the province was associated with a very small decline in z-score of 0.004 standard deviations, and the difference is not statistically significant.

Without knowing separately the urban and non-urban distributions of height in the population at risk for migration, we must be circumspect in drawing conclusions from these patterns;[84] but further within-province evidence from among the passengers is informative. In column (2), the coefficient on the urban indicator suggests that urban passengers were only 0.001 standard deviations shorter than their non-urban peers within their provinces, and this difference is not statistically significant; however, we cannot determine whether this reflects the lack of an urban height penalty or simply different selection across sectors that compensates for such a penalty. In column (3), the measure of urban over-representation is taken at the district level, allowing us to include province-fixed effects. The coefficient is practically zero, suggesting that individuals from districts with greater urban over-representation did not have different z-scores from other districts in the same province. Furthermore, as seen in column (4), controlling for the degree of urban over-representation does not change the baseline differential result from Table 3. Moreover, our baseline results are present in each of the two sectors of urban and rural passengers (Table A.9). Finally, in column (5) we exclude provinces containing cities of more than 100,000 inhabitants in order to account for the possibility that the urban height premium (or penalty) existed only in very large cities.[85] Again, urban over-representation does not show a meaningful relationship with the degree of self-selection, nor is the baseline result affected.

Taken together with the mixed historical evidence as to whether an urban height penalty was present in Italy, we interpret the evidence as suggesting that the correlations between the height of the population, urban over-representation, and local self-selection do not indicate that any potential urban height penalty accounts for the higher z-score of the shorter province-cohorts.

[84] For example, it could be the case that height is distributed independently of the urban status; in such case, variations in urban over-representation cannot account for variations in the degree of selection.

[85] We omit passengers from the provinces of Bologna, Catania, Florence, Genoa, Messina, Milan, Naples, Palermo, Rome, Turin, and Venice.

6.2.2 Occupations

Could variations in the occupational composition of migrants account for the selection patterns of Italian passengers? We briefly review evidence from the subsample for which we transcribed occupations to study this question. We classify occupations into four categories—professional, skilled or artisan, farm worker, and unskilled or unproductive. As a first step, Table A.10 verifies that the occupational status is correlated with our measure of selection. Professionals lead with a large gap, having at least one-half standard deviation greater z-score than agricultural workers and farmers. Skilled workers and artisans are distant seconds, with a z-score advantage of more than 0.1 standard deviations over farm workers. These gaps are consistent across specifications, and persist when controlling for province fixed-effects or when estimated separately for the South and the North.

The descriptive statistics do not suggest that the Southern advantage in local self-selection can be accounted for by a difference in the occupational composition of the passengers.[86] As shown in Table A.11, the shares of professional and skilled workers among passengers are comparable in the two regions, but the North had a much smaller share of farm workers and a greater share of unskilled and unproductive. In a simple comparison of means, South Italians had a greater z-score than Northerners in each occupational category, with the gap narrowest among farm workers (0.065 standard deviations). Regardless of these differences, the occupational composition of migrants cannot be used to explain why the selection was stronger from the South. The data cannot tell us whether the fact that Southern agricultural workers had a relatively narrow selection advantage over Northern ones was due to a relatively weaker selection into migration among them, or because in the South they were relatively worse off compared to the their non-farming peers within their provinces. Nonetheless, the pattern of differential selection across provinces persists when the sample is divided to farm and non-farm workers, as shown in Table A.12.

6.3 Substitution Between Destinations

One of the major differences between the streams of emigration from Southern and Northern Italy was that the latter was remarkably diversified in its destinations, including major movements to countries in South America and Western Europe. According to the Italian emigration statistics, on average more than 87 percent of the Northern emigrants chose destinations other than the United States, whereas only 45 percent

[86] This is not to say that there was no difference in local selection on the masis of occupation, as we have not compared migrants to the occupation distributions of their places of origin. Only the distributions of occupation among migrants are somewhat similar across regions.

of the Southerners did so (Tables A.4 and A.1).[87] If the selection into migration to non-US destinations was different from the selection into migration to the United States, and in particular, if these non-US places were relatively more attractive to the higher-quality passengers, then it is possible that the differential selection pattern that we observe may be accounted for not by selection into migration in general, but rather by selection into migration to a particular destination. For example, it could be that emigrants from all provinces were similarly selected, but other destinations drained the high-quality emigrants coming from the North, leaving the US with relatively poorly selected residual migration compared to cohorts coming from the South. This idea is consistent, for example, with the difference in experiences of Italian immigration in Argentina and the US. In the former, Italians performed better in terms of access to skilled labor and land, home, and business ownership than did Italian migrants to the United States (Klein, 1983), possibly offering better returns to skill as compared to the United States and thus drawing higher quality migrants.[88] It should be noted that the phenomenon of migration to multiple destinations does not threaten the validity of our results. Only our interpretation of these results is affected. Our results must be interpreted as the degree of migrant self-selection to the United States rather than the selection into emigration in total. We investigate below to what extent the former is indicative of the latter.

Although we do not have the appropriate data to approach this issue rigorously, we look for suggestive evidence regarding substitution of migrant quantity and quality across destinations using the Italian official statistics on emigration. The emigration data in the *Statistica della Emigrazione Italiana per l'Estero, 1876–1925* record the yearly number of emigrants from each province to each country.[89] In Table 9 we use these data to test for a correlation between exposure to non-US destinations—represented by the share of non-US emigrants out of all emigrants—and passengers' z-score. In column (1), we regress the province-cohort z-score on the log of the fraction of emigrants in a province-year migrating to a non-US destination, controlling for birth year and province fixed effects. The coefficient is negative, but not statistically significant. Taken

[87] Accordingly, either as a result or as cause, it could be that fewer Southerners faced the option to migrate to alternative destinations, as they were not directly linked to migration networks in countries other than the US.

[88] In the case of the Italian migration, Gomellini and Ó Gráda (2013, p. 274) argue, without formally testing the claim, that the rise in migration to Argentina in the 1900s was in part the result of a decline of migration to Brazil. Ardeni and Gentili (2014) estimated the Italian emigration separately by primary destinations. Hatton and Williamson (1998, ch. 6) estimate a rudimentary three-destination model for Italy (with the US, Argentina, and Brazil as destinations). Borjas, Bronars, and Trejo (1992) estimate a model of destination selection in interstate migration in the United States with multiple destinations, analyzing the selection of migrants to each destination. Abramitzky and Braggion (2006) study location choice with respect to migrant quality in a setting with two destinations (in particular, among British indentured servants in the 17th and 18th century). We are not aware of studies of the Age of Mass Migration that test substitution effects in the quality of migrants (i.e, the effect of conditions in, or option to migrate to, one country on the selection of immigration to another country). A handful of recent papers provide a framework to estimate such substitution effects using discrete choice or general equilibrium models of contemporary internal migration (Armenter and Ortega, 2010; Dahl, 2002; Gemici, 2011; Kennan and Walker, 2011) and of international migration (Bertoli, Moraga, and Ortega, 2013).

[89] The records are based on the required passport applications. See discussion of the origins and limitations of these data in Foerster (1919, ch. 1). They were previously used by Ardeni and Gentili (2014).

at face value, the magnitude of the coefficient (-0.034) implies that the difference in location choice between the North and the South accounts for 0.026 standard deviations of the z-score advantage in favor of the South, a small yet non-negligible share of the South's actual advantage. Controlling for province fixed effects in column (2) enables us to test whether within provinces, years of greater exposure to non-US migration are associated with lower z-score. The coefficient is unchanged, and it is still statistically insignificant. In column (3) we test whether the differential selection relative to height is affected by adding the exposure to non-US migration; we still find a similar coefficient on height as in Table 3, but the coefficient on the share of non-US migration switches signs, becoming positive and statistically significant. We conclude that we cannot rule out that to some degree, exposure to migration to other destinations caused more negative selection among the taller provinces. However, the evidence is mixed at best, and more conclusive statements would require both better experimental design and more precisely estimated coefficients. We therefore caution that our results should be interpreted as the degree of self-selection into migration to the United States, rather than into emigration from Italy in total.

6.4 Inequality

Finally, we evaluate how the leading hypotheses on the determinants of migrant self-selection are reflected in the height data. As discussed in Section 2.2, the relative inequality model predicts that greater returns to skill in the sending economy (often proxied by greater inequality) cause more negative self-selection, holding constant the returns to skill in the receiving economy.[90] Other views highlight the role of migration costs, liquidity constraints, and alternative forms of utility in generating positively selected migration, with network effects potentially mitigating the effects of costs and liquidity constraints. The results discussed above support both of these theories. The relatively high returns to skill in the United States compared to those in Italy support the positive selection that we have found. However, the large income differences between Italy and the United States also predict positive selection according to Grogger and Hanson's (2011) model, and the more positive selection from poorer areas that we have found agree with this model.

In column (1) of Table 10, we provide another test of the performance of the relative inequality theory by proxying for inequality using the coefficient of variation of height,[91] while the share of property owners

[90] A more rigorous examination of this model would require knowledge of the returns to skill among migrants from each province in the United States. As such data are not available, we approximate the returns to skill for migrants in the United States by some (unspecified) distribution that is constant across provinces of origin, permitting us to focus only on differences in inequality across provinces of origin.

[91] The use of the coefficient of variation of height as a measure of inequality follows Blum (2013) and Stolz and Baten (2012). In the absence of direct data on local variations in returns to skill, we consider the evidence gleaned from these regressions as suggestive and inconclusive.

in the province is meant to capture the ability to overcome liquidity constraints (though it may also capture aspects of inequality).[92] The coefficient of variation is indeed negatively correlated with z-score, as the relative inequality model predicts, but the relationship is statistically insignificant. The share of property owners is positively, strongly, and statistically significantly correlated with z-score.[93]

In column (2), we add controls for the average height of the province-cohort. The coefficient on the inequality measure is still negative, but is closer to zero than in column (1) and is still statistically insignificant. The differential selection with respect to height is robust to controlling for inequality, and the negative correlation is consistent with an interpretation that would regard improved standards of living as a proxy for the ability to overcome the liquidity constraints. In columns (3) and (4) we include individual characteristics that would suggest the network status of the passengers. Controlling for province fixed-effects in column (4), passengers paying for their own passage were much taller whereas those linked to an immediate family member were significantly shorter (though the significance is only at the 10 percent level). This is strongly consistent with the view that thicker networks reduce the quality of passengers: the migration of disadvantaged individuals is more dependent on each other's support.

7 Conclusions

The finely disaggregated data that we use in this paper enable the investigation of selection on the local level, breaking from the common tendency in the migration literature to focus on migrant selection with respect to national averages. It reveals that the seemingly disadvantaged Southern Italian immigrants were, indeed, "the best of their class," and sheds light on the meaning of the Italian migration for the receiving and the sending economies. It suggests that South Italy experienced a human capital drain, one that may have contributed to the contemporaneously widening North-South divide within Italy.[94] While our investigation focuses on the case of Italy, we believe that the cross-regional trend of increasingly positive selection from disadvantaged provinces and the evidence on weaker quality of immigrants with close network support provide an important lesson that contributes to the understanding of the economics of migration in general, by lending further empirical support for theories that assign important roles for fixed costs and network effects. Moreover, the aftermath of the 1917 literacy requirement shows that a policy measure that restricted migration based on one crude measure of quality was quite effective in improving the selection of immigrants based on other

[92] Hatton (2010) shows that controlling for the presence of liquidity constraints is essential to performing a proper test of the relative inequality model.

[93] A one-standard deviation increase in share of proprietors is associated with a 0.067 standard deviation increase in z-score.

[94] This is in line with Mokyr and Ó Gráda (1982), who suggested that positively selected emigration could have been partly responsible for hindering Irish development.

measures.

The main lesson that we take away is a call for attention to the quality of migrants relative to their local environment, and not only in absolute terms or relative to a larger national pool. We argue that the tendency to focus on self-selection with respect to national distributions of quality masks an important part of the potential transfer in human capital between countries, and we show that the greatest flow of human capital may, in fact come from the poorest areas, and among migrants who would appear to be negatively selected in conventional analyses. This lesson is particularly important when debating the benefits of immigration from large and widely diverse countries, such as Mexico, China, and India. Policy makers may do well to notice that the greatest gains in human capital might come from those among whom it is least expected.

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.

——— (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 and Fabio Braggion (2006). “Migration and Human Capital: Self-Selection of Indentured Servants to the Americas.” *The Journal of Economic History* 66:4, pp. 882–905.

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.* Giovanni Vecchi (ed.). 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.

Ardeni, Pier Giorgo and Andrea Gentili (2014). “Revisiting Italian Emigration before the Great War: A Test of the Standard Economic Model.” *European Review of Economic History* 18:4, pp. 452–471.

Armenter, Roc and Francesc Ortega (2010). “Credible Redistributive Policies and Migration across US States.” *Review of Economic Dynamics* 13:2, pp. 403–423.

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.

Baten, Jörg and John E. Murray (2000). “Heights of Men and Women in 19th-Century Bavaria: Economic, Nutritional, and Disease Influences.” *Explorations in Economic History* 37:4, pp. 351–369.

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.

Beine, Michel, Frédéric Docquier, and Çağlar Özden (2011). “Diasporas.” *Journal of Development Economics* 95, pp. 30–41.

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

Bertoli, Simone, Jesús Fernández-Huertas Moraga, and Francesc Ortega (2013). “Crossing the Border: Self-Selection, Earnings and Individual Migration Decisions.” *Journal of Development Economics* 101, pp. 75–91.

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.

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.

Blum, Matthias and Jörg Baten (2011). “Anthropometric within-country Inequality and the Estimation of Skill Premia with Anthropometric Indicators.” *Jahrbuch für Wirtschaftswissenschaften (Review of Economics)* 62:2, pp. 107–138.

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

——— (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.

——— (2014). "Caveat Lector: Sample Selection in Historical Heights and the Interpretation of Early Industrializing Economies." NBER Working Paper 19955.

——— (2015). "Sample-Selection Biases and the 'Industrialization Puzzle'." NBER Working Paper 21249.

Bonifazi, Corrado and Frank Heins (2000). "Long-term Trends of Internal Migration in Italy." *International Journal of Population Geography* 6, pp. 111–131.

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

——— (2014). *Immigration Economics.* Cambridge: Harvard University Press.

Borjas, George J., Stephen G. Bronars, and Stephen J. Trejo (1992). "Self-Selection and Internal Migration in the United States." *Journal of Urban Economics* 32, pp. 159–185.

Borjas, George J., Ilpo Kauppinen, and Panu Poutvaara (2015). "Self-Selection of Emigrants: Theory and Evidence on Stochastic Dominance in Observable and Unobservable Characteristics." NBER Working Paper 21649.

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. (1978). "The Effect of Americanization on the Earnings of Foreign-born Men." *Journal of Political Economy* 86:5, pp. 897–921.

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

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 Önder (2014). “The Research Productivity of New PhDs in Economics: The Surprisingly High Non-Success of the Successful.” *Journal of Economic Perspectives* 28:3, pp. 205–216.

Connor, Dylan Shane (2015). “The Cream of the Crop? Selectivity and Local Determinants of Migration from Ireland to North American in the Early 20th Century.” Mimeo., UCLA.

Costa, Dora L. and Richard H. Steckel (1997). “Long-Term Trends in Health, Welfare, and Economic Growth in the United States.” In *Health and Welfare During Industrialization.* Richard H. Steckel and Roderick Floud (ed.). Chicago: University of Chicago Press. Chap. 2, pp. 47–90.

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.

Curtis, Daniel (2013). “Is there an ‘Agro-Town’ Model for Southern Italy? Exploring the Diverse Roots and Development of the Agro-Town Structure through a Comparative Case Study in Apulia.” *Continuity and Change* 28:3, pp. 377–419.

Dahl, Gordon B. (2002). “Mobility and the Return to Education: Testing a Roy Model with Multiple Markets.” *Econometrica* 70:6, pp. 2367–2420.

Daniels, Roger (2004). *Guarding the Golden Door: American Immigration Policy and Immigrants since 1882.* New York: Hill and Wang.

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.

Docquier, Frédéric and Abdeslam Marfouk (2006). “International Migration by Educational Attainment, 1990–2000.” In *International Migration, Remittances, and the Brain Drain.* Çağlar Özden and Maurice Schiff (ed.). Washington, D.C.: The World Bank and Palgrave McMillan, pp. 151–200.

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

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.

——— (2013). “Understanding Different Migrant Selection Patterns in Rural and Urban Mexico.” *Journal of Development Economics* 103, pp. 182–201.

Floud, Roderick (1985). “Measuring the Transformation of the European Economies: Income, Health, and Welfare.” *Historical Social Research* 33, pp. 25–41.

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*. Stanley L. Engerman and Robert E. Gallman (ed.). Chicago: University of Chicago Press, pp. 439–556.

——— (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, Roderick Floud, Gerald Friedman, Robert A. Margo, Kenneth Sokoloff, Richard H. Steckel, T. James Trussell, Georgia Villaflor, and Kenneth W. Wachter (1983). "Secular Changes in American and British Stature and Nutrition." *The Journal of Interdisciplinary History* 14:2, pp. 445–481.

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.

Früwirth-Schnatter, Sylvia (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

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.

Gemici, Ahu (2011). "Family Migration and Labor Market Outcomes." Mimeo., New York University.

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*. Claudia Goldin and Gary D. Libecap (ed.). 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*. Gianni Toniolo (ed.). New York: Oxford University Press. Chap. 10, pp. 271–302.

Gould, Eric D. and Omer Moav (2015). "Does High Inequality Attract High Skilled Immigrants." *The Economic Journal* Forthcoming.

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.

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.

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.

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.

Haines, Michael R., Lee A. Craig, and Thomas Weiss (2003). "The Short and the Dead: Nutrition, Mortality, and the "Antebellum Puzzle" in the United States." *The Journal of Economic History* 53:2, pp. 382–413.

Haines, Michael R. and Richard H. Steckel (2000). "Childhood Mortality and Nutritional Status as Indicators of Standard of Living: Evidence from World War I Recruits in the United States." *Jahrbuch für Wirtschaftsgeschichte* 43:1.

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

Hanushek, Eric A. and Lei Zhang (2009). "Quality-Consistent Estimates of International Schooling and Skill Gradients." *Journal of Human Capital* 3:2, pp. 107–143.

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. (2010). "The Cliometrics of International Migration: A Survey." *Journal of Economic Surveys* 24:5, pp. 941–969.

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.

——— (2005). *Global Migration and the World Economy: Two Centuries of Policy and Performance*. Cambridge: MIT Press.

Hing, Bill Ong (2004). *Defining America: Through Immigration Policy*. Philadelphia: Temple University Press.

Horrell, Sara, David Meredith, and Deborah Oxley (2009). "Measuring Misery: Body Mass, Ageing and Gender Inequality in Victorian London." *Explorations in Economic History* 46, pp. 93–119.

Horrell, Sara and Deborah Oxley (2015). "Gender Discrimination in 19th Century England: Evidence from Factory Children." University of Oxford Discussion Papers in Economic and Social History.

Humphries, Jane and Timothy Leunig (2009a). "Cities, Market Integration, and Going to Sea: Stunting and the Standard of Living in Early Nineteenth-Century England and Wales." *The Economic History Review* 62:2, pp. 458–478.

——— (2009b). "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*. George J. Borjas (ed.). Chicago: University of Chicago Press. Chap. 5, pp. 159–192.

Jayachandran, Seema and Rohini Pande (2015). "Why are Indian Children so Short?" NBER Working Paper 21036.

Kaestner, Robert and Ofer Malamud (2014). "Self-Selection and International Migration: New Evidence from Mexico." *The Review of Economics and Statistics* 96:1, pp. 78–91.

Kennan, John and James R. Walker (2011). "The Effect of Expected Income on Individual Migration Decisions." *Econometrica* 79:1, pp. 211–251.

Klein, Herbert S. (1983). "The Integration of Italian Immigrants into the United States and Argentina: A Comparative Analysis." *The American Historical Review* 88:2, pp. 306–329.

Komlos, John (1987). "The Height and Weight of West Point Cadets: Dietary Change in Antebellum America." *The Journal of Economic History* 47:4, pp. 897–927.

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

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 (2014). "Who Crossed the Border? Self-Selection of Mexican Migrants in the Early 20th Century." *The Journal of Economic History* 74:4, pp. 1015–1044.

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.

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.

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

Martínez-Carrión, José-Miguel and Javier Moreno-Lázaro (2007). "Was there an Urban Height Penalty in Spain, 1840–1913." *Economics and Human Biology* 5:1, pp. 144–164.

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

Massey, Catherine (2012). "Immigration Quotas and Immigrant Skill Composition: Evidence from the Pacific Northwest." University of Colorado at Boulder Working Paper 12-07.

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 (2007). "Network Effects and the Dynamics of Migration and Inequality: Theory and Evidence from Mexico." *Journal of Development Economics* 84, pp. 1–24.

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

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

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

——— (1996). "Height and Health in the United Kingdom 1815–1860: Evidence from the East India Company Army." *Explorations in Economic History* 33, pp. 141–168.

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.

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.

Orrenius, Pia M. and Madeline Zavodny (2005). "Self-Selection among Undocumented Immigrants from Mexico." *Journal of Development Economics* 78, pp. 215–240.

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.

Prinzo, Zita Weise (2000). “Pellagra and its Prevention and Control in Major Emergencies.” Mimeo., World Health Organization, Department of Nutrition for Health and Development.

Psacharoupolos, George and Harry Anthony Patrinos (2004). “Returns to Investment in Education: A Further Update.” *Economics of Education* 12:2, pp. 111–134.

Reis, Jaime (2009). “Urban Premium or Urban Penalty? The Case of Lisbon, 1840–1912.” *Historia Agraria* 47.

Robertson, Craig (2010). *The Passport in America: The History of a Document.* New York: Oxford University Press.

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.

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

Snowden, Frank M. (1995). *Naples in the Time of Cholera, 1884–1911.* Cambridge: Cambridge University Press.

——— (2006). *The Conquest of Malaria: Italy, 1900–1962.* New Haven: Yale University Press.

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.

——— (2003). “Changing Times, Changing Faces: Franz Boas’s Immigrant Study in Modern Perspective.” *American Anthropologist* 105:2, pp. 333–337.

Spitzer, Yannay (2015a). “Pogroms, Networks, and Migration: The Jewish Migration from the Russian Empire to the United States, 1881–1914.” Mimeo., Hebrew University of Jerusalem.

——— (2015b). “The Dynamics of Mass Migration: Estimating the Effect of Income Differences on Migration in a Dynamic Model with Diffusion.” Mimeo., Hebrew University of Jerusalem.

Steckel, Richard H. (1986). “A Peculiar Population: The Nutrition, Health, and Mortality of American Slaves from Childhood to Maturity.” *The Journal of Economic History* 46:3, pp. 721–741.

——— (1995). “Stature and the Standard of Living.” *Journal of Economic Literature* 33:4, pp. 1903–1940.

——— (2008). “Biological Measures of the Standard of Living.” *Journal of Economic Perspectives* 22:1, pp. 129–152.

Steckel, Richard H. (2009). "Heights and Human Welfare: Recent Developments and New Directions." *Explorations in Economic History* 46, pp. 1–23.

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.

Twarog, Sophia (1997). "Heights and Living Standards in Germany, 1850–1939: The Case of Württemberg." In *Health and Welfare During Industrialization*. Richard H. Steckel and Roderick Floud (ed.). Chicago: University of Chicago Press, pp. 285–330.

US Congress (1911). *Reports of the Immigration Commission*. Vol. 1. Washington: Government Printing Office, 61st Congress, 3rd Session, Document No. 747.

Vecchi, Giovanni and Michela Coppola (2006). "Nutrition and Growth in Italy, 1861–1911: What Macroeconomic Data Hide." *Explorations in Economic History* 43:3, pp. 438–464.

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.

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

——— (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.

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*. R. L. Hall (ed.). Praeger.

Zimran, Ariell (2015). "Does Sample-Selection Bias Explain the Antebellum Puzzle? Evidence from Military Enlistment in the Nineteenth-Century United States." Mimeo., Northwestern University.

Zolberg, Aristide R. (2006). *A Nation by Design: Immigration Policy in the Fashioning of America*. New York: Russell Sage Foundation and Harvard University Press.

Tables

Table 1: Summary statistics.

		First-Timers Only		
	(1)	(2)	(3)	(4)
Variable	All	All	South	North
Age	31.402	29.858	30.049	29.047
	(8.061)	(7.724)	(7.835)	(7.181)
Married	0.703	0.617	0.639	0.526
	(0.457)	(0.486)	(0.480)	(0.499)
Italian Port	0.867	0.857	0.924	0.573
	(0.340)	(0.350)	(0.265)	(0.495)
Post 1917	0.218	0.204	0.198	0.228
	(0.413)	(0.403)	(0.399)	(0.420)
Southern	0.829	0.809		
	(0.376)	(0.393)		
Urban	0.320	0.342	0.389	0.142
	(0.467)	(0.474)	(0.487)	(0.349)
	[22,848]	[12,457]	[10,016]	[2,441]
Height (cm)	163.943	163.874	163.490	165.493
	(7.148)	(7.468)	(7.528)	(6.984)
Repeater	0.441			
	(0.497)			
Imm. Fam. Conn.	0.319	0.312	0.317	0.287
	(0.466)	(0.463)	(0.465)	(0.453)
Any Conn.	0.947	0.967	0.973	0.944
	(0.224)	(0.178)	(0.162)	(0.230)
Paid for Self	0.887	0.916	0.911	0.938
	(0.317)	(0.277)	(0.285)	(0.240)
Observations	23,386	12,755	10,237	2,518

Notes: The sample covered in this table includes transcribed and successfully geolocated male migrants, ages 22–65. Standard deviations are in parentheses. Sample sizes are the minimum with observations for all variables, except for urban, which is available only for a subset of more precisely geolocated individuals; the square brackets under urban denote the number of individuals with data for all observations, including urban. Urban is defined using population counts from the 1901 Italian census, defining an urban locality as one with a population of 10,000 or more; it is an individual-level indicator. Imm. Fam. Conn. is Immediate Family Connection; Any Conn. is Any Connection; Southern is based on the location to which the individual was matched, as is the division in columns (3) and (4).

Table 2: All-Italian selection.

Variables	(1)	(2)	(3)
Southern		-0.333^a (0.023)	-0.330^a (0.023)
Constant	-0.105^a (0.009)	0.164^a (0.021)	
Observations	12,881	12,881	12,881
R-squared	0.000	0.017	0.028
Arrival Year FE	No	No	Yes
Birth Year FE	No	No	Yes
Constant + Southern		-0.169^a (0.010)	

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Dependent variable is height, standardized by all-Italy-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. All standard errors are clustered on the family level. The lower section presents the sums of estimated coefficients and their standard errors. Constants are not reported in the presence of fixed effects.

Table 3: Local selection.

Variables	(1)	(2)	(3)	(4)
Southern		0.146^{a}		-0.145^{b}
		(0.026)		(0.060)
Average Height (cm)			-0.064^{a}	-0.102^{a}
			(0.005)	(0.021)
Southern × Average Height (cm)				0.031
				(0.022)
Constant	0.037^{a}	-0.081^{a}		
	(0.011)	(0.023)		
Observations	12,881	12,881	12,881	12,881
R-squared	0.000	0.003	0.024	0.025
Arrival Year FE	No	No	Yes	Yes
Birth Year FE	No	No	Yes	Yes
Constant + Southern		0.065^{a}		
		(0.012)		
Average Height + Southern × Average Height				-0.071^{a}
				(0.008)

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Dependent variable is height, standardized by province-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. All standard errors are clustered on the province-birth cohort level. Aveage Height is of the province-birth cohort and is demeaned. The lower section presents the sums of estimated coefficients and their standard errors. Constants are not reported in the presence of fixed effects.

Table 4: Selection and migration probabilities.

	(1)	(2)	(3)
Variables	All	South	North
Standardized Height × 0.01	0.360	0.838	−0.119
	$[0.205, 0.515]$	$[0.567, 1.110]$	$[-0.205, -0.032]$
Constant	0.077	0.116	0.020
	$[0.077, 0.078]$	$[0.116, 0.117]$	$[0.020, 0.021]$
Observations	12548	10073	2475
Scaled Standardized Height	0.018	0.042	−0.006

Notes: The dependent variable is migration probability, conditional on province-birth cohort-standardized height as determined by Bayes's Theorem. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. Estimation is weighted by the probability of migration. Bootstrap 95% confidence intervals in square brackets. The standardized height is standardized by the province-birth cohort mean and standard deviation. The Scaled Standardized Height coefficient is 0.05 times that of standardized height × 0.01 (that is, five times the coefficient on standardized height).

Table 5: Balancing tests for height.

Variables	(1) National	(2)	(3)	(4) South	(5) North
Geolocated	−0.029 (0.026)	−0.002 (0.027)	−0.019 (0.027)	−0.034 (0.030)	0.057 (0.144)
Surname-Implied Average Height (cm)			-0.088^{a} (0.013)	-0.105^{a} (0.016)	-0.090^{b} (0.044)
Surname-Implied Average Height (cm) × Geolocated			0.009 (0.014)	−0.007 (0.017)	0.012 (0.048)
Observations	14,309	14,309	14,309	12,631	1,678
R-squared	0.013	0.014	0.030	0.036	0.061
Arrival Year FE	Yes	Yes	Yes	Yes	Yes
Birth Year FE	Yes	Yes	Yes	Yes	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Dependent variable is height standardized by surname-implied province-birth cohort mean and standard deviation in every column except column (1), where it is height normalized by the national mean and standard deviation. The sample covered in this table consists of male migrants aged 22–65 making a first arrival who could be matched to a province by their surname. Standard errors are clustered by surname-implied province-birth cohort. Average Height is that of the surname-implied province-birth cohort, and is demeaned. North and South refer to surname implied provinces. Constants are not reported in the presence of fixed effects.

Table 6: Robustness to geolocation errors.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables	Ethnicity	Ethnicity	Ethnicity	Ethnicity	Surname	Surname	Surname	Surname
Southern		0.032		-0.330^{a}		-0.009		-0.304^{b}
		(0.028)		(0.069)		(0.054)		(0.146)
Average Height (cm)			-0.052^{a}	-0.124^{a}			-0.045^{a}	-0.086^{c}
			(0.006)	(0.023)			(0.010)	(0.045)
Southern × Average Height (cm)				0.047^{c}				0.011
				(0.024)				(0.046)
Constant	0.051^{a}	0.024			0.049^{b}	0.057		
	(0.011)	(0.025)			(0.020)	(0.049)		
Observations	11,852	11,852	11,852	11,852	3,320	3,320	3,320	3,320
R-squared	0.000	0.000	0.021	0.025	0.000	0.000	0.036	0.040
Arrival Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Birth Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Constant + Southern		0.056^{a}				0.048^{b}		
		(0.012)				(0.021)		
Average Height + Southern × Average Height				-0.076^{a}				-0.075^{a}
				(0.008)				(0.013)

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Dependent variable is height, standardized by province-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. All standard errors are clustered on the province-birth cohort level. Columns with the header Ethnicity exclude individuals whose ethnicity conflicts with the province to which they were geomatched (e.g., North Italians matched to a southern province). Columns with the header Surname include only individuals whose surname-implied province matches their actual province. Average Height is of the province-birth cohort and is demeaned. The lower section presents the sums of certain estimated coefficients and their standard errors.

Table 7: Selection before and after 1917.

Variables	(1)	(2)	(3)	(4)
Post-1917	0.147^a (0.024)	0.055 (0.052)		
Southern		0.123^a (0.030)		
Post-1917 × Southern		0.123^b (0.059)		
Average Height (cm, demeaned)			-0.060^a (0.006)	
Post-1917 × Average Height (cm)			−0.017 (0.012)	
Post-1917 × Male Literacy Rate				-0.331^b (0.150)
Constant	0.007 (0.012)	-0.094^a (0.027)		
Observations	12,881	12,881	12,881	12,881
R-squared	0.003	0.007	0.024	0.034
Arrival Year FE	No	No	Yes	Yes
Birth Year FE	No	No	Yes	Yes
Province FE	No	No	No	Yes
Constant + Post-1917	0.154^a (0.021)	−0.039 (0.045)		
Constant + Southern		0.029^b (0.014)		
Post-1917 + Post-1917 × Southern		0.178^a (0.026)		
Constant + Southern + Post-1917 + Post-1917 × Southern		0.207^a (0.023)		
Average Height + Post-1917 × Average Height			-0.078^a (0.010)	

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Dependent variable is height, standardized by province-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. Standard errors are clustered by province-birth cohort, except in columns (4)–(6), in which they are clustered by province. Average Height is of the province-birth cohort and is demeaned. Post-1917 includes 1917 arrivals. The division by regions in columns (5) and (6) is based on geolocation. The male literacy rate is from the 1911 Italian census and is on the province level. The lower section presents the sums of certain estimated coefficients and their standard errors. Constants are not reported in the presence of fixed effects.

Table 8: Urban overrepresentation and selection.

	(1)	(2)	(3)	(4)	(5)
Variables					Under 100k
Urban Over.	−0.042			−0.039	−0.025
	(0.083)			(0.054)	(0.056)
Urban		−0.001			
		(0.024)			
Urban Over. (District)			−0.011		
			(0.035)		
Average Height (cm)				-0.064^a	-0.064^a
				(0.009)	(0.010)
Observations	12,831	12,577	10,344	12,831	8,986
R-squared	0.012	0.035	0.037	0.024	0.029
Arrival Year FE	Yes	Yes	Yes	Yes	Yes
Birth Year FE	Yes	Yes	Yes	Yes	Yes
Province FE	Yes	No	Yes	No	No

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Dependent variable is height, standardized by province-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. The column with the header Under 100k excludes individuals from provinces with cities of over 100,000 individuals. The column with the header Surname includes only individuals whose surname-implied province is the same as their location-implied province. Urban is defined using population counts from the 1901 Italian census, defining an urban locality as one with population 10,000 or more. Urban is an individual-level indicator. Urban overrepresentation is the fraction of 1907–1925 migrants from a province or district coming from an urban locality divided by the fraction of the population of that province or district living in an urban locality in 1901. Standard errors are clustered on the province level, except in column (2), in which they are clustered on the family level, and column (3), in which they are clustered on the district level. Constants are not reported in the presence of fixed effects.

Table 9: Multiple destinations and selection.

Variables	(1)	(2)	(3)
log(Fraction Emigrating)		0.108[c]	0.053
		(0.059)	(0.032)
log(Fraction of Emigrants to Non-US)	−0.034	−0.045	0.054[b]
	(0.026)	(0.064)	(0.023)
Average Height (cm)			−0.068[a]
			(0.009)
Observations	11,185	11,185	11,185
R-squared	0.010	0.033	0.022
Arrival Year FE	Yes	Yes	Yes
Birth Year FE	Yes	Yes	Yes
Province FE	No	Yes	No

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

Notes: Dependent variable is height standardized by province-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. Average Height is of the province-birth cohort and is demeaned. Standard errors are clustered on the province level. The variable log(Fraction Emigrating) is the log of the ratio of total emigration from a province in a particular year of arrival to its most recent population measure (from the 1901 or 1911 Italian census). The variable log(Fraction of Emigrants to non-US) is the log of the ratio of the number of emigrants from a province in a particular year of arrival to non-US destinations to total emigration from that province in that year. The sample is restricted to pre-1920 (inclusive) arrivals. Constants are not reported in the presence of fixed effects.

Table 10: Inequality, liquidity, and selection.

Variables	(1)	(2)	(3)	(4)
Stature CV	−0.041 (0.072)	−0.026 (0.051)	−0.026 (0.049)	
Fraction Owning Property	1.280[a] (0.351)	0.512 (0.318)	0.486 (0.319)	
Average Height (cm)		−0.060[a] (0.009)	−0.060[a] (0.009)	
Urban			−0.016 (0.025)	0.002 (0.024)
Imm. Fam. Conn.			−0.039[c] (0.021)	−0.037[c] (0.021)
Paid for Self			0.104[b] (0.043)	0.114[a] (0.042)
Observations	12,881	12,881	12,510	12,510
R-squared	0.015	0.024	0.026	0.036
Arrival Year FE	Yes	Yes	Yes	Yes
Birth Year FE	Yes	Yes	Yes	Yes
Province FE	No	No	No	Yes

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

Notes: Dependent variable is height, standardized by province-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. Standard errors are clustered on the province level, except in column (4) in which they are clustered on the province-birth cohort level. Average Height is of the province-birth cohort and is demeaned. Urban is defined using population counts from the 1901 Italian census, defining an urban locality as one with a population of 10,000 or more. Urban is an individual-level binary variable. Stature CV is 100 times the ratio of the province-birth cohort standard deviation and mean. Property ownership is from the 1901 Italian census. Imm. Fam. Conn. is Immediate family connection. Constants are not reported in the presence of fixed effects.

Figures

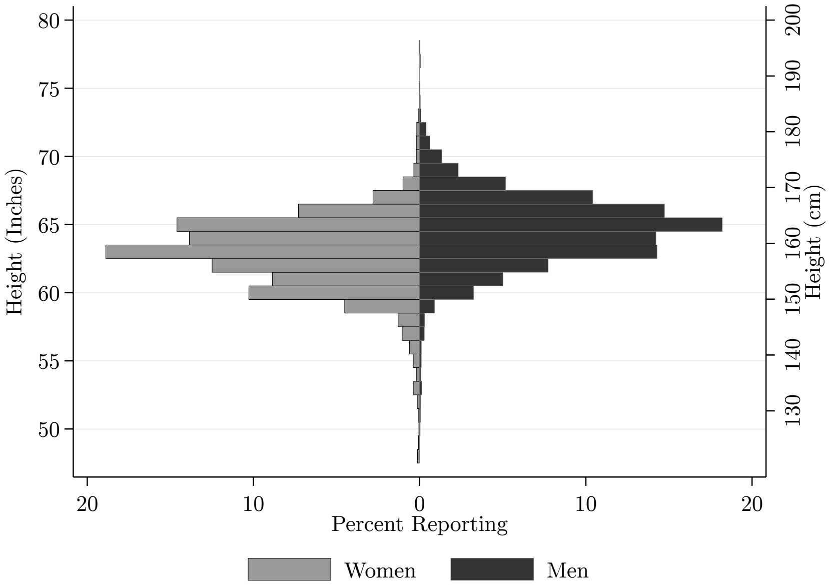


Figure 1: Passenger height distributions.

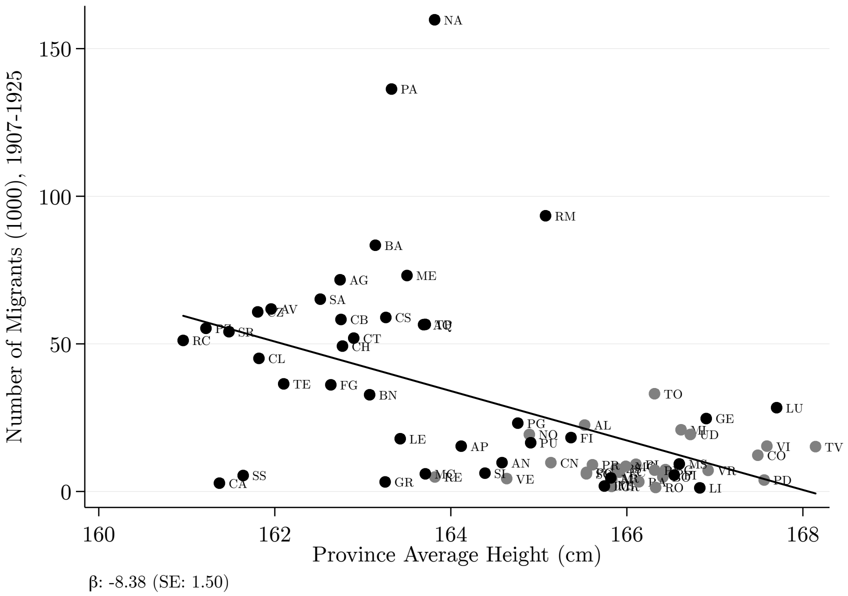
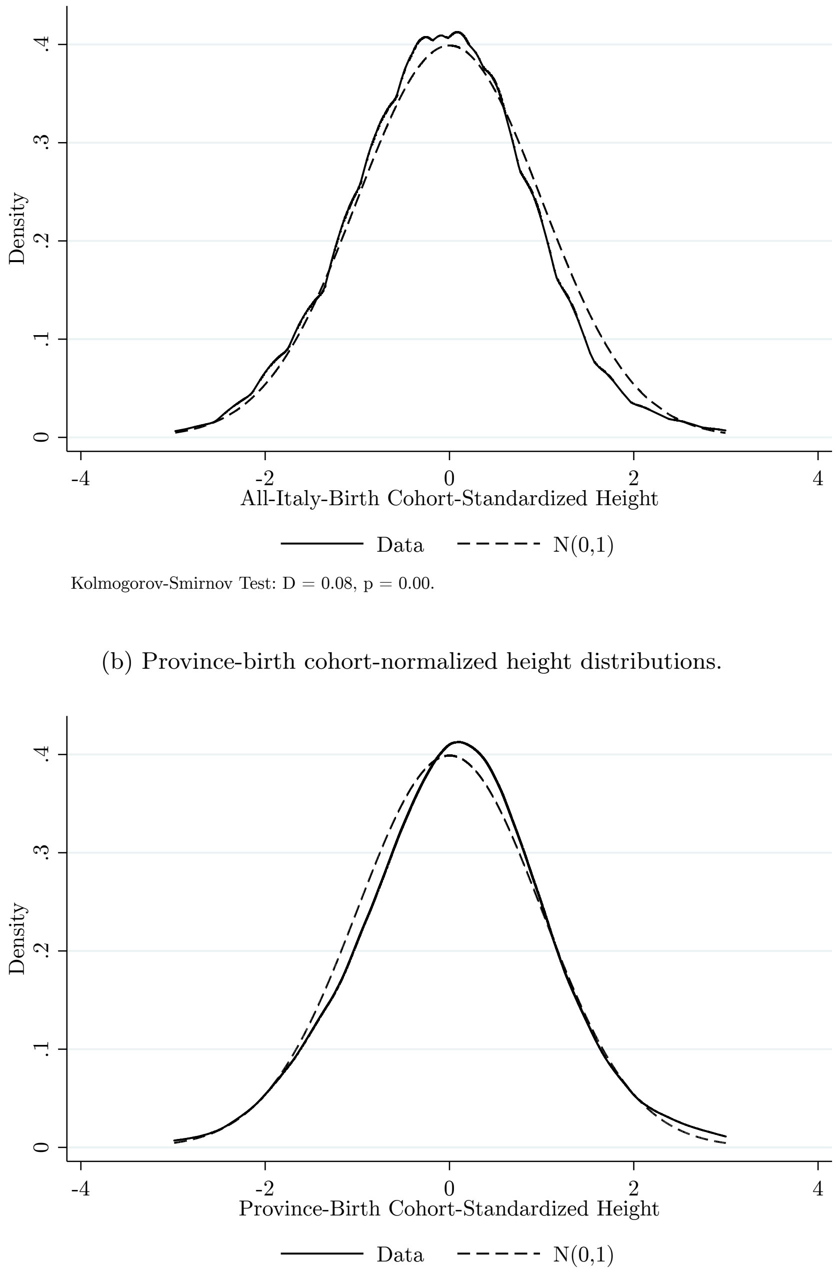


Figure 2: Migration and average heights by province.

Note: Northern provinces in gray, southern provinces in black. The number of migrants is based on the results of our geolocation algorithm. Average height is weighted across birth years by the number of migrants in our sample.



(a) Countrywide birth cohort-normalized height distributions.

Kolmogorov-Smirnov Test: D = 0.08, p = 0.00.

(b) Province-birth cohort-normalized height distributions.

Kolmogorov-Smirnov Test: D = 0.04, p = 0.00.

Figure 3: 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. Kolmogorov-Smirnov tests are conducted to test whether these distributions are different from a hypothetical $N(0,1)$ distribution.

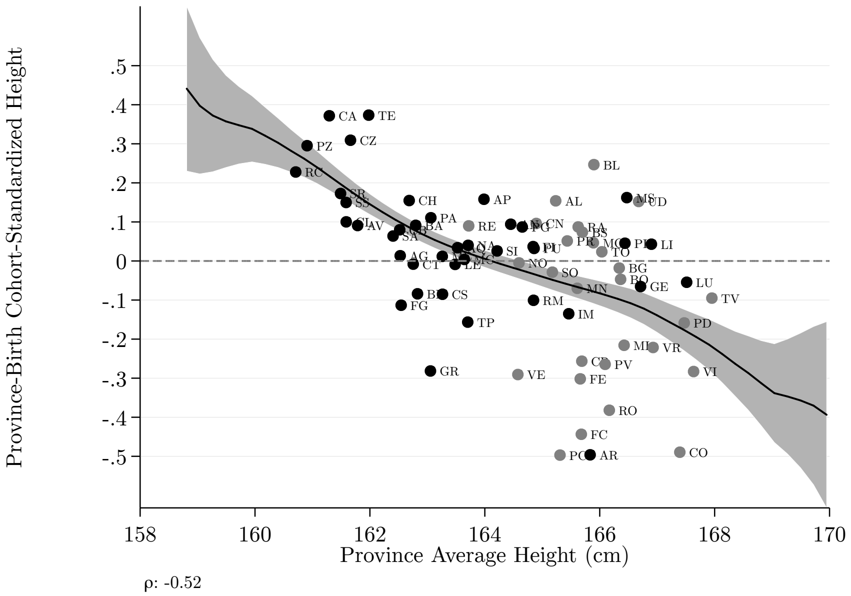


Figure 4: Province-birth cohort z-score and average province height.

Note: Northern provinces in gray, southern provinces in black. Average height is weighted within province across birth cohorts by the number of migrants in our sample. The line is the local polynomial regression of the same relationship using individual-level data. The shaded region is the 95% confidence interval for that regression.

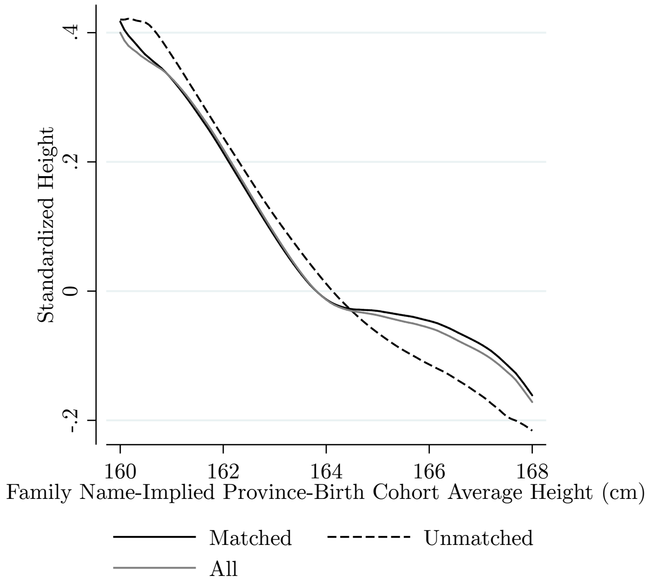


Figure 5: Differential imbalance from errors in geolocation.

Note: Standardized height is based on the surname-implied province-birth cohort mean and standard deviation. The imputation is described in Appendix F.

A Appendix Tables

Table A.1: Emigration by destination according to official statistics.

Destination	(1) All	(2) North	(3) South
United States	0.380	0.091	0.582
Canada	0.014	0.008	0.018
Argentina	0.108	0.079	0.128
Brazil	0.048	0.026	0.064
Germany	0.092	0.178	0.031
France	0.123	0.181	0.082
Switzerland	0.109	0.230	0.024
Other Europe	0.096	0.189	0.031
Other Americas	0.010	0.005	0.013
Other	0.020	0.011	0.027

Notes: These counts are taken from Table V of the *Statistica della Emigrazione Italiana per l'Estero*, and represent the fraction of all emigration going to each destination.

Table A.2: Sample size by province.

Province	Abbreviation	*N*	Province	Abbreviation	*N*	Province	Abbreviation	*N*	Province	Abbreviation	*N*
Lombardy			**Sicily**			**Piedmont**			**Campania**		
Bergamo	BG	91	Agrigento	AG	453	Alessandria	AL	153	Avellino	AV	346
Brescia	BS	93	Caltanissetta	CL	256	Cuneo	CN	147	Benevento	BN	202
Como	CO	158	Catania	CT	381	Novara	NO	150	Napoli	NA	917
Cremona	CR	42	Messina	ME	400	Torino	TO	267	Salerno	SA	366
Mantova	MN	26	Palermo	PA	718	**Marches**			**Sardinia**		
Milano	MI	149	Siracusa	SR	420	Ancona	AN	56	Cagliari	CA	47
Pavia	PV	81	Trapani	TP	301	Ascoli Piceno	AP	89	Sassari	SS	79
Sondrio	SO	51	**Emilia**			Macerata	MC	53	**Tuscany**		
Venetia			Bologna	BO	57	Pesaro e Urbino	PU	138	Arezzo	AR	25
Belluno	BL	60	Ferrara	FE	17	**Abruzzi**			Firenze	FI	109
Padova	PD	49	Forlí-Cesena	FC	41	Campobasso	CB	322	Grosseto	GR	18
Rovigo	RO	36	Modena	MO	51	Chieti	CH	329	Livorno	LI	3
Treviso	TV	158	Parma	PR	56	l'Aquila	AQ	290	Lucca	LU	183
Udine	UD	207	Piacenza	PC	51	Teramo	TE	277	Massa-Carrara	MS	63
Venezia	VE	22	Ravenna	RA	10	**Apulia**			Pisa	PI	33
Vernoa	VR	62	Reggio Emilia	RE	72	Bari	BA	517	Siena	SI	21
Vicenza	VI	189	**Umbria**			Foggia	FG	321	**Calabria**		
Liguria			Perugia	PG	222	Lecce	LE	175	Catanzaro	CZ	350
Genova	GE	209	**Latium**			**Basilicata**			Cosenza	CS	315
Imperia	IM	28	Roma	RM	631	Potenza	PZ	351	Reggio Calabria	RC	378

Notes: Boldface areas are regions. Underlying provinces fall into these regions. Sample sizes are the numbers of men with usable height data making a first arrival in the United States.

Table A.3: Summary statistics.

	All Passengers			Transcribed Only							
						Males					
							First-Timers Only				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Variable	All	Males	Females	All	Females	All	All	South	North	Pre-1917	Post-1917
Age	31.945	31.722	32.658	31.502	31.841	31.402	29.858	30.049	29.047	30.280	28.207
	(9.026)	(8.554)	(10.371)	(8.368)	(9.327)	(8.061)	(7.724)	(7.835)	(7.181)	(7.763)	(7.342)
Married	0.696	0.697	0.693	0.701	0.694	0.703	0.617	0.639	0.526	0.648	0.500
	(0.460)	(0.459)	(0.461)	(0.458)	(0.461)	(0.457)	(0.486)	(0.480)	(0.499)	(0.478)	(0.500)
Italian Port	0.877	0.868	0.904	0.874	0.899	0.867	0.857	0.924	0.573	0.853	0.871
	(0.329)	(0.339)	(0.294)	(0.332)	(0.302)	(0.340)	(0.350)	(0.265)	(0.495)	(0.354)	(0.335)
Post 1917	0.236	0.213	0.312	0.241	0.319	0.218	0.204	0.198	0.228		
	(0.425)	(0.409)	(0.463)	(0.428)	(0.466)	(0.413)	(0.403)	(0.399)	(0.420)		
Male	0.763			0.771							
	(0.425)			0.420							
Southern	0.848	0.849	0.845	0.828	0.823	0.829	0.809			0.815	0.786
	(0.359)	(0.358)	(0.362)	(0.377)	(0.382)	(0.376)	(0.393)			(0.389)	(0.410)
Urban	0.381	0.367	0.426	0.330	0.362	0.320	0.342	0.389	0.142	0.346	0.324
	(0.486)	(0.482)	(0.494)	(0.470)	(0.481)	(0.467)	(0.474)	(0.487)	(0.349)	(0.476)	(0.468)
	[1,085,928]	[825,918]	[260,010]	[30,832]	[7,984]	[22,848]	[12,457]	[10,016]	[2,441]	[9,861]	[2,596]
Height (cm)				162.989	159.721	163.943	163.874	163.490	165.493	163.575	165.036
				(11.492)	(19.921)	(7.148)	(7.468)	(7.528)	(6.984)	(7.696)	(6.373)
Repeater				0.376	0.155	0.441					
				(0.484)	(0.362)	(0.497)					
Imm. Fam. Conn.				0.410	0.719	0.319	0.312	0.317	0.287	0.296	0.372
				(0.492)	(0.449)	(0.466)	(0.463)	(0.465)	(0.453)	(0.456)	(0.483)
Any Conn.				0.948	0.955	0.947	0.967	0.973	0.944	0.970	0.958
				(0.223)	(0.208)	(0.224)	(0.178)	(0.162)	(0.230)	(0.171)	(0.201)
Paid for Self				0.833	0.652	0.887	0.916	0.911	0.938	0.907	0.953
				(0.373)	(0.476)	(0.317)	(0.277)	(0.285)	(0.240)	(0.291)	(0.211)
Literate				0.620	0.560	0.638	0.626	0.576	0.835	0.538	0.978
				(0.485)	(0.496)	(0.481)	(0.484)	(0.494)	(0.371)	(0.499)	(0.148)
				[12,561]	[3,230]	[9,331]	[5,086]	[4,083]	[1,003]	[4,041]	[1,045]
Farm				0.306	0.067	0.369	0.362	0.383	0.277	0.402	0.206
				(0.461)	(0.249)	(0.483)	(0.481)	(0.486)	(0.448)	(0.490)	(0.405)
Skilled or Artisan				0.107	0.065	0.118	0.127	0.121	0.152	0.106	0.210
				(0.309)	(0.246)	(0.323)	(0.333)	(0.326)	(0.360)	(0.308)	(0.407)
Professional				0.028	0.015	0.032	0.030	0.029	0.034	0.021	0.066
				(0.166)	(0.120)	(0.176)	(0.171)	(0.169)	(0.182)	(0.144)	(0.249)
Unskilled or Unproductive				0.558	0.854	0.481	0.480	0.467	0.537	0.471	0.518
				(0.497)	(0.353)	(0.500)	(0.500)	(0.499)	(0.499)	(0.499)	(0.500)
Observations	1,126,433	858,209	268,224	31,590	8,204	23,386	12,755	10,237	2,518	10,089	2,666

Notes: The sample covered in this table is all geolocated arrivals between 1907 and 1925. Standard deviations in parentheses. Sample sizes are the minimum with data for all variables, except for urban, occupation, and literacy, which are available only for subsets of the sample. The square brackets under urban are the numbers with data on all variables except occupation and literacy. The square brackets under literacy show the number with data on all variables except urban. Urban is defined using population counts from the 1901 Italian census, defining an urban place as one with population 10,000 or more; it is a binary variable. Imm. Fam. Conn. is Immediate Family Connection; Any Conn. is Any Connection.

Table A.4: Summary statistics for province-level variables.

Variable	(1) All	(2) South	(3) North
Average Height (cm)[†]	164.530	163.525	166.002
	(1.835)	(1.604)	(0.942)
Stature CV[†]	3.720	3.758	3.663
	(0.273)	(0.289)	(0.242)
Southern	0.594		
	(0.495)		
Male Literacy Rate (1911)	0.672	0.558	0.832
	(0.177)	(0.130)	(0.087)
Fraction Urban (1901, 10,000+)	0.396	0.472	0.285
	(0.226)	(0.231)	(0.166)
Fraction Emigrating[‡]	0.019	0.020	0.018
	(0.012)	(0.009)	(0.014)
Fraction of Emigrants to Non-US[‡]	0.621	0.450	0.871
	(0.297)	(0.268)	(0.077)
Fraction Owning Property (1901)	0.127	0.129	0.123
	(0.055)	(0.043)	(0.070)
Population (1901, 10,000)	68.124	71.019	63.885
	(44.814)	(49.410)	(37.562)
Observations	68	40	28

Notes: The unit of observation in this table is a province. Averages are weighted across provinces by 1901 population, except for literacy, which is weighted by 1911 population. All variables marked with a census year (i.e., 1901, 1911) are taken from Italian Census records.

[†]: averaged over birth cohorts, weighting by the number of observations in our sample in each cohort.

[‡]: averaged over years of arrival within province weighting by the number of arrivals in that year in our sample. These should be considered annual variables (e.g., approximately 2% emigration per year).

Table A.5: Balancing tests for all individual characteristics.

	All				Transcribed First-Timers			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dep. Variable	All	Unspecified	Northern	Southern	All	Unspecified	Northern	Southern
Recorded Ethnicity								
Unspecified	-0.065^a				-0.053^a			
	(0.001)				(0.011)			
Northern	-0.014^a				-0.033^a			
	(0.001)				(0.009)			
Southern	0.079^a				0.086^a			
	(0.001)				(0.012)			
Italian Port	0.081^a	0.128^a	0.108^a	0.013^a	0.101^a	0.151^a	0.117^a	0.015^a
	(0.001)	(0.002)	(0.003)	(0.001)	(0.010)	(0.021)	(0.028)	(0.006)
Age	-0.081^a	-0.156^a	-0.458^a	−0.023	0.130	0.001	−0.558	0.260
	(0.021)	(0.040)	(0.056)	(0.027)	(0.173)	(0.324)	(0.427)	(0.233)
Birthyear	0.351^a	0.517^a	1.107^a	0.310^a	0.030	0.261	1.165^b	−0.111
	(0.023)	(0.046)	(0.062)	(0.030)	(0.208)	(0.410)	(0.494)	(0.276)
Married	−0.002	0.006^b	-0.012^a	-0.013^a	0.011	0.021	−0.000	−0.009
	(0.001)	(0.002)	(0.004)	(0.001)	(0.012)	(0.022)	(0.030)	(0.015)
Height (cm)					−0.090	-0.693^b	0.494	0.304
					(0.162)	(0.305)	(0.407)	(0.211)
Any Conn.					0.008^c	−0.002	−0.018	0.018^a
					(0.005)	(0.008)	(0.012)	(0.006)
Imm. Fam. Conn.					0.009	−0.006	−0.000	0.015
					(0.011)	(0.021)	(0.026)	(0.015)
Paid for Self					0.007	0.005	0.010	0.011
					(0.007)	(0.012)	(0.014)	(0.010)
Observations	1,055,709	253,819	119,285	682,605	14,900	3,592	2,053	9,255
Num. Geolocated	858,181	195,926	94,467	567,788	12,741	2,974	1,696	8,071

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: The sample covered in this table consists of male migrants aged 22–65 making a first arrival, whether successfully geolocated or not. The reported coefficients are from univariate regression of an individual dependent variable 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 Unspecified Italians than those in the non-geolocated sample. Sample sizes are the minimum number of observations with data for all variables. Divisions in the first three rows and in columns (3), (4), (7), and (8) are based on recorded ethnicity, and not on province of geolocation.

Table A.6: Countrywide selection before and after 1917.

	(1)	(2)	(3)	(4)	(5)
Variables				South	North
Post-1917	0.156^a (0.021)	0.077^c (0.045)			
Southern		-0.347^a (0.027)			
Post-1917 × Southern		0.087^c (0.051)			
Post-1917 × Male Literacy Rate			−0.208 (0.141)	0.028 (0.170)	−0.525 (0.744)
Constant	-0.137^a (0.011)	0.146^a (0.024)			
Observations	12,881	12,881	12,881	10,341	2,540
R-squared	0.004	0.020	0.056	0.041	0.095
Arrival Year FE	No	No	Yes	Yes	Yes
Birth Year FE	No	No	Yes	Yes	Yes
Province FE	No	No	Yes	Yes	Yes
Constant + Post-1917	0.018 (0.018)	0.223^a (0.038)			
Constant + Southern		-0.201^a (0.012)			
Post-1917 + Post-1917 × Southern		0.164^a (0.024)			

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Dependent variable is height, standardized by all-Italy-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. Standard errors are clustered by family, except in columns (3)–(5), in which they are clustered by province. Post-1917 includes 1917 arrivals. The division by regions in columns (4) and (5) is based on geolocation. The male literacy rate is from the 1911 Italian census and is on the province level. The lower section presents the sums of certain estimated coefficients and their standard errors. Constants are not reported in the presence of fixed effects.

Table A.7: Literacy and selection.

Variables	(1)	(2)	(3)	(4) South	(5) North	(6)
Literate	0.151^a (0.033)	0.125^a (0.036)	0.163^a (0.038)	0.118^a (0.041)	0.421^a (0.101)	0.170^a (0.038)
Literate × Post-1917						−0.339 (0.316)
Constant	−0.044^c (0.026)					
Observations	5,133	5,133	5,133	4,120	1,013	5,133
R-squared	0.005	0.024	0.048	0.043	0.162	0.048
Arrival Year FE	No	Yes	Yes	Yes	Yes	Yes
Birth Year FE	No	Yes	Yes	Yes	Yes	Yes
Province FE	No	No	Yes	Yes	Yes	Yes
Literate + Literate × Post-1917						−0.169 (0.315)

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Dependent variable is height, standardized by province-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival for whom literacy information was also transcribed. Standard errors are clustered on the family level. Post-1917 includes 1917 arrivals. The division by regions in columns (4) and (5) is based on geolocation. The lower section presents the sums of certain estimated coefficients and their standard errors. Constants are not reported in the presence of fixed effects.

Table A.8: Literacy by occupation and period.

	Pre-1917		Post-1917		Diff.
Sector	(1) Mean	(2) Share	(3) Mean	(4) Share	(5)
Professional	0.877	0.024	1.000	0.068	−0.046
Skilled or Artisan	0.780	0.107	0.977	0.214	−0.125
Farm	0.414	0.402	0.992	0.200	−0.032
Unskilled or Unproductive	0.571	0.467	0.972	0.519	−0.237
Weighted Total	0.546		0.975		−0.429
Observations	4011		1035		

Notes: The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival form whom literacy information was transcribed. Means are of literacy for men aged 22–65 making first arrivals. Shares are the factions in each occupational group. The difference is between the products of mean and weight in yeach time period.

Table A.9: Sectoral decomposition.

Variables	(1) Urban	(2) Urban	(3) Urban	(4) Urban	(5) Rural	(6) Rural	(7) Rural	(8) Rural
Southern		0.121^b		−0.099		0.152^a		-0.151^b
		(0.055)		(0.138)		(0.030)		(0.066)
Average Height (cm)			-0.050^a	−0.083			-0.069^a	-0.102^a
			(0.010)	(0.051)			(0.006)	(0.023)
Southern × Average Height (cm)				0.033				0.017
				(0.052)				(0.024)
Constant	0.041^b	−0.071			0.037^a	-0.077^a		
	(0.018)	(0.052)			(0.013)	(0.025)		
Observations	4,268	4,268	4,268	4,268	8,309	8,309	8,309	8,309
R-squared	0.000	0.001	0.030	0.030	0.000	0.004	0.031	0.032
Arrival Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Birth Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Constant + Southern		0.051^a				0.074^a		
		(0.019)				(0.015)		
Average Height + Southern × Average Height				-0.050^a				-0.084^a
				(0.012)				(0.010)

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Dependent variable is height, standardized by province-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival. All standard errors are clustered on the province-birth cohort level. Urban is defined using population counts from the 1901 Italian census, defining an urban locality as one with population 10,000 or more. The lower section presents the sums of certain estimated coefficients and their standard errors. Constants are not reported in the presence of fixed effects.

Table A.10: Selection by occupation.

Variables	(1)	(2)	(3)	(4) South	(5) North
Professional	0.423^a	0.382^a	0.376^a	0.366^a	0.566^c
	(0.083)	(0.090)	(0.091)	(0.090)	(0.297)
Skilled or Artisan	0.072	0.064	0.056	0.055	0.005
	(0.049)	(0.049)	(0.049)	(0.056)	(0.101)
Farm	-0.113^a	-0.096^a	-0.126^a	-0.140^a	−0.083
	(0.035)	(0.036)	(0.038)	(0.043)	(0.082)
Constant	0.066^a				
	(0.023)				
Observations	5,045	5,045	5,045	4,046	999
R-squared	0.009	0.027	0.052	0.050	0.151
Arrival Year FE	No	Yes	Yes	Yes	Yes
Birth Year FE	No	Yes	Yes	Yes	Yes
Province FE	No	No	Yes	Yes	Yes

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1
Notes: Dependent variable is height, standardized by province-birth cohort mean and standard deviation. The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival for whom occupation information was transcribed. All standard errors are clustered on the family level. Excluded group is Unskilled or Unproductive. The regional decomposition in columns (4) and (5) is based on geolocation.

Table A.11: Decomposition of North-South differences by occupation.

	South		North		Diff.
Sector	(1) Mean	(2) Share	(3) Mean	(4) Share	(5)
Professional	0.506	0.032	0.426	0.036	0.001
Skilled or Artisan	0.190	0.122	−0.039	0.157	0.029
Farm	−0.038	0.383	−0.103	0.267	0.013
Unskilled or Unproductive	0.094	0.463	−0.038	0.540	0.064
Weighted Total	0.072		−0.021		0.093
Observations	4046		999		

Notes: The sample covered in this table consists of successfully geolocated male migrants aged 22–65 making a first arrival for whom occupation information was transcribed. Means are of province-birth cohort standardized height for males aged 22–65 making first arrivals. Shares are the fraction in each occupational group. The difference is between the product of mean and weight in each region. North and South are based on geolocation.

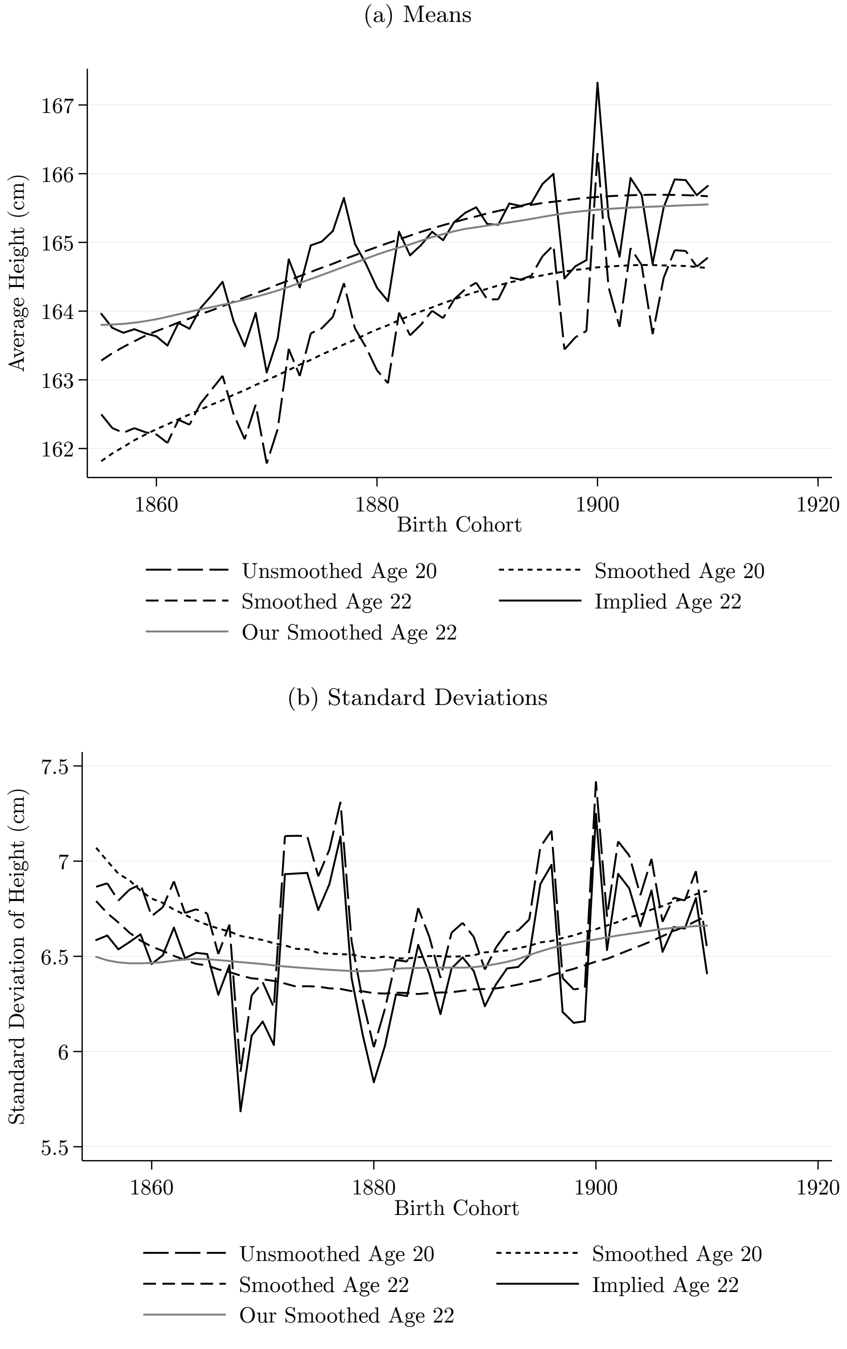
Table A.12: Sectoral decomposition.

Variables	(1) Non-Farm	(2) Non-Farm	(3) Non-Farm	(4) Non-Farm	(5) Farm	(6) Farm	(7) Farm	(8) Farm
Southern		0.150^{a}		−0.110		0.065		-0.466^{a}
		(0.047)		(0.107)		(0.073)		(0.155)
Average Height (cm)			-0.064^{a}	-0.088^{b}			-0.066^{a}	-0.210^{a}
			(0.010)	(0.035)			(0.016)	(0.059)
Southern × Average Height (cm)				0.016				0.131^{b}
				(0.037)				(0.062)
Constant	0.100^{a}	−0.017			-0.047^{c}	−0.103		
	(0.021)	(0.040)			(0.026)	(0.067)		
Observations	3,227	3,227	3,227	3,227	1,818	1,818	1,818	1,818
R-squared	0.000	0.003	0.041	0.042	0.000	0.000	0.055	0.059
Arrival Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Birth Year FE	No	No	Yes	Yes	No	No	Yes	Yes
Constant + Southern		0.133^{a}				−0.038		
		(0.025)				(0.029)		
Average Height + Southern × Average Height				-0.072^{a}				-0.079^{a}
				(0.014)				(0.021)

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Dependent variable is height, standardized by province-birth cohort mean and standard deviation. All standard errors are clustered on the province-birth cohort level. The lower section presents the sums of certain estimated coefficients and their standard errors. Constants are not reported in the presence of fixed effects.

B Appendix Figures



(a) Means

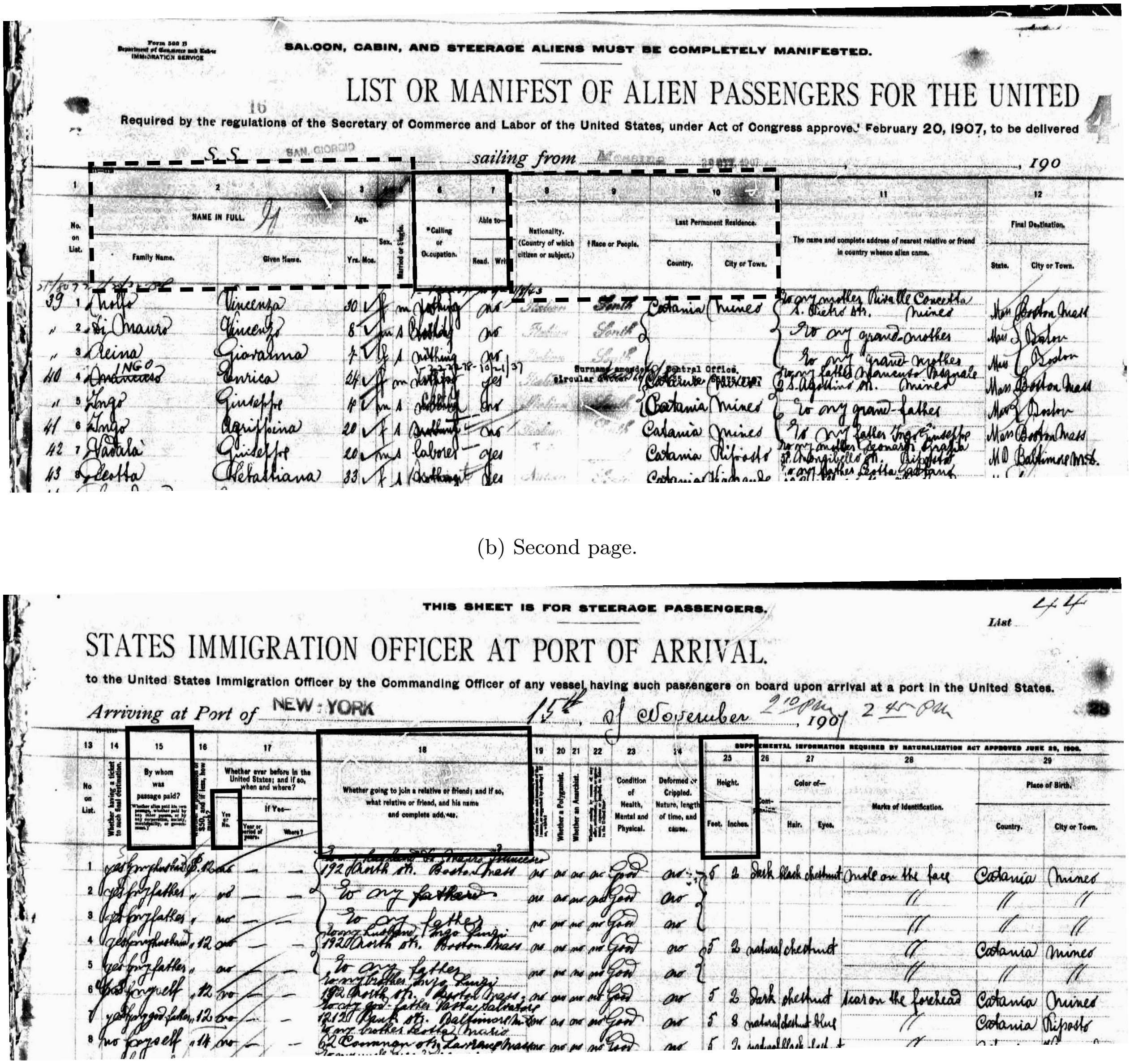
(b) Standard Deviations

Figure B.1: Moments of the height distributions: an example.

Note: These graphs are for the province of Roma.

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

(a) First page.



Form 500 B
Department of Commerce and Labor
IMMIGRATION SERVICE

SALOON, CABIN, AND STEERAGE ALIENS MUST BE COMPLETELY MANIFESTED.

LIST OR MANIFEST OF ALIEN PASSENGERS FOR THE UNITED

Required by the regulations of the Secretary of Commerce and Labor of the United States, under Act of Congress approved February 20, 1907, to be delivered

S.S. SAN GIORGIO sailing from Messina [illegible] 1907, 190

No. on List.	Family Name.	Given Name.	Age. Yrs.	Mos.	Sex.	Married or Single.	*Calling or Occupation.	Able to— Read.	Write.	Nationality. (Country of which citizen or subject.)	†Race or People.	Last Permanent Residence. Country.	City or Town.	The name and complete address of nearest relative or friend in country whence alien came.	Final Destination. State.	City or Town.
39 1	Nolfo	Vincenza	30	✓	f	m	nothing	no		Italian	South	Catania	Mineo	To my mother Rivalle Concetta S. Pietro str. Mineo	Mass	Boston Mass
〃 2	Di Mauro	Vincenzo	8	✓	m	s	nothing	no		Italian	South			To my grand-mother	Mass	Boston
〃 3	Reina	Giovanna	7	✓	f	s	nothing	no		Italian	South			To my grand-mother	Mass	Boston
40 4	~~Amancaso~~ INGO	Enrica	24	✓	f	m	nothing	yes		Italian		Catania	Mineo	To my father Infancuso Pasquale 2 S. Agostino str. Mineo	Mass	Boston Mass
〃 5	Ingo	Giuseppe	4	✓	m	s	nothing	no		Italian		Catania	Mineo	To my grand-father	Mass	Boston
41 6	Ingo	Agrippina	20	✓	f	s	nothing	no		Italian		Catania	Mineo	To my father Ingo Giuseppe	Mass	Boston Mass
42 7	Vadalà	Giuseppe	20	✓	m	s	laborer	yes				Catania	Riposto	To my mother Leonardi Grazia 57 Mongibello str. Riposto	MD	Baltimore Md.
43 8	Leotta	Sebastiana	33	✓	f	s	nothing	yes		[illegible]		Catania	[illegible]	To my father Leotta Gaetano		

1-222278-10/21/37

Surname amended Circular Letter [illegible] Central Office No. 55875/[illegible]

(b) Second page.

THIS SHEET IS FOR STEERAGE PASSENGERS.

List 44

STATES IMMIGRATION OFFICER AT PORT OF ARRIVAL.

to the United States Immigration Officer by the Commanding Officer of any vessel having such passengers on board upon arrival at a port in the United States.

Arriving at Port of NEW YORK 15th of November 2 PM 1907 2 45 PM

SUPPLEMENTAL INFORMATION REQUIRED BY NATURALIZATION ACT APPROVED JUNE 29, 1906.

No. on List.	Whether having a ticket to such final destination.	By whom was passage paid? (Whether alien paid his own passage, whether paid by any other person, or by any corporation, society, municipality, or government.)	Whether in possession of $50, and if less, how much?	Whether ever before in the United States; and if so, when and where? Yes or No.	If Yes— Year or period of years.	Where?	Whether going to join a relative or friend; and if so, what relative or friend, and his name and complete address.	19	Whether a Polygamist.	Whether an Anarchist.	22	Condition of Health, Mental and Physical.	Deformed or Crippled. Nature, length of time, and cause.	Height. Feet.	Inches.	Complexion.	Color of— Hair.	Eyes.	Marks of Identification.	Place of Birth. Country.	City or Town.
1	yes	for my husband	$.12	no	—	—	To my husband Di Mauro Francesco 192 North str. Boston Mass	no	no	no	no	Good	no	5	2	dark	black	chestnut	mole on the face	Catania	Mineo
2	yes	for my father	〃	no	—	—	To my father	no	no	no	no	Good	no						〃	〃	〃
3	yes	for my father	〃	no	—	—	To my father	no	no	no	no	Good	no						〃	〃	〃
4	yes	for my husband	〃 12	no	—	—	To my husband Ingo Luigi 192 North str. Boston Mass	no	no	no	no	Good	no	5	2	natural	chestnut		〃	Catania	Mineo
5	yes	for my father	〃	no	—	—	To my father	no	no	no	no	Good	no						〃	〃	〃
6	yes	for myself	〃 12	no	—	—	To my brother Ingo Luigi 192 North str. Boston Mass	no	no	no	no	Good	no	5	2	dark	chestnut		scar on the forehead	Catania	Mineo
7	yes	for my god-father	〃 12	no	—	—	To my god-father Vadalà Salvatore 1212 Bank str. Baltimore Md.	no	no	no	no	Good	no	5	8	natural	chestnut	blue	〃	Catania	Riposto
8	no	for myself	〃 14	no	—	—	To my brother Leotta Mario 62 Common str. Lawrence Mass	no	no	no	no	Good	no	5	2	natural	black	[illegible]	〃		

Figure B.2: Sample manifests.

Note: Fields in dashed boxes are available in the SOLEIF files. We transcribed the fields in solid boxes.

Source: SOLEIF

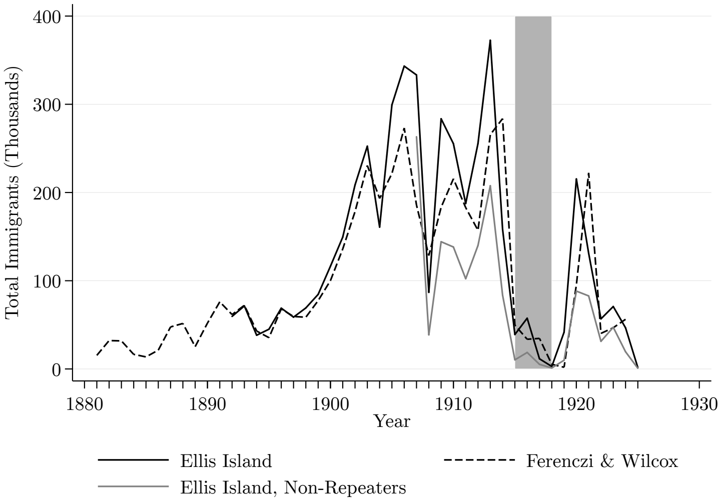


Figure B.3: Italian immigration by year.

Note: The shaded region is World War I.

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

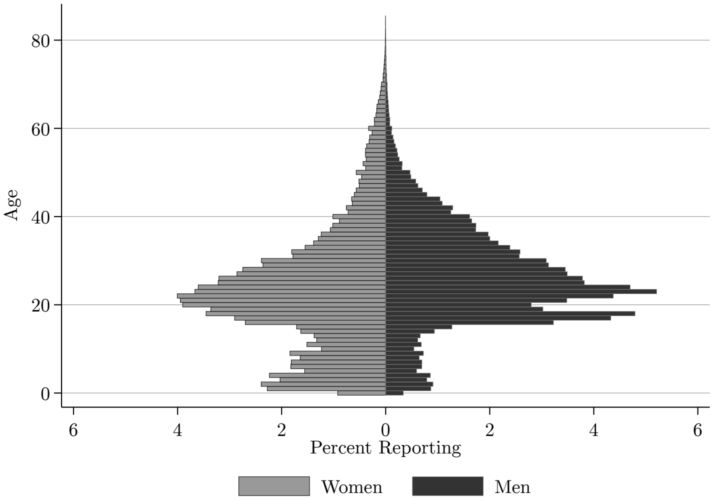


Figure B.4: Age distributions of migrants.

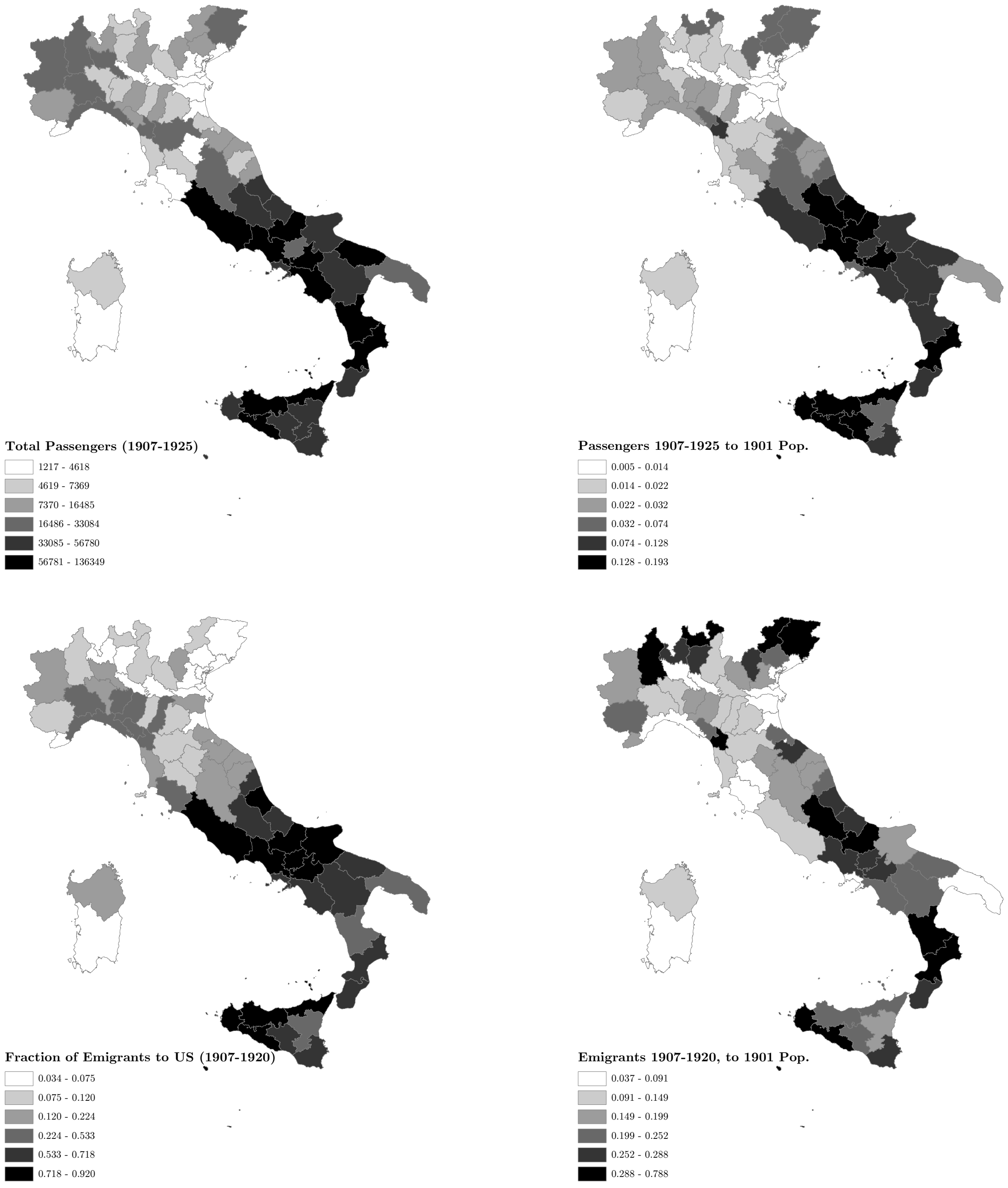
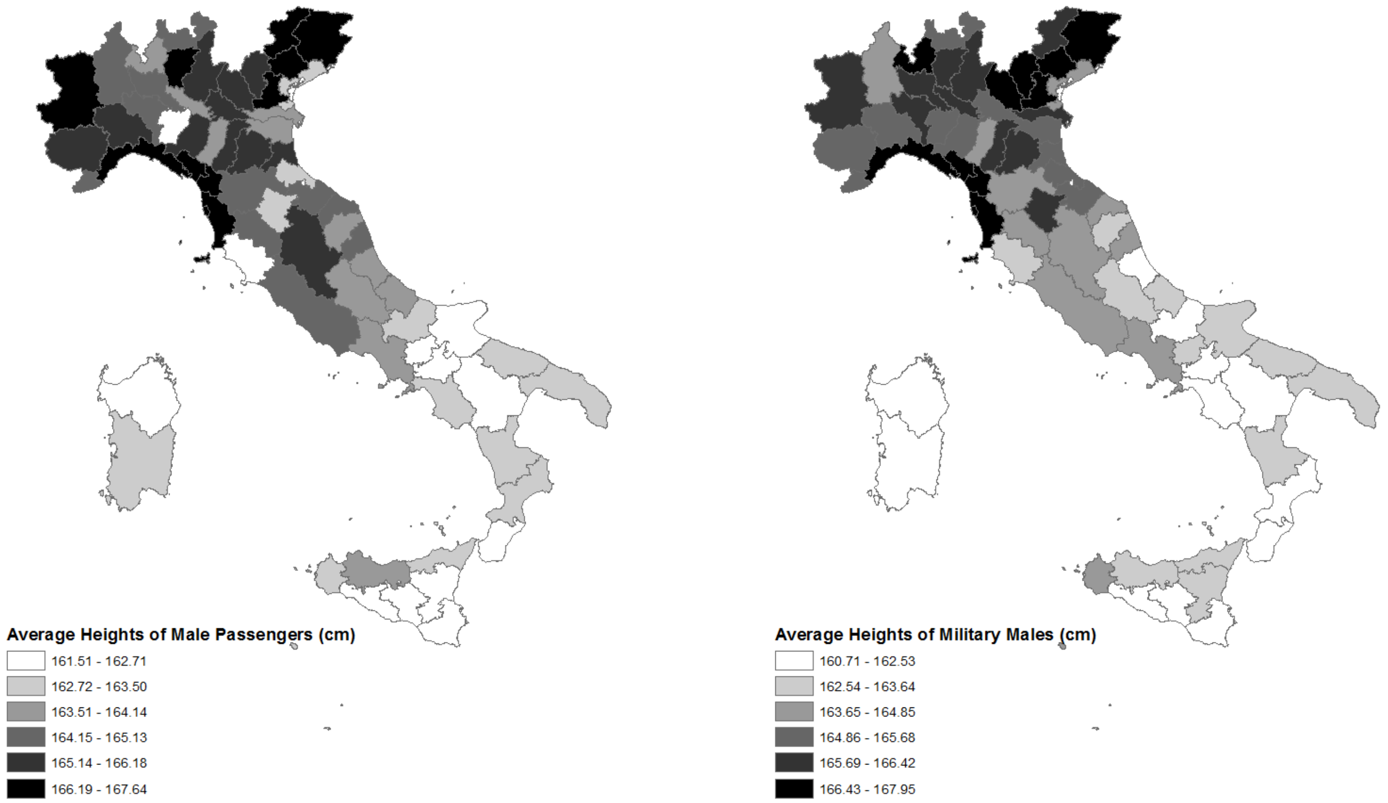


Figure B.5: Origins of Italians traveling to the United States.

Note: Fraction of Emigrants to US is the fraction of all migrants from a particular province who traveled to the United states rather than to some other destination. Emigrants 1907–1920 as a share of 1901 population encompasses emigration to all destinations, and may include return migrants. Passengers 1907–1925 and the ratio of that figure to 1901 population does not account for individuals who could not be geolocated.

Source: Our elaborations on the SOLEIF data and the *Statistica della Emigrazione Italiana*.



(a) Average heights of passengers, by province.

(b) Average heights by province.

Figure B.6: Average heights of passengers and populations.

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

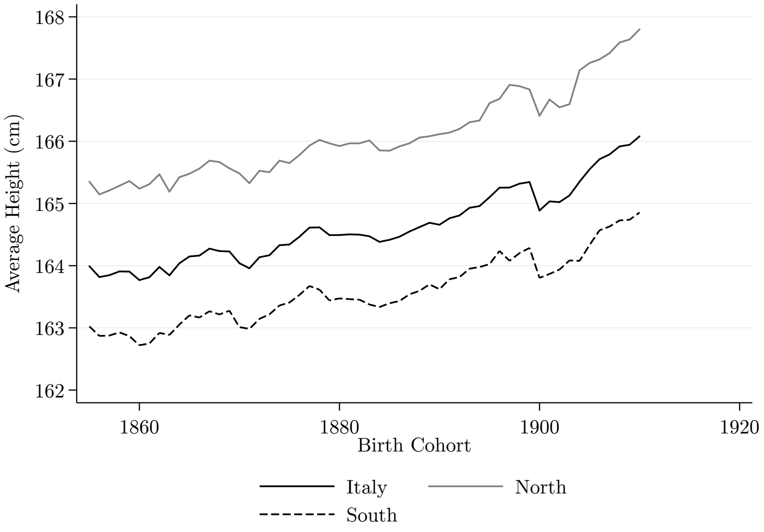


Figure B.7: Trends in average height of Italian men.

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).

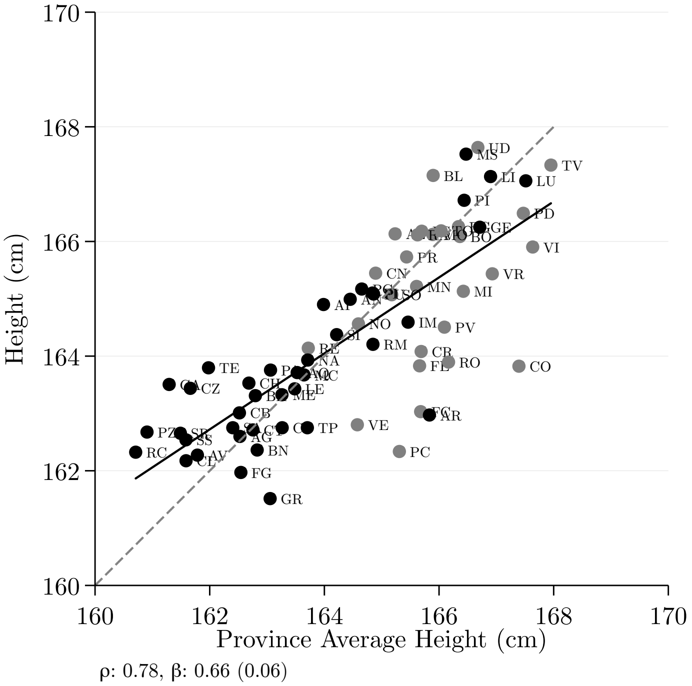
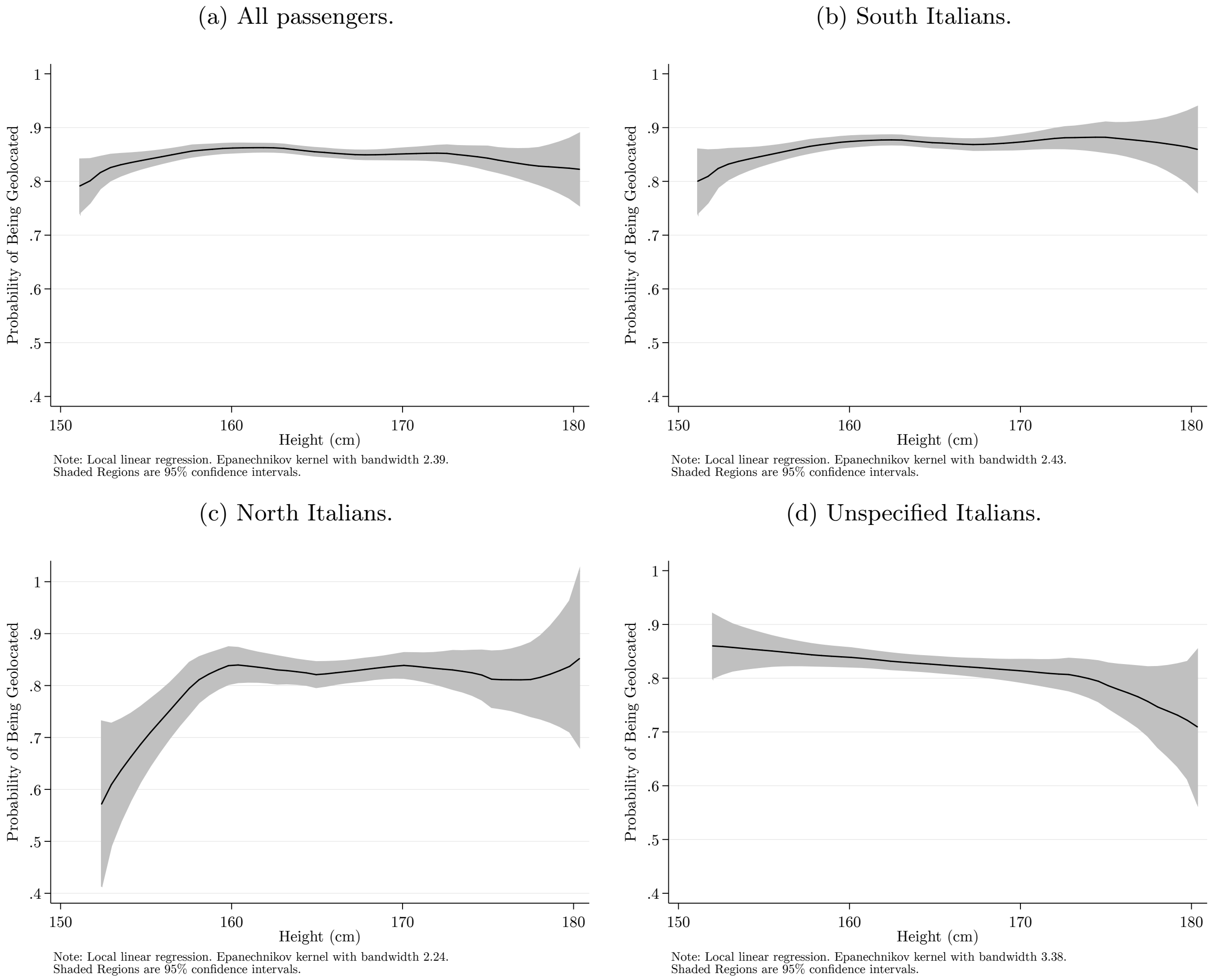


Figure B.8: Heights of migrants by province-birth cohort average height.

Note: Northern provinces in gray, southern provinces in black. Heights are weighted within provinces across birth cohorts by the number of migrants from each province in our sample.


(a) All passengers.

Note: Local linear regression. Epanechnikov kernel with bandwidth 2.39.
Shaded Regions are 95% confidence intervals.

(b) South Italians.

Note: Local linear regression. Epanechnikov kernel with bandwidth 2.43.
Shaded Regions are 95% confidence intervals.

(c) North Italians.

Note: Local linear regression. Epanechnikov kernel with bandwidth 2.24.
Shaded Regions are 95% confidence intervals.

(d) Unspecified Italians.

Note: Local linear regression. Epanechnikov kernel with bandwidth 3.38.
Shaded Regions are 95% confidence intervals.

Figure B.9: 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. Shaded regions are 95% confidence intervals.

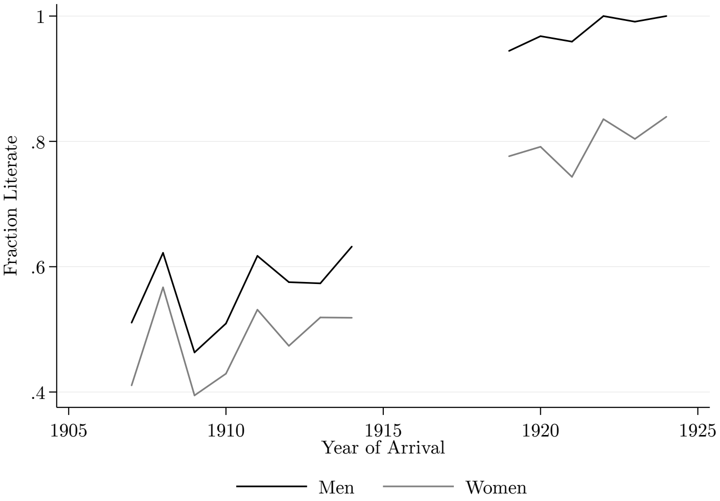
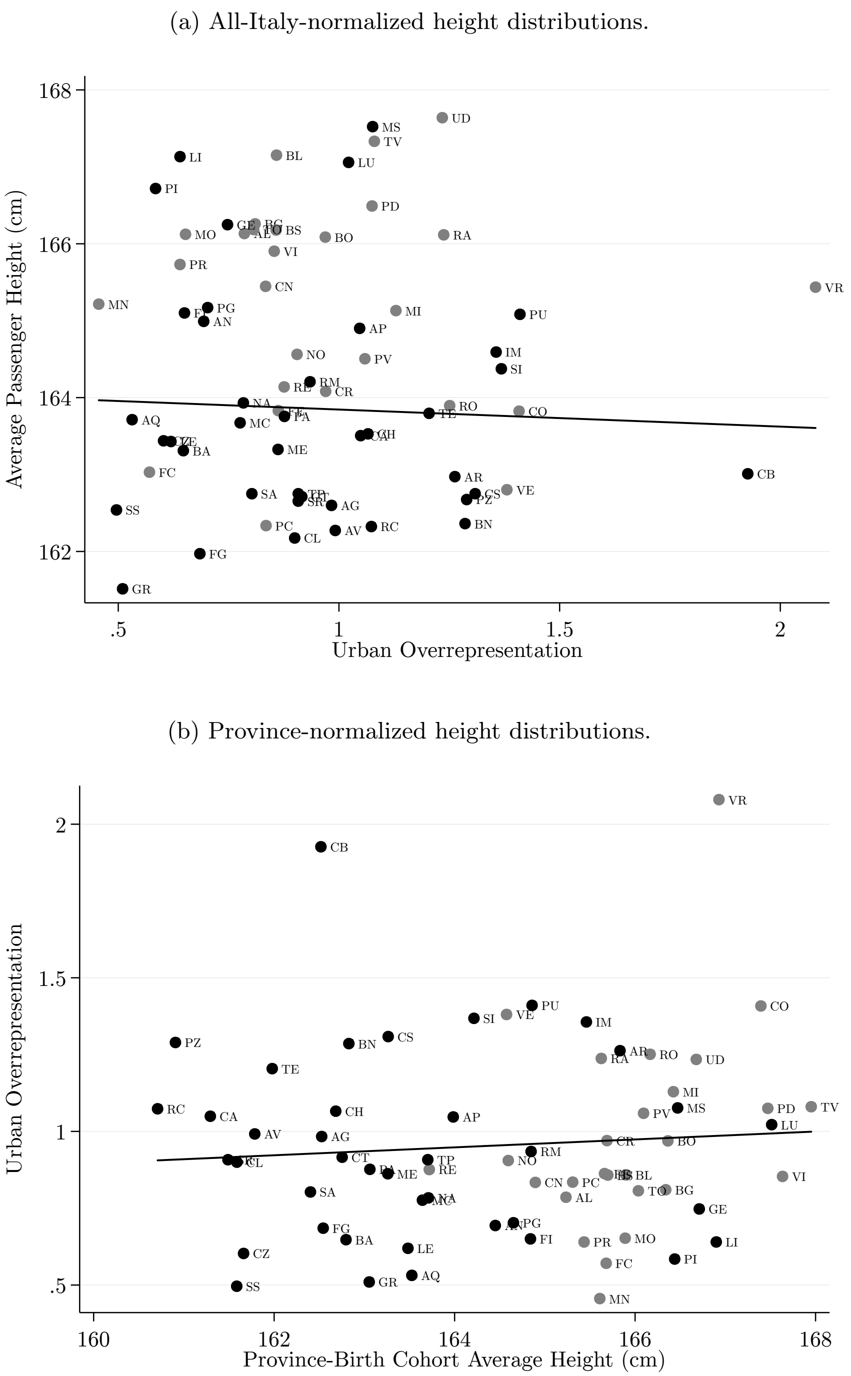


Figure B.10: Literacy by year.

Note: The sample is restricted to first arrivals aged 22–65. World War I years are excluded.


(a) All-Italy-normalized height distributions.

(b) Province-normalized height distributions.

Figure B.11: Urban overrepresentation and selection.

Note: Urban overrepresentation is calculated relative to the 1901 Italian census. Southern provinces in black, northern provinces in gray.

]

C 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. 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 C.1.
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.

For the individuals who were randomly selected to be transcribed, a more time-consuming and rigorous search was conducted, which restricted the type of match to be a *comune*, and took extra pains to ensure that no match was made in the case of an ambiguous location name.

Approximately 3.2 million immigrants (arriving between 1892 and 1925) could be matched to a province by this algorithm. Among the remaining 1.6 million, about half were not searched for because either no location was provided, or the string was determined to indicate a non-Italian location (usually somewhere in the United States or Argentina). The other half of the 1.6 million 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. In total, 79.4 percent of all migrants for whom a search was conducted were matched to a province of origin; among those arriving in 1907 or later, the figure is 81.6 percent; among men aged 22–65 making a first arrival and providing usable height data, it is 85.5 percent

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 C.2 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 in 1904 or later.[95] 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 C.2 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.[96] Moreover, it appears that the failure to decompose

[95] This appears to be the first year in which the distinction between North and South Italians was made rigorously.

[96] In particular, 21.8 percent of passengers arriving in 1904 or later and matched to a southern province are classified as general Italians, as compared to 27.7 percent of those matched to a northern province.

Italians into north and south is primarily driven by certain ships who do not decompose Italians at all. Of the 30,217 voyages in our data, 68 percent do not decompose Italians at all. If we limit consideration to passengers classified as North Italian or South Italian, 91.5 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 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 arriving 1904 or later 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,[97] and limit the sample to individuals for whom a location was found. The remaining passengers constitute 47.5 percent of the 511,838 passengers leaving Palermo for whom a location search was performed. 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 98.3 percent of these South Italian passengers to locations in southern Italy,[98] and 92.7 percent to Sicily specifically.[99]

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.[100]

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

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

[97] That is, we remove all passengers aboard ships that had at least one "Italian", without the North-South distinction.

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

[99] This rate was not driven by a tendency to blindly assign the port of departure as their last place of residence; only 14.5 percent of these passengers reported Palermo as their last place of residence.

[100] 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.

Clearly, N^S and $N^{\neg S}$ are unknown, but they can be reasonably approximated in several ways. We use the share of current Italian communes located in Sicily; these amount to 4.8 percent of the 8,100 Italian communes.

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^*:

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

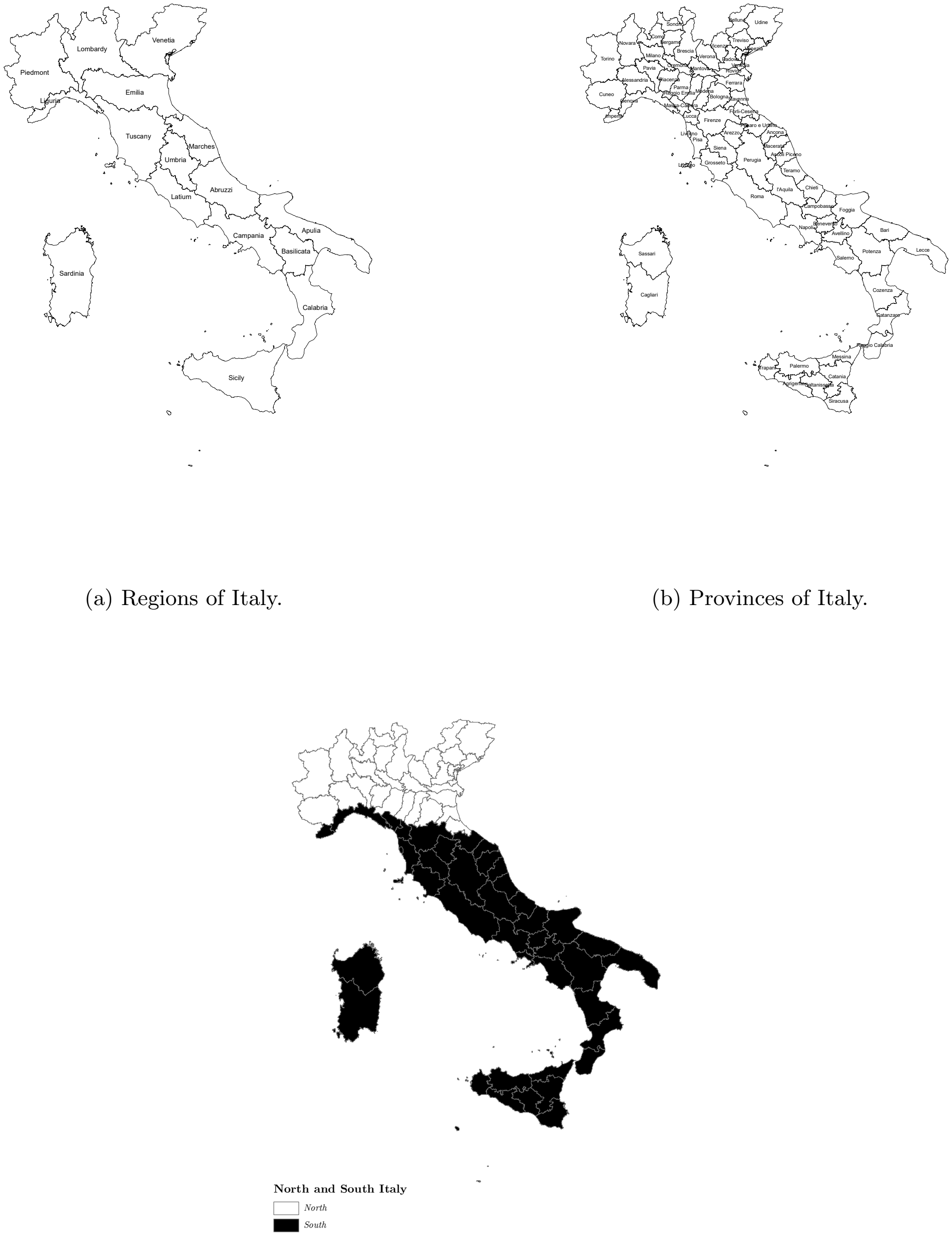
Note that this probability is decreasing in S^*, and thus a lower bound for the rate of successful matching is given when $S^* = 1$.[101] This gives

$$p \geq \frac{S - s}{1 - s} = \frac{0.927 - 0.048}{1 - 0.048} = 0.923,$$

meaning that with probability of at least 92.3 percent, the matching algorithm successfully matches a passenger to his correct last place of residence.

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

Figures



(a) Regions of Italy.

(b) Provinces of Italy.

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

Figure C.1: Geographic divisions of Italy.

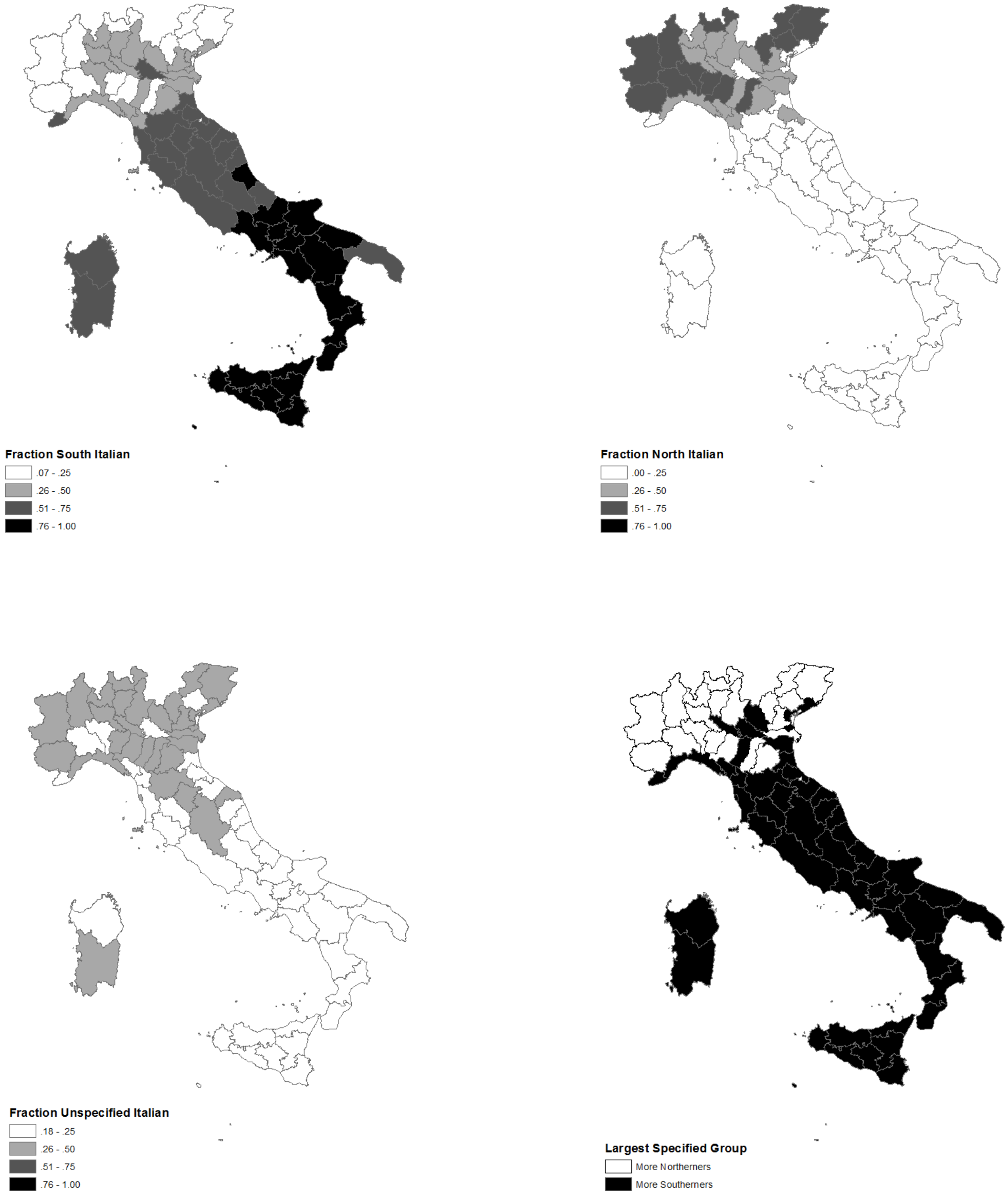


Figure C.2: Ethnicity of geolocated Italian passengers.

D Importance of Local vs. National Selection

This appendix provides the formal details behind our argument in section 4.3.1 that local self-selection can be valuable in predicting migrant outcomes in the receiving country, conditional on the level of absolute (country-level) selection.

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. For simplicity, we abstract from noise of genetic and other varieties so that, conditional on the effect of the local environment and individual quality, height is deterministic. The same conclusions would follow if such noise were taken into account.

Consider 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-cohort standard deviation of height, or at least its non-genetic component). Then μ_{jt} is the average height in province j and birth cohort t, and the national degree of self-selection is represented simply by the height (after subtracting the average national height) $\tilde{h}_{ijt} = h_{ijt} - \mu_t$. 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 (or the absolute measure of quality) is to characterize the function $\hat{w}$ that predicts the outcome w_{ijt} conditional on the two measures 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 denote $\text{MRTS}^w(\mu, z)$ similarly. The claim to be proven is the following.

Claim. Greater local selection predicts better outcomes conditional on the level of national selection (that is, $\frac{\partial \hat{w}(\tilde{h},z)}{\partial z} > 0$) if and only if

$$\text{MRTS}^h(\mu, z) > \text{MRTS}^w(\mu, z). \tag{D.1}$$

Proof. Let $h(\mu, z)$ be continuously differentiable and 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 $\tilde{h}$. That is, it is the level of μ required to achieve a particular $\tilde{h}$ conditional on z. More formally, $\mu_{\tilde{h}}(z)$ is defined as the function that for a given $\tilde{h}$ satisfies the equality

$$\tilde{h} = h(\mu_{\tilde{h}}(z), z) - \mu_t. \tag{D.2}$$

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 $\text{MRTS}^h(\mu, z) = \frac{\partial h(\mu,z)}{\partial \mu} / \frac{\partial h(\mu,z)}{\partial z}$, and similarly denote $\text{MRTS}^w(\mu, z)$.

By the implicit function theorem and the definition of $\mu_{\tilde{h}}(z)$ in equation (D.2), we have that

$$\frac{\partial \mu_{\tilde{h}}(z)}{\partial z} = -\frac{\frac{\partial h(\mu,z)}{\partial z}}{\frac{\partial h(\mu,z)}{\partial \mu}}. \tag{D.3}$$

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

$$\begin{aligned}
\frac{\partial \hat{w}(\tilde{h}, z)}{\partial z} &= \frac{\partial w(\mu_{\tilde{h}}(z), z)}{\partial z} \\
&= \frac{\partial w(\mu, z)}{\partial \mu} \times \frac{\partial \mu_{\tilde{h}}(z)}{\partial z} + \frac{\partial w(\mu, z)}{\partial z} \\
&= -\frac{\partial w(\mu, z)}{\partial \mu} \times \frac{\frac{\partial h(\mu,z)}{\partial z}}{\frac{\partial h(\mu,z)}{\partial \mu}} + \frac{\partial w(\mu, z)}{\partial z} \qquad \text{(D.4)} \\
&= \frac{\partial w(\mu, z)}{\partial \mu} \left(-\frac{1}{\text{MRTS}^h(\mu, z)} + \frac{1}{\text{MRTS}^w(\mu, z)} \right),
\end{aligned}$$

where expression (D.4) follows from equation (D.3). 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 $\text{MRTS}^h(\mu, z) > \text{MRTS}^w(\mu, z)$. □

Thus, if two individuals have the same absolute quality or national level of selection, $\tilde{h}$, this claim provides a condition under which the one who is more positively self-selected on the local level will have a better outcome in the receiving country.

E Marginal Effects of Height and Living Standards on Migration Probabilities

This appendix provides the formal details for the exercise performed in section 4.3.2, determining, from the available data, the marginal effects of differences in height and in the biological standard of living on migration probabilities. Let z_{ijt} denote the standardized height of individual i from province j and birth cohort t. Suppose that it can be decomposed into a portion associated with the biological standard of living, α_{ijt}, and a random (primarily genetic) component, ε_{ijt}, both with mean zero (since z_{ijt} is a z-score, it must have mean zero). Assuming that

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

and that ε_{ijt} is independent of α_{ijt} satisfies the classical measurement error assumptions invoked in section 4.3.2. Since z_{ijt} is a z-score, it must have variance of one. Letting ξ^2 denote the variance of α_{ijt} and ψ^2 denote the variance of ε_{ijt}, it must be (from the independence assumption) that $\xi^2 + \psi^2 = 1$.

The first step of the exercise in section 4.3.2 is to determine conditional migration probabilities. Let y_{ijt} be an indicator for whether individual i migrates, taking a value of one if he does and zero otherwise. Thus, the probability of migrating conditional on a particular z-score is $P(y_{ijt} = 1|z_{ijt})$. We do not observe this object; but, according to Bayes's theorem, it can be written as

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

where $\tau(\cdot)$ is a density. The density $\tau(z_{ijt}|y_{ijt} = 1)$—the distribution of z-scores among migrants—is given by our data. The density $\tau(z_{ijt})$—the population distribution of z-scores—is, by construction, a standard normal if heights are normally distributed in each province-cohort. The population probability of migration, $P(y_{ijt} = 1)$, can be learned from external data or from our total count of migrants in our data combined with population counts.

This inversion makes it possible to learn the marginal effect of differences in z-score on the probability of migration. The next step of the exercise is to determine the marginal effects of variations in living standards (measured with error by the z-score) on migration probability. In particular, we assume that the relationship is given by

$$P(y_{ijt} = 1|z_{ijt}) = c + \delta\alpha_{ijt} + \eta_{ijt}, \tag{E.1}$$

where η_{ijt} is an error term uncorrelated with α_{ijt}. At this point, standard measurement error arguments can be invoked. In particular, the fact that we observe z_{ijt} and not α_{ijt} prevents estimation of equation (E.1). Instead, it is only possible to estimate

$$P(y_{ijt} = 1|z_{ijt}) = c + \delta z_{ijt} + \nu_{ijt},$$

where $\nu_{ijt} = -\delta\varepsilon_{ijt} + \eta_{ijt}$. A standard measurement-error argument yields the classic attenuation bias result that

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

Thus, the true value of δ is given by

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

Bounding ψ^2 between 0 and 0.8 (Silventoinen, 2003) allows us to conclude that

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

providing bounds for δ based on our estimated $\hat{\delta}$.

F Surname-Based Province Imputation Algorithm

The goal of this algorithm is to create a linkage of passengers to provinces that is unrelated to that passenger's outcome in the geolocation algorithm. Thus, this algorithm makes it possible to create a rough estimate of the place of origin for passengers for whom geolocation fails. It also provides alternative information on place of origin for those for whom geolocation was successful, allowing us to identify cases in which the location is confirmed by two methods. The algorithm is based on the fact that Italian surnames are informative of geographic origin (Guglielmino and De Silvestri, 1995). Intuitively, it assumes that our geolocation algorithm described in Appendix C is accurate on average and uses the modal location to which individuals with a passenger's surname (or similar surname), other than he and his traveling companions are matched. The procedure is as follows.

1. We match each transcribed passenger to each geolocated passenger according to the similarity of their surnames. We require that, if an intra-Italian ethnicity is reported in both records, the two agree with one another (to avoid matching southerners to northerners), and that if both surnames end in a vowel other than "U" that these vowels match.[102] Moreover, to ensure that the surname-based match is not driven by errors in matching that individual and his family by the geolocation algorithm, matches between the passenger and any passengers on the same voyage (including himself) are removed.

2. For each transcribed individual, we tabulated the number of geolocated individuals from each province that were matched to that individual. The surname-implied province is determined based on the province to which the plurality of the matches are geolocated. For example, if an individual links to 20 passengers from Palermo, 10 passengers from Messina, and 5 passengers from Caltanisetta, their surname-implied province is Palermo.[103] In case of ties, we average across the tying provinces.

Table F.1 presents the results of regressions of the geolocation algorithm-implied province-cohort average height on the surname-implied province-cohort average height, providing a test of how representative surnames are of provinces of origin. In column (1), the relationship between the two measures is positive and statistically significant, indicating that surnames are informative regarding province of origin. Columns (2) and (3) divide this analysis by region (North and South) in order to eliminate the gains from requiring ethnicity agreement in the matching.[104] These regressions show that the strong relationship between the province-cohort average stature implied by the two algorithms holds also within regions. Columns (4)–(6) repeat the analysis of columns (1)–(3), but restrict the sample to individuals whose surname-implied province disagrees with the province of geolocation according to the algorithm of Appendix C. The relationship between the average province-cohort height implied by the two procedures is still positive and statistically significant even among these individuals, indicating that even when surnames do not provide an exact province match, they tend to provide a similar match to that given by the geolocation algorithm.

The relationship between the province-cohort average height implied by each algorithm is depicted graphically in Figure F.1, which presents a histogram approximation of the joint distribution of the predictions of the two algorithms. The strong agreement of the two algorithms—evidenced by the large mass on the diagonal—further supports the informativeness of surnames. It should be noted that because the matching algorithm described above does not permit an individual to match to anyone on the same voyage this result indicates that individuals with similar surnames who are not traveling together tend to match to the same or similar province.

[102] Our geolocation algorithm shows that the last letter of the surname is a good predictor of place of origin.

[103] We have also performed the same exercise penalizing provinces with relatively more arrivals. The results are similar to those presented here, but we prefer the current results because there is a stronger relationship between the surname-implied province and the geomatched province.

[104] Even if surnames were totally uninformative, requiring ethnicity agreement would lead to some agreement between the two algorithms.

Tables

Table F.1: Surname imputation regressions.

Variables	(1) All	(2) South	(3) North	(4) All	(5) South	(6) North
Surname-Implied Average Height (cm)	0.519^{a}	0.408^{a}	0.135^{a}	0.302^{a}	0.166^{a}	0.068^{a}
	(0.011)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)
Constant	78.836^{a}	96.459^{a}	143.867^{a}	114.392^{a}	135.958^{a}	154.900^{a}
	(1.736)	(2.153)	(2.058)	(2.053)	(2.167)	(2.276)
Observations	12,577	10,117	2,460	9,257	7,218	2,039
R-squared	0.211	0.159	0.052	0.065	0.026	0.012

Significance levels: ^{a} p<0.01, ^{b} p<0.05, ^{c} p<0.1

Notes: Dependent variable is province-birth cohort mean height. The sample covered in this table consists of transcribed and successfully geolocated male migrants aged 22–65 making a first arrival. Average Height is of the province-birth cohort. Columns titled North and South include only passengers geolinked to each region. Columns (4)–(6) include only those whose surname-implied province is not the same as their geolinked province.

Figures

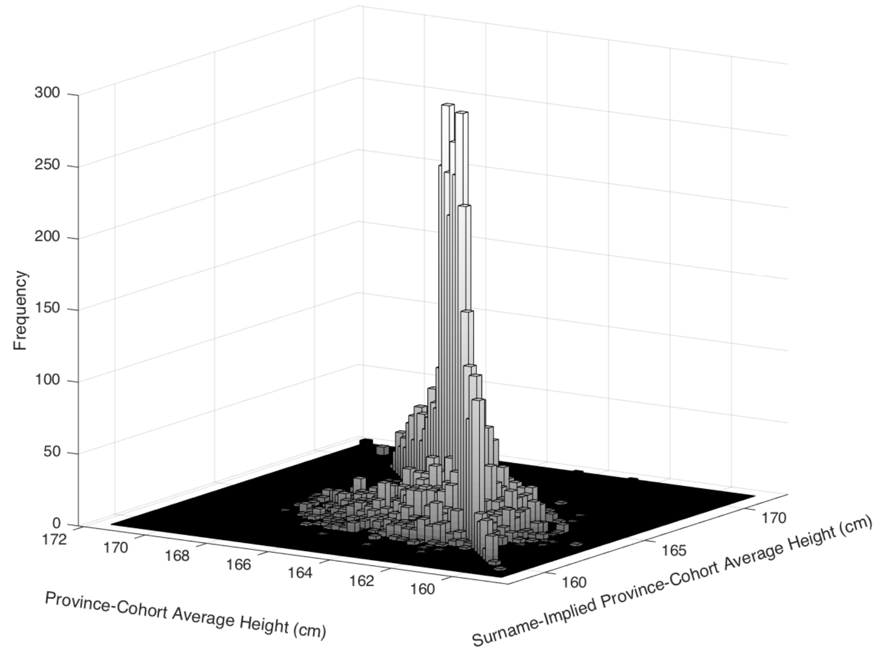


Figure F.1: Geomatched and surname-implied province-cohort average height.

Note: This histogram divides the range of province-cohort average heights into 50 equally spaced bins each (a step size of about 0.25 cm) and plots the frequency falling into each bin. The “on-diagonals” here thus capture both exact agreement between the two algorithms and small disagreements (i.e., matching to a province with very similar average height.

G Bias Induced by Errors in Geolocation

This appendix formalizes our arguments in section 5.3, where we discuss the possibility that the systematic variation that we uncover of selection with respect to average stature may be driven by errors in our geolocation algorithm. 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 population at risk for migration, 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 constitutes our main results. 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

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

where $\beta_0 + \mu_t$ is the mean of the all-Italy distribution of migrant heights for birth cohort t.[105] 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

$$E(z_{ijt}) = \frac{\beta_0}{\sigma_{jt}} + (1-p)\left(\frac{\mu_t - \mu_{jt}}{\sigma_{jt}}\right).$$

We first consider the effect of incorrect geolocation on our results of negative national self-selection, presented in Table 2. 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 positive local self-selection result in Table 3. 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 3, 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 in birth cohort t 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

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

[105] In fact, the probability of matching to the correct province is slightly greater than the probability of a correct match, p. This is because it is possible to accidentally locate the migrant to the correct province of origin, even if the particular location is in error. We disregard this possibility, which biases the analysis against our findings because it understates the probability of a correct match.

Then

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

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 (G.1) in our data, and our estimate in column (1) of Table 3, the value of $(1-p)$ that would be required to produce the results in column (1) of Table 3 spuriously if the true β_0 is zero (i.e., if there is no self-selection at all) is 0.243.[106] 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 (G.1). 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 (G.1) 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, and more importantly, we consider the effects of incorrect geolocation on our findings of differential self-selection in column (3) of Table 3. 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. Recall that, 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. The argument could also be made in standardized height, although the derivations are somewhat more complex due to heteroskedasticity of height distributions across provinces and birth cohorts. In this case, the intuition is the same, as are, roughly, the implied probabilities of mismeasurement required in order to spuriously produce our results. We focus on the regression of height rather than of z-score for simplicity. Under these assumptions, generating differential selection that results in a coefficient of p in a regression of heights on province means simply through incorrect geocoding (and no change in the degree of selection across provinces) requires that individuals be mismeasured with probability $(1-p)$. In the context of our data, generating a differential self-selection result solely through measurement error requires that approximately 39.7 percent of migrants be incorrectly assigned.[107] Thus, although some mismatching likely occurs, it would have had to have been implausibly large in order to spuriously generate our differential self-selection result.

[106] This figure is calculated as follows. The value of the summation on the right-hand side in equation (G.1) is 0.152. The value of the self-selection found in Table 3 is 0.037. If the true value of the local self-selection were zero, then the value of $(1-p)$ that would generate an observed selection of 0.037 is $0.037/0.152 = 0.243$.

[107] This figure is derived as follows. A regression of observed heights on a constant and province-cohort average height yields a coefficient of 0.603 (standard error 0.030). Thus, the value of $(1-p)$ that would generate this slope if there were in fact no differential selection is $1 - 0.603 = 0.397$.