

Renner, Laura; Krieger, Tim; Ruhose, Jens

**Conference Paper**

## Culture, Selection, and International Migration

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik  
- Session: International Trade, Inequality, and Migration, No. E17-V4

**Provided in Cooperation with:**

Verein für Socialpolitik / German Economic Association

*Suggested Citation:* Renner, Laura; Krieger, Tim; Ruhose, Jens (2014) : Culture, Selection, and International Migration, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: International Trade, Inequality, and Migration, No. E17-V4, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:

<https://hdl.handle.net/10419/100434>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# Culture, Selection, and International Migration

Tim Krieger\*

Laura Renner<sup>†</sup>

Jens Ruhose<sup>‡</sup>

PRELIMINARY VERSION.

Please do not cite without permission.

January 2014

## Abstract

This paper looks at the effect of cultural barriers on the skill selection of international migration. The data covers bilateral migration stocks by skill level in 2000 from about 99 sending countries to the main 15 destination countries. We use genetic distance as a proxy for cultural distance and exploit exogenous variation in genetic distance in 1500 to show that a higher genetic distance leads to a higher selectivity of migrants. This reveals that cultural traits are an important determinant of the skill mix of current migrant populations.

JEL-Code: F22, J61, Z1

Keywords: Culture, Genetic Distance, Selection, International Migration

---

We are grateful to Oliver Falck and Paola Giuliano for helpful comments and discussion. We also thank Ingo Isphording and Sebastian Otten for sharing their language distance data with us.

\*Department of Economics, University of Freiburg, Wilhelmstr. 1b, 79085 Freiburg i. Br., Germany; E-mail: tim.krieger@vwl.uni-freiburg.de; Phone: +49 761 203 67651

<sup>†</sup>Department of Economics, University of Freiburg, Wilhelmstr. 1b, 79085 Freiburg i. Br., Germany; E-mail: laura.renner@vwl.uni-freiburg.de; Phone: +49 761 203 67652

<sup>‡</sup>Ifo Institute - Leibniz Institute for Economic Research at the University of Munich, Poschingerstr. 5, 81679 Munich, Germany; E-mail: ruhose@ifo.de; Phone: +49 89 9224 1388

# 1 Introduction

It is a well-known fact, that higher-skilled individuals are more likely to migrate (see e.g. Grogger and Hanson (2011) and Figure 1). A large literature has evolved to explain why and how people select into migration. However, most papers try to explain differences in the selectivity of migrants with differences in returns to skills and schooling (Abramitzky, 2009; Belot and Hatton, 2012; Borjas, 1987; Chiquiar and Hanson, 2005; Grogger and Hanson, 2011; Stolz and Baten, 2012; Kaestner and Malamud, 2013). Thus, this paper offers a new channel to explain international migrant selection patterns in a cross-section of countries with differences in the cultural distance between the sending and the receiving country.<sup>1</sup> Papers by Falck et al. (2012), Belot and Ederveen (2011), and Mayda (2009) already show that the size of migration flows is affected by cultural traits. However, the influence of cultural barriers on the skill content of migration has been neglected so far.

We argue that a higher cultural distance increases migration costs especially for low-skilled migrants. As a consequence, high-skilled migrants should be more able to cope with different cultural traits and to overcome cultural barriers to migration more easily. Hence, we should observe a more positively selected group of migrants from countries with a larger cultural distance.

We use data from Spolaore and Wacziarg (2009) on genetic distance between countries as a proxy for cultural distance.<sup>2</sup> Explaining bilateral migrant stocks, we see that low-skilled individuals are less likely to move to countries which are farther in cultural distance. For high-skilled individuals, we observe the opposite. However, these two relationships are not significant. At the same time, we find a significant positive effect of genetic distance on the skill mix, i.e. the selection, of international migrants. The effect is robust to the inclusion of several control variables and to an instrumental variables approach which uses exogenous variation in genetic distance in 1500 to correct for endogeneity bias.

In recent years, the concept of culture has attracted the attention of many researchers in explaining economic outcomes (Ottaviano and Peri, 2005; Guiso et al., 2006; Tabellini, 2010; Ashraf and Galor, 2013; Burchardi and Hassan, 2013).<sup>3</sup> Cultural traits are especially successful in explaining the size and the direction of economic exchange, such as income differences between countries (Spolaore and Wacziarg, 2009), migration flows (Falck et al., 2012; Belot and Ederveen, 2011; Dahl and Sorenson, 2010; Mayda, 2009), the diffusion of technology (Comin et al., 2012), trade patterns (Guiso et al., 2009; Felbermayr and Toubal, 2010), or investment behavior (Guiso et al., 2009).<sup>4</sup>

---

<sup>1</sup>Bauernschuster et al. (2013) show that high-skilled individuals are more likely to cross cultural borders within Germany. However, our focus is on international migrant selection.

<sup>2</sup>This does not assume that genes determine culture. It is the genetic distance due to the fact that populations have been separated that offers the possibility to proxy cultural distance in that way.

<sup>3</sup>See Guiso et al. (2006); Spolaore and Wacziarg (2013).

<sup>4</sup>In a recent contribution, Spring and Grossmann (2013) show that bilateral trust might not predict economic exchange as well as Guiso et al. (2009) suggest. They use genetic distance as an instrument

Spolaore and Wacziarg (2009) show that genetic distance, i.e. the time since two populations have shared common ancestors, is able to predict the income differences between countries. They interpret culture in the same way as Guiso et al. (2006, p. 23) by arguing that culture represents *those customary beliefs and values that ethnic, religious, and social groups transmit fairly unchanged from generation to generation*. Thus, the genetic distance between two countries identifies the differences in these transmitted characteristics and values.

This paper contributes to the literature on international migrant selection by offering an additional, so far neglected, selection mechanism. It also adds to the literature on culture as a determinant of economic exchange by showing that culture has a strong effect on the skill mix of migrants in current destination countries.

The remainder of the paper is organized as follows. Section 2 describes the data and introduces the scale and selection measures. Section 3 provides the econometric setup and explains the identification strategy. In section 4, we provide the results of our analysis. Section 5 concludes.

## 2 Data

To investigate the relationship between selection of migrants and cultural differences, we need bilateral migration data by skill level between countries. Hence, we use the 2000 cross-section from Docquier et al. (2007). Their data provides information on emigrant stocks and residents by source and destination countries, including education level (primary, secondary and tertiary) and gender for 30 destination and 195 source countries in 1990 and 2000. We only look at total migration flows, without considering the gender dimension. Since we follow Grogger and Hanson (2011) in their estimation strategy, we restrict the destination country sample to high-income countries that are also the main immigrant receiving countries, i.e. 15 OECD countries: Australia, Austria, Canada, Denmark, Finland, France, Germany, Ireland, the Netherlands, New Zealand, Norway, Spain, Sweden, the UK and the US. Due to data availability in the variables of interest, the sample of source countries is limited to 99.

Departing from utility maximization and assuming that the error structure follows an i.i.d. extreme value distribution, it can be shown that the log odds of migrating to destination country  $d$  versus staying in source country  $s$  is equal to  $\ln \frac{E_{sd}^j}{E_s^j}$  for skill group  $j \in \{H(igh), L(ow)\}$  (McFadden, 1974).  $E_s^j$  is the population of skill level  $j$  in  $s$  that remains in  $s$ .  $E_{sd}^j$  is the population of skill level  $j$  from  $s$  that has migrated to  $d$ . Hence,  $\frac{E_{sd}^j}{E_s^j}$  gives the share of the migrants in  $d$  of skill group  $j$  from country  $s$ .

Figure 1 plots the log odds of emigration for tertiary educated versus the log odds for trust. Thus, by using genetic distance we avoid the critique of Spring and Grossmann (2013) and capture, beside trust, also broader aspects of cultural differences between countries.

of emigration for primary educated for each source country in our sample. The first thing to notice is that all log odds for the primary educated migrants are below zero which indicates that the low-skilled migrant population is always lower than the low-skilled population left behind. This pattern is not true for tertiary-educated migrants in countries like Trinidad and Tobago, Jamaica, or Guyana. These countries indicate that more high-skilled migrants live outside the country compared to the resident population within the country. The red 45°-line in Figure 1 indicates that the log odds of migration are equal between the two skill groups. Countries above the line show a higher propensity of high-skilled than low-skilled migration. The USA is a notable exception.

[Figure 1 here]

Table 1 provides in Panel A summary statistics for emigration shares by skill level and the migrant skill mix. The emigration share of the primary educated population has a mean of 0.003. That means that, on average, 0.3 percent of the source country low-skilled population lives abroad. That share is equal to 2.58 percent for the high-skilled. This again reveals the positive selection in international migration already seen from Figure 1. The log shares are our measures of the scale of international migration by skill group.

[Table 1 here]

The migrant skill mix,  $\left(\ln \frac{E_{sd}^H}{E_{sd}^L} - \ln \frac{E_s^H}{E_s^L}\right)$ , will be the outcome of interest when we look how international migrants select themselves into different destinations. The idea is that migrants are positively selected when the share of migrants from country  $s$  is disproportionally high-skilled, i.e. when the scale of high-skilled migrants is larger than the scale of low-skilled migrants,  $\ln \frac{E_{sd}^H}{E_{sd}^L} > \ln \frac{E_s^H}{E_s^L}$ . Table 1 shows the sample mean of the selection variable. The log odds interpretation indicates that it is, on average, 85 percent more likely to see high-skilled emigration versus low-skilled emigration.

Our proxy for cultural differences, genetic distance, is taken from Spolaore and Wacziarg (2009).<sup>5</sup> The measure comes from genetics and is based on the analysis of differences in alleles across populations. An allele is a form a gene can take. While a gene determines a certain trait, e.g. the blood group, the allele, specifies which blood group an individual has (Cavalli-Sforza, 2001). The so-called  $F_{ST}$  distance is the probability that the alleles of two individuals randomly taken from two populations are different. If alleles are identically distributed across two populations, the  $F_{ST}$  distance is zero. A higher  $F_{ST}$  distance implies that the alleles of the two populations have developed differently, which can be ascribed to a separate development over time. The alleles considered are 'neutral', which means that they are only changing due to random drift and not due to natural selection. The earlier the populations separated, the earlier has random drift taken their

---

<sup>5</sup>Spolaore and Wacziarg (2009) refer to the genetic distance data by Cavalli-Sforza et al. (1994).

development in different directions. Therefore, combining the latter argumentation with the understanding of culture as transmitted characteristics and values, the genetic distance can be interpreted as a proxy for the time the two populations have been separated and developing both, differences in genes and in culture.

The weighted  $F_{ST}$  distance is provided for populations and then matched from populations to countries and weighted by the share of the respective populations in a country (Spolaore and Wacziarg, 2009). In addition a  $F_{ST}$  genetic distance as of 1500 is provided by matching countries with the majority population in 1500. Further information on the dataset is provided by Spolaore and Wacziarg (2009). Table 1, Panel B shows summary statistics for the genetic distance data. One standard deviation in genetic distance is represented by 564 points. Based on the genetic distance between the USA and Germany (352), one standard deviation indicates a shift to the genetic distance between the USA and Mexico (904), the USA and Thailand (920), or the USA and Turkey (927).

Figure 2 gives a first impression of the scale of high- and low-skilled migration between pairs of countries in connection with the genetic distance of the two countries. To ease interpretation we use non-parametric binned scatter plots. Instead of showing all country pairs, we bin genetic distance into 20 equal sized bins and then plot only the means within each bin for genetic distance and the scale of high- or low-skilled migration.<sup>6</sup> We see a clear negative relationship between the migration propensity of the low-skilled and genetic distance, meaning that a higher genetic distance is associated with lower migration of low-skilled individuals. There is not such a clear cut relationship between high-skilled migration and genetic distance. However, if at all, the relationship is positive.

[Figure 2 here]

Regarding selection, figure 3 shows the skill mix of migrants from a certain source country within a destination country together with the genetic distance. Again, the figure is a non-parametric binned scatter plot as described above. The relationship is strongly positive. On first sight, a higher genetic distance is again associated with more high-skilled migrants than low-skilled migrants.

[Figure 3 here]

Cultural distance is complemented with data on linguistic distance. The language data is provided by Isphording and Otten (2013) and relies on the Levensthein distance. This method compares the pronunciation of a set of words with the same meaning across languages and can be understood as the number of cognates, that is common ancestries, between two languages. The final Levensthein distance is achieved by averaging over the set of words and normalizing and gives a percentage measure of similartiy.<sup>7</sup> The

---

<sup>6</sup>In this procedure, we follow Chetty et al. (2014).

<sup>7</sup>When languages do not even have random similarities the value can be above 100 per cent, e.g. Vietnamese to English (104,06) (Isphording and Otten, 2013).

closer the languages of source and destination country are, the smaller the Levensthein distance. The smallest language distance in our sample is between Finland and Estonia, while Denmark and Jordan have the maximum value.

We include several control variables that are likely to affect migration flows. The summary statistics for all variables are documented in Table 1.

To capture the influence of skill premia, we include the wage data by Grogger and Hanson (2011). They provide wage measures for the different skill groups for each source and destination country in our sample. They estimate of the 20th and 80th percentiles of income using data from the World Development Indicators and from the WIDER World Income Inequality Database. The difference between income for low-skill wages for destination and source country is then given by  $W_d^{20^{th}} - W_s^{20^{th}}$ , the difference for high-skill wages and the differentials within high and low-skill wages in a country are constructed correspondingly.

We control for destination countries where English is the primary language as this is likely to ease communication since English is so widely taught in schools.

If both source and destination countries are signatories of the Schengen agreement migration and travel costs are lower, so a dummy is included for this case. Even if countries are not Schengen members, sharing a border may make migration easier, therefore contiguity is controlled for. The latter is taken from the gravity dataset by Head et al. (2010) as well as the geographic distance between destination and source country in km (weighted by population) and a dummy for a (former) colonial relationship between the two countries. While a greater distance in km is increasing migration costs, a colonial relationship implies a shared heritage which might lower costs. Finally, two geographical variables are included: First, the difference in absolute longitude of the two countries and second the difference in absolute latitude. The latter also controls for differences in climate which is connected with agricultural productivity Ashraf and Galor (2013).

Differences in the structure of the economies within a country pair might play a role, so differences between a country pair in the value added of the agricultural, industrial and service sector is included. Data has been collected from the World Development Indicators. We control for heterogeneity in religions across a country pair by including differences in the share of Protestants, Muslims and other religions in the population. The data is taken from Ashraf and Galor (2013). The openness of a destination country is addressed by including the inflow of foreigners and asylum-seekers, retrieved from the International Migration Dataset from the OECD. The skill mix of residents in source and destination country are controlled for by the difference in years of schooling and the share of people with completed tertiary schooling, both taken from Barro and Lee (2013). Finally, the difference in ethnic fractionalization controls for the degree of homogeneity, or to be more precise, the difference in homogeneity across a country pair is included.<sup>8</sup>

---

<sup>8</sup>We use the data of Ashraf and Galor (2013) who provide the index calculated by Alesina et al. (2003).

## 3 Econometric Setup

### 3.1 Estimation

In the econometric setup of this study, we follow Grogger and Hanson (2011) and estimate *scale* and *selection* regressions of bilateral migration flows. We already argued that the log emigration shares by skill group can be interpreted as log odds of migrating to destination country  $d$  versus staying in source country  $s$ , equal to  $\left(\ln \frac{E_{sd}^j}{E_s^j}\right)$  for skill group  $j \in \{H(igh), L(ow)\}$  when the error structure follows an i.i.d. extreme value distribution (McFadden, 1974). Equation (1) then shows the basic regression setup for explaining the scale of migration by skill level.

$$\ln \frac{E_{sd}^j}{E_s^j} = \alpha_0 + \alpha_1 \text{Genetic distance}_{sd} + \mathbf{X}_{sd}'\gamma + \epsilon_{sd} \quad (1)$$

In our baseline specifications, our set of control variables contains the language distance, the log geographic distance, the difference in wage levels, the difference in the absolute longitude and the difference in the absolute latitude. Note that we include the language distance into the regressions. Both measures, genetic and language distance, are positively correlated ( $r = 0.23$ ). Thus, we want to avoid that genetic distance only picks up differences in languages which are certainly a part of genetic distance. We also add dummies for an anglophone destination country, whether the two countries participate in the Schengen agreement, whether they are contiguous, and whether they have a common colonial origin. All these variables are very common in the literature to capture basic determinants of trade and migration flows.

In extended models, we also control for further variables which should capture factors which could be correlated with both, genetic distance and migrant flows and/or migrant selection. Thus, we use differences in the industry structure by using the differences in the shares of the agricultural, service, and industry sector. Countries that have a similar industry structure might be also more connected over migration flows and might show a lower genetic distance. The same argumentation holds for differences in religious orientations. Here we include the shares of Catholics, Protestants, Muslims, and other religions in the regressions.

To measure the general openness of a country towards immigration, we include the log of the aggregate inflow of foreigners and the log number of asylum-seekers into the country. The next controls are differences in the general skill structure between two countries. The reason is that countries with a similar skill structure are more likely to interact with each other. We therefore include the difference in the average years of schooling and the difference in the share of high-skilled. The last control uses the difference in the ethnic fractionalization. Countries that show a similar fractionalization could find it easier to interact with each other.



The error term  $\epsilon_{sd}$  of regression (1) is clustered at the destination country level to allow for arbitrary correlation within destination countries.<sup>9</sup>

In a second step, we use  $\ln \frac{E_{sd}^H}{E_{sd}^L} - \ln \frac{E_s^H}{E_s^L}$  as a measure of the skill mix of migrants from  $s$  in  $d$ . Whenever  $\ln \frac{E_{sd}^H}{E_{sd}^L} > \ln \frac{E_s^H}{E_s^L}$ , then migrants are positively selected from the source country population. Equation 2 shows the regression that we are using for estimating the effect of genetic distance on the skill mix of migrants. We exploit the same set of controls as in the scale regressions.

$$\ln \frac{E_{sd}^H}{E_{sd}^L} - \ln \frac{E_s^H}{E_s^L} = \beta_0 + \beta_1 \text{Genetic distance}_{sd} + \mathbf{X}'_{sd}\phi + \mu_{sd} \quad (2)$$

### 3.2 Identification

Identifying a causal effect of genetic distance on migrant selection is difficult, mainly due to omitted variables that could drive both, migrant selection and genetic distance. Even though genetic distance is measured in 1990 and migrant stocks are measured in 2000, we can also not rule out reverse causality. This can be the case, when persistent migration flows have led to the genetic distance, that we observe today.

To solve this problem, we use an instrumental variables approach. Thus, we need an instrument that is correlated with the genetic distance in 1990 but has no direct effect on migrant selection. We will follow Spolaore and Wacziarg (2009) and use the genetic distance in 1500 as an instrument for the genetic distance in 1990. Importantly, Spolaore and Wacziarg (2009) argue that the genetic distance in 1500 is measured before large migration waves have happened. Thus, persistent migration flows should not be a major concern by using this instrument (see Spolaore and Wacziarg (2009) for a detailed discussion).

Depending on slight changes in the control vector, the first stage regression of the instrumental variables regression is the same in all applications (see equation 3).

$$\text{Genetic distance}_{sd} = \lambda_0 + \lambda_1 \text{Genetic distance}_{sd}^{1500} + \mathbf{X}'_{sd}\omega + \nu_{sd} \quad (3)$$

Once we have the predicted genetic distance from the first stage, we can include the fitted values into the second stages of the scale (equation 4) and the selection (equation 5) regression. The variation is now only driven by the arguably exogenous variation in genetic distance in 1500.

$$\ln \frac{E_{sd}^j}{E_s^j} = \alpha_0 + \alpha_1 \widehat{\text{Genetic distance}_{sd}} + \mathbf{X}'_{sd}\gamma + \epsilon_{sd} \quad (4)$$

$$\ln \frac{E_{sd}^H}{E_{sd}^L} - \ln \frac{E_s^H}{E_s^L} = \beta_0 + \beta_1 \widehat{\text{Genetic distance}_{sd}} + \mathbf{X}'_{sd}\phi + \mu_{sd} \quad (5)$$

---

<sup>9</sup>Using two-way clustering (Cameron et al., 2011) at the destination and the source country level do not affect our results.

Note that we still cluster the standard errors at the destination country level and that we estimate the first and the second regression together to get the standard errors in the second regression right. This is necessary because the predicted values of the genetic distance are only estimates from the first stage and not real observations.

## 4 Results

### 4.1 The Scale of International Migrants

We start our exposition of the results by describing the scale effects of genetic distance before we move on to describe the selection of international migrants caused by genetic distance between countries.

Tables 2 and 3 show the scale regressions for primary and tertiary education, respectively. Table 2, which deals with primary education, we show in Column (1) the simple correlation between genetic distance and the scale of primary educated migrants. Note that we have rescaled genetic distance and language distance by their own standard deviations to, first, give an easier interpretation to the coefficients, and, second, to make the comparison between language distance and genetic distance much more easier. The coefficient is negative and highly significant. This indicates that country pairs which show a higher genetic distance are more likely to show a low share of primary educated migrants. The reason for this correlation are most likely to be high migration costs for low-skilled migrants which hold them back from migration.

**[Tables 2 and 3 here]**

However, this correlation might also be explained by other factors. Thus, the next important variable that we include is the language distance. We have seen earlier that language distance and genetic distance are positively correlated. Therefore, we see that in Column (2), where we include language distance, that the coefficient on genetic distance decreases slightly but remains highly significant. This indicates that language is a part of genetic distance but does not explain the whole story. By adding controls, we actually see that genetic distance decreases to almost zero and is not significant anymore. Language distance decreases as well but remains significantly negative.

However, as discussed above, OLS might be biased by omitted variables and/or reversed causality. The direction of the bias is unknown a priori because we do not know which omitted variables are missing in the regression. Thus, we deploy our instrumental variables estimation in Columns (4) to (7). Column (4) shows the first stage. As expected, the genetic distance in 1500 is a highly significant predictor for the genetic distance in 1990. The corresponding Kleibergen-Paap F statistic is equal to 138.2 (see Column (6)).<sup>10</sup>

---

<sup>10</sup>In case of non-i.i.d. errors, the Cragg-Donald F statistic is not valid anymore. Instead, we have to rely on the Kleibergen-Paap F statistic (Kleibergen and Paap, 2006).

This statistic should indicate whether the IV suffers from a weak instrumental variables problem. The empirical F statistic is well above the critical values tabulated by Stock and Yogo (2005). Thus, we do not have to fear a bias from weak instruments even though the critical values are only valid under i.i.d. errors.<sup>11</sup>

Column (5) shows the reduced form regression. We find a negative relationship between the genetic distance and the scale of primary educated migrants. That already indicates that there is, if at all, only a weak link between the scale of primary educated migrants and genetic distance. This notion is confirmed by looking at the second stage in Column (6). Again, the effect of genetic distance is negative but not significant. However, the coefficient is much larger in absolute terms than in the OLS regression of column (3). Recall that we have rescaled the genetic and language distance by their own standard deviations. Thus, one standard deviation increase in genetic distance decreases the scale of low-skilled migrants by  $1 - e^{-0.179} \cdot 100 = 16.4$  percent. Thus, the effect seems to be sizeable in economic terms but is not very precisely estimated. Language distance shows a much larger coefficient, even though, it has no causal meaning. There, increasing language distance by one standard deviation is associated with a decrease in the scale of low-skilled migrants by  $e^{0.472} - 1 \cdot 100 = 60.3$  percent. Looking at Column (7) where we add a bunch of other controls does not change the picture. The coefficient on genetic distance is slightly more negative and language distance becomes slightly more positive. Even though it is not significant, genetic distance seems to hold back low-skilled migrants from international migration.

Table 3 shows the results for the scale of tertiary educated migrants. There, contrary to the case of primary educated migrants, we see in Column (1) that genetic distance is positive but not significantly related to the scale of high-skilled migrants. However, Column (2) reveals that language distance acts in a similar way for high-skilled migrants than for low-skilled migrants. The coefficients on language distance are almost identical between the two regressions. Adding more controls in Column (3), the coefficient on genetic distance becomes even significantly positive whereas the coefficient on language distance is very close to that of primary educated migrants.

But of course, the same caveat applies as in the case for low-skilled migrants, that the correlation cannot be interpreted causally. Therefore, Column (4) repeats the first stage which looks slightly different because we included  $W_d^{80th} - W_s^{80th}$  instead of  $W_d^{20th} - W_s^{20th}$ . However, that has not changed the predictive power of the genetic distance in 1500. The Kleibergen-Paap F statistic is still very high at 135.9. The reduced form in Column (5) is positive but also not significant. This again induces a non-significant second stage with a positive coefficient on genetic distance. The coefficient is slightly larger than in

---

<sup>11</sup>At the moment, there are no critical values for the case of non-i.i.d. errors available. Baum et al. (2007) suggest to use the critical values with caution. However, the empirical F statistic is well above the critical values.

the OLS case which was significant before. Thus, the imprecise estimation of the IV is also prevalent in the high-skilled scale regressions. The effect size would implicate that a one standard deviation increase in genetic distance increase the scale of tertiary educated migrants of about 49.7 percent. However, including further controls in Column (7) decreases the coefficient down to 12.7 percent again.<sup>12</sup> Interestingly, the coefficient on language distance is still in the same ballpark as in the regressions for the low-skilled migrants.

## 4.2 The Selection of International Migrants

The scale regressions have shown that low- and high-skilled migrants react differently to genetic distance. Low-skilled migrants are hold back from countries which are genetically farer away from their home country and high-skilled migrants might even react positively to a higher genetic distance. Even though this was not significant in the scale regressions, it could lead to significant differences when we put both parts together.

Table 4 shows the OLS selection regressions. Columns (1) to (3) perform a horse race between genetic distance and language distance. Taking Columns (1) and (2) both distances enter significantly positive. Thus, a higher genetic or language distance is associated with a higher positive selectivity of migrants. In Column (3), when we include both distances together, we see that only genetic distance has predictive power and language distance is almost zero. In Column (4) we add control variables. The coefficient of genetic distance drops by half but remains positive and significant. The drop is likely to be caused by the primary educated migrants where we have seen a similar drop in the coefficient once we included the controls. Including further controls drive down the coefficient again substantially. However, it is very precisely estimated and therefore is still highly significant.

[Tables 4 here]

To check a potential bias, we also perform IV estimations for the selection regressions. In Table 5 we find the corresponding IV regression results. The first column replicates the OLS results for comparison. Column (2) shows that the first stage is again very strong with an Kleibergen-Paap F statistic of 148.3. The reduced form now shows up to be highly significant and has the expected positive sign. This already indicates that there is a causal impact of genetic distance on the selection of migrants.

[Tables 5 here]

---

<sup>12</sup>This reduction is not due to the sample size. If we run the regression in Column (6) on the sample in Column (7), we receive a point estimate of 0.472 on genetic distance which is significant on the 10 percent level.

Columns (4) to (7) show the IV estimation results. Jumping directly to Column (6), we observe that the coefficient is rather large. Thus, increasing genetic distance by one standard deviation increases the selection of migrants by 82.9 percent (coefficient equal to 0.604). Adding further controls decreases the coefficient to 0.348 or 42.6 percent. However, this is partly driven by sample selection. The implied effect size on the reduced sample with the controls from Column (6) is equal to 56.5 percent (coefficient equal to 0.448). However, taking the most conservative estimate of 42.6 percent is still an economically highly significant effect.

Summarizing the evidence, we see that genetic distance significantly increases the skill mix of migrants that a country attracts.

### 4.3 Extensions and Robustness Checks

A crucial assumption of the model is the assumption of irrelevant alternatives. That means that the estimates should not be influenced by the presence of an alternative destination. We can check this assumption by omitting one destination country and see whether the coefficients in the model remain stable over the different samples. In Table 6, we show the outcome of this exercise. The upper column shows the country that is left out. All specifications indicate very similar coefficients and significance levels. Thus, the assumption of irrelevant alternatives should not be a problem for our model.

We argue in the introduction that the genetic distance is a persistent process that has established over decades and is not changed easily. Thus genetic distance should only be able to explain the cross-sectional variation in migrant selectivity between countries and not the variation within countries over time. Table 7 shows regressions where we control for the migrant skill mix in 1990, i.e. the lagged dependent variable. We see that this variable is a highly significant predictor of the skill mix in 2000. Over the different specifications, we observe that the coefficient on genetic distance becomes very small and insignificant. This confirms that genetic distance is only able to capture long-run differences and is not able to explain short-run changes in migrant selectivity. Interestingly, the difference in wage inequality,  $W_d^{80/20} - W_s^{80/20}$ , is a highly significant predictor of the change in the migrant skill mix between 2000 and 1990.

## 5 Conclusion

This paper provides evidence on the role of cultural distance between countries for the selectivity of international migration. We are using genetic distance as a proxy for cultural distance to show that country pairs with a higher genetic distance experience a higher selection of migrants. This conclusion is robust to the introduction of several control variables as well as to an instrumental variables approach. Running regressions on the scale of migration, separately by skill group, we find (weak) evidence that low-skilled

migrants shy away from migrating to countries which are culturally farther and that high-skilled migrants act in the opposite way. This is in line with the argument that low-skilled migrants exhibit larger migration costs from cultural barriers than high-skilled migrants. Actually high-skilled migrants seem to be attracted from higher genetic diversity.

## References

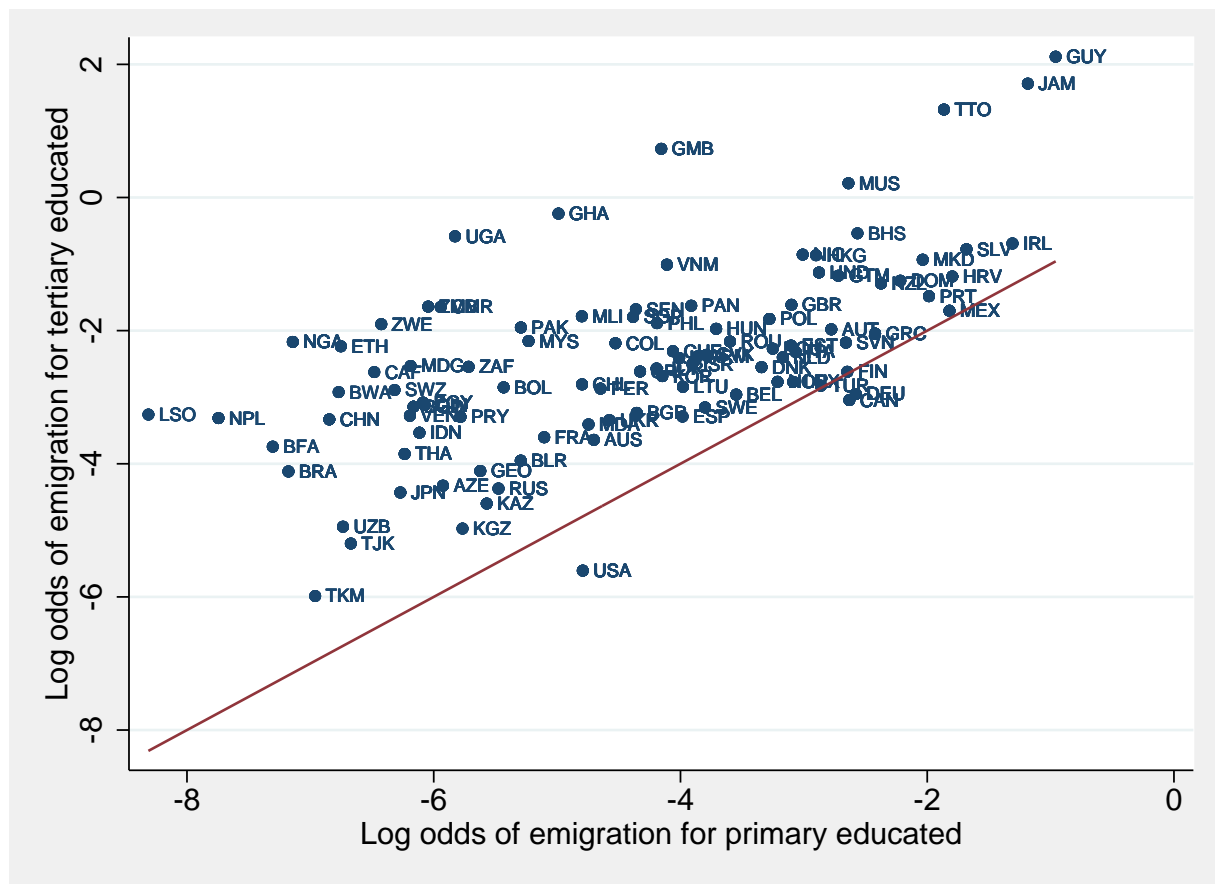
- Abramitzky, R. (2009). The effect of redistribution on migration: Evidence from the Israeli kibbutz. *Journal of Public Economics*, 93(3-4):498–511.
- Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., and Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth*, 8:155–194.
- Ashraf, Q. and Galor, O. (2013). The "Out of Africa" Hypothesis, Human Genetic Diversity, and Comparative Economic Development. *American Economic Review*, 103(1):1–46.
- Barro, R. and Lee, J.-W. (2013). A new data set of educational attainment in the world, 1950-2010. *Journal of Development Economics*, 104:184–198.
- Bauernschuster, S., Falck, O., Heblich, S., Suedekum, J., and Lameli, A. (2013). Why Are Educated and Risk-Loving Persons More Mobile Across Regions? *Journal of Economic Behavior and Organization*, forthcoming.
- Baum, C. F., Schaffer, M. E., and Stillman, S. (2007). Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal*, 7(4):465–506.
- Belot, M. and Ederveen, S. (2011). Cultural barriers in migration between OECD countries. *Journal of Population Economics*, 25(3):1077–1105.
- Belot, M. V. K. and Hatton, T. J. (2012). Immigrant Selection in the OECD. *Scandinavian Journal of Economics*, 114(4):1105–1128.
- Borjas, G. J. (1987). Self-Selection and the Earnings of Immigrants. *American Economic Review*, 77(4):531–553.
- Burchardi, K. B. and Hassan, T. A. (2013). The Economic Impact of Social Ties: Evidence from German Reunification. *Quarterly Journal of Economics*, 128(3):1219–1271.
- Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.
- Cavalli-Sforza, L. L. (2001). *Genes, Peoples and Languages*. Penguin Group.
- Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.
- Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. NBER Working Paper Series.

- Chiquiar, D. and Hanson, G. H. (2005). International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the United States. *Journal of Political Economy*, 113(2):239–281.
- Comin, D. A., Dmitriev, M., and Rossi-Hansberg, E. (2012). The Spatial Diffusion of Technology. NBER Working Paper 18534.
- Dahl, M. S. and Sorenson, O. (2010). The migration of technical workers. *Journal of Urban Economics*, 67(1):33–45.
- Docquier, F., Lowell, B. L., and Marfouk, A. (2007). A gendered assessment of the brain drain. Mimeo.
- Falck, O., Heblich, S., Lameli, A., and Südekum, J. (2012). Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, 72(2-3):225–239.
- Felbermayr, G. J. and Toubal, F. (2010). Cultural proximity and trade. *European Economic Review*, 54(2):279–293.
- Grogger, J. and Hanson, G. H. (2011). Income maximization and the selection and sorting of international migrants. *Journal of Development Economics*, 95(1):42–57.
- Guiso, L., Sapienza, P., and Zingales, L. (2006). Does Culture Affect Economic Outcomes? *Journal of Economic Perspectives*, 20(2):23–48.
- Guiso, L., Sapienza, P., and Zingales, L. (2009). Cultural Biases in Economic Exchange. *Quarterly Journal of Economics*, 124(3):1095–1131.
- Head, K., Mayer, T., and Ries, J. (2010). The erosion of colonia trade linkages after independence. *Journal of International Economics*, 81:1–14.
- Isphording, I. E. and Otten, S. (2013). The costs of babylon-linguistic distance in applied economics. *Review of International Economics*, 21(2):354–369.
- Kaestner, R. and Malamud, O. (2013). Self-Selection and International Migration: New Evidence from Mexico. *Review of Economics and Statistics*, forthcoming.
- Kleibergen, F. and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1):97–126.
- Mayda, A. M. (2009). International migration: a panel data analysis of the determinants of bilateral flows. *Journal of Population Economics*, 23(4):1249–1274.
- McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics*, 3(4):303–328.



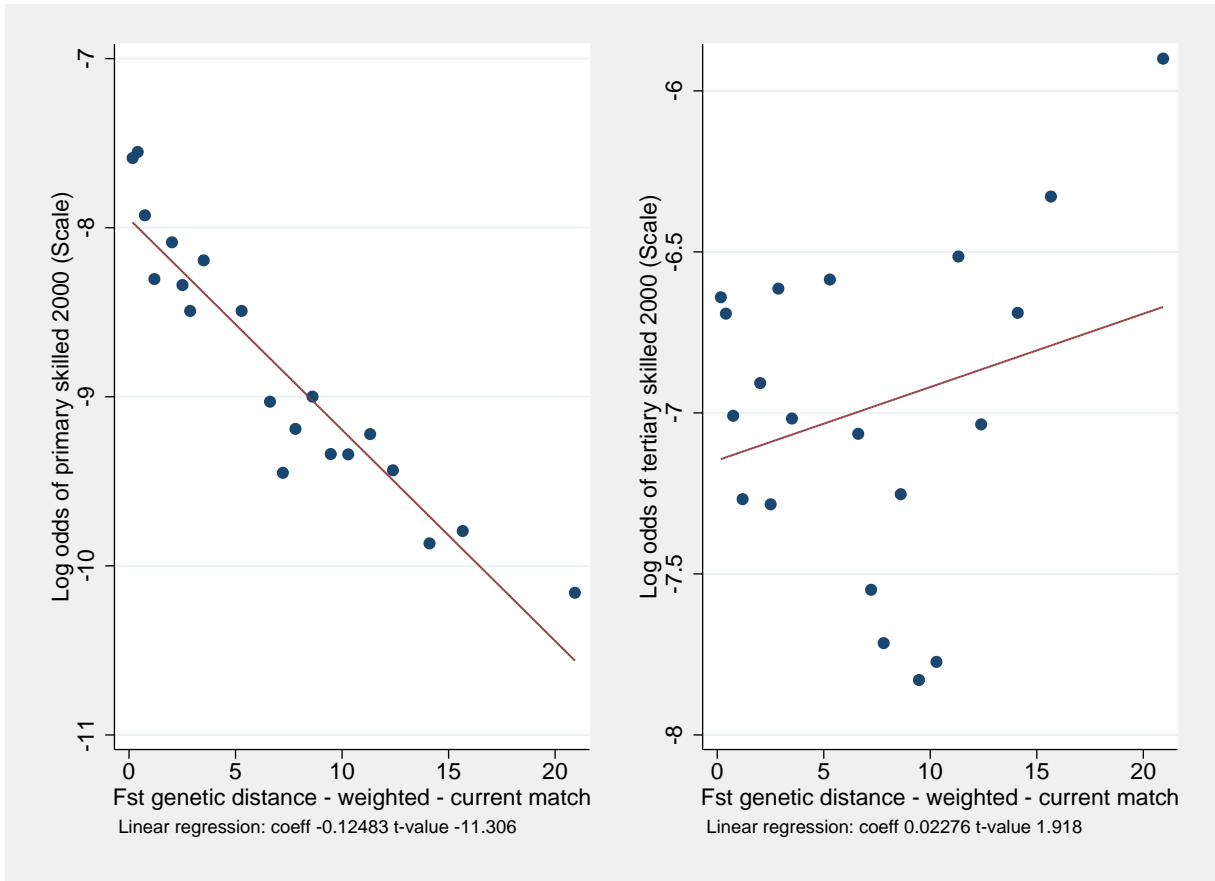
- Ottaviano, G. I. and Peri, G. (2005). The economic value of cultural diversity: evidence from US cities. *Journal of Economic Geography*, 6(1):9–44.
- Spolaore, E. and Wacziarg, R. (2009). The Diffusion of Development. *Quarterly Journal of Economics*, 124(2):469–529.
- Spolaore, E. and Wacziarg, R. (2013). How Deep Are the Roots of Economic Development? *Journal of Economic Literature*, 51(2):325–369.
- Spring, E. and Grossmann, V. (2013). Does Bilateral Trust Affect International Movement of Goods and Labor? IZA Discussion Paper 7385.
- Stock, J. H. and Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In Stock, J. H. and Andrews, D. W., editors, *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*. Cambridge University Press.
- Stolz, Y. and Baten, J. (2012). Brain Drain in the Age of Mass Migration: Does Relative Inequality Explain Migrant Selectivity? *Explorations in Economic History*, 49(2):205–220.
- Tabellini, G. (2010). Culture and Institutions: Economic Development in the Regions of Europe. *Journal of the European Economic Association*, 8(4):677–716.

Figure 1: *Emigration odds (primary and tertiary-educated) by source country, 2000*



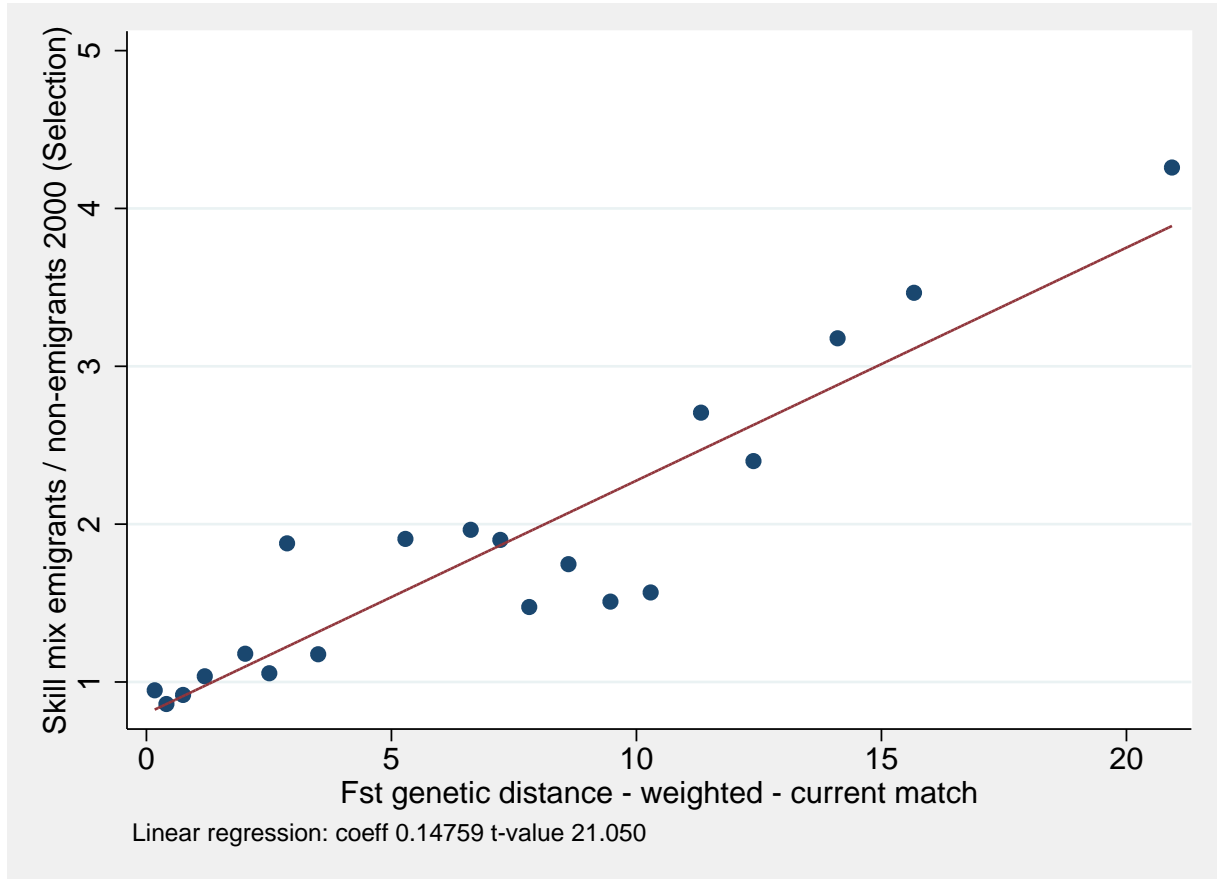
Notes: The figure shows the log odds of emigration for primary and tertiary educated migrants. Source: Docquier et al. (2007). The graph is replicated from Grogger and Hanson (2011).

Figure 2: *Genetic distance and emigration odds by skill level, 2000*



*Notes:* These non-parametric binned scatter plots show the relationship between the log odds of emigration and genetic distance. The figure on the left shows the relationship between the log odds of emigration for tertiary educated migrants and genetic distance and the figure on the right shows the relationship between the log odds of emigration for primary educated migrants and genetic distance. For both figures, the coefficients and t-statistics are from OLS regressions with the microdata. We bin genetic distance into 20 bins of equal size and then the means of genetic distance and log emigration odds are obtained within each bin. Data of migrant stocks by skill level is from Docquier et al. (2007) and the genetic distance data is from Spolaore and Wacziarg (2009).

Figure 3: *Genetic distance and emigration selection, 2000*



*Notes:* The non-parametric binned scatter plot shows the relationship between the skill mix of migrants and genetic distance. For the figure, the coefficients and t-statistics are from OLS regressions with the microdata. We bin genetic distance into 20 bins of equal size and then the means of genetic distance and log emigration odds are obtained within each binn. Data of migrant stocks by skill level is from Docquier et al. (2007) and the genetic distance data is from Spolaore and Wacziarg (2009).

Table 1: *Summary Statistics*

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
<i>Panel A: Migrant Selection Measures</i>					
Emigration share of primary-educated population, $(E_d^L/E_s^L)$	0.003	0.0152	0	0.2515	1,286
Emigration share of tertiary-educated population, $(E_{sd}^H/E_s^H)$	0.0258	0.2117	0	4.7979	1,286
Migrant skill mix $\left(\ln \frac{E_{sd}^H}{E_d^L} - \ln \frac{E_s^H}{E_s^L}\right)$	1.8549	1.5824	-3.0409	8.0772	1,286
<i>Panel B: Genetic and Language Distance Data</i>					
$F_{ST}$ genetic distance	715	564	0	2,695	1,286
$F_{ST}$ genetic distance, 1500	978	652	0	3,557	1,286
Language distance	87	23	0	105	1,286
<i>Panel C: Controls</i>					
$W_d^{80^{th}} - W_s^{80^{th}}$	14.873	11.6478	-24.2247	33.3972	1,286
$W_d^{20^{th}} - W_s^{20^{th}}$	5.6668	4.4973	-11.5201	13.1767	1,286
$W_d^{80/20} - W_s^{80/20}$	23.054	10.2995	-7.9336	48.0755	1,286
Anglophone destination	0.4067		0	1	1,286
Schengen	0.1431		0	1	1,286
Contiguous	0.0295		0	1	1,286
Distance (km)	7,076	4,788	161	19539	1,286
Colony	0.0599		0	1	1,286
$\Delta$ Share Agriculture	-9.7646	11.7269	-54.9525	8.6881	1,111
$\Delta$ Share Industry	-1.3291	9.7702	-29.7185	29.5948	1,115
$\Delta$ Share Service	11.1053	13.7527	-31.5867	47.9596	1,101
$\Delta$ Share Catholics	-0.6215	49.9455	-96.8000	96.9000	1,286
$\Delta$ Share Protestants	29.449	40.2538	-97.7	97.8000	1,271
$\Delta$ Share Muslims	-15.5989	29.9303	-99.2	3	1,286
$\Delta$ Share Other Religion	-13.5222	38.6729	-98.2	68.7	1,271
Inflow of foreign population, 1999	167,468	214,409	7,937	673,873	1,286
Inflow of asylum seekers, 1999	27,123	26,541	1,528	95,113	1,286
$\Delta$ Absolute latitude	-7.0933	67.8	-172	172	1,286
$\Delta$ Absolute longitude	18.2821	20.6433	-39	62.6333	1,286
$\Delta$ Years of schooling	2.5643	2.9771	-4.5709	11.984	1,132
$\Delta$ Share tertiary	5.9055	8.8887	-22.8357	30.4371	1,132
$\Delta$ Ethnic fractionalization	-0.1897	0.3114	-0.8716	0.7104	1,286

Notes:  $\Delta$  represents the simple difference between destination and source country, i.e.  $\Delta X = X_d - X_s$ .

Table 2: *Scale Regressions for Primary Education*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Scale Primary	Scale Primary	Scale Primary	Genetic Distance	Scale Primary	Scale Primary	Scale Primary
	OLS	OLS	OLS	FS	RF	IV	IV
$F_{ST}$ genetic distance	-0.703*** (0.143)	-0.539*** (0.147)	-0.037 (0.166)			-0.179 (0.350)	-0.212 (0.271)
$F_{ST}$ genetic distance, 1500				0.779*** (0.066)	-0.139 (0.277)		
Language distance		-0.723*** (0.113)	-0.494*** (0.108)	0.129*** (0.035)	-0.495*** (0.104)	-0.472*** (0.101)	-0.358*** (0.090)
<i>Controls:</i>							
$W_d^{20^{th}} - W_s^{20^{th}}$			0.027 (0.060)	0.003 (0.013)	0.028 (0.061)	0.028 (0.060)	0.003 (0.023)
Anglophone destination			1.081 (0.953)	-0.019 (0.146)	1.069 (0.934)	1.066 (0.904)	-0.522 (0.384)
Schengen			-0.009 (0.390)	-0.178*** (0.056)	0.023 (0.406)	-0.009 (0.377)	-0.400 (0.320)
Contiguous			1.232*** (0.401)	-0.045 (0.092)	1.271** (0.452)	1.263*** (0.432)	1.064** (0.452)
Log distance			-0.743** (0.304)	-0.233*** (0.053)	-0.658 (0.383)	-0.699** (0.323)	-0.664*** (0.195)
Colony			1.884*** (0.599)	0.152 (0.119)	1.923*** (0.570)	1.950*** (0.557)	1.705*** (0.471)
$\Delta$ Absolute latitude			-0.001 (0.005)	-0.000 (0.001)	-0.001 (0.006)	-0.001 (0.005)	0.007* (0.004)
$\Delta$ Absolute longitude			-0.027** (0.011)	0.012*** (0.003)	-0.026** (0.009)	-0.024** (0.010)	0.011 (0.007)
<i>Further Controls:</i>							
$\Delta$ Share Industry							-0.013* (0.007)
$\Delta$ Share Service							-0.016*** (0.006)
$\Delta$ Share Protestants							-0.002 (0.003)
$\Delta$ Share Muslims							-0.001 (0.003)
$\Delta$ Share Other Religion							-0.002 (0.003)
Log inflow foreigners							0.760** (0.309)
Log inflow asylum-seekers							0.356 (0.323)
$\Delta$ Years of schooling							-0.257*** (0.049)
$\Delta$ Share tertiary							0.070*** (0.026)
$\Delta$ Ethnic fractionalization							0.631** (0.298)
$R^2$	0.079	0.159	0.322	0.685	0.323	0.320	0.613
Observations	1,286	1,286	1,286	1,286	1,286	1,286	959
Cluster	15	15	15	15	15	15	14
Kleibergen-Paap F statistic						138.2	90.9

Notes: *Scale Primary* indicates that the log odds of primary educated migrants is used as the dependent variable, i.e.  $\ln(E_{sd}^L/E_s^L)$ .  $F_{ST}$  genetic distance,  $F_{ST}$  genetic distance, 1500, and language distance are divided by their own standard deviation to provide an interpretation in standard deviations.  $\Delta$  represents the simple difference between destination and source country, i.e.  $\Delta X = X_d - X_s$ . Robust standard errors in parentheses clustered at the destination country level. Significance levels: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Table 3: *Scale Regressions for Tertiary Education*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Scale Tertiary	Scale Tertiary	Scale Tertiary	Genetic Distance	Scale Tertiary	Scale Tertiary	Scale Tertiary
	OLS	OLS	OLS	FS	RF	IV	IV
$F_{ST}$ genetic distance	0.128 (0.205)	0.287 (0.195)	0.393** (0.173)			0.403 (0.297)	0.120 (0.167)
$F_{ST}$ genetic distance, 1500				0.776*** (0.067)	0.313 (0.245)		
Language distance		-0.702*** (0.114)	-0.522*** (0.124)	0.118*** (0.038)	-0.476*** (0.133)	-0.524*** (0.120)	-0.426*** (0.095)
<i>Controls:</i>							
$W_d^{80^{th}} - W_s^{80^{th}}$			0.040 (0.029)	0.004 (0.006)	0.042 (0.030)	0.040 (0.029)	0.028*** (0.009)
Anglophone destination			2.008** (0.878)	-0.020 (0.128)	2.002** (0.886)	2.010** (0.842)	0.207 (0.311)
Schengen			0.377 (0.301)	-0.152** (0.059)	0.315 (0.333)	0.376 (0.291)	0.082 (0.190)
Contiguous			0.719** (0.304)	-0.030 (0.092)	0.705* (0.367)	0.717** (0.338)	0.276 (0.326)
Log distance			-0.633** (0.259)	-0.228*** (0.053)	-0.728* (0.361)	-0.637** (0.297)	-0.623*** (0.129)
Colony			1.943*** (0.493)	0.150 (0.119)	1.999*** (0.516)	1.938*** (0.463)	1.715*** (0.422)
$\Delta$ Absolute latitude			0.005 (0.006)	-0.000 (0.001)	0.005 (0.006)	0.005 (0.006)	0.010*** (0.002)
$\Delta$ Absolute longitude			-0.004 (0.012)	0.012*** (0.003)	0.000 (0.011)	-0.004 (0.012)	0.022*** (0.006)
<i>Further Controls:</i>							
$\Delta$ Share Industry							0.017** (0.008)
$\Delta$ Share Service							-0.003 (0.006)
$\Delta$ Share Protestants							-0.009** (0.004)
$\Delta$ Share Muslims							-0.002 (0.002)
$\Delta$ Share Other Religion							-0.004 (0.003)
Log inflow foreigners							0.776*** (0.225)
Log inflow asylum-seekers							0.381* (0.195)
$\Delta$ Years of schooling							-0.085* (0.044)
$\Delta$ Share tertiary							0.078*** (0.018)
$\Delta$ Ethnic fractionalization							0.671*** (0.242)
$R^2$	0.002	0.080	0.347	0.686	0.340	0.347	0.723
Observations	1,286	1,286	1,286	1,286	1,286	1,286	959
Cluster	15	15	15	15	15	15	14
Kleibergen-Paap F statistic						135.9	89.6

Notes: *Scale Tertiary* indicates that the log odds of primary educated migrants is used as the dependent variable, i.e.  $\ln(E_{sd}^H/E_s^H)$ .  $F_{ST}$  genetic distance,  $F_{ST}$  genetic distance, 1500, and language distance are divided by their own standard deviation to provide an interpretation in standard deviations.  $\Delta$  represents the simple difference between destination and source country, i.e.  $\Delta X = X_d - X_s$ . Robust standard errors in parentheses clustered at the destination country level. Significance levels: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Table 4: *OLS Results of Migrant Selection*

	(1)	(2)	(3)	(4)	(5)
$F_{ST}$ genetic distance	0.832*** (0.104)		0.827*** (0.098)	0.440*** (0.088)	0.178*** (0.056)
Language distance		0.209** (0.071)	0.021 (0.057)	0.009 (0.043)	-0.015 (0.032)
<i>Controls:</i>					
$W_d^{80/20} - W_s^{80/20}$				0.027** (0.011)	0.017* (0.010)
Anglophone destination				0.812** (0.276)	0.637*** (0.170)
Schengen				0.281 (0.174)	0.457** (0.201)
Contiguous				-0.564** (0.220)	-0.791*** (0.167)
Log distance				0.097 (0.091)	0.076 (0.103)
Colony				0.116 (0.175)	0.107 (0.131)
$\Delta$ Absolute latitude				0.006*** (0.002)	0.003 (0.002)
$\Delta$ Absolute longitude				0.025*** (0.004)	0.016*** (0.003)
<i>Further Controls:</i>					
$\Delta$ Share Industry					0.032*** (0.005)
$\Delta$ Share Service					0.018*** (0.004)
$\Delta$ Share Protestants					-0.007*** (0.002)
$\Delta$ Share Muslims					-0.001 (0.002)
$\Delta$ Share Other Religion					-0.003 (0.002)
Log inflow foreigners					0.034 (0.166)
Log inflow asylum-seekers					0.022 (0.201)
$\Delta$ Years of schooling					0.165*** (0.040)
$\Delta$ Share tertiary					0.013 (0.014)
$\Delta$ Ethnic fractionalization					-0.151 (0.129)
$R^2$	0.276	0.017	0.275	0.532	0.642
Observations	1,286	1,286	1,286	1,286	959
Cluster	15	15	15	15	14

Notes: The dependent variable is the migrant skill mix in 2000, i.e.  $\ln(E_{sd}^H/E_{sd}^L) - \ln(E_s^H/E_s^L)$ .  $F_{ST}$  genetic distance and language distance are divided by their own standard deviation to provide an interpretation in standard deviations.  $\Delta$  represents the simple difference between destination and source country, i.e.  $\Delta X = X_d - X_s$ . Robust standard errors in parentheses clustered at the destination country level. Significance levels: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.



Table 5: *IV Results for Genetic Distance*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Selection	Genetic Distance	Selection	Selection	Selection	Selection	Selection
	OLS	FS	RF	IV	IV	IV	IV
$F_{ST}$ genetic distance	0.440*** (0.088)			1.123*** (0.181)	1.129*** (0.182)	0.604*** (0.165)	0.348** (0.145)
$F_{ST}$ genetic distance, 1500		0.778*** (0.064)	0.470*** (0.102)				
Language distance	0.009 (0.043)	0.117*** (0.036)	0.055 (0.046)		-0.047 (0.070)	-0.016 (0.044)	-0.045 (0.039)
<i>Controls:</i>							
$W_d^{80/20} - W_s^{80/20}$	0.027** (0.011)	0.007 (0.007)	0.030*** (0.010)			0.026** (0.011)	0.017 (0.010)
Anglophone destination	0.812** (0.276)	-0.056 (0.106)	0.804** (0.277)			0.838*** (0.269)	0.662*** (0.167)
Schengen	0.281 (0.174)	-0.156** (0.055)	0.187 (0.168)			0.281* (0.167)	0.437** (0.184)
Contiguous	-0.564** (0.220)	-0.030 (0.086)	-0.618** (0.223)			-0.600*** (0.227)	-0.839*** (0.173)
Log distance	0.097 (0.091)	-0.228*** (0.052)	-0.092 (0.099)			0.046 (0.089)	0.028 (0.082)
Colony	0.116 (0.175)	0.162 (0.113)	0.135 (0.159)			0.037 (0.177)	0.039 (0.114)
$\Delta$ Absolute latitude	0.006*** (0.002)	-0.000 (0.001)	0.006*** (0.002)			0.006*** (0.002)	0.003 (0.002)
$\Delta$ Absolute longitude	0.025*** (0.004)	0.012*** (0.002)	0.028*** (0.004)			0.021*** (0.005)	0.012*** (0.004)
<i>Further Controls:</i>							
$\Delta$ Share Industry							0.032*** (0.005)
$\Delta$ Share Service							0.017*** (0.003)
$\Delta$ Share Protestants							-0.006*** (0.002)
$\Delta$ Share Muslims							-0.001 (0.002)
$\Delta$ Share Other Religion							-0.002 (0.001)
Log inflow foreigners							0.024 (0.168)
Log inflow asylum-seekers							0.041 (0.201)
$\Delta$ Years of schooling							0.174*** (0.044)
$\Delta$ Share tertiary							0.008 (0.017)
$R^2$	0.532	0.689	0.526	0.242	0.241	0.526	0.637
Observations	1,286	1,286	1,286	1,286	1,286	1,286	959
Cluster	15	15	15	15	15	15	14
Kleibergen-Paap F statistic				134.5	137.2	148.3	118.0

Notes: *Selection* indicates that the skill mix of migrants is used as the dependent variable, i.e.  $\ln(E_{sd}^H/E_{sd}^L) - \ln(E_s^H/E_s^L)$ .  $F_{ST}$  genetic distance,  $F_{ST}$  genetic distance, 1500, and language distance are divided by their own standard deviation to provide an interpretation in standard deviations.  $\Delta$  represents the simple difference between destination and source country, i.e.  $\Delta X = X_d - X_s$ . Robust standard errors in parentheses clustered at the destination country level. Significance levels: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Table 6: *Omitting Destination Countries*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	AUS	AUT	CAN	DNK	FIN	FRA	DEU	IRL	NLD	NZL	NOR	ESP	SWE	USA	GBR
$F_{ST}$ genetic distance	0.481*** (0.150)	0.624*** (0.167)	0.559*** (0.177)	0.630*** (0.171)	0.648*** (0.170)	0.609*** (0.183)	0.630*** (0.170)	0.623*** (0.166)	0.630*** (0.170)	0.526*** (0.160)	0.642*** (0.171)	0.635*** (0.171)	0.639*** (0.168)	0.490*** (0.135)	0.627*** (0.181)
Language distance	0.008 (0.045)	-0.017 (0.045)	-0.025 (0.048)	-0.029 (0.042)	0.019 (0.029)	-0.027 (0.045)	-0.015 (0.044)	-0.019 (0.047)	-0.010 (0.044)	-0.020 (0.050)	-0.008 (0.041)	-0.005 (0.047)	-0.021 (0.044)	-0.029 (0.056)	-0.018 (0.049)
<i>Controls:</i>															
$W_d^{80/20} - W_s^{80/20}$	0.026*** (0.010)	0.026** (0.011)	0.026** (0.011)	0.020*** (0.011)	0.019** (0.008)	0.027** (0.011)	0.027** (0.012)	0.029*** (0.011)	0.025** (0.011)	0.020* (0.012)	0.023** (0.011)	0.032*** (0.012)	0.026** (0.011)	0.035** (0.015)	0.025** (0.011)
Anglophone destination	0.778*** (0.266)	0.830*** (0.274)	0.814*** (0.288)	0.772*** (0.274)	0.722*** (0.254)	0.912*** (0.282)	0.891*** (0.280)	0.601*** (0.194)	0.827*** (0.283)	0.807*** (0.270)	0.897*** (0.284)	0.946*** (0.277)	0.883*** (0.277)	0.896*** (0.286)	1.064*** (0.318)
Schengen	0.257 (0.170)	0.224 (0.170)	0.227 (0.162)	0.292 (0.178)	0.364** (0.166)	0.357** (0.166)	0.313* (0.178)	0.368** (0.166)	0.258 (0.179)	0.252 (0.169)	0.203 (0.164)	0.284 (0.177)	0.309* (0.186)	0.242 (0.156)	0.297* (0.177)
Contiguous	-0.596*** (0.215)	-0.645** (0.265)	-0.570*** (0.220)	-0.586** (0.238)	-0.652*** (0.213)	-0.589** (0.263)	-0.807*** (0.179)	-0.546** (0.221)	-0.596** (0.247)	-0.576*** (0.213)	-0.609** (0.258)	-0.513** (0.253)	-0.541** (0.226)	-0.471** (0.213)	-0.690*** (0.236)
Log distance	0.054 (0.096)	0.040 (0.091)	0.047 (0.092)	0.050 (0.089)	0.042 (0.095)	0.034 (0.094)	0.016 (0.085)	0.121* (0.073)	0.046 (0.090)	0.128 (0.098)	0.029 (0.090)	-0.008 (0.079)	0.059 (0.085)	0.058 (0.085)	0.004 (0.100)
Colony	0.134 (0.155)	0.033 (0.181)	0.032 (0.189)	-0.027 (0.159)	-0.014 (0.178)	0.007 (0.221)	0.054 (0.185)	0.114 (0.157)	0.050 (0.185)	0.043 (0.165)	0.038 (0.182)	-0.136 (0.163)	0.044 (0.184)	0.040 (0.193)	0.252* (0.149)
$\Delta$ Absolute latitude	0.006*** (0.001)	0.006*** (0.002)	0.006*** (0.002)	0.005*** (0.001)	0.006*** (0.002)	0.006*** (0.002)	0.006*** (0.002)	0.007*** (0.001)	0.005*** (0.002)	0.007*** (0.002)	0.005*** (0.002)	0.006*** (0.001)	0.005*** (0.002)	0.006*** (0.002)	0.004*** (0.002)
$\Delta$ Absolute longitude	0.024*** (0.006)	0.021*** (0.005)	0.021*** (0.006)	0.021*** (0.005)	0.024*** (0.004)	0.023*** (0.005)	0.020*** (0.005)	0.021*** (0.005)	0.020*** (0.005)	0.023*** (0.005)	0.020*** (0.005)	0.024*** (0.005)	0.020*** (0.005)	0.022*** (0.005)	0.021*** (0.005)
$R^2$	0.515	0.516	0.504	0.524	0.566	0.531	0.520	0.542	0.514	0.534	0.525	0.556	0.526	0.518	0.520
Observations	1,190	1,229	1,191	1,192	1,201	1,187	1,195	1,236	1,187	1,202	1,190	1,214	1,216	1,187	1,187
Cluster	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
Kleibergen-Paap F statistic	117.3	140.0	142.0	137.4	133.1	144.3	137.8	139.3	137.4	218.6	133.4	133.6	137.0	264.1	138.3

Notes: The dependent variable is the migrant skill mix in 2000, i.e.  $\ln \left( \frac{E_{sd}^H}{E_{sd}^L} \right) - \ln \left( \frac{E_s^H}{E_s^L} \right)$ .  $\Delta$  represents the simple difference between destination and source country, i.e.  $\Delta X = X_d - X_s$ . Robust standard errors in parentheses clustered at the destination country level. Significance levels: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.

Table 7: *Controlling for Migrant Selection in 1990*

	(1)	(2)	(3)
	OLS		IV
Skill mix, 1990	0.773*** (0.040)	0.716*** (0.046)	0.700*** (0.054)
$F_{ST}$ genetic distance	0.045 (0.054)	0.023 (0.046)	0.108 (0.094)
Language distance	0.059** (0.024)	0.036 (0.026)	0.024 (0.023)
<i>Controls:</i>			
$W_d^{80/20} - W_s^{80/20}$		0.018*** (0.004)	0.018*** (0.004)
Anglophone destination		0.384** (0.142)	0.400*** (0.130)
Schengen		0.023 (0.064)	0.026 (0.065)
Contiguous		-0.363*** (0.093)	-0.383*** (0.104)
Log distance		-0.089 (0.054)	-0.107** (0.044)
Colony		0.044 (0.105)	0.009 (0.100)
$\Delta$ Absolute latitude		-0.001 (0.001)	-0.001 (0.001)
$\Delta$ Absolute longitude		0.004** (0.002)	0.003 (0.002)
$R^2$	0.823	0.849	0.847
Observations	1,228	1,228	1,228
Cluster	15	15	15
Kleibergen-Paap F statistic			92.8

*Notes:* The dependent variable is the migrant skill mix in 2000, i.e.  $\ln(E_{sd}^H/E_{sd}^L) - \ln(E_s^H/E_s^L)$ .  $\Delta$  represents the simple difference between destination and source country, i.e.  $\Delta X = X_d - X_s$ . Robust standard errors in parentheses clustered at the destination country level. Significance levels: \*\*\* p<0.01, \*\* p<0.05, \* p<0.1.