Income Redistribution and Self-Selection of Immigrants

Preliminary draft! Please, do not cite without permission!

Giacomo Corneo1  Guido Neidhöfer2*

1 Freie Universität Berlin, CEPR London, CESifo Munich, IMK Düsseldorf, IZA Bonn

2 ZEW Mannheim

14th September 2018

Abstract: The Roy-Model of income maximization predicts that countries with a more progressive tax and transfer system attract low skilled immigrants, while low redistribution is associated with inflows of high skilled immigrants. We test the model using administrative data on almost all Italian citizens registered as living abroad. Our results confirm the predictions of the model, even accounting for individual and country level covariates, migration costs, and testing for stochastic dominance between the skill distributions of migrants and stayer. Finally, we run a discrete choice model that shows substantial effects of potential net income returns on the choice of the destination country. Our analysis also sheds light on the determinants of migration decisions and selection of immigrants.

Keywords: Roy-Model, Self-selection, Migration, Administrative Data

*Corresponding Author: Guido Neidhöfer, ZEW Centre for European Economic Research, L7 1, 68161 Mannheim, Germany. guido.neidhoefer@zew.de
1 Introduction

Theoretical and empirical studies on the selection process of migrants agree on one fundamental aspect: Migrants are not a random draw from the population of their home country neither are they undistinguishable in their observable and unobservable characteristics from the native population of their host country. Besides this basic consensus, different theories and findings on the patterns characterizing the self-selection of immigrants coexist.

For instance, Borjas (1987) applies the Roy-Model of self-selection and argues that the returns to human capital in source and destination country determine whether high or low skilled individuals migrate: the higher are the returns to human capital in a country, the more high skilled individuals will tend to migrate in this particular country, particularly from countries with lower returns to skills. This hypothesis has been empirically confirmed by Moraga (2010) and Parey et al. (2017), among others. On the other side, Chiswick (1999) argues in favour of a general positive selectivity of migrants in line with the predictions of standard human capital theory since Sjaastad (1962); another hypothesis that has as well been confirmed in empirical studies like Liebig and Sousa-Poza (2004) and Chiquiar and Hanson (2005).

The empirical measurement of returns to skills has also not been uniform in this branch of literature. Some studies focus on the earnings or income distribution through macro measures like the degree of inequality of market or disposable incomes, while others use ad-hoc measures for education or skill premia. For the first time, we approximate the relative level of returns to skills of a country by the progressiveness of its tax and transfer system. This choice has two main advantages: First, it gives a more accurate approximation of the net returns to migration, in the spirit of the Roy-Model. Second, it considers the role of public policy to influence returns to skills and, hence, in- and out-migration patterns.

Another crucial issue is the availability of suitable data sources to test the model. Most studies that tested the self-selection of immigrants, so far, rely on macro-data containing aggregate information about the characteristics of migrants, census data that allows to analyse flows and stocks
of migrants from one particular country to the other (mostly from Mexico to the US), or survey
data reporting pre-migration earnings or migration intentions. Further evaluations using novel data
sources seem therefore necessary to deepen our understanding about the process of immigrants’
self-selection.

The strength of our analysis is that we are able to test the selection process on an administrative
dataset that contains almost all Italian families living outside of Italy, one of the countries in the
world with the highest absolute number of emigrants. The data at our disposal encompasses 88
% of Italians registered abroad. Overall, our sample comprises more than four million Italian
citizens living in 13 foreign countries: Argentina, Australia, Belgium, Brazil, Canada, France,
Great-Britain, Germany, The Netherlands, New-Zealand, Switzerland, the US, and Venezuela. Of
these, about 1.3 million have an own migration experience (i.e. were born in Italy). To the best of
our knowledge, the only other study testing the selection process of migrants with administrative
data on almost the entire population of emigrants is Borjas et al. (2018) on Danish migration register
data.\(^1\)

Studies analysing the self-selection of migrants with micro-data mostly observed flows from
developing countries to developed countries, hence from typically poor and unequal to rich and
less unequal countries (as pointed out by Hatton, 2014), or as mentioned above from Denmark, a
country with a relatively flat income distribution, to the rest of the world. In contrast, we observe
migrants from a country with medium level of inequality and redistribution, Italy, who migrated
countries with less progressive tax systems like Argentina, Brazil, Venezuela, New Zealand and the
US, and to more progressive ones like Belgium, France, Germany and the Netherlands. Hence, an
important contribution of our study is that with this powerful data source at our disposal we are
able to test both sides of the relationship hypothesized by the Roy-Model: positive and negative
selection, namely high returns to human capital associated with the immigration of high skilled in-
dividuals and low returns to skills associated with inflows of low skilled immigrants. Furthermore,

---

\(^1\)The fifteen countries with the highest number of emigrants are Russia, Mexico, India, Bangladesh, Ukraine, China,
UK, Germany, Kazakhstan, Pakistan, Philippines, Italy, Turkey, Afghanistan, and Morocco. Denmark places 122th on
this ranking with about 240,000 emigrants (Global Migrant Origin Database v4, Migration DRC).
the data at our disposal enables us to measure selection carefully with respect to the sub-national region of origin, instead of the national average.

We complement our administrative dataset with Italian household survey data and measure selection on observable characteristics predicting the counterfactual labour earnings of immigrants have they stayed in Italy. Furthermore, we estimate the relative educational position of every migrant with respect to his or her reference group, i.e. non-migrants born in the same year living in the Italian region where the individual resided before moving to the foreign country. Selection on unobservable characteristics is measured by the probability to be unemployed or to have a high occupational position given the level of education.

All estimations confirm the predictions of the Roy-Model: lower redistribution is associated with higher levels of skill selectivity of Italian immigrants. This relationship holds controlling for individual characteristics of the migrants, country characteristics like GDP per capita and the unemployment rate, migration costs – approximated by the distance of the country of residence to the Italian border, the existence of migration agreements between the two countries and the share of migrants from the same Italian province residing in the same country as indicator of network effects –, and finally even country fixed effects. Furthermore, the skill distribution of Italian migrants in countries with relatively low levels of redistribution stochastically dominates the distributions of Italian migrants in countries with high levels of redistribution, and of the population of non-migrants in Italy.

Last, we use harmonized household survey data from the Luxembourg Income Study to estimate the net monetary returns to migration in the destination country, as well as the counterfactual potential returns in all other possible destination countries as well as in Italy. We run a discrete choice model of the decision to migrate as a function of the predicted net income gains in the country of destination, including demographic characteristics and place characteristics of the country of destination as controls. Also this exercise confirms the predictions of the Roy-Model of income maximization as driver of migration decisions.
2. Returns to Skills and Self-selection of Migrants

In economic models of human migration, the point of departure is usually given by human capital theory. Here, the migration decision of rational agents is a function of the expected lifetime utility gains of the move, net to the costs of migration (Sjaastad, 1962). Following this framework the act of moving from one place to another is nothing more and nothing less than an investment in human capital. Another important aspect is that not all individuals have the capabilities to estimate their costs and benefits of an eventual move, or the right incentives to make the investment. Hence, migrants are a self-selected sample among the populations of host and source countries.

The main dispute in the literature dealing with the self-selection of immigrants is the answer to the question whether immigrants have higher or lower observed and unobserved skills in comparison to non-migrants. The two alternative scenarios are known as positive and negative self-selection, meaning a higher or lower than average degree of skills among the immigrants, respectively. Borjas (1987) argues, applying the Roy-Model (Roy, 1951), that the decision to migrate to one country or another (or stay in the country of origin) depends on the comparative advantages of individuals to obtain the highest possible earnings for their particular level of skills (see also Borjas et al., 1992). Hence, countries with high returns to skills attract individuals at the top of the skill distribution from countries with lower returns to skills. Vice versa, individuals at the bottom of the skill distribution in their country of origin should have an incentive to move to more egalitarian countries, where the expected earnings for their given skill level are less far away from the earnings of high skilled individuals.
Formally, consider individuals living in country I deciding to migrate either to country G or to country U or staying in their home country. The potential log earnings \( w \) in country \( i \in \{I, G, U\} \) depend on the individual’s skill level \( s \), as well as the returns to skills \( \rho \) and the social minimum \( \omega \) in \( i \):

\[
w_i = \omega_i + \rho_i s
\]  

(1)

We assume that the degree of returns to skills is highest in U and lowest in G \( \rho_U > \rho_I > \rho_G \), while the reverse applies to the level of social minimum \( \omega_U < \omega_I < \omega_G \). Individuals will migrate if their expected net gains in the foreign country will exceed the net gains in their home country. Hence individuals will migrate from I to U if for their given skill level

\[
w_U - c_U > \max(w_I, w_G - c_G),
\]  

(2)

and to country G if

\[
w_G - c_G > \max(w_I, w_U - c_U),
\]  

(3)

where \( c_i \) are the costs of migration to G or U. Consequently, the threshold skill levels that define whether an individual migrates or not and to which country, are

\[
s_1 = \frac{\omega_U - c_U - \omega_I}{\rho_I - \rho_U},
\]  

(4)

\[
s_2 = \frac{\omega_G - c_G - \omega_I}{\rho_I - \rho_G}.
\]  

(5)

Figure 1 shows this relationship graphically for the case of migration to both foreign countries. Returns to skills are lower (higher) in G (in U) than in I, hence individuals migrate from I to G (to U) if their skill level is below (above) the threshold \( s_1 \) \( (s_2) \). For individuals with skill levels in between the two thresholds \( s_1 \) and \( s_2 \) there is no incentive to migrate. Hence, there is a situation of negative self-selection from country I to G and positive self-selection from I to U.
Furthermore, as it is easy to see, if the relative returns to skills between the source and the host country change, this affects the average degree of migrants’ selection. For instance, if the relative returns between U and I become lower – for example as consequence of a marginal rise of $\rho_U$ with constant $\rho_I$ – the threshold $s_2$ shifts to the left, i.e. the last stayer has a lower skill level than before.\footnote{The effect of highly skilled individuals leaving their country of origin and consequently lowering the stock of skills is known as brain drain in the migration literature. Some models show that the brain drain might also contribute to human capital formation in the sending countries if the higher returns to education that cause the brain drain also constitute an incentive for people with migration intentions to invest in education (Beine et al., 2008).} This shift has no effect on the general pattern of positive self-selection of immigrants from I to U, but since the average skill level in the source country is endogenously determined by the migration process, it has an effect on the average degree of self-selection. The population of stayer has a lower average degree of skills and, under regularity assumption on the distribution of skills (e.g. log-normal distribution), the average skill level of the immigrants in U rises in comparison to the stayers. The same applies to the average degree of selection of immigrants in G with rising $\rho_G$, as...
long as $C_G < \omega_G - \omega_f$. In contrast, when $\rho_G$ declines, $s_1$ shifts to the right: the average skill level of the immigrants in $G$ becomes lower in comparison to the stayer.

The implications of this theoretical framework were firstly confirmed by empirical analyses of international and internal migrants in the US: Borjas (1987) shows that the degree of income inequality in the home country, a proxy measure for the returns to skills, is a predictor for the type of selectivity of migrants, while Borjas et al. (1992) show that interstate variations in the returns to skills affect the skill structure of migration inflows. In contrast, Chiswick (1999) states that immigrants tend always to be favourably self-selected also in presence of higher levels of income inequality and Borjas (1987)’s empirical results only show that income inequality attenuates the degree of selection, but not the generally positive selection pattern. Liebig and Sousa-Poza (2004) confirm this hypothesis applying an empirical analysis using cross-country survey data on migration intentions. Other empirical papers testing the selection of migrants from one or more source countries to one or more destination countries, with individual level and aggregated data, obtained contrasting results; recent studies that review the empirical literature and also test this theory are Parey et al. (2017) and Patt et al. (2017).

One common explanation for the differing results so far has been argued to lie in the important role of migration costs for the migration decision and the selection process. For instance, Chiquiar and Hanson (2005) show that if migration costs are negatively correlated with skills, both positive and negative self-selection into countries with higher returns are possible solutions. Apart from transportation costs and the value of friends, family, and culture left behind, other factors are determinant for the costs of migration; for instance, immigration policies and migration networks (Hatton, 2014). McKenzie and Rapoport (2010) show indeed that self-selection is more likely to be positive in places with low migration networks and more likely to be negative in places where many migrants from the same origin countries reside.

Another important issue in the measurement of self-selection is the associated reference population. Spitzer and Zimran (2018) analyse the self-selection in stature of Italians that migrated to the US between 1907 and 1925. Their findings show that Italian immigrants were negatively selec-
ted with respect to the Italian national height average, but positively selected among their province of origin. This highlights the importance of evaluating selection patterns with respect to the correct reference group, for instance at the sub-national level.

3 Data & Measurement

3.1 Administrative Data on Italians Abroad

The micro-data basis of our empirical tests is the Registry of Italians resident abroad (Anagrafe degli italiani residenti all’estero, AIRE), an administrative registry dataset provided us by the Italian Ministry of Foreign Affairs through the Italian Embassy in Germany. All Italians who are at least one year abroad or born outside of Italy are required to register to the AIRE by law. Furthermore, some bureaucratic tasks, e.g. renewing an Italian passport or ID card, and voting, are only possible being recorded in the AIRE. Furthermore, for Italians residing abroad the registration in the AIRE is a necessary condition to avoid to pay the income tax in Italy.

The dataset at our disposal contains individual information on 4,079,646 registered Italian citizens in 13 different foreign countries between 2014 and 2015 as well as information on their spouses and children without Italian citizenship. Our data encompasses approximately 88% of all Italian citizens registered in the AIRE worldwide. The number of registered people in the AIRE data at our disposal.

Table 1 shows the number of registered people

For our analysis, we are only interested in individuals with own migration experience. Hence we focus only on those that were born in Italy and exclude the foreign-born children and grandchildren of immigrants (so-called second and third generation immigrants) from our final sample. Furthermore, since we want to capture the selection mechanism, we assure that individuals in our sample had already finished their educational career when they moved to the foreign country. We

---

3Only Italian civil servants working abroad, for instance at embassies or consulates, diplomats, and Italian military in service at NATO facilities located abroad are not required to register to the AIRE.

4The number of Italians registered in AIRE on January 1, 2015, is 4,636,647 (Fondazione Migrantes, 2015). The only country missing in our sample of the ten countries with the highest concentration of Italian immigrants worldwide in 2015 is Spain. A comparison of the number of Italian migrants in AIRE and the International Migration Database of the OECD is included in the Supplemental Material.
Table 1: Number of registered individuals in the registry of Italians resident abroad (AIRE)

<table>
<thead>
<tr>
<th>Country</th>
<th>Total</th>
<th>Italian citizenship</th>
<th>Born in Italy</th>
<th>Born 1940-1985</th>
<th>Arrival 1960-2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>1,191,059</td>
<td>893,974</td>
<td>119,008</td>
<td>50,044</td>
<td>7,271</td>
</tr>
<tr>
<td>AUS</td>
<td>221,292</td>
<td>149,246</td>
<td>53,900</td>
<td>35,277</td>
<td>18,444</td>
</tr>
<tr>
<td>BEL</td>
<td>333,235</td>
<td>273,415</td>
<td>95,438</td>
<td>67,160</td>
<td>34,249</td>
</tr>
<tr>
<td>BRA</td>
<td>597,232</td>
<td>450,939</td>
<td>36,804</td>
<td>19,334</td>
<td>7,454</td>
</tr>
<tr>
<td>CAN</td>
<td>188,289</td>
<td>137,289</td>
<td>72,909</td>
<td>43,645</td>
<td>24,318</td>
</tr>
<tr>
<td>CH</td>
<td>695,081</td>
<td>607,084</td>
<td>220,133</td>
<td>168,060</td>
<td>117,636</td>
</tr>
<tr>
<td>FRA</td>
<td>214,512</td>
<td>170,023</td>
<td>70,736</td>
<td>45,468</td>
<td>23,785</td>
</tr>
<tr>
<td>GBR</td>
<td>308,077</td>
<td>263,916</td>
<td>130,100</td>
<td>87,127</td>
<td>60,511</td>
</tr>
<tr>
<td>GER</td>
<td>813,254</td>
<td>694,694</td>
<td>300,863</td>
<td>258,315</td>
<td>165,929</td>
</tr>
<tr>
<td>NLD</td>
<td>48,895</td>
<td>41,346</td>
<td>16,767</td>
<td>12,782</td>
<td>4,696</td>
</tr>
<tr>
<td>NZL</td>
<td>5,056</td>
<td>4,052</td>
<td>1,497</td>
<td>1,100</td>
<td>757</td>
</tr>
<tr>
<td>USA</td>
<td>334,093</td>
<td>250,176</td>
<td>133,498</td>
<td>97,212</td>
<td>67,086</td>
</tr>
<tr>
<td>VEN</td>
<td>218,351</td>
<td>143,492</td>
<td>28,801</td>
<td>14,097</td>
<td>4,972</td>
</tr>
<tr>
<td>Total</td>
<td>5,168,426</td>
<td>4,079,646</td>
<td>1,209,454</td>
<td>899,621</td>
<td>537,108</td>
</tr>
</tbody>
</table>

Notes: Subsequent columns show the respective subtotal.

do so excluding all individuals that registered to AIRE when being younger than 20 years old. Finally, to avoid bias deriving from, first, individuals who did not finished their educational career, and, second, differential mortality rates between people with different educational levels, we restrict our sample to the age range 30 to 75, i.e. to the cohorts 1940 to 1985. The individual information contained in the registry data are: Date of birth, date of arrival in the host country, sex, place of birth, place of residence, education, profession and the last municipality of residence in Italy before migration. Due to the availability of cross-country macro data from other sources, we further restrict the sample to people that migrated after 1960. Table 2 shows descriptive statistics of individual and country characteristics from the final sample.

3.2 Relative Skills and Predicted Net Returns

3.2.1 Counterfactual Earnings

To verify the level of individual selection of movers (people registered to AIRE) relative to the population of the stayers (Italians residing in Italy), we complement this dataset with an established
Table 2: Descriptive Statistics of the Sample – Country Averages

<table>
<thead>
<tr>
<th>Region</th>
<th>Year of birth</th>
<th>Year of arrival</th>
<th>Share of female</th>
<th>Years of education</th>
<th>Rural origin</th>
<th>Internal migrant</th>
<th>GDP p.c.</th>
<th>GDP growth</th>
<th>Unemployment rate</th>
<th>Distance (in km)</th>
<th>Bilateral agreement</th>
<th>Policy toward high skilled</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG</td>
<td>1948.48</td>
<td>1994.83</td>
<td>0.53</td>
<td>6.79</td>
<td>0.45</td>
<td>0.21</td>
<td>0.11</td>
<td>6.53</td>
<td>4.66</td>
<td>6.41</td>
<td>20.200</td>
<td>1</td>
</tr>
<tr>
<td>AUS</td>
<td>1958.68</td>
<td>1988.86</td>
<td>0.40</td>
<td>9.79</td>
<td>0.53</td>
<td>0.00</td>
<td>0.13</td>
<td>24.19</td>
<td>5.70</td>
<td>5.18</td>
<td>22.000</td>
<td>1</td>
</tr>
<tr>
<td>BEL</td>
<td>1960.24</td>
<td>1992.26</td>
<td>0.42</td>
<td>10.05</td>
<td>0.26</td>
<td>0.06</td>
<td>0.15</td>
<td>24.13</td>
<td>2.47</td>
<td>7.57</td>
<td>0.550</td>
<td>0</td>
</tr>
<tr>
<td>BRA</td>
<td>1981.08</td>
<td>2001.67</td>
<td>0.20</td>
<td>12.68</td>
<td>0.15</td>
<td>0.01</td>
<td>0.21</td>
<td>7.07</td>
<td>3.23</td>
<td>8.21</td>
<td>7.000</td>
<td>1</td>
</tr>
<tr>
<td>CAN</td>
<td>1972.83</td>
<td>1992.98</td>
<td>0.45</td>
<td>8.74</td>
<td>0.50</td>
<td>0.02</td>
<td>0.09</td>
<td>14.75</td>
<td>3.08</td>
<td>6.36</td>
<td>3.000</td>
<td>1</td>
</tr>
<tr>
<td>CH</td>
<td>1981.66</td>
<td>1994.45</td>
<td>0.79</td>
<td>10.28</td>
<td>0.32</td>
<td>0.01</td>
<td>0.16</td>
<td>49.65</td>
<td>1.71</td>
<td>7.05</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>ELT</td>
<td>1954.65</td>
<td>1990.93</td>
<td>0.46</td>
<td>17.27</td>
<td>0.17</td>
<td>0.32</td>
<td>0.18</td>
<td>30.62</td>
<td>1.80</td>
<td>9.18</td>
<td>0.000</td>
<td>0</td>
</tr>
<tr>
<td>GBR</td>
<td>1973.28</td>
<td>2004.40</td>
<td>0.42</td>
<td>13.17</td>
<td>0.17</td>
<td>0.06</td>
<td>0.17</td>
<td>35.29</td>
<td>2.18</td>
<td>6.68</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>GBR</td>
<td>1954.65</td>
<td>1996.00</td>
<td>0.57</td>
<td>8.42</td>
<td>0.31</td>
<td>0.04</td>
<td>0.11</td>
<td>24.13</td>
<td>2.00</td>
<td>5.35</td>
<td>0.070</td>
<td>1</td>
</tr>
<tr>
<td>NLD</td>
<td>1973.91</td>
<td>2007.96</td>
<td>0.37</td>
<td>14.45</td>
<td>0.14</td>
<td>0.28</td>
<td>0.19</td>
<td>46.05</td>
<td>1.10</td>
<td>7.46</td>
<td>0.640</td>
<td>1</td>
</tr>
<tr>
<td>NZL</td>
<td>1980.64</td>
<td>2008.08</td>
<td>0.42</td>
<td>17.72</td>
<td>0.15</td>
<td>0.14</td>
<td>0.18</td>
<td>30.39</td>
<td>2.57</td>
<td>5.24</td>
<td>18.000</td>
<td>0</td>
</tr>
<tr>
<td>USA</td>
<td>1960.65</td>
<td>1967.88</td>
<td>0.45</td>
<td>11.58</td>
<td>0.20</td>
<td>0.01</td>
<td>0.14</td>
<td>30.13</td>
<td>2.67</td>
<td>6.12</td>
<td>6.100</td>
<td>0</td>
</tr>
<tr>
<td>VEN</td>
<td>1960.67</td>
<td>1963.97</td>
<td>0.40</td>
<td>19.36</td>
<td>0.24</td>
<td>0.47</td>
<td>0.16</td>
<td>3.88</td>
<td>3.52</td>
<td>8.49</td>
<td>7.700</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>1962.10</td>
<td>1994.36</td>
<td>0.40</td>
<td>10.55</td>
<td>0.27</td>
<td>0.11</td>
<td>0.14</td>
<td>30.94</td>
<td>2.54</td>
<td>5.81</td>
<td>1.952</td>
<td>1</td>
</tr>
</tbody>
</table>

Sources: Individual characteristics from AIRE. Sample is restricted to people born in Italy, 30-64 years old, and who registered in AIRE after the age of 20. Unemployment rate, GDP growth, GDP per capita from World Bank Data. Distance to Italy measured from border to border (Google Maps). Migration agreement is equal to one if there is/have been bilateral migration agreements between Italy and the country of destination; information retrieved from different sources. Policies oriented towards high skilled is equal to one if the policies of the country in the last 50 years have been more oriented at attracting high skilled immigrants or disincentive low skilled immigrants; information retrieved from the DEMIG Policy Database.

Italian household survey collected by the Bank of Italy: the Survey on Household Income and Wealth (SHIW). The SHIW collects since 1960 information on Italian families including individual characteristics for each single household member. For the present study, we use the comparable survey waves 1977 to 2014, normalizing the sampling weights for each single year if more than one single wave is used for the analysis.

We use this data on Italians living in Italy to measure the relative skills of movers, running an augmented Mincer regression of log labour earnings on the survey sample of stayer in 2014 and using the coefficients of the regression to predict the counterfactual log labour earnings that mover would have obtained in Italy. The variables included in the regression are: years of schooling, age, quadratic age, sex, Italian region of origin, Italian region of birth and an indicator on whether the individual is an internal migrant (i.e. does not live in the same region or did not migrate from the same region where he or she was born). For this exercise, we exclude individuals older than 65. Table 3 shows the OLS estimates in column (1).

To account for the selection of mover in the earnings equation that could bias OLS estimates we apply the Heckman selection procedure in two stages. In the first stage, the population weighted probability to emigrate is estimated on the pooled sample of stayer and mover, i.e. pooling SHIW and AIRE data.\(^5\) Hereby, the number of emigrants born in the same year and in the same Italian

\[^5\]For SHIW we use the data design weights. For AIRE we compute weights that counterbalance the observations with missing information on educational attainment.
### Table 3: Augmented Mincer regressions to predict log labour earnings

<table>
<thead>
<tr>
<th>Sample</th>
<th>Dependent variable</th>
<th>(1) OLS log Labour Earnings</th>
<th>(2) Heckman log Labour Earnings</th>
<th>(3) Selection equation Stayer (0/1)</th>
</tr>
</thead>
</table>
| Italians in Italy (SHIW) | Female | -0.374***  
|                 | (0.0161) | -0.372***  
|                 | (0.0162) | 0.0523**  
|                 | (0.0224) | 0.296***  
| Italians worldwide (AIRE+SHIW) | Age | 0.0506***  
|                 | (0.00933) | 0.0622***  
|                 | (0.0147) | (0.00860)  
|                 | (0.0224) | 0.296***  
|                 | (0.00352) | -0.00330***  
|                 | (0.0000878) | 0.00352  
|                 | (0.0000878) | -0.00704**  
|                  | Age × Age | -0.000399***  
|                 | (0.000101) | -0.000530***  
|                 | (0.00163) | -0.00330***  
|                 | (0.0000878) | 0.00352  
|                 | (0.00352) | -0.00704**  
|                 | (0.0000878) | -0.00330***  
|                  | Years of education | 0.0624***  
|                 | (0.00243) | 0.0621***  
|                 | (0.00244) | (0.00352)  
|                 | (0.00352) | -0.209***  
|                 | (0.0401) | 0.00352  
|                 | (0.0401) | -0.00704**  
|                  | Internal migrant | -0.00622  
|                 | (0.0300) | -0.0137  
|                 | (0.0309) | (0.0401)  
|                 | (0.0401) | -0.00704**  
|                  | Inverse Mills Ratio | 0.319  
|                 | (0.313) | 0.00352  
|                  | Number of emigrants born in the same year and region | -0.000490***  
|                 | (0.0000481) | -0.000490***  
|                  | _cons | 7.769***  
|                 | (0.219) | 7.511***  
|                 | (0.335) | (0.214)  
|                  | Region of origin and birth controls | Yes  
|                 | Yes | Yes  
|                  | Observations | 3965  
|                 | 3965 | 308518  
|                  | R² | 0.276  
|                 | 0.276 | 0.276  

Notes: Weighted regressions using survey design waves from SHIW and constructed population weights for AIRE. Statistical significance level * 0.1 ** 0.05 *** 0.01. Source: SHIW and AIRE, own estimations.

The assumption is that this amount captures the predictive power of network effects and diasporas on individual migration decisions without exerting a direct effect on earnings in 2014. Then, the inverse Mills Ratio for each observation, estimated in the first stage, is included in the earnings regression as further control variable to obtain unbiased estimates. Table 3 shows the coefficients of this application in column (2) and the first-stage estimates in column (3).

The estimates in Table 3 show that the coefficients change only slightly and the coefficient of the Mills ratio is not significantly different from zero. This means that the observable characteristics included in the regression account for selection properly and we can safely adopt the predictions from the OLS estimates. All the applications using the counterfactual labour earnings estimated with the Heckman procedure does not yield significantly different results.

---

6To approximate the total amount of emigrants for each birth cohort and Italian region of birth we use the AIRE data for the countries at our disposal, that cover almost 90 percent of all Italians registered worldwide.
3. DATA & MEASUREMENT

3.2.2 Relative Educational Position

Our estimates of counterfactual earnings do not correct for the effect of out-migration flows on the Italian earnings distribution. However, it would be impossible to estimate the counterfactual wage distribution in Italy abstracting from the distorting effects of migration without posing arbitrary assumptions. Since in our analysis counterfactual earnings are just used as a measure for the relative skills of immigrants, we can safely keep our estimates deriving from the actual Italian earnings distribution in 2014. As further sensitivity analysis, we estimate the relative educational position with respect to the reference group and use it as an additional measure for the individual degree of self-selection.

The relative educational position of individual $i$ born in year $b$ is defined as the relative difference of his or her years of education $y$ with respect to the average of stayers born in the same year residing in his or her Italian region of origin $j$:

$$s_{ijb}^e = \frac{(y_{ijb} - \bar{y}_{jb})}{\bar{y}_{jb}}.$$  \hfill (6)

The average of stayers of each birth cohort from 1940 to 1985 in the 20 Italian regions is computed using the Italian household survey SHIW. Most results applying this specification of relative skills are included in the Supplemental Material.

3.2.3 Predicted Net Returns to Migration

We use harmonized microdata from the Luxembourg Income Study (LIS) to estimate the net incomes in the country of residence for mover, their counterfactual net incomes in Italy, and in every other of the possible destination countries. Furthermore, we estimate as well the net incomes of stayer in Italy, and their counterfactual incomes in all possible destinations. To measure this, we use the survey samples of every single destination country around the year 2014. Unfortunately, LIS data is not available for Argentina, New Zealand, and Venezuela. Furthermore, the last available survey for Belgium dates back to the year 2000. Therefore, the parts of the analysis that analyse
predicted net income returns are restricted to the remaining nine possible destination countries and Italy.

We again estimate an augmented Mincer regression on disposable household income including the variables sex, age, quadratic age, education, and an indicator on whether at least one child lives in the household. Then, we predict the disposable incomes (in international US Dollars applying Purchasing Power Parity) of Italian immigrants in all possible destinations and in Italy using the coefficients of this regression. Returns to migration are then defined as the difference between predicted incomes in the destination country and the counterfactual income in Italy.\footnote{In most country-surveys of the LIS a country of origin variable is not available. In the few where it is available the number of observations for (first generation) Italian immigrants is too low to provide consistent estimates. In principle, in most countries it would be possible to restrict the sample just to immigrants. However, without further specification, the ample heterogeneity within the group of immigrants leads to a very dispersed income distribution among this group in every country. Hence, we do not restrict the country-samples and assume that the disposable incomes of Italian immigrants in the destination countries are not substantially different from the incomes of natives and other immigrants.}

For our main application, we use total household income net of taxes and transfers. To account for cross-country differences in assortative mating and household compositions we also equivalize the incomes applying the square root scale. Results using the latter are included in the Supplemental Material.

### 3.3 Country characteristics

Finally, we collect data on country characteristics from different sources: for instance, the Standardized World Income Inequality Database (SWIID), the World Income Inequality Database (WIID), World Bank Macro Data (WB-Data), and the Andrew Young School World Tax Indicators (WTI). From the SWIID we retrieve net and market income inequality indices and compute measures for absolute and relative redistribution, the first measured by the difference between the market and net Gini index, the second by this difference divided by the market Gini (see Reynolds and Smolensky, 1977; Solt, 2016). The level of redistribution is our primary measure for the returns to skills and therefore our main variables of interest.
Figure 2: Levels of Inequality Pre- and Post-Redistribution

Source: SWIID, own calculations. Average over all available years.

Figure 2 shows the average level of pre and post-redistribution inequality for the countries in our sample, and for Italy. The distance from the 45 degree line shows the contribution (in Gini points) of taxes and transfers to the reduction of market income inequality. Figure 3 ranks the countries by their average level of redistribution over all observed years. On the x-axis redistribution is defined in absolute terms while on the y-axis in relative terms as defined above. The country ranking is rather similar applying both measures and depicts basically three main groups of countries with high, medium and low levels of redistribution. The first group comprises Belgium, the Netherlands, Germany and France. The second, besides Italy, also Great-Britain, Australia, Switzerland, Canada and the US. The third, New-Zealand, Brazil, Venezuela and Argentina. Figure 4 shows the time trends in the evolution of net income inequality with respect to Italy. As is evident, the countries in our sample are quite heterogeneous in their levels of inequality and redistribution, in particular with respect to Italy, and with an interesting amount of variation over time.
To cross-check the results obtained using SWIID we perform the same analysis using WIID and obtain no significant differences in results. The latter is a dataset with sensibly less observations for each country than the former, but has been argued in past to rely on a more consistent methodology and data basis than SWIID (see Jenkins, 2015). Anyway, for the countries in our sample, including Italy, we find a very high degree of congruence between SWIID and WIID data with a correlation of about 0.95.

From different sources we retrieve other country characteristics which could also affect the type and level of selection of immigrants and hence act as control variables in the regressions; see the notes of Table 2. Last, we perform the analysis also with the Marginal Tax Rate Progression retrieved from WTI data as further proxy for returns to skills.
Figure 4: Redistribution trends with respect to Italy

Source: SWIID, own calculations.
4 Descriptive Evidence

4.1 Selection on Educational Attainment

The comparison of the education of Italian immigrants in different countries already gives some important hints about differential self-selection patterns. Figure 5 shows the average years of education of individuals in our administrative data on Italians abroad (Mover), compared to the average of Italians living in Italy estimated from Italian household survey data (Stayer). On average male and female Italians born from 1940 to 1961 and living abroad in 2015 have lower educational attainment than the average of the Italian population of people born in the same years. After this year, subsequent cohorts of migrants have a relatively higher average education by about one year of education. It shows up that especially the share of individuals with a completed tertiary education degree experienced a dramatic rise. Furthermore, we observe a similar development in gender differences among movers and stayers: older men have higher education than women, while younger women are more educated than men. Figure 6 shows the differences in average educational attainment by Italian macro-region of origin over time. Mover from the Centre and North of Italy have constantly a higher education than stayer, while mover from South Italy and the Islands have lower or rather similar average education than stayer. These findings highlight the crucial importance to evaluate the degree of selection of people with respect to their region of origin rather than with respect to the national average.

The pattern that emerges evaluating the selection of Italian immigrants by their place of origin is very conspicuous. Interestingly, the Italian North-South divide is mirrored and even enhanced by Italians living abroad. Figure 6 shows a map of Italy that displays the average relative educational position of Italians registered in AIRE by their province of origin. The relative educational position of each migrant is estimated, as explained in Section 3.2.2, on the reference group of Italians living in Italy in the same region of origin and born in the same year. Values higher than zero stand

\[^8\]Years of schooling are coded following this scheme: No school degree, 0 years. Uncompleted compulsory schooling, 5 years. Compulsory schooling, 8 years. Beyond compulsory education, 13 years. Tertiary degree, 16 years.
Figure 5: Average Years of Education of Italian Mover and Stayer by Year of Birth. Italians of all regional origins and subdivided by gender.

Source: Averages for Mover are own calculations using AIRE, averages for Stayer are own calculations using SHIW.

for positive selection on education, i.e. the individual is more educated than the average of his or her reference group, while values lower than zero for negative selection. Not only that people in Southern Italy have on average lower education than people in the North, and that this pattern is also observed for Italians abroad. Italians abroad original from South Italy are also negatively selected among the population of their region of origin.

Figure 8 shows different patterns by year of arrival and country of destination, highlighting some interesting contrasts. Italians living in countries like Brazil, New-Zealand, and Venezuela are almost constantly more educated than their peers living in Italy. In contrast, the average education of Italian immigrants in Germany is persistently lower than the average education of stayer.

4.2 Income Redistribution and Self-Selection of Immigrants

Figure 9 show the result of a first, stylized cross-country analysis on the relationship between redistribution and the degree of self-selection of immigrants. The average amount of relative redis-
4. DESCRIPTIVE EVIDENCE

Figure 6: Average Years of Education of Italian Mover and Stayer by Year of Birth and Geo-graphic Macro-Region of Origin.

Source: Averages for Mover are own calculations using AIRE, averages for Stayer are own calculations using SHIW.

tribution over all available years in SWIID is plotted in relation to the average degree of selection of Italian immigrants residing in that particular country. In the upper part of the figure, selection is measured by the relative educational position, in the lower part by the predicted counterfactual log earnings.

Both measures show the same pattern of association. The population weighted correlations are -0.38 and -0.65 respectively. Hence, this stylized analysis shows the first evidence in favour of the Roy-Model: Countries with less progressive tax and transfer system, and hence higher returns to skills, attract immigrants from the upper part of the skill distribution.
Figure 7: Relative educational position of Italian Mover by their province of origin.

Notes: Map shows the province level average of relative educational position of all individuals in our AIRE sample. The individual relative educational position is estimated with respect to the population of stayer born in the same year and residing in the region of origin. Averages for Stayer are own calculations using SHIW.
Figure 8: Average Relative Educational Position of Italian Mover by Year of Arrival and Country of Residence

Notes: Relative educational position computed with respect to stayer born in the same year and resident in the region of origin. A value of zero is equivalent to the average of the reference group, values lower than zero show a negative selection, and values higher than zero a positive selection, on average. Source: AIRE, own estimates. Regional averages for every birth cohort are estimated using SHIW.
Figure 9: Returns to skills and self-selection of immigrants

Notes: All variables are averages over the complete observation period. Sources: AIRE, SWIID, own calculations.
5 Empirical Set-Up and Results

5.1 Redistribution and degree of selection

To measure the association between the level of redistribution as a proxy measure for the returns to skills and the self-selection of immigrants, we run the following linear regression:

\[ s_{ijtc}^* = \alpha + \beta T_{tc} + \delta Z_{tc} + \gamma X_{ijtc} + \zeta M_c + \lambda_j + \tau_t + \varphi_c + \epsilon_{ijtc} \]  

(7)

\( s^* \) are the relative skills measured by the predicted counterfactual log labour earnings that individual \( i \) from the Italian region of origin \( j \) who registered in year \( t \) to the registry of Italians living abroad in country \( c \) would have obtained had he stayed in Italy.\(^9\) \( T \) are the returns to skills in country \( c \) in year \( t \). The coefficient \( \beta \) shows whether returns to skills are associated with the relative skill level of immigrants’. In our main specification, we measure returns to skills by the level of relative redistribution. To smooth the time variation and avoid bias resulting from temporary shocks and measurement error in the inequality measures, the value associated to each year \( t \) is the average from \( t \) to the last year with available information (mostly 2015). The assumption behind this procedure is that rational agents take future developments into account when choosing to stay in the foreign country of residence.\(^10\)

The same procedure is applied to the other macroeconomic variables included in \( Z \). \( Z \) is a vector of controls for country characteristics that vary by \( t \): i) unemployment rate, ii) GDP-growth, iii) GDP per capita. These variables are indicators that may shape the expectations of individuals willing to migrate about the overall conditions in a particular country. \( X \) is a vector of controls for individual characteristics that have not been used for the prediction of the counterfactual income: year of birth (polynomial of second degree), month of birth, year of arrival of the individual and the first immigrated member of the same household, a dummy indicating whether the individual was the first household member to arrive in the country of destination, rural or urban location of origin

\(^9\)Applications measuring \( s^* \) by the relative educational position are included in the Supplemental Material.

\(^{10}\)Results associating the macroeconomic variables only in the year of arrival show the same patterns and do not differ significantly.
(definition: < 150 inhabitants/km²), and a dummy variable indicating if the individual lives in the capital of the host country. \( M \) is a vector of controls for non-varying country characteristics that act as proxies for the costs of migrating to a particular country: i) distance from the Italian border, ii) a dummy signalizing whether this country and Italy signed an immigration agreement and iii) the share of migrants from the same Italian province residing in that particular country. \( \lambda_j, \tau_t \) and \( \varphi_c \) are fixed effects for the Italian region of origin, the year of arrival and country of destination. We restrict the coefficients of the control variables to be zero in some of the estimations. Standard errors are clustered at the country-year level.

Table 4 shows the results of estimating equation 7 on the sample of Italian immigrants worldwide. The dependent variable are the predicted counterfactual log labour earnings in Italy as measure of skills. The coefficient of the variable that indicates the relative level of redistribution in the country of destination is negative and highly significant in all specification of the model. For instance, our most conservative estimates obtained including country fixed effects show that an increase of relative redistribution by 10 percentage points is associated to a 5 percent decrease in the degree of self-selection in education of Italian immigrants. Hence, returns to skills, measured here by the amount of redistribution, seems to be associated with the degree of immigrants’ selectivity, as well as their likelihood to be positively self-selected from their population of origin.

Two variables capturing so called network effects show the expected negative relationship with the selectivity of immigrants. Past research argued that networks lower the cost of migration making it more attractive for low skilled individuals to migrate, and hence lowering, on average, the pattern of positive self-selection of particular immigrant groups (e.g. McKenzie and Rapoport, 2007; McKenzie and Rapoport, 2010). Our results confirm this findings. The presence of people from the same province of origin is negatively and significantly associated with the skill level of migrants. Moreover, the presence of a family member in the country of residence is associated with around 20 percent lower degree of skill selection. We observe furthermore that Italian immigrants originating from rural places have a lower degree of selection, while those residing in the capitals of their country of destination are more likely to be positively selected.
### 5. empirical set-up and results

Table 4: Redistribution and degree of selection

<table>
<thead>
<tr>
<th>Variable</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Redistribution</td>
<td>-0.861***</td>
<td>-0.580***</td>
<td>-0.682***</td>
<td>-0.549*</td>
</tr>
<tr>
<td></td>
<td>(0.0866)</td>
<td>(0.0657)</td>
<td>(0.0716)</td>
<td>(0.316)</td>
</tr>
<tr>
<td>Share of people from the same province in country of residence</td>
<td>-0.339***</td>
<td>-0.328***</td>
<td>-0.245***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0412)</td>
<td>(0.0407)</td>
<td>(0.0459)</td>
<td></td>
</tr>
<tr>
<td>Other family members migrated earlier (0/1)</td>
<td>-0.187***</td>
<td>-0.185***</td>
<td>-0.181***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00791)</td>
<td>(0.00776)</td>
<td>(0.00774)</td>
<td></td>
</tr>
<tr>
<td>Rural place of origin (0/1)</td>
<td>-0.128***</td>
<td>-0.127***</td>
<td>-0.122***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00506)</td>
<td>(0.00496)</td>
<td>(0.00481)</td>
<td></td>
</tr>
<tr>
<td>Resident in the capital (0/1)</td>
<td>0.0938***</td>
<td>0.0907***</td>
<td>0.0741***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00552)</td>
<td>(0.00559)</td>
<td>(0.00489)</td>
<td></td>
</tr>
<tr>
<td>Distance of country of residence from Italian border (in 1000 km)</td>
<td>0.00148*</td>
<td>-0.000795</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000796)</td>
<td>(0.00101)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Migration agreement between Italy and country of residence (0/1)</td>
<td>0.00198</td>
<td>0.0118*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00598)</td>
<td>(0.00655)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Migration policies oriented towards high skilled (0/1)</td>
<td>0.0158*</td>
<td>0.0174*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00857)</td>
<td>(0.00999)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>0.0123***</td>
<td>-0.00349</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00215)</td>
<td>(0.00279)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDP growth</td>
<td>0.0181***</td>
<td>0.0177***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00630)</td>
<td>(0.00449)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDP per capita</td>
<td>0.000623***</td>
<td>0.00457***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000209)</td>
<td>(0.000597)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>9.962***</td>
<td>-2059.8***</td>
<td>-2072.5***</td>
<td>-2046.4***</td>
</tr>
<tr>
<td></td>
<td>(0.0288)</td>
<td>(86.78)</td>
<td>(87.87)</td>
<td>(86.38)</td>
</tr>
<tr>
<td>Demographic controls</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Country F.E.</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>218236</td>
<td>217571</td>
<td>217571</td>
<td>218236</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.038</td>
<td>0.250</td>
<td>0.252</td>
<td>0.259</td>
</tr>
</tbody>
</table>

Notes: Dependent variable in the regressions is the predicted counterfactual log labour earnings in Italy as individual measure of relative skills. Demographic controls include year of birth (polynomial of second degree), month of birth, year of arrival of the individual and the first immigrated member of the same household, a dummy indicating whether the individual was the first household member to arrive in the country of destination, rural or urban location of origin (definition: < 150 inhabitants/km2), and a dummy variable indicating if the individual lives in the capital of the host country. Standard errors clustered at the country-year level. Statistical significance level * 0.1 ** 0.05 *** 0.01. Source: AIRE, own estimations.
Among country characteristics, the distance of the country of destination from Italy is very weakly associated with selectivity. The same applies to the presence of bilateral migration agreements between Italy and the country of destination. Last, unemployment, growth and GDP per capita in the country of destination are positively associated with the degree of self-selection of Italian immigrants.

5.2 Stochastic Dominance

The regression coefficients of relative redistribution shown in Table 4 measure the average association between returns to skills and the self-selection of immigrants. However, as shown by Borjas et al. (2018) the Roy-Model of self selection also implies a first order stochastic dominance relationship between the distributions of mover and stayer. We test these predictions using our administrative data on Italian immigrants and the survey data on Italians in Italy.

As argued for instance by Chiquiar and Hanson (2005), the complete group of stayer might not be a suitable comparison group for mover because of the selection on unobserved characteristics that shapes the income distributions as well. Hence, we subdivide the group of stayers in two separate groups: i) internal migrants, i.e. individuals that migrated within Italy between regions, and ii) non-migrants, i.e. individuals that reside in their region of birth. Mover, i.e. individuals in our administrative data, are also divided in two groups according to the amount of redistribution they experienced in the country of destination in comparison to the redistribution in Italy: i) Italian migrants in countries with higher redistribution, and ii) Italian migrants in countries with lower redistribution. Figure 10 plots the cumulative distributions of the predicted counterfactual earnings in Italy of these four groups.

This test confirms again the hypotheses of the Roy-Model of self-selection. The distribution of skills of migrants in countries with lower redistribution dominates the distribution of stayer and migrants in countries with higher redistribution. At the same time, the distribution of immigrants in countries with higher redistribution is rather similar to the distribution of non-migrants, but stochastically dominated by the distribution of internal migrants. A Kolmogorov-Smirnov test of
Figure 10: Cumulative distributions of counterfactual earnings

Notes: Counterfactual earnings of individuals have they stayed in Italy are log labor earnings predicted for all individuals in AIRE after running a Mincer-regression on SHIW data. Distributions for migrants are own estimations using AIRE. Distribution of non-migrant and internal migrants are own estimations using SHIW.
the equality of distributions shows that all the differences between the distributions of the four groups visualized in Figure 10 are statistically significant.

5.3 Selection on unobservable skills

To test the relationship between returns to skills and the selection on unobservable skills of migrants (e.g. abilities or motivation) we analyse if the likelihood to attain certain occupations changes with the level of redistribution, holding observable skills constant. We apply a Probit model on a binary variable indicating the occupation status \( h \) of individual \( i \) who registered in year \( t \) to the registry of Italians living abroad in country \( c \). We adopt two different specifications for \( h \): i) one if the individual is unemployed or inactive, and zero if in an employment situation, ii) one if the individual is an executive or manager, and zero if unemployed or in another type of occupation. Furthermore, educational attainment \( y^* \) are included in the equation as binary variable that is one if the individual attained beyond compulsory education and zero otherwise.

The following model is estimated where the level of redistribution is interacted with educational attainment:

\[
Prob(h_{ijtc} = 1) = \Phi(\vartheta T_{tc} \cdot y^*_{ijtc} + \beta T_{tc} + \theta y^*_{ijtc} + \gamma' X_{ijtc} + \delta' Z_{tc})
\]  

(8)

Individual level covariates are included in \( X \), country level covariates in \( Z \). The marginal effect of the interaction term \( \vartheta \) computed at different values of \( T \) shows how the likelihood to be unemployed or to have a high occupational status (executive or manager) varies with redistribution for immigrants with high and low education, respectively.

While it is safe to assume that educational achievements is a (quasi) time-invariant characteristic in adulthood, this does not apply to occupation. Hence, we have to restrict our sample further to people available to the labour market. For instance, as Figure 11 highlights, there are substantial cross-country difference in female labour participation across countries. To avoid that these differences affect our estimates we restrict this part of the analysis on male immigrants.

Table 5 shows the estimated coefficients of the Probit models. In the first and third columns of both specifications of the dependent variable, the coefficient of the interaction term is restricted to
Figure 11: Occupation and Education of Italian immigrants by country

[Bar charts showing occupation and education levels for Italian immigrants by country, differentiated by gender (men and women).]
5. EMPIRICAL SET-UP AND RESULTS

Figure 12: Selection on unobservable skills

Notes: Dots show the predicted probabilities for different levels of relative redistribution. Left figure shows the marginal effects of the interaction term in model (2) on Table 5, right figure shows the marginal effects of the interaction term in model (6).

zero. The marginal effects of the interaction terms for different levels of relative redistribution are plotted in Figure 5. Again, our findings confirm the hypothesis of the Roy-Modell, pointing at a positive relationship between returns to skills and the self-selection of immigrants on unobservable characteristics. Controlling for education, redistribution is associated with a higher likelihood of being unemployed or inactive, and a lower likelihood to be an executive or manager. However, in the case of the latter dependent variable, including country fixed effects the coefficient is not statistically significant from zero in both applications. The reason for this could be that the largest variation in the level of redistribution takes place between countries, while within countries this variable changes only marginally.

The evidence in favour of the Roy-Modell is reinforced looking at the likelihoods to attain certain occupation types for individuals with different educational attainments. As is evident from the marginal effects shown in in Figure 5, the likelihood to be unemployed is significantly lower for individuals with higher educational attainments. Interestingly, this likelihood rises for people regardless of their educational level with rising redistribution. The difference in the likelihood between educational levels is higher, the higher is the level of redistribution. The same pattern is observed for the likelihood to be executive or manager. Especially among individuals with higher
Table 5: Selection on unobservable skills

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Probit estimates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beyond compulsory education (0/1)</td>
<td>-0.299***</td>
<td>-0.380***</td>
<td>-0.288***</td>
<td>-0.415***</td>
<td>1.184***</td>
<td>0.507***</td>
<td>1.163***</td>
<td>0.657***</td>
</tr>
<tr>
<td></td>
<td>(0.0193)</td>
<td>(0.0867)</td>
<td>(0.0192)</td>
<td>(0.0870)</td>
<td>(0.0237)</td>
<td>(0.0805)</td>
<td>(0.0233)</td>
<td>(0.0847)</td>
</tr>
<tr>
<td>Relative Redistribution</td>
<td>1.522***</td>
<td>1.416***</td>
<td>4.651***</td>
<td>4.513***</td>
<td>-1.284***</td>
<td>-3.015***</td>
<td>1.944</td>
<td>0.857</td>
</tr>
<tr>
<td></td>
<td>(0.183)</td>
<td>(0.221)</td>
<td>(1.548)</td>
<td>(1.553)</td>
<td>(0.196)</td>
<td>(0.239)</td>
<td>(1.324)</td>
<td>(1.365)</td>
</tr>
<tr>
<td>Beyond compulsory - Relative Redistribution</td>
<td>0.223</td>
<td>0.352</td>
<td>2.041***</td>
<td>1.545***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.225)</td>
<td>(0.229)</td>
<td>(0.246)</td>
<td>(0.270)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share of people from the same province in country of residence</td>
<td>0.233***</td>
<td>0.242***</td>
<td>0.134***</td>
<td>0.139***</td>
<td>-0.771***</td>
<td>-0.740***</td>
<td>-0.469***</td>
<td>-0.462***</td>
</tr>
<tr>
<td></td>
<td>(0.0573)</td>
<td>(0.0586)</td>
<td>(0.0676)</td>
<td>(0.0676)</td>
<td>(0.0620)</td>
<td>(0.0612)</td>
<td>(0.0635)</td>
<td>(0.0638)</td>
</tr>
<tr>
<td>Other family members migrated earlier (0/1)</td>
<td>0.0589***</td>
<td>0.0585*</td>
<td>0.0590**</td>
<td>0.0588*</td>
<td>-0.0661</td>
<td>-0.0689</td>
<td>-0.101**</td>
<td>-0.104**</td>
</tr>
<tr>
<td></td>
<td>(0.0318)</td>
<td>(0.0318)</td>
<td>(0.0318)</td>
<td>(0.0319)</td>
<td>(0.0456)</td>
<td>(0.0456)</td>
<td>(0.0452)</td>
<td>(0.0453)</td>
</tr>
<tr>
<td>Rural place of origin (0/1)</td>
<td>-0.0752***</td>
<td>-0.0754***</td>
<td>-0.0815***</td>
<td>-0.0817***</td>
<td>-0.192***</td>
<td>-0.194***</td>
<td>-0.185***</td>
<td>-0.186***</td>
</tr>
<tr>
<td></td>
<td>(0.0165)</td>
<td>(0.0165)</td>
<td>(0.0166)</td>
<td>(0.0166)</td>
<td>(0.0179)</td>
<td>(0.0180)</td>
<td>(0.0181)</td>
<td>(0.0181)</td>
</tr>
<tr>
<td>Internal migration experience before emigration (0/1)</td>
<td>0.0178</td>
<td>0.0180</td>
<td>0.0151</td>
<td>0.0152</td>
<td>0.0727***</td>
<td>0.0739***</td>
<td>0.0761***</td>
<td>0.0769***</td>
</tr>
<tr>
<td></td>
<td>(0.0232)</td>
<td>(0.0232)</td>
<td>(0.0235)</td>
<td>(0.0235)</td>
<td>(0.0158)</td>
<td>(0.0158)</td>
<td>(0.0159)</td>
<td>(0.0160)</td>
</tr>
<tr>
<td>Resident in the capital (0/1)</td>
<td>-0.0450</td>
<td>-0.0483</td>
<td>-0.0166</td>
<td>-0.0192</td>
<td>0.0389</td>
<td>0.0253</td>
<td>0.118**</td>
<td>0.111***</td>
</tr>
<tr>
<td></td>
<td>(0.0363)</td>
<td>(0.0367)</td>
<td>(0.0384)</td>
<td>(0.0385)</td>
<td>(0.0247)</td>
<td>(0.0247)</td>
<td>(0.0241)</td>
<td>(0.0241)</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>-0.0340***</td>
<td>-0.0332***</td>
<td>0.00630</td>
<td>0.00720</td>
<td>0.0669***</td>
<td>0.0682***</td>
<td>-0.00495</td>
<td>-0.00185</td>
</tr>
<tr>
<td></td>
<td>(0.0115)</td>
<td>(0.0115)</td>
<td>(0.0160)</td>
<td>(0.0161)</td>
<td>(0.0122)</td>
<td>(0.0122)</td>
<td>(0.0130)</td>
<td>(0.0131)</td>
</tr>
<tr>
<td>GDP growth</td>
<td>-0.0137</td>
<td>-0.0133</td>
<td>-0.0183</td>
<td>-0.0185</td>
<td>-0.109**</td>
<td>-0.106***</td>
<td>0.00377</td>
<td>0.00236</td>
</tr>
<tr>
<td></td>
<td>(0.0210)</td>
<td>(0.0212)</td>
<td>(0.0238)</td>
<td>(0.0241)</td>
<td>(0.0294)</td>
<td>(0.0296)</td>
<td>(0.0250)</td>
<td>(0.0251)</td>
</tr>
<tr>
<td>GDP per capita</td>
<td>-0.0113***</td>
<td>-0.0114***</td>
<td>-0.0125***</td>
<td>-0.0117***</td>
<td>0.00573***</td>
<td>0.00571***</td>
<td>-0.00490</td>
<td>-0.00025</td>
</tr>
<tr>
<td></td>
<td>(0.00112)</td>
<td>(0.00118)</td>
<td>(0.00332)</td>
<td>(0.00334)</td>
<td>(0.00154)</td>
<td>(0.00154)</td>
<td>(0.00336)</td>
<td>(0.00340)</td>
</tr>
<tr>
<td>Constant</td>
<td>33.09***</td>
<td>33.17***</td>
<td>32.93***</td>
<td>33.06***</td>
<td>60.01***</td>
<td>61.01***</td>
<td>58.67***</td>
<td>59.45***</td>
</tr>
<tr>
<td></td>
<td>(2.768)</td>
<td>(2.783)</td>
<td>(2.762)</td>
<td>(2.777)</td>
<td>(2.717)</td>
<td>(2.797)</td>
<td>(2.783)</td>
<td>(2.848)</td>
</tr>
<tr>
<td>Demographic controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Country F.E.</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>106121</td>
<td>106214</td>
<td>106214</td>
<td>106214</td>
<td>106137</td>
<td>106137</td>
<td>106137</td>
<td>106137</td>
</tr>
<tr>
<td>Pseudo $R^2$</td>
<td>0.056</td>
<td>0.057</td>
<td>0.060</td>
<td>0.060</td>
<td>0.234</td>
<td>0.235</td>
<td>0.250</td>
<td>0.251</td>
</tr>
</tbody>
</table>

Notes: Dependent variable in columns (1) to (4) is one if the individual is unemployed or inactive, and zero if in an employment situation. Dependent variable in columns (5) to (8) is one if the individual is an executive or manager, and zero if unemployed or in another type of occupation. Demographic controls include year of birth, year of arrival, italian region of origin, a dummy indicating whether the individual was the first household member to arrive in the country of destination, rural or urban location of origin (definition: < 150 inhabitants/km2), and dummy variables indicating if the individual has an internal migration experience prior to emigration and if he or she lives in the capital of the host country. Standard errors clustered at the country-year level. Statistical significance level * 0.1 ** 0.05 *** 0.01. Source: AIRE, own estimations.
educational levels, the likelihood to be executive or manager is substantially higher when the level of redistribution is rather low.

5.4 Monetary Returns and the Choice of Migration

5.4.1 Net Returns to Migration

Figure 13 shows the average net returns to migration estimated with LIS data for the Italian migrants in our sample. Net returns are measured as the difference between the predicted yearly disposable household incomes (net of taxes and transfers) in the country of destination and the counterfactual disposable income in Italy. All values are in international US-Dollars using purchasing power parity at 2011 prices.

We observe that the predicted monetary returns for highly educated migrants are almost 4,000 Dollars per month in the US and Switzerland, and substantially lower in the other countries. In contrast, low educated Italian migrants in Australia have about 2,900 Dollars more at disposition than they would have in their home country. In the US, highly educated Italian migrants earn more than twice as much as low educated migrants. Net returns of Italian migrants in the four EU countries in our sample are rather similar between 800 and 1,600 Dollars per month. Migration to Brazil has on average only positive returns for highly educated migrants.

We bring the Roy Model of self-selection to our data in a graphical representation. Figure 14 plots the predicted net income in the host country on the y-axis, and the predicted counterfactual log labour earnings in Italy on the x-axis as a measure of skills. The slope of the single curves is indicative for the degree of returns to skills of each country. As expected, the curves are rather steep for the US and Brazil, which are countries with high returns to skills and relatively low levels of redistribution. The curves of countries like Germany, Canada, and the Netherlands show a lower gradient, comparable to the slope of Italy. It conspicuous that for all countries except Brazil individuals with the lowest level of skills would have lower net incomes in Italy than in any other destination country.
5. EMPIRICAL SET-UP AND RESULTS

Figure 13: Predicted monetary returns to migration

Notes: The coefficients used to predict the net incomes and counterfactual net incomes of immigrants in AIRE are estimated using LIS data, running an augmented Mincer regression for each single country on disposable household income including the variables sex, age, quadratic age, education, and an indicator on whether at least one child lives in the household. Disposable incomes (in international US Dollars applying Purchasing Power Parity) of Italian immigrants and their counterfactual in Italy are predicted using the coefficients of this regression. Returns to migration are then defined as the difference between predicted incomes in the destination country and the counterfactual income in Italy. Source: AIRE and LIS, own estimates.
5. EMPIRICAL SET-UP AND RESULTS

Figure 14: Self selection of Italian immigrants

Notes: The coefficients used to predict the net incomes and counterfactual net incomes of immigrants in AIRE are estimated using LIS data, running an augmented Mincer regression for each single country on disposable household income. The coefficients used to predict counterfactual labour earnings in Italy are estimated using SHIW data.
5.4.2 Choice of Migration and Destination Country

We estimate a discrete choice model of the decision to migrate as a function of personal characteristics and place characteristics of the country of destination. These sort of models are usually applied to estimate the determinants of the migration decision (e.g. Davies et al., 2001; O’Keefe, 2004; Vigdor, 2002). To test the Roy-Model of income maximization, our main interest lies in estimating if net income returns determine the choice of migration. A similar set up has been adopted, for instance, by Grogger and Hanson (2011) on aggregate data to test the selection of international migrants.

We pool our administrative data on Italian migrants with the Italian survey data on stayer and run an alternative specific conditional logit model (McFadden, 1974). The explanatory variables are either alternative specific or case specific. The former vary among countries and individuals, while the latter only among individuals. The model is motivated by a random utility framework, which models the potential utility of migrant \( i \) in country \( c = 1,\ldots,C \) as a function of the obtainable net income returns \( Y_{\text{net}} \) in this country, which vary for each individual depending on his or her level of education and other individual characteristics, and some other country specific characteristics that vary for each individual depending on his or her year of migration. The model can be expressed as

\[
U_{ic} = \gamma'Z_{ic} + \alpha'_c A_{ic} + \epsilon_{ic} \tag{9}
\]

where \( U_i \) is the utility associated to the potential choice of each alternative country of destination, including the possible decision to stay in Italy. The country actually chosen by \( i \) is the one that maximizes his or her utility. The vector \( Z \) includes \( Y_{\text{net}} \) as well as the other alternative specific characteristics. \( A \) is a vector that contains dummy variables for each country and individual specific characteristics that do not change across alternatives. These must be interacted with each potential choice, yielding coefficients \( \alpha_c \) for each potential country of destination. Hereby, we must set one of the countries as the baseline alternative, setting \( \alpha_k = 0 \) for this baseline country \( k \). We set Italy – i.e. the choice to stay – as baseline when the full decision set is evaluated, and Switzerland
when the estimation are run just on the sample of mover. $\varepsilon_{ic}$ is the random component, which is assumed to be independently and identically distributed, with an extreme-value distribution. Under this assumption the probability that $i$ chooses destination country $c$ is

$$Prob(D_{ic} = 1) = \frac{e^{\gamma Z_{ic} + \alpha_i' A_{ic}}}{\sum_{j=1}^C e^{\gamma Z_{ij} + \alpha_j' A_{ij}}}. \quad (10)$$

Where $D_{ic}$ is an indicator of $i$’s decision regarding the country of destination. Each individual chooses among the 10 countries in the choice set, including Italy. Hence, the dataset is expanded to encompass 10 observations for each individual where $D$ is equal to one if $c$ is the actual country of residence and zero otherwise. $Z$ captures the circumstances the individual faces in the actual country of residence and that he or she would face in the other potential destinations, for instance the different amounts of disposable household income. Individual level control variables included in $A$ are indicators for age, sex and the Italian geographic region of origin. The model is estimated on the whole sample, as well as separately for each education group. Population weights are applied.\footnote{As before, for SHIW we use the data design weights and for AIRE we compute weights that counterbalance the observations with missing information on educational attainment.}

Table 6 shows the estimated coefficients of the conditional logit model including only predicted disposable income as alternative specific variable and the associated marginal effects of net income returns on the choice of the country of destination; the first including Italy in the choice set, the second excluding it. We observe that the coefficient of predicted net income is positive and highly significant. This pattern holds excluding Italy among the possible destinations and hence focusing on the population of migrants. The marginal effects show that a yearly net income raise by 10,000 international Dollars PPP in Italy rises the probability of low and high educated individuals to stay in Italy by around 2 and 6 percent, respectively. To give another example, the average yearly returns to migration of high educated Italian immigrants in Switzerland, 45,000 USD, are associated to a
Table 6: Conditional Logit Estimates I

<table>
<thead>
<tr>
<th>Choice: Destination Country</th>
<th>w/ Italy</th>
<th>w/o Italy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Net Income (absolute)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education Level</td>
<td>All</td>
<td>Low</td>
</tr>
<tr>
<td>Low</td>
<td>1.316***</td>
<td>5.041***</td>
</tr>
<tr>
<td>(0.0361)</td>
<td>(0.270)</td>
<td>(0.207)</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.0361)</td>
<td>(0.270)</td>
<td>(0.207)</td>
</tr>
</tbody>
</table>

Log-lik. | -1303369.9 | -420752.0 | -369768.2 | -396590.7 | -332271.9 | -94933.0 | -98755.8 | -123248.2 |
Observations | 1519960 | 558070 | 437330 | 524560 | 1326195 | 481896 | 378963 | 465336 |
Cases | 151996 | 55807 | 43733 | 52456 | 147355 | 53544 | 42107 | 51704 |
Alternatives | 10 | 10 | 10 | 10 | 9 | 9 | 9 | 9 |

Marginal effect 100: w/ Italy / w/o Italy

<table>
<thead>
<tr>
<th>Country</th>
<th>ITA</th>
<th>AUS</th>
<th>BRA</th>
<th>CAN</th>
<th>CH</th>
<th>FRA</th>
<th>GBR</th>
<th>GER</th>
<th>NLD</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>Low</td>
<td>1.91 / -</td>
<td>0.01 / 0.39</td>
<td>0.02 / 0.13</td>
<td>0.00 / 0.40</td>
<td>0.5 / 4.63</td>
<td>0.05 / 0.54</td>
<td>0.26 / 1.72</td>
<td>1.04 / 6.02</td>
<td>0.00 / 0.07</td>
</tr>
<tr>
<td>High</td>
<td>5.84 / -</td>
<td>0.09 / 0.57</td>
<td>0.11 / 0.38</td>
<td>0.02 / 0.53</td>
<td>1.19 / 3.80</td>
<td>0.78 / 2.38</td>
<td>2.03 / 4.91</td>
<td>0.89 / 2.98</td>
<td>0.12 / 0.60</td>
<td>0.61 / 3.88</td>
</tr>
</tbody>
</table>

Notes: Probability to stay in Italy or chose one of the 9 destination countries as a function of predicted net income returns to migration. Controls for age, sex and the Italian geographic region of origin are included. Marginal effects (multiplied by 100) of a yearly net income rise by 10,000 international USD PPP in the country on the probability to stay/migrate in/to this country. Standard errors clustered at the individual level. Statistical significance level * 0.1 ** 0.05 *** 0.01. Source: AIRE and SHIW, own estimations.

The patterns of the results do not change when the household disposable income is equivalized by the square root scale.

higher likelihood by 5.4 percent to leave Italy and move to Switzerland. These probabilities are substantially lower for low educated individuals.¹²

Table 7 shows the estimated coefficients of the conditional logit model including the full set of alternative specific control variables and excluding Italy as possible destination. The coefficient of net income is positive and significant. Furthermore, the probability to chose a country over the other is positively associated with language relatedness and GDP per capita, and negatively with the distance of the country from the Italian border. The effect of the unemployment rate is positive and uniform across education groups. However, the unemployment rate is relatively low (between 2 and 10 percent) and changes only slightly within countries over the observation period. The positive association between the share of migrants from the same province of origin and the likelihood to reside in a particular country is merely mechanical and serves here just as control variable.

A crucial assumption of most discrete choice models is the independence of irrelevant alternatives. The violation of this assumption cannot be excluded here, neither intuitively nor by a Hausmann test. Possible ways to circumvent the problem would be to order the choice alternatives or apply a Multinomial Probit. The former is clearly not applicable to our set-up, while the latter is

¹² The patterns of the results do not change when the household disposable income is equivalized by the square root scale.
### 6. Conclusions

In this study we tested the predictions of the Roy-Model about the self-selection of immigrants using an administrative dataset including about 90% of Italians living abroad. Our results confirm the predictions of the model: high returns to human capital, which we measure by the degree of income redistribution, are significantly associated with a positive skill selection of Italian immigrants. On the contrary, countries with more progressive tax and transfer systems largely attract immigrants from the lower part of the skill distribution. These patterns are confirmed by a multitude of distinct exercises and test procedures.

Currently, the coordination of migration policies is a widely debated topic. However, the discussion is mostly focused on migration quotas and takes less into account that particular country

<table>
<thead>
<tr>
<th>Table 7: Conditional Logit Estimates II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Level</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Choice: Destination Country Predicted Net Income (absolute)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Share of migrants from the same province of origin</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Distance of country of residence from Italian border (in 1000 km)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Language relatedness</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Unemployment rate</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>GDP per capita</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Log-lik.</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>Cases</td>
</tr>
<tr>
<td>Alternatives</td>
</tr>
</tbody>
</table>

Notes: Probability to choose one of the 9 destination countries as a function of predicted net income returns to migration and other country characteristics. Controls for age, sex and the Italian geographic region of origin are included. Standard errors clustered at the individual level. Statistical significance level * 0.1 ** 0.05 *** 0.01. Source: AIRE, own estimations.

computed infeasible with so many options and observations as in our case. Hence, we adopt a different approach. We subdivide the sample into close and far locations and run the estimations. For both sub-samples the results confirm our main findings. Overall, the results of this exercise again confirm the predictions of the Roy-Model: income maximization is a substantial determinant of the migration decision.
characteristics might attract certain types of immigrants. Our empirical results show again that migrants are likely to move to countries in order to maximize their income and, hence, that the selection of immigrants substantially depends on the relative returns to skills. Since these patterns of selection seem to apply not only to observable skills, like education, but also to unobservable skills, policy strategies that establish migration quotas based on qualifications are not sufficient for policy makers aimed to attract immigrants from the top of the skill distribution.

One of the few ways in which the skill composition of migrants can be influenced by public policy, is through the tax and transfer system as a tool to change the returns to skills. As a consequence, countries have an incentive to reduce their redistributive effort to attract high skilled immigrants and discourage the immigration of low-skilled immigrants. This situation might lead to a race to the bottom where eventually most countries have lower than optimal levels of taxation and redistribution. An international coordination of income redistribution seems therefore necessary to face the free movement of individuals.

References


