GRAPE Working Paper #102

Linguistic proximity and inequality in returns to migrant skills

Jonas Feld and Joanna Tyrowicz

FAME | GRAPE, 2025

Foundation of Admirers and Mavens of Economics
Group for Research in Applied Economics

Linguistic proximity and inequality in returns to migrant skills

Jonas Feld
Trier University and IAAEU

Joanna Tyrowicz
FAME|GRAPE, University of Augsburg,
University of Warsaw, and IZA

Abstract

We provide novel evidence on the inequality of returns to immigrant skills in hosting economies. Although migrant wage gaps are well established in the literature, less is known about the origins of their heterogeneity. We propose a potential rationale for this gap related to the linguistic proximity between the destination and origin countries. We exploit individual-level data from nine diverse destination countries, with migrants from a highly heterogeneous group of origin countries, for both recent and long-term migrants. We find that lower linguistic proximity between origin and destination is associated with a higher average wage penalty for highly skilled migrants and a substantially lower position in the wage distribution.

Keywords:

migration, linguistic proximity, returns to education

JEL Classification

F22, I23, I26, Z13

Corresponding author

Joanna Tyrowicz, j.tyrowicz@grape.org.pl

Acknowledgements

The authors are grateful for inspiring comments to Michel Beine, Ben Elsner, Tomasso Frattini, Harry Ganzeboom, Laszlo Goerke, Sven Hartmann, Konstantin Homolka, Pawel Kaczmarczyk, Yoshihiro Kitamura, Eva Markowsky, David McKenzie, Alberto Palermo, Nazareno Panichella, Solomon Polachek, Panu Poutvaara, Mariola Pytlikova, Laura Renner, Dominik Sachs, Yannik Schenk, Gabriel Schultze, Katrin Sommerfeld, Nicolas Ziebarth, and the anonymous referees. We thankfully acknowledge the feedback from the audiences at Paris School of Economics, WEAI 2019, VfS Annual Conference 2020, ESCR Workshop 2020, ESPE 2021 and Waseda University Tokyo. Joanna Tyrowicz gratefully acknowledges the support of National Science Center (grant # 2021/43/B/HS4/03241). Any remaining errors are ours.

Published by: FAME | GRAPE
ISSN: 2544-2473

Foundation of Admirers and Mavens of Economics
Koszykowa 59/7
00-660 Warszawa
Poland

W | grape.org.pl
E | grape@grape.org.pl
TT | GRAPE_ORG
FB | GRAPE.ORG
PH | +48 799 012 202

1 Introduction

We study the inequality in returns to tertiary education for migrants from various origin countries in nine diverse destination countries. A rich array of studies documents migrant wage gaps that remain even after adjusting for socioeconomic characteristics. In line with the human capital framework, numerous studies have emphasized the role of skill portability from the origin to the destination economy for immigrant wage assimilation (e.g. Friedberg 2000, Bazzi et al. 2016, both these studies also provide a thorough overview of the empirical literature in the field).[1] Considerably less attention has been devoted to the *inequality* of the immigrant-native wage gap between destinations and origins. Our study aims to fill this gap.

To conceptualize the dispersion of returns to tertiary education among migrants, we emphasize the costs associated with assessing migrants' qualifications. We leverage information asymmetry in employment to hypothesize that employers face costs inspecting the qualifications of job candidates with foreign education and certification (imposing a friction). Before actual employment, employers cannot fully assert job candidates' qualifications due to this information asymmetry, and they typically rely on references, diplomas, and certificates to bridge the gap. Linguistic proximity facilitates the comparability of this information for migrants and natives, reducing friction. When linguistic proximity is high, it is easier for employers to assess qualifications, leading to fewer costs. Conversely, lower linguistic proximity increases these costs, potentially influencing hiring decisions. We expect linguistic proximity to be particularly relevant for individuals with tertiary degrees, as their credentials often require closer scrutiny (e.g. Peri and Sparber 2009).

Thus, we obtain a testable empirical hypothesis: *linguistic proximity reduces the wage gaps of high-skilled migrants*. Note that our hypothesis is not a foregone conclusion. First, it has been empirically demonstrated that linguistic proximity drives migration flows (e.g. Chiswick and Miller 1994, Caragliu et al. 2013, Adsera and Pytlikova 2015). This selectivity of migration patterns could imply that there is no variation in linguistic proximity in our data, or that it is insufficient to estimate any meaningful relations. Second, the frictions for the employers related to linguistic proximity can be too low for any relation to emerge from observational data. Finally, these frictions may be quantitatively dominated by mechanisms that are not related to linguistic proximity, such as network effects.

We address our hypothesis by providing a comprehensive set of measures of returns to human capital (tertiary education) across a diverse group of origin and destination countries. We harmonize individual-level data for nine popular migrant destination countries: Argentina, Brazil, Mexico, Germany, UK, France, Israel, the U.S. and Canada. These countries are located on four continents, cover six language families, and feature a strong representation of migrants from many different origin countries. Together, they host about 36 percent of the total international migrant population worldwide.[2] Our analysis draws on origin- and destination-specific data on returns to tertiary education for migrants, derived from rich, harmonized wage information. Both natives and migrants

[1] Speaking the language of the host economy is a prerequisite for being able to communicate one's knowledge and thus transfer the existing human capital, which is why language skills have been at the core of empirical research on the immigrant wage gap. Empirical research on the migrant wage gap is based on the assimilation hypothesis (Chiswick 1986), the skill portability hypothesis (e.g. Rivera-Batiz 1990, Chiswick 1991, Dustmann 1994, Chiswick and Miller 1995, Berman et al. 2003, Bleakley and Chin 2004, 2010), as well as the segmentation / discrimination of the labor market (Piore 1972, Reich et al. 1973, Doeringer and Piore 1985). The existing literature typically finds that in various hosting economies, competence in the language of the hosting economy mitigates – but does not eradicate – the wage penalty (e.g. McManus et al. 1983, McManus 1985, Kossoudji 1988, Rivera-Batiz 1990, Chiswick 1991, Dustmann 1994, Chiswick and Miller 1995, Dustmann and Fabbri 2003, Ferrer et al. 2006, Goldmann et al. 2015). In addition, learning the language of the host economy is associated with lower wage penalties (Berman et al. 2003).

[2] Data sourced from "International Migrant Stock", which is periodically disseminated by The United Nations.

report wages in our dataset, enabling us to estimate the wage gap between migrants and natives in each destination country, both with and without a tertiary degree. These estimates are obtained in different origin countries and for each destination country separately. They are adjusted for individual characteristics such as age, gender, marital status, and time since arrival in the destination country, as well as for job-related characteristics, such as occupation and industry.

Once we establish the dispersion of returns to tertiary education for migrants in each country pair, we relate it to the bilateral measures of linguistic proximity for these country pairs. Our measures of linguistic proximity draw on the experience of international economics and capture the *ease of communication* between random individuals from two countries (Melitz and Toubal 2014). They reflect the inherent similarity of languages in these countries *per se*, as well as the knowledge of common languages among the two populations. Our findings indicate that, *ceteris paribus*, greater linguistic proximity between origin and destination countries is associated with an improved position of migrants in the wage distribution of the destination country. Furthermore, we observe that the (adjusted) wage gap for highly skilled migrants is smaller for origin countries that are linguistically closer, *ceteris paribus*.

Our paper contributes to two distinct strands of the literature. First, we build on the rich literature that studies the role of linguistic proximity in explaining bilateral worker flows. Earlier studies, such as Chiswick and Miller (1994), Caragliu et al. (2013), and Adsera and Pytlikova (2015), show that migrants tend to select destination countries that are culturally *and* linguistically closer to their origin (see also Sprenger 2024), complementing migration networks (Bredtmann et al. 2020). Second, we contribute to the literature highlighting that linguistic proximity not only helps securing employment (Wong 2023, Ghio et al. 2023), and a type of job obtained by migrants (Adserà and Ferrer 2021). We show that the wage gaps and position in the wage distribution in the hosting economy are lower for immigrants from linguistically close countries. Our result is not explained away by language skills of the immigrants (Chiswick 1991, Chiswick and Miller 1995, Dustmann and Fabbri 2003, Berman et al. 2003, Chiswick and Miller 2012, among others).

Against this literature, our study offers several important innovations. First, we focus on highly skilled migrants, whereas most of the existing literature focuses on low-skilled workers. This is relevant because highly skilled migrants often speak some of the most important global languages as second languages (such as English, Spanish, and French; as documented by e.g. LaLonde and Topel 1992, Dustmann 1994, Dustmann and Fabbri 2003). Also, most existing studies on immigrant-native wage gaps relate to destination countries with one of the main global languages as a native language, whereas our sample also comprises less popular destination languages. Second, unlike existing studies that typically focus on a single destination country, our analysis spans many destination countries. This broader scope allows for the disentanglement of linguistic proximity from destination-specific conditions, and is complemented by our application of a rich, multidimensional measure of linguistic proximity.[3] Third, we document the remarkable dispersion of returns to tertiary education among migrants across diverse origin and destination countries and show that this dispersion correlates with linguistic proximity.

Our paper is structured as follows. Section 2 summarizes the literature on immigrant-native wage gaps in the context of language proximity. In section 3 we discuss our methodology and in section 4 we cover data and stylized facts emerging from this study. Section 5 presents the results. Finally, the last section offers a summary of the results, discusses policy implications, and outlines avenues for further research.

[3]Wong (2023) explores the random assignment of asylum seekers to cantons in Switzerland and takes advantage of the fact that cantons differ in terms of dominant language.

2 Related literature

Even after adjusting for individual characteristics, wage gaps persist between natives and migrants (Chiswick 1978, Borjas 1985, Chiswick 1986, Jasso and Rosenzweig 1988, Borjas 1992, 1994). These gaps are attributed to migrants' challenges as job seekers (e.g., limited knowledge of the local labor market) and the costly, time-consuming process of skill transfer and assimilation. [4] Chiswick and Miller (2013) argue that what matters is not merely host country language skills, but the match between an immigrant's language proficiency and the language requirements of the job. Imai et al. (2019) find that in Canada, migrants with limited language skills tend to self-select into manual occupations, even if they held jobs demanding cognitive skills in their origin countries. Yao and van Ours (2015) show that in a rarely learned language (Dutch), where migrants usually lack prior exposure, wage gaps prevail particularly among disadvantaged groups.

In addition, Bleakley and Chin (2004, 2010) demonstrate that labor market success hinges on an immigrant's ease of learning the destination country's language; familiarity with similar languages facilitates this learning process (Beenstock et al. 2001). Moreover, extent literature suggests that not just language proficiency but also linguistic proximity to the host country's language significantly impacts migrants' labor market success (e.g., Crystal 1987, Espenshade and Fu 1997, Chiswick and Miller 1998, 2001, Isphording and Otten 2013, Isphording 2014, Wong 2023).

Measuring language proximity is challenging, though. Early studies, such as Chiswick and Miller (2005), focused on proximity to English, based on how easily Americans could learn a language. They classified migrants into English speakers and non-native speakers, further dividing the latter into groups whose languages are close to, distant from, or intermediate to English. This admittedly crude measurement has been refined with advances in linguistic science. *Ethnologue* measures language proximity on a six-level scale based on linguistic lineage, ranging from full proximity (identical languages) to zero (completely different languages) (e.g., Adsera and Pytlikova 2015, Adserà and Ferrer 2021). Dyen's lexicostatistical measure assesses Indo-European languages' similarity based on 200 common words (Dyen et al. 1992, Belot and Ederveen 2012). The Levenshtein distance, applied to the expert-judged Automatic Similarity Judgment Program (ASJP) data, compares languages using the wording, length, and phonetics of 40 basic words. We report an overview of measures used in economic literature in Table A1.

Many of these economic studies exploit heterogeneity across origin countries, but within a single destination country. They thereby confound the role of language proximity with the role of characteristics of the labor market, and cultural factors in a given destination country with factors for a given origin-destination country pair. To obtain meaningful information about the role of linguistic proximity in isolation from these other factors, it is imperative to take advantage of the heterogeneity of the destination countries in addition to the heterogeneity of origin countries.

Language learning difficulties in the host economy may be less relevant for highly skilled migrants (those with tertiary degrees from their origin countries) for several reasons: First, they often possess proficiency in the languages of major host economies before immigrating.[5] Second, their educational background typically involves exposure to multiple foreign languages, facilitating easier acquisition of the host country's language (Beenstock et al. 2001). Finally, highly educated migrants are more likely than low-educated ones to select destination countries with a common or similar language

[4] Early studies on the US include Borjas (1987), LaLonde and Topel (1992), Borjas and Friedberg (2009), Borjas (2015), while wage gaps have also been documented for the Netherlands (Kee 1995), Israel (Friedberg 2000), and Canada (Schaafsma and Sweetman 2001, Fortin et al. 2016).

[5] For example, university entry or high school exit exams in many countries require knowledge of a major global second language.

(Clark et al. 2007, Pedersen et al. 2008, Beine et al. 2011, Grogger and Hanson 2011, Belot and Ederveen 2012, Belot and Hatton 2012, Adsera and Pytlikova 2015).

Despite the unique characteristics of highly skilled migration and persistent wage gaps among this group, the systematic drivers of these gaps remain underexplored. For highly skilled migrants, factors like cultural distance or limited language proficiency are unlikely to be exclusive or dominant explanations. A crucial but overlooked factor may lie with employers: while highly skilled migrants often communicate well in the host country's language, employers might find it costly to assess credentials or educational quality from linguistically distant countries of origin.

This mechanism is especially relevant for occupations without standardized international certification. In these fields, skills are often highly portable (e.g., analysts, engineers, and creative professionals), but verifying them can be challenging for employers. Unlike fields with strict certification, such as medicine[6] or IT[7], many other professions require employers to evaluate university curricula or portfolios. When these are in linguistically distant languages, assessing skill content and educational quality becomes costly. This can lead to wage gaps for migrants from linguistically distant countries, as the cost of verifying their qualifications may outweigh the perceived benefit.

To sum up, empirical literature shows that linguistic proximity influences migration flows. Building on this, we hypothesize that linguistic proximity between origin and destination economies affects screening costs (e.g. acquiring information in a distant language). This issue is particularly relevant in the case of jobs for which verifying qualifications is essential and language proficiency *per se* is not a barrier. We empirically test the hypothesis that linguistic proximity contributes to migrant wage gaps. By leveraging multiple destination languages and destination-by-origin fixed effects, we isolate the role of linguistic proximity, distinct from simple language proficiency.

3 Methods

To study the role of language in explaining immigrant wage gaps, we utilize a measure of linguistic proximity based on *ease of communication*. Linguistic proximity varies between country pairs. It can also vary over time within a country pair, as an effect of migration flows and increased command of foreign languages in either of the country pairs. This measure has previously been used in trade literature (Melitz and Toubal 2014). It is suitable for capturing the proposed mechanism: the costs to an employer of assessing the quality of a job seeker's credentials. Our measure of *ease of communication* begins with the Automated Similarity Judgment Program (ASJP), which indexes language similarities between 200 common words between country pairs, based on an assessment by ethnologists and ethnostatisticians. In subsequent steps, this measure is adjusted for common spoken and official languages and standardized to produce a common language index (CLI), which has a bounded distribution within our sample: between 0 for the most linguistically distant country pair and 1 for the highest linguistic proximity (see section 4 in this paper as well as the original paper of Melitz 2008, Melitz and Toubal 2014, for a detailed description on how to construct the measure).[8] More details are presented in Appendix B.

To investigate the determinants of immigrant wage gaps, we pursue two empirical strategies. First, we work directly with the individual-level data as explained in the section 3.1. We refer to this

[6]Medical doctors are subject to strict certification, which makes skills portability an issue, but linguistic proximity is less relevant given the high degree of standardization of certification in this occupation.

[7]IT specialists rely on internationally recognized certificates rather than national ones, so neither skill portability nor linguistic proximity poses a significant obstacle.

[8]The online appendix to their study also distributes the data on the ASJP, as well as on common spoken and official languages around the world.

approach as the *one-step approach*. We exploit the richness of 40 million individual-level observations, but this choice poses some econometric challenges. Therefore, we present a second empirical strategy the *two-step approach*: we obtain measures of the immigrant wage gaps across countries of origin and destination from individual-level data (the first step), which we then relate to linguistic proximity for origin-destination country pairs (the second step). The first step, the process of obtaining the immigrant wage gaps, is described in section 3.2, whereas the second step, the method for linking wage gaps and linguistic proximity, is described in section 3.3.

3.1 Individual-level analysis

This section outlines our one-step approach. The benefit of this procedure is that it takes direct advantage of the richness of our data at the individual-level. We convert the individual measure of wages into a percentile measure (denoted by $\hat{w}$) within each destination country to maintain comparability of the estimates across the samples. We then estimate the contribution of language proximity measured by common language index (CLI) from:

$$\hat{w}_i = \alpha + \beta \mathbf{X}_i + \gamma TE_i \times CLI_{d,o} + \mu \mathbf{M}_i + \delta_Z \mathbf{Z}_o + \delta_{d,o} \mathbf{Z}_{d,o} + \delta_d + \delta_o + \epsilon_i, \quad (1)$$

where the superscript d denotes destination country, o superscript denotes origin country, index i denotes individual. TE_i identifies the educational status of the individual i: it is a dummy variable that takes on a value of 1 if an individual has tertiary education and 0 otherwise. $\mathbf{M}_i$ identifies the migrant status of an individual i: it is a dummy variable taking on the value of 1 if an individual is an immigrant in a given destination country, and 0 if an individual is a native.[9] $\mathbf{X}_i$ is a vector of individual demographic control variables (personal, job and household characteristics, including information on years since immigration). The country-level controls included in $\mathbf{Z}_o$ consist of population size, GDP per capita, fertility, and mortality rates. Country pair controls included in $\mathbf{Z}_{d,o}$ consist of geographical variables (geographical distance, contiguity) and historical variables (years at war, common colonizer, common religion, common legal system). These controls are required by Head et al. (2010), because they adjust for cultural barriers important for selectivity of migrations and integration patterns (e.g., O'Rourke and Sinnott 2006, Mayda 2010, Adsera and Pytlikova 2015, Wong 2023).

The one-step procedure represented by equation (1) estimates a well-known Mincerian wage regression, jointly with the role of linguistic proximity in determining wage inequality. The specification in equation (1) allows the returns to tertiary education to vary by country of origin, but is restricted to be common between destination countries. Hence, the term $CLI_{d,o}$ is the only variable specific to a country of origin *jointly* with a country of destination. Our main research hypothesis implies a positive estimate of γ, that is, the wages of migrants increase with language proximity. Note that given the specification of the explained variable – percentile of the wage distribution within the destination country – the interpretation of the γ estimate refers to immigrant wage gaps in relative terms (position in income distribution) rather than in absolute terms (percent of the wage).

The strategy proposed in equation (1) has a significant advantage: the entire variation in our sample is exploited in one step. But it also has two limitations. First, there is no way to obtain correct estimates of the standard errors for characteristics related to several levels of variation: individual, country of destination, country of origin, and origin-destination country pairs. Second, the destination countries differ in the level of development and structure of the labor market, and the destination countries' data sets differ in sampling design and sample sizes. We introduce destination country fixed effects (δ_d) to account for country-level characteristics, and we introduce sample weights from

[9] We consider individuals who obtained education in the destination economy as natives, regardless of their ancestry. See section 4 for detailed coverage of the variable definitions and data sources.

the original sampling design in the estimation. Thus, the estimates reflect the relative size of the populations in the destination countries, not the sample sizes for the data utilized.

Given the limitations of the one-step approach, we also propose a two-step approach. In the first step, we estimate the returns to tertiary education by country of origin, separately for each destination country. Specifically, we obtain deviations of returns to tertiary education specific to a given country of origin. This step is described in section 3.2. In the second step, we relate these estimated deviations to linguistic proximity at the country-pair level. This step is described in section 3.3.

3.2 Estimating the returns to skills by country of origin

We begin our two-step approach by estimating the country-of-origin-specific returns to skills, separately for each destination country. In other words, for each destination country, we estimate the systematic wage deviation by education level that results from an individual's origin in a given country. The goal is to find out, for instance, whether the returns to skills for Norwegian migrants in Germany differ from those experienced by Swedes in Germany. We seek to obtain a data set that lists the returns for skills for all destination countries *and* all origin countries that are sufficiently covered in our micro-level data (i.e., we seek to obtain an $m \times n$ matrix in which migrants' destinations are in the rows and their places of origin are in the columns, and the entries in the matrix represent the estimated wage variances for each of these combinations).

For this purpose, we first estimate destination-specific Mincerian wage regressions using individual-level data. We denote by $log(w_i)$ the log of the hourly wage of individual i:

$$log(w_i) = \alpha + \boldsymbol{\beta X}_i + \eta TE_i + \boldsymbol{\mu M\_COO_i} + \boldsymbol{\rho TE_i} \times \boldsymbol{M\_COO_i} + \epsilon_i, \tag{2}$$

where $\mathbf{X}_i$ is a vector of individual demographic control variables (personal, job, and household characteristics, including information on migrant status and years since immigration). In this notation, TE_i is a dummy variable for the tertiary education status of individual i (taking on the value of 1 if completed, and 0 otherwise), and $\boldsymbol{M\_COO_i}$ is a sequence of dummy variables taking on the value of 1 if an individual i is an immigrant arriving from the given country of origin (COO) and zero otherwise. Accordingly, parameter η measures the average returns to tertiary education for natives in that destination country. Analogously, the vector of parameters μ captures the average effect of a immigrant from a given origin in a given destination for individuals without tertiary education. Finally, the vector of parameters $\boldsymbol{\rho}$ captures the additional effect (positive or negative) of tertiary education for migrants migrating from a particular COO to a particular destination.[10]

We estimate equation (2) with two samples of migrants: recent migrants with less than five years since immigration (the first sample), and non-recent migrants, i.e. those who have resided in the destination country for at least five years (the second sample). We identify samples using subscript t. This modeling choice has two advantages. Splitting the estimates for recent migrants and non-recent migrants allows us to tackle the potential bias stemming from selectivity in return migration: we cannot capture the size of the bias, but when estimated among individuals with similar duration of stay in the hosting economy, the bias should be homogeneous (Dustmann and Görlach 2016). Second, estimating the two subsamples allows us to control for general differences of recent migrants relative to those who had the chance to fully explore the local conditions, as described by

[10]We explain in detail in section 4 that for tertiary-educated individuals if the age at arrival in the destination country is below the customary graduation age, we re-code those individuals to be natives in a sense that their highest achieved education had been obtained in the destination country.

Chiswick and Miller (2012).[11]

Using the estimates of $\hat{\mu}$ and $\hat{\rho}$ we create a database of estimates for each country of origin o in each destination country d, and for every available sample t. In analogy to the approach in section 3.1, we construct a vector $\hat{\boldsymbol{\gamma}}_{\boldsymbol{d,o,t}}$. This vector contains estimates for individuals without tertiary education in the vector of $\hat{\mu}$ and with tertiary education from the vector of $\hat{\rho}$. This set of estimates becomes a vector $\hat{\boldsymbol{\gamma}}_{\boldsymbol{d,o,t}}$ used as a dependent variable in the second stage, in Section 3.3 below.

To ensure that each $\hat{\boldsymbol{\gamma}}_{\boldsymbol{d,o,t}}$ element can be reliably estimated, sufficient degrees of freedom are needed. Given the number of variables in equation (2), we impose a sample restriction that in each destination country, at least ten individuals from a given origin country are available for each level of education and duration of stay in the destination country.

3.3 Immigrant wage gaps and language proximity

Section 3.2 described the first step of our two-step approach, i.e. the methodology for determining average returns to tertiary education by destination and country of origin. The second step in this approach, laid out in the following, tests the relevance of linguistic proximity in explaining differences in these returns. For this purpose, we estimate the following equation:

$$\hat{\gamma}_{d,o,t} \quad = \quad \boldsymbol{\lambda} TE_{d,o} \times CLI_{d,o} + \delta_Z \mathbf{Z}_{d,o} + \delta_d + \delta_o + \delta_t + \epsilon_{d,o,t}, \tag{3}$$

where the dependent variable $\hat{\gamma}_{d,o,t}$ is the estimated systematic dispersion in returns to tertiary education for migrants from equation (2), for a given destination (d) and origin (o) country and time since migration (t, a dummy variable which determines if a sample comprised recent migrants or long-term migrants); TE dummy identifies whether a given estimate concerns migrants with a tertiary degree from a given country of origin in a given destination country; $\mathbf{Z}$ is a vector of controls for both the country of origin (population size and GDP per capita) and country-pair unique factors (geographical distance, contiguity, common legal system, religion, and historic past), and δ_d, δ_o and δ_t capture fixed effects for destination, origin and migrant sample type, respectively. The key variable of interest is $\boldsymbol{\lambda}$ vector of parameters for linguistic proximity ($CLI_{d,o}$).

We estimate equation (3) with several specifications. First, we account for the sample size used to estimate a given $\hat{\gamma}_{d,o,t}$ estimate. Not only do larger immigrant groups in a given destination country give more confidence in the robustness of the $\hat{\gamma}_{d,o,t}$ estimates, but the size of the migration flows is also typically proportional to the sample size of a population of migrants from a given country of origin in a given destination country. In our case, the sample sizes differ between destination countries for methodological reasons (for instance, the sampling size of the US data is much larger than that of Germany), so the sample weights are inappropriate. We thus use migrant stocks for country pairs from the United Nations as weights. Second, since $\hat{\gamma}_{d,o,t}$ is an estimated parameter, we bootstrap standard errors in estimating equation (3). An alternative way to address the issue that $\hat{\gamma}_{d,o,t}$ is an estimate with a measurement error is to re-weight the regressions with e.g. $t - statistic$ or the inverse of standard errors obtained in estimating the equation (2). However, such re-weighting for precision excludes re-weighting for migration flows. We adopt both re-weighting and bootstrapping to test the robustness of our results.

[11] Also Lubotsky (2007) offers several arguments for why to estimate immigrant-native wage gaps for recent and non-recent migrants separately.

4 Data

We acquire individual-level data from around the world. The following criteria are applied: First, the data source has to report wages and individual characteristics relevant to estimating the Mincerian regression. Second, the data source has to refer to a popular migrant destination country. Third, the data sources should be linguistically diversified, to exploit variation in linguistic proximity across country pairs, holding constant the destination country fixed effects and the origin country fixed effects. Data sources for nine migration destination countries meet these basic criteria: Encuesta Permanente de Hogares (*EPH*) for Argentina, IPUMS-published census data for Brazil, Canada, Israel, and Mexico (IPUMS-I 2020), Labor Force Surveys for France and the UK, Socio-Economic Panel for Germany and American Community Survey (ACS) data from the US. These destinations collectively host about 36 percent of the total migrant population worldwide.[12] We report in Table A2 details of data coverage.

A migrant is defined in the data as a person whose reported country of birth differs from the survey country. Unfortunately, the data does not allow for distinguishing reasons for migration, meaning refugees, asylum seekers, and family-reunification migrants cannot be separated from other groups. Additionally, the datasets do not include information on migrants' command of the language(s) spoken in the host economy. We construct a variable measuring years since migration to this destination country.[13] Individuals who arrived as children are classified as natives, while those reporting their year of arrival no earlier than five years prior to the survey are categorized as recent migrants. Those with longer stays are classified long-term migrants. Individuals, for whom the implied age of arrival is after obtaining a tertiary degree are considered to have a foreign diploma. The data does not permit to identify the situations, in which an individual left their country of birth before obtaining a tertiary degree, studied in a third country, and later moved to the current destination. If this phenomenon were frequent in the sample, it could bias CLI estimates toward zero due to the absence of informational frictions for employers.

The data sets are harmonized to ensure comparability across countries. For each destination country, waves of data spanning several periods are pooled together. Wages are adjusted for inflation using the consumer price index (CPI) from the World Development Indicators database by the World Bank. Occupations are categorized into five levels: jobs not requiring skills, jobs requiring primary skills, specialists, high-skilled jobs, and managers. Similarly, industries are grouped into four levels: agriculture, manufacturing, construction, and services. More detailed classifications for both industry and occupation could not be obtained in a harmonized manner. Educational attainment is standardized using ISCED categories. Age is measured as a continuous variable, while marital status is classified into four levels: single, married, in an informal relationship, and widowed/divorced.

We restrict our full sample to salaried workers aged 18 to 64. We use hourly wages, calculating them from weekly or monthly wage data, with average hours in the respective period used whenever available. Observations with negative incomes, missing data on hours worked, or hours strictly equal to 0 or exceeding 100 per week are dropped (similar to e.g., Mishel et al. 2012). In cases where earned income or hours worked are reported in intervals, the middle value of these intervals is used. This is the case for the data from Israel. In the case of Brazil and the USA, the middle value for hours was used whenever hours were reported in intervals.

As controls for country of origin, we include population, mortality, and fertility as well as GDP per capita (adjusted for purchasing power parity), which we take from the World Development Indicators database by the World Bank. We merge country-level variables with the individual-level data using

[12] We obtain this value using "International Migrant Stock" data disseminated by The United Nations.

[13] For natives, this variable takes the value of the individual's age.

the year relevant for the individual immigrant. For example, if an individual originates from Nigeria and is identified as a resident in the UK in the 2005 Labor Force Survey, with (self-reported) eleven years since migration, we merge the individual-level observation with the 1994 (= 2005 - 11) for the World Bank data on Nigeria.

Some controls unique to a given country pair do not vary over time. These controls include geographical distance between the two countries, contiguity, common legal system, common religion, and measures for common historic past (colonizer, war, etc.). Data on the distance between two countries and their contiguity originally comes from the CEPII database. Common colonization history data is taken from Head et al. (2010). Data on common legal systems comes from *JuriGlobe*, the University of Ottawa's world legal systems database. The measure on common religion is based on the *CIA World Factbook*, enhanced with information obtained from the International Religious Freedom Report[14], The WorldChristianDatabase.org, and the Pew Research Center[15]. Finally, the number of years at war comes from the *CorrelatesOfWar.org* archive. Bilateral migration flows and stocks are taken from the United Nations.

We use the same variables as suggested by Head et al. (2010). To obtain the common language index (CLI), which is a measure of the *ease of communication* between individuals from two different countries, we follow the same steps as Melitz and Toubal (2014). We start from the raw data on the ASJP linguistic proximity measure from Bakker et al. (2009) and combine it with measures of common official language (COL), common spoken language (CSL), and common native language (CNL) from *CIA World Factbook* and European Commission (2006) data.[16] The logic behind the approach is intuitive: If a COL exists for a given pair of countries, individuals from the two countries are likely to communicate fairly easily (and assess one another's credentials effectively). Similarly, suppose that there is some positive probability that two randomly selected individuals, one from each country, are able to communicate in a CNL (for example, many people in the United States report Spanish as their native language) or in another CSL (e.g., via a *lingua franca*). In essence, CNL and CSL can roughly be interpreted as measuring the likelihood that two randomly selected individuals, one from each country, can communicate in their native language (CNL) or another language they both speak (CSL). See Melitz and Toubal (2014) for a detailed description of the underlying methodology. To obtain the CLI, we standardize the original language proximity index (*ease of communication*) by our sample, bounded by the values of 0 and 1 for the lowest and highest linguistic proximity in our data, respectively.[17]

Table 1 reports the summary statistics of our data. Note that we use several sources of data, each with multiple variables. In the interest of brevity, we report in Table 1 key variables of interest for three groups of variables: data that we are able to recover from individual data in our sample, data on migration from the United Nations which are matched at the country level, and economic data matched at the country or country pair level.

Our sample consists of a large number of observations from Brazil (32 percent of observations) and

[14]See `https://www.state.gov/international-religious-freedom-reports/`.

[15]The data comes from "Mapping the Global Muslim Population. A Report on the Size and Distribution of the World's Muslim Population" available at `https://www.pewforum.org/2009/10/07/mapping-the-global-muslim-population/`.

[16]The missing data on CNL and CSL fractions were hand collected by Melitz and Toubal (2014) and can be found in the online appendix of their paper.

[17]Refer to the appendix B for a technical description of the CLI. Note that our measure of linguistic proximity departs from Melitz and Toubal (2014) in one minor detail, namely we standardize the ASJP measures by common spoken language rather than common native language; see Figure A1 for comparison. This form of standardization is closer to the mechanisms we propose in our model: the ability of an employer to inquire about the qualifications of a candidate employee with a foreign diploma. In the interest of transparency, we estimate our model also with the original common language index by Melitz and Toubal (2014), as reported in Appendix D.

the United States (48 percent of observations). This is because these two data sources are essentially censuses. Meanwhile, data for Israel, Germany, France, and the UK come from representative population surveys. To adjust for these effects in the individual-level regressions, the observations are weighed by the inverse of the destination country sample size.

Table 1: Summary statistics

	Total	Argentina	Brazil	Canada	France	Germany	Israel	Mexico	UK	USA
	Individual-level data									
# of obs.	42,852,038	1,117,908	13,555,036	1,076,098	402,225	301,969	192,622	5,095,568	568,277	20,542,335
# of natives	39,359,351	1,075,475	13,523,179	910,651	369,750	253,229	109,488	5,081,264	528,039	17,508,276
# of migrants	3,492,687	42,433	31,857	165,447	32,475	48,740	83,134	14,304	40,238	3,034,059
# of recent migrants	532,987	38,523	11,278	28,820	2,976	3,352	15,891	10,884	9,830	411,433
% of recent migrants	0.153	0.908	0.354	0.174	0.092	0.069	0.191	0.761	0.244	0.136
Years since immigration	19.620	3.495	9.315	18.902	28.247	19.959	23.279	3.195	19.951	19.868
% of TE natives	0.368	0.333	0.145	0.619	0.308	0.313	0.329	0.195	0.169	0.588
% of TE migrants	0.505	0.196	0.446	0.657	0.237	0.160	0.352	0.449	0.235	0.518
Age of natives	37.25	38.72	34.19	37.49	39.42	40.45	32.96	33.53	39.60	40.46
Age of migrants	40.97	43.90	42.24	41.88	43.88	40.31	43.36	35.30	38.76	40.83
	Country-level data: migration (from United Nations)									
# of migrants in DC	-	1,749,013	781,583	4,487,372	5,969,137	7,058,662	1,632,704	516,352	3,766,946	27,571,091
(s.d. of OCs in DC)	-	(136,592)	(48,342)	(89,099)	(282,479)	(225,947)	(50,543)	(6,751)	(96,272)	(242,630)
	Country-level data (World Bank, Barro and Lee (2013) and Melitz and Toubal (2014)									
common language index (CLI)	-	0.686	0.476	0.555	0.629	0.416	0.348	0.688	0.541	0.462
	-	(0.351)	(0.331)	(0.300)	(0.206)	(0.267)	(0.169)	(0.343)	(0.299)	(0.284)
common official language	-	0.462	0.071	0.391	0.316	0.044	0.056	0.516	0.338	0.291
	-	(0.503)	(0.259)	(0.493)	(0.471)	(0.206)	(0.232)	(0.504)	(0.475)	(0.455)
common spoken language	-	0.458	0.076	0.392	0.429	0.239	0.138	0.520	0.372	0.302
	-	(0.454)	(0.197)	(0.299)	(0.236)	(0.251)	(0.073)	(0.454)	(0.339)	(0.324)
common native language	-	0.371	0.043	0.100	0.047	0.025	0.054	0.423	0.101	0.112
	-	(0.429)	(0.175)	(0.183)	(0.106)	(0.124)	(0.074)	(0.412)	(0.257)	(0.233)
log of distance	-	8.704	8.891	8.831	6.998	7.920	7.448	8.547	8.257	8.858
	-	(1.052)	(0.653)	(0.653)	(0.952)	(1.031)	(0.773)	(0.782)	(0.982)	(0.562)
% of DC-contiguous OCs	-	0.192	0.190	0.043	0.316	0.103	0.111	0.065	0.014	0.015
% of common colonizer OCs	-	0.000	0.000	0.000	0.000	0.000	0.056	0.000	0.000	0.000
common religion	-	0.499	0.373	0.199	0.349	0.147	0.057	0.447	0.122	0.151
GDP pc in OC (as % of DC)	-	267%	269%	43%	76%	31%	33%	357%	44%	29%

Notes: Individual-level descriptive statistics obtained for sample of salaried workers aged 18-64 years old. Migrants defined as recent if resident in hosting economy for less than 5 years. The linguistic proximity index following Melitz and Toubal (2014), except that we standardize the distribution by common spoken language rather than common native language indicators. All estimations were obtained also for the original Melitz and Toubal (2014) specification standardized by common native language, these robustness results are reported in Appendix D. Standard errors in parentheses (where applicable).

The sample of destination countries allows us to study quite diverse cases in terms of immigration prevalence, years since migration, and sociodemographic characteristics such as education and age. We have countries with long and short traditions of immigration (high average value of years since immigration indicator), attracting mostly highly skilled migrants and quite the opposite. Even in terms of age, there are countries where migrants are younger than native salaried workers, as well as the opposite. Perhaps more relevant for this study, the sample of destination countries is characterized by quite diverse migration patterns in terms of geographic proximity (e.g. contiguity or geographical distance) and common culture (proxied here by religion and common colonial past).

5 Results

We present results in four substantive parts: the dispersion of returns to tertiary education among migrants, the estimates at individual level, which reflect the link between linguistic proximity and the position of migrants with tertiary education in the wage distribution, and the estimates at country level, which reflect the link between linguistic proximity and the magnitude of deviation of return to tertiary education among migrants (enhanced with a robustness check for effect sizes across the wage distribution). These analyses look only at the population of salaried workers, aged 18-64. Note that individuals who obtained their education in the hosting economy are considered native workers, regardless of their country of birth. Our estimates therefore purposefully identify the foreign origin of a tertiary degree.

5.1 The dispersion of returns to tertiary education among migrants ($\hat{\gamma}_{d,o,t}$ estimates)

The first step in our two-step approach, i.e. the estimation of equation (2), yields estimates of returns to skills for individuals with tertiary education, across origin countries, relative to natives in the respective destination countries. Since we estimate the wages in logs, the coefficients have a clear interpretation: It is the percentage wage deviation of migrants from the average wage of native university graduates, broken down by country of origin. Tertiary education per se is estimated relative to individuals without tertiary education. For example, consider that a tertiary-educated native earns 20 percent more than a native without a tertiary degree. Further, consider that for a specific country of origin in the given destination country, the estimate $\hat{\gamma} = -10\%$ is equivalent to stating that an immigrant earns 10 percent less than a native with the same education. We do not interpret the estimates of $\gamma_{d,o,t}$ as causal measures of wage discrimination against particular groups of migrants - we treat them as what they are analytically: measures of the dispersion of wages across the migrants' origin countries.

Table 2 reports the basic features of the obtained $\hat{\gamma}_{d,o,t}$ estimates from equation (2) on nine utilized individual-level datasets. We obtain in total 1,570 estimates of $\hat{\gamma}_{d,o,t}$ for 153 different countries of origin. Overall, roughly 89 percent of these $\hat{\gamma}_{d,o,t}$ estimates are statistically significant, i.e. the estimated $\hat{\gamma}_{d,o,t}$ coefficient has a p-value lower than 10 percent. Note that there are two reasons for an insignificant estimate: it is either insufficiently precisely estimated (true $\gamma_{d,o,t} \neq 0$, but we fail to reject the null hypothesis due to insufficient power relative to the variation in the data) or it is actually zero (true $\gamma_{d,o,t} = 0$). The latter implies that there is no wage premium/penalty for workers (of certain education level) from a given country of origin in a given destination country. The former implies that estimations with significant $\hat{\gamma}_{d,o,t}$ will be missing distribution mass around 0. This bias may not be large, though, as indicated by the overwhelming majority of estimates that are statistically significant. The outcome of the first step in our two-step procedure thus suggests

that indeed for the vast majority of origins and destinations, there seem to be systematic deviations of migrant wages relative to the wages of native workers.

Table 2: Number of $\hat{\gamma}_{d,o,t}$ estimates by country of destination

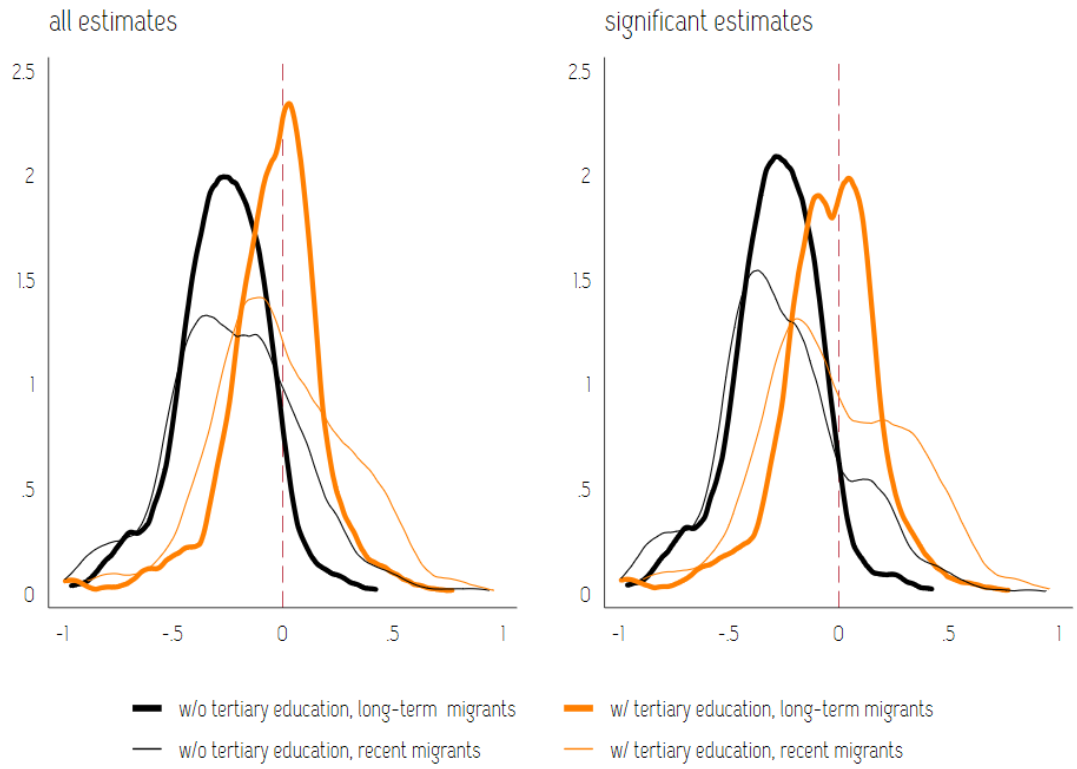
Destination	# of estimates			# of significant estimates			# of countries
	Total	Positive	Negative	Total	Positive	Negative	of origin
Argentina	151	36	115	96	11	85	28
Brazil	245	121	124	227	112	115	44
Canada	146	70	76	141	68	73	25
France	117	46	71	92	30	62	20
Germany	396	85	311	255	34	221	72
Israel	118	55	63	106	50	56	22
Mexico	164	24	140	143	17	126	33
UK	454	159	295	437	150	287	77
USA	839	143	696	831	138	693	140
Total	2630	739	1891	2328	610	1718	153

Notes: This table reports the number of estimates of the $\hat{\gamma}_{d,o,t}$ for all the countries of destination listed in Table A2. The reference group is a native worker with a tertiary degree, thus a positive coefficient signifies that a migrant worker with a tertiary degree earns more than an otherwise identical native worker. We report jointly the estimates for recent migrants and long-term migrants. In principle, the total number of estimates should be six times the number of the countries of origin (with and without a tertiary degree, for recent migrants, long-term migrants, and for both groups together). For example, in the case of Brazil, 44 countries of origin imply 264 estimates in total. For some countries of origin and education levels, the coefficient $\hat{\gamma}_{d,o,t}$ could not be identified (insufficient degrees of freedom), which reduces the final sample to 245 estimated coefficients.

The deviations we find are not only statistically significant, but also economically relevant. Figure 1 visualizes the raw distribution of the $\hat{\gamma}_{d,o,t}$ estimates obtained from equation (2). In line with the results reported in Table 2, the distributions are remarkably dispersed and skewed toward negative values. We report two distributions: the one on the left-hand side provides the distribution of all estimates we obtained, while the graph on the right-hand side visualizes the significant estimates of $\hat{\gamma}_{d,o,t}$ only. Consequently, the estimates in the graph on the right-hand side are more dispersed since there are fewer estimates to be found in the range close to zero. Our results confirm that the vast dispersion in terms of wages between origin countries is not a unique feature of the US (Butcher 1994), but is a fairly general phenomenon.

Incidentally, Figure 1 also documents that the dispersion of $\hat{\gamma}_{d,o,t}$ is smaller for migrants without tertiary education than for migrants with tertiary education: kernel density estimates for migrants without tertiary education (marked in black) are thinner than those for migrants with tertiary education (marked in blue). We quantify this observation with the use of ANOVA analysis. We perform two independent variance decompositions: on the coefficient of the migrant dummy (M_COO) and on the estimates $\hat{\gamma}_{d,o,t}$. In both cases, we use the same set of components: origin country fixed effects, destination country fixed effects, and migrant sample (long-term versus recent versus jointly estimated for all migrants). We report these results in Table A4, which illustrates the following observations. First, the variance of immigrant dummy estimates is considerably lower than for the migrants with tertiary education, but it is also more idiosyncratic. In fact, there is a high fraction in the variation of the estimates of $\hat{\gamma}_{d,o,t}$ originating from destination countries, almost 56 percent (compared to 34 percent on a lower total variation for the immigrant dummy). We also show that destination-by-origin variation essentially trumps the effects attributable to destination countries, but not due to the origin countries. We interpret this analysis of variance as an indication that even if some hosting economies have particular tastes for migrants from some origin countries, this

Figure 1: The returns to tertiary education for migrants ($\hat{\gamma}_{d,o,t}$) are dispersed



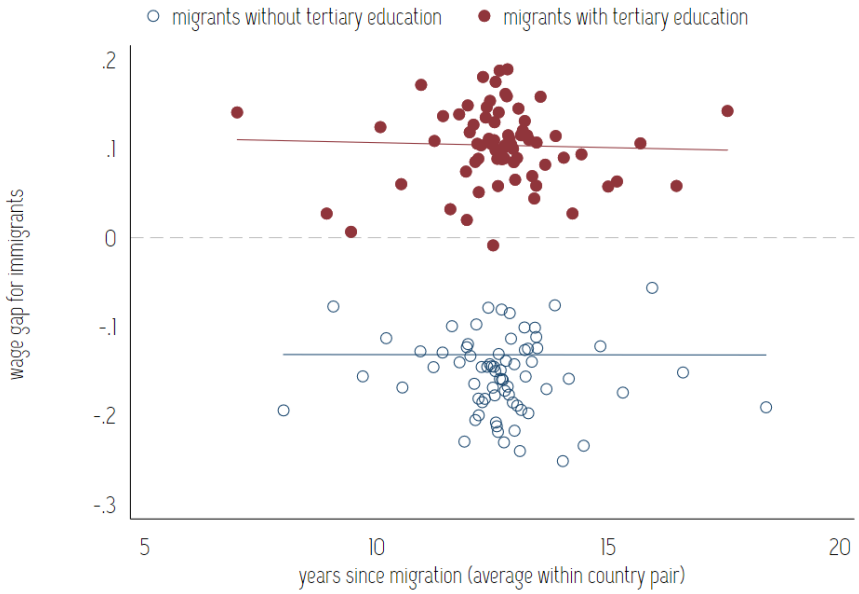
Notes: The figures report the kernel density estimates for the distribution of $\hat{\gamma}_{d,o,t}$ from equation (2) for the full sample reported in Table A2. Blue lines show the distribution of estimates for highly (i.e. tertiary) educated migrants, and black lines for non-tertiary educated migrants. Bold lines represent the estimates for non-recent migrants, thin lines the ones for recent migrants. Recent migrants are individuals who report residing in the country of destination for less than 5 years. Long term migrants refer to individuals residing for 5 years or longer in the given destination country.

preference is certainly not enough to explain the dispersion of the estimates obtained.

Our destination countries differ in their historical patterns of migration. In a similar manner, data from destination countries span periods of diverse duration. One potential explanation for migrant wage gap is their command of local language, with research showing that successful integration into the hosting labor market usually involves learning the local language. This channel is independent of the ones proposed in our study: linguistic proximity generating lower frictions for the employers to inspect the qualification of migrant job applicants. In Figure 2 we scatter the estimates of the migrant wage gap as discussed above against the characteristic of the migrant population in a given destination country from a given country of origin. We particularly focus on the duration of stay in the destination country. We rely on binned scatters, residualizing both the wage gaps and the average year since migration on country pair fixed effects. There appears to be no correlation on average between the average length of stay of migrants from a given origin country in a given destination country and the wage gaps experienced in the job market by these migrants. In other words, there can be important effects of duration of stay on individual wages, but there is no correlation between the average duration of stay in a country pair and the estimated migrant wage gap.

Stereotyping and mental accounting lead to a form of a "halo effect", where migrants from a given country of origin are lumped together to represent one stereotype – distinct from natives as

Figure 2: There is no link between migrant wage gap and duration of stay



Notes: The figure reports the binned scatter of $\hat{\gamma}_{d,o,t}$ from equation (2) for the full sample reported in Table A2 against a measure of average number of years since migration to a destination country for each country of origin. The scatters are residualized for country pairs.

well as migrants from another origin (e.g. Guiso et al. 2009, Rydgren and Ruth 2013). In Figure 3 we portray on the horizontal axis the coefficient on the M_COO migrant dummy (i.e., the estimated effect for migrants without a tertiary degree for a given country pair) and on the vertical axis the estimated $\hat{\gamma}_{d,o,t}$ that is the estimated effects for tertiary educated migrants from the same pair. In fact, we find no significant correlation for long-term migrants and a weak positive correlation of 0.17 for recent migrants (with a standard error of 0.039).[18] The strength of the correlation is heterogeneous between the nine countries of destination studied.

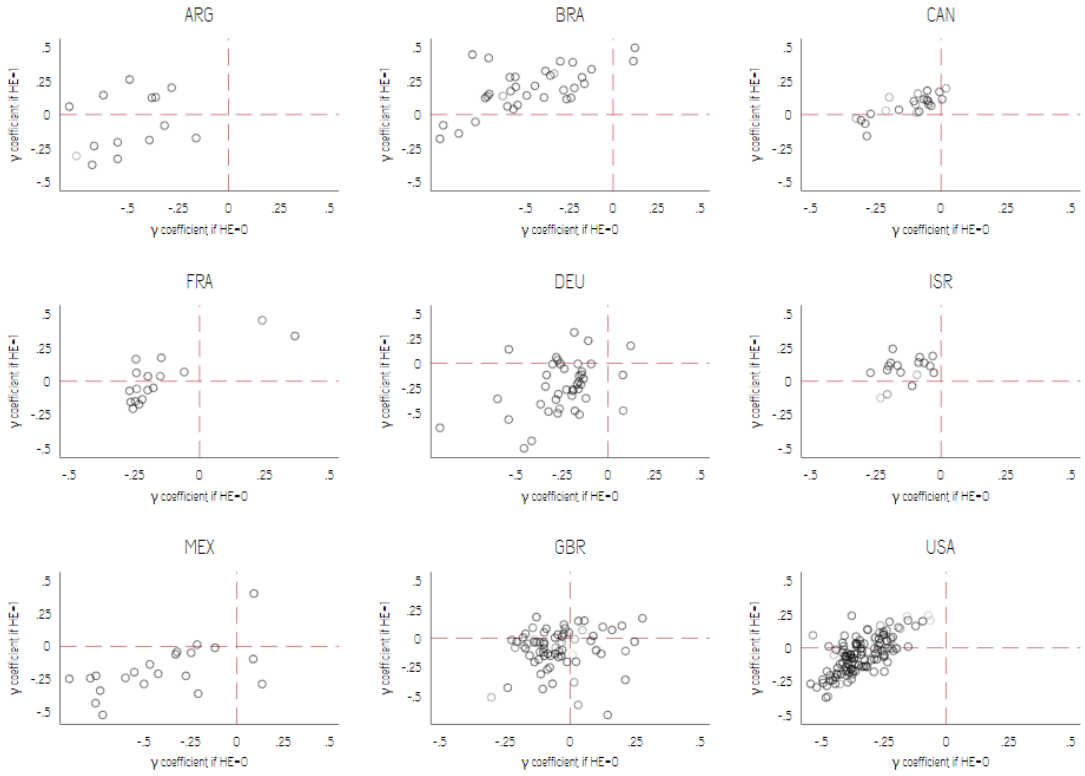
5.2 Individual-level analysis

We estimate equation (1) with a linear model. $TE_i \times$ linguistic proximity$_{d,o}$ is the variable of interest in our study. While the variation in TE_i occurs at the individual level, CLI$_{d,o}$ has only the variation at the country pair level. To trust the standard errors of the parameter estimates for CLI$_{d,o}$ we need to cluster standard errors at the level of country pair, but then the standard errors are overstated for all the parameters that are characterized by variation at the level of the individual, the origin country, or the destination country. Note that this modeling choice does not affect the estimates of the parameters, but only their standard errors. Table 3 reports our results.

We find that among native workers, tertiary education is associated with a wage percentile roughly 9.4 points higher than with secondary education or less, after adjusting for occupation and sector of employment. This result is roughly consistent with standard estimates of Mincerian wage regressions, which typically report approximately 20 percent wage gain due to tertiary education. Column (2) shows that among migrants, linguistic proximity improves the position in wage distribution. The effect is stronger for migrants with tertiary education. This result is robust to clustering standard errors at

[18] This is an unconditional correlation coefficient, with fixed effects for the destination country.

Figure 3: The correlation between wage gaps for TE migrants and migrants without TE is weak



Notes: The figures report the significant estimates $\hat{\gamma}_{d,o,t}$ from equation (2) for the full sample reported in Table A2, obtained jointly for recent and long-term migrants. The reference level for the two reported dummy coefficients is a tertiary-educated native. The shade of the circles denotes linguistic proximity (darker shade akin to more similar languages).

a country-pair level. In columns (3)-(4) we add a battery of origin and destination country controls, and the results remain of the same economic magnitude and statistical significance. In the bottom panel of Table 3, we compare migrants, using natives without tertiary education as the reference group. Note that in our specifications, the **M** dummy is insignificant, which can be attributed to the inclusion of adjustment for the number of years since immigration. This variable equals age for natives and measures time since arrival in the destination country for migrants, capturing differences in wage distribution positions between migrants and natives. Linguistic proximity proves significant in the most demanding specification in column (4) of the bottom panel. We find positive effects for migrants without tertiary education, and additional positive effects for migrants with tertiary education. Note that our estimates adjust for industry and occupation, which implies that the interpretation of these two estimates is *within* the same occupation and industry rather than on average.

We find that greater linguistic proximity between the origin and destination country is associated with a significantly higher percentile position in the wage distribution, in particular for migrants with a tertiary degree (though the effects are not very precisely estimated).

To interpret the magnitude of 7.5 percentiles for the linguistic proximity (as measured by CLI) from our preferred specification in column (4), we perform the following exercise: With a wave of migration from Russia to Israel in the 1990s, Russian has become a significant common spoken

Table 3: One-step individual level regression

VARIABLES	(1)	(2) (1) + $CLI_{d,o}$	(3) (2) + Z_0	(4) (3) + $Z_{d,o}$
	Panel A: Sample of immigrants only			
TE	9.987***	8.781***	8.923***	8.963***
	(0.051)	(0.858)	(0.853)	(0.847)
Linguistic proximity				
for migrants w/o TE		5.668***	5.823***	7.453***
		(1.402)	(1.404)	(1.739)
additionally for migrants w/ TE		3.451**	3.221**	3.097**
		(1.388)	(1.391)	(1.385)
Observations	1,490,240	1,490,240	1,490,240	1,490,240
R-squared	0.385	0.388	0.389	0.390
	Panel B: Full sample			
TE	9.423***	9.423***	9.519***	9.708***
	(0.013)	(0.735)	(0.735)	(0.711)
Migrants	-0.374***	-0.374	-0.274	-0.050
	(0.012)	(0.468)	(0.465)	(0.485)
Linguistic proximity				
for migrants w/o TE		0.016	0.232	3.326**
		(1.795)	(1.800)	(1.625)
additionally for migrants w/ TE		2.400**	2.276**	1.543+
		(1.133)	(1.129)	(1.006)
Observations	42,513,808	42,513,808	42,513,808	42,513,808
R-squared	0.364	0.364	0.365	0.367
Marital status, YSI, gender	Yes	Yes	Yes	Yes
COO FE	Yes	Yes	Yes	Yes
COD FE	Yes	Yes	Yes	Yes
Origin X's	No	No	Yes	Yes
Pair X's	No	No	No	Yes

Notes: Linguistic proximity measured as common language index, following Melitz and Toubal (2014), with the exception that we standardize the final measure by common spoken language rather than common native language. Sensitivity analysis with common native language as standardizing variable reported in Table A6 in the Appendix D. Standard errors clustered at the level of destination-by-origin country pairs in parentheses. $\hat{w}$ is the dependent variable, which is the destination country-specific wage percentile for individual i as per equation (1). Individual controls (available upon request) include age, age squared, gender, marital status, no. of children, occupation, industry, and years since immigration. Origin country controls (available upon request) include GDP per capita (PPP adjusted), fertility, mortality, and population size, merged by the year of arrival at a destination country for each individual i. Country-pair controls (available upon request) include the geographical distance between origin and destination countries, as well as contiguity dummy, common colonizer dummy, years at war measure, common religion, and common legal system (constant over time). Conventional levels of statistical significance are denoted by asterisk: ***, **, *, and + denote $p < 0.01$, $p < 0.05$, $p < 0.10$, and $p < 0.15$, respectively.

language (CSL) for the country-pair: an estimated 12.25 percent of the population in Israel can communicate effectively in Russian as a spoken language.[19] The CLI value we observe for the country pair is 0.138. We evaluate a counterfactual value of CLI in our sample, as if this wave of immigration had not occurred, i.e. as if the share of Russian speakers in Israel would be essentially close to zero. In other words, we calculate the country-pair CLI value for a scenario in which Russian

[19]The Russian diaspora in Israel accounts for almost all of the reported CSL value of the country pair.

would be an irrelevant language in Israel. We obtain the value of $CLI = 0.018$. Our estimates in Table 3 on the influence of CLI then imply that, on average, migrants from Russia with a tertiary degree can expect to be about 1 percentile higher in Israel's wage distribution in Israel nowadays, once one in eight citizens in Israel can easily communicate with them. Note that in this regard, our estimates are in line with the study of Eckstein and Weiss (2004).

Another intuitive example is provided by the differences among Nordic languages: we compare Norwegian, Swedish, and Danish relative to contemporaneous German. The ASJP measure of linguistic proximity reports a relative linguistic proximity of 0.43 for Norwegian, and roughly 0.36 for Swedish and Danish. If Swedish or Danish were linguistically as similar to German as Norwegian is, then the average wages of migrants with a university degree from those two countries would rank roughly 1 percentile higher in the relative wage distribution in Germany, *ceteris paribus*.

The migrant dummy continues to be negative in the more comprehensive specifications, but in terms of magnitude, it declines by roughly a factor of two. The same holds for the relative wages of migrants with a tertiary degree, relative to natives without a tertiary degree. Naturally, these estimates should be evaluated against standard errors of the magnitudes similar to those reported in column (1), because these are individual-level dummy variables, whereas columns (2)-(4) cluster standard errors at origin-by-destination country level.

5.3 Country-level analysis

The country-level analysis is based on a two-step procedure: we first obtained estimates of returns to tertiary education in each destination country for each observed origin country (see section 5.1) and subsequently utilize these obtained measures of dispersion in rewards to human capital as explained variables. Whereas in Table 3 the explained variable was in percentiles of (within the destination country) wage distribution, in our two-step country analysis the explained variable is the immigrant wage gap, expressed in percent of average wages. The results are reported in Table 4. The estimated coefficients report the effects relative to natives without tertiary education. We find that migrants with the same level of educational attainment earn less than natives if they come from countries with higher linguistic proximity, whereas migrants with tertiary education earn more. The premium for linguistic proximity among tertiary educated migrants does not fully compensate the gap between tertiary educated natives and tertiary educated migrants, but it helps to narrow it.

Table 4: Migrant wage gaps: two-step country-level regression

Specification			$p-value<0.15$	$w=[t-stat]$		$w=[migr_{o,d}]$	
	(1)	(1a)	(2)	(3)	(3a)	(4)	(4a)
TE	0.206***	0.206***	0.232***	0.251***	0.251***	0.216***	0.216***
	(0.021)	(0.023)	(0.022)	(0.016)	(0.024)	(0.020)	(0.026)
Linguistic proximity							
migrants w/o TE	-0.025	-0.025	0.001	-0.052	-0.052	0.003	0.003
	(0.045)	(0.053)	(0.048)	(0.048)	(0.072)	(0.044)	(0.066)
migrants w/ TE	0.108***	0.108***	0.098***	0.103***	0.103***	0.068***	0.068*
	(0.036)	(0.040)	(0.038)	(0.028)	(0.044)	(0.033)	(0.044)
Observations	1,447	1,447	1,260	1,447	1,447	1,447	1,447
R-squared	0.521	0.521	0.576	0.767	0.767	0.552	0.552
OC X's	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DC FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country pair X's	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country pair FE's	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Linguistic proximity measured as CLI (common language index), following Melitz and Toubal (2014), with the exception that we standardize the final measure by common spoken language rather than common native language, a robustness check with common native language standardization reported in Table A7 in the Appendices. Standard errors clustered at the level of the destination country. $\hat{\gamma}_{o,d,t}$ is the dependent variable, which is the destination-by-origin country-specific wage gap for migrants, with and without tertiary education, relative to natives without tertiary education, as per equation (3). Origin country controls (available upon request) include GDP per capita (PPP adjusted), and population size, merged by the year of arrival at a destination country for each individual i. Country-pair controls (available upon request) include the geographical distance between origin and destination countries, as well as contiguity dummy, common colonizer dummy, years at war measure, common religion, and common legal system (constant over time). Conventional levels of statistical significance are denoted by asterisk: ***, **, and * denote $p<0.05$, $p<0.10$, and $p<0.15$, respectively. Standard errors in parentheses.

Since the explained variable itself is estimated from individual-level data, we apply bootstrapping to adjust the estimators, these bootstrapped specifications are denoted by a in Table 4. The procedure of bootstrapping addresses the random nature of the explained variable, but still treats all observations equally. Meanwhile, highly statistically significant estimates of $\gamma_{d,o,t}$ are naturally more reliable than the insignificant ones. In a similar spirit, the estimates of $\gamma_{d,o,t}$ which refer to large migration flows may be of more policy relevance than those that refer to rare and quantitatively negligible migration flows. We thus introduce two separate weighting mechanisms: by $t-statistic$ of the obtained $\gamma_{d,o,t}$ estimator and by the size of the bilateral migration flow (scaled by the size of the sending country).

While weighting by $t-statistic$ leverages our certainty about our estimates, weighting by migration flows yields a picture of reality that adjusts for the importance of migration flows: it weights our estimates by the relative number of migrants from a certain COO in a COD, i.e. mirroring the actual migrant counts reported by the United Nations. This should further minimize any potential biases due to differences in the sampling techniques used in the underlying datasets – further increasing overall reliability. Our preferred specification is reported in column (4a) of Table 4. The other specifications permit identifying the role of weighing and bootstrapping against the raw estimation in column (1). All specifications include control factors for the origin countries, fixed effects for the destination countries, and control factors for origin-by-destination country pairs. Finally, in all specifications, we cluster standard errors by origin country.

Generally, migrants with a tertiary degree earn approximately 20 percent more than migrants without it, as reported in the first row of Table 4 (the reference level for both groups is native workers). There is no effect of linguistic proximity for the estimates of the wage gaps of migrants without tertiary education and a large positive correlation for individuals with tertiary education. In other words, wages of migrants with a tertiary degree increase in linguistic proximity (relative to native workers). The effect is relatively large, approximately 10 percent. The estimates fall to 7 percent, when we reweigh the estimation by the size of bilateral migration flows. To put this number

in perspective, an immigrant with a tertiary degree from a linguistically close country (CLI close to 1) will earn 7 percent more than an immigrant with an equal educational achievement obtained in a linguistically distant country (CLI close to 0), according to our preferred specification (4a).

6 Discussion and conclusions

Immigrant wage gaps are well documented in the literature, but their variation across origin countries has been less studied. Two hypotheses were offered in the literature, related to skill portability and knowledge of the hosting country language. We propose to distinguish between skills (which can be positively correlated with the duration of stay in the hosting economy) and the linguistic proximity between the origin and destination country. We propose that employers may find it harder to assess foreign credentials of job candidates from linguistically distant countries. We tested this hypothesis empirically. We exploit individual-level data from nine destination countries for many origin countries and document the inequality in returns to education between origin countries. We relate this inequality to linguistic proximity between the origin and destination countries, finding that lower linguistic proximity between origin and destination correlates with higher wage gaps for skilled migrants.

Our results demonstrate that linguistic proximity partially explains the dispersion of returns to foreign tertiary education among migrants. In other words, the costs to employers of inspecting the qualifications of job candidates who obtained their education in a foreign country prove to be non-negligible, and their variation correlates with linguistic proximity. Employers may struggle to assess the qualifications of job candidates educated abroad. This friction is separate from previously considered frictions such as language skills of job candidates or the pure portability of skills across borders.[20]

Our study is not without other caveats. First, we work with worker-level data, unable to study employers and their behavioral patterns. We work with wages, thus employment contracts where the friction was not prohibitively high. Further research would help validating the role of the frictions on the side of the employer. Second, *migrant selectivity* could be relevant on several levels: self-selection in destination countries (Adsera and Pytlikova 2015, Bredtmann et al. 2020), return migration decision (Dustmann and Görlach 2016), and task specialization (Peri and Sparber 2009). Addressing these selectivity patterns is notoriously difficult, as it requires data from both sending and origin country for the same individuals. Third, we cannot address the hypothesis of *enclaves*, because our data is insufficient for geographical location. Fourth, although our two-step approach delivers estimates of the migrant wage gap, the gaps used in this study do not have a direct causal interpretation. We merely argue that methodologically comparable gaps adjusted for individual characteristics are lower for migrants arriving in linguistically closer countries *ceteris paribus*. So long as the destination countries' selectivity bias is constant across the sending countries, our estimates remain reliable. If destination country selectivity was pair-wise rather than destination-wise, our estimates could potentially confound selectivity with linguistic proximity. Migrants may also self-select into production tasks that tend to be less linguistically intensive than typical tasks preformed by

[20] Our mechanism is distinct also from discrimination per se, or cultural and social norms. Some earlier literature has argued that migrants' employment prospects depend on natives' trust towards origin countries (Keita and Valette 2019), which is typically lower the greater the geographic distance (Cettolin and Suetens 2019). Earlier literature argues that social, ethnic, and cultural distance shape natives' attitudes towards migrants from certain origin countries (O'Rourke and Sinnott 2006). In the labor context, the authors report discrimination in hiring decisions based on the geographic origin of migrants (e.g. Pager 2007, Bertrand and Duflo 2017, Neumark 2018, Lancee 2021). Our study proposes an additional source of immigrant wage gaps.

native workers (Peri and Sparber 2009). This division of labor may be the basis for the migrant wage gap per se (D'Amuri and Peri 2014).

Our results lend tentative support to several policy implications. For sending countries, in order to improve emigration outcomes of their citizens, it is useful to provide rigorous and well-structured curriculum in international language to the graduates. For receiving countries, labor market efficiency could be improved if information friction related to inspecting skills of foreign educated individuals is eliminated. Identifying effective ways to do so remains a fruitful area of further research.

References

Adserà, A. and Ferrer, A.: 2021, Linguistic proximity and the labour market performance of immigrant men in canada, *Labour* **35**(1), 1–23.

Adsera, A. and Pytlikova, M.: 2015, The role of language in shaping international migration, *Economic Journal* **125**(586), 49–81.

Bakker, D., Müller, A., Velupillai, V., Wichmann, S., Brown, C. H., Brown, P., Egorov, D., Mailhammer, R., Grant, A. and Holman, E. W.: 2009, Adding typology to lexicostatistics: A combined approach to language classification, *Linguistic Typology* **13**(1), 169–181.

Bar-Haim, E. and Birgier, D. P.: 2024, Language distance and labor market integration of migrants: Gendered perspective, *Plos one* **19**(4), e0299936.

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.

Bazzi, S., Gaduh, A., Rothenberg, A. D. and Wong, M.: 2016, Skill transferability, migration, and development: Evidence from population resettlement in Indonesia, *American Economic Review* **106**(9), 2658–98.

Beenstock, M., Chiswick, B. R. and Repetto, G. L.: 2001, The effect of linguistic distance and country of origin on immigrant language skills: Application to Israel, *International Migration* **39**(3), 33–60.

Beine, M., Docquier, F. and Ozden, C.: 2011, Diasporas, *Journal of Development Economics* **95**(1), 30 – 41.

Belot, M. and Ederveen, S.: 2012, 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.

Berman, E., Lang, K. and Siniver, E.: 2003, Language-skill complementarity: Returns to immigrant language acquisition, *Labour Economics* **10**(3), 265 – 290.

Bertrand, M. and Duflo, E.: 2017, Field experiments on discrimination, *in* A. V. Banerjee and E. Duflo (eds), *Handbook of Economic Field Experiments*, Vol. 1, Elsevier, Amsterdam, pp. 309–393.

Bleakley, H. and Chin, A.: 2004, Language skills and earnings: Evidence from childhood immigrants, *The Review of Economics and Statistics* **86**(2), 481–496.

Bleakley, H. and Chin, A.: 2010, Age at arrival, English proficiency, and social assimilation among US immigrants, *American Economic Journal: Applied Economics* **2**(1), 165–192.

Borjas, G. J.: 1985, Assimilation, changes in cohort quality, and the earnings of immigrants, *Journal of Labor Economics* **3**(4), 463–489.

Borjas, G. J.: 1987, Self-selection and the earnings of immigrants, *American Economic Review* **77**(4), 531–553.

Borjas, G. J.: 1992, National origin and the skills of immigrants in the postwar period, *in* G. J. Borjas and R. B. Freeman (eds), *Immigration and the Workforce: Economic Consequences for the United States and Source Areas*, University of Chicago Press, Chicago, IL, pp. 17–48.

Borjas, G. J.: 1994, The economics of immigration, *Journal of Economic Literature* **32**(4), 1667–1717.

Borjas, G. J.: 2015, The slowdown in the economic assimilation of immigrants: Aging and cohort effects revisited again, *Journal of Human Capital* **9**(4), 483–517.

Borjas, G. J. and Friedberg, R. M.: 2009, Recent trends in the earnings of new immigrants to the United States. NBER Working Paper No. 15406.

Bredtmann, J., Nowotny, K. and Otten, S.: 2020, Linguistic distance, networks and migrants' regional location choice, *Labour Economics* **65**, 101863.

Brown, C. H., Holman, E. W., Wichmann, S. and Velupillai, V.: 2008, Automated classification of the world's languages: A description of the method and preliminary results, *STUF–Language Typology and Universals* **61**(4), 285–308.

Butcher, K. F.: 1994, Black immigrants in the United States: A comparison with native blacks and other immigrants, *ILR Review* **47**(2), 265–284.

Caragliu, A., Del Bo, C., de Groot, H. L. and Linders, G.-J. M.: 2013, Cultural determinants of migration, *Annals of Regional Science* **51**(1), 7–32.

Cettolin, E. and Suetens, S.: 2019, Return on trust is lower for immigrants, *The Economic Journal* **129**(621), 1992–2009.

Chen, M. K.: 2013, The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets, *American Economic Review* **103**(2), 690–731.

Chiswick, B.: 1978, The effect of Americanization on the earnings of foreign-born men, *Journal of Political Economy* **86**(5), 897–921.

Chiswick, B. R.: 1986, Is the new immigration less skilled than the old?, *Journal of Labor Economics* **4**(2), 168–192.

Chiswick, B. R.: 1991, Speaking, reading, and earnings among low-skilled immigrants, *Journal of Labor Economics* **9**(2), 149–170.

Chiswick, B. R. and Miller, P. W.: 1994, Language choice among immigrants in a multi-lingual destination, *Journal of Population Economics* **7**(2), 119–131.

Chiswick, B. R. and Miller, P. W.: 1995, The endogeneity between language and earnings: International analyses, *Journal of Labor Economics* **13**(2), 246–288.

Chiswick, B. R. and Miller, P. W.: 1998, English language fluency among immigrants in the United States, *Research in Labor Economics* **17**, 151–200.

Chiswick, B. R. and Miller, P. W.: 2001, A model of destination-language acquisition: Application to male immigrants in Canada, *Demography* **38**(3), 391–409.

Chiswick, B. R. and Miller, P. W.: 2005, Linguistic distance: A quantitative measure of the distance between English and other languages, *Journal of Multilingual and Multicultural Development* **26**(1), 1–11.

Chiswick, B. R. and Miller, P. W.: 2012, Negative and positive assimilation, skill transferability, and linguistic distance, *Journal of Human Capital* **6**(1), 35–55.

Chiswick, B. R. and Miller, P. W.: 2013, The impact of surplus skills on earnings: Extending the over-education model to language proficiency, *Economics of Education Review* **36**, 263–275.

Clark, X., Hatton, T. J. and Williamson, J. G.: 2007, Explaining U.S. Immigration, 1971-1998, *Review of Economics and Statistics* **89**(2), 359–373.

Crystal, D.: 1987, *The Cambridge Encyclopedia of Language*, Cambridge University Press, Cambridge.

Dale-Olsen, H. and Finseraas, H.: 2020, Linguistic diversity and workplace productivity, *Labour Economics* **64**, 101813.

Desmet, K., Weber, S. and Ortuño-Ortín, I.: 2009, Linguistic diversity and redistribution, *Journal of the European Economic Association* **7**(6), 1291–1318.

Doeringer, P. B. and Piore, M. J.: 1985, *Internal labor markets and manpower analysis*, ME Sharpe.

Dustmann, C.: 1994, Speaking fluency, writing fluency and earnings of migrants, *Journal of Population Economics* **7**(2), 133–156.

Dustmann, C. and Fabbri, F.: 2003, Language proficiency and labour market performance of immigrants in the UK, *Economic Journal* **113**(489), 695–717.

Dustmann, C. and Görlach, J.-S.: 2016, The economics of temporary migrations, *Journal of Economic Literature* **54**(1), 98–136.

Dyen, I., Kruskal, J. B. and Black, P.: 1992, An Indoeuropean classification: A lexicostatistical experiment, *Transactions of the American Philosophical society* **82**(5), iii–132.

D'Amuri, F. and Peri, G.: 2014, Immigration, jobs, and employment protection: Evidence from Europe before and during the Great Recession, *Journal of the European Economic Association* **12**(2), 432–464.

Eckstein, Z. and Weiss, Y.: 2004, On the wage growth of immigrants: Israel, 1990–2000, *Journal of the European Economic Association* **2**(4), 665–695.

Espenshade, T. J. and Fu, H.: 1997, An analysis of English-language proficiency among US immigrants, *American Sociological Review* pp. 288–305.

European Commission: 2006, Europeans and their languages. Report, Special Eurobarometer 243.

Ferrer, A., Green, D. A. and Riddell, W. C.: 2006, The effect of literacy on immigrant earnings, *Journal of Human Resources* **41**(2), 380–410.

Fortin, N., Lemieux, T. and Torres, J.: 2016, Foreign human capital and the earnings gap between immigrants and Canadian-born workers, *Labour Economics* **41**, 104–119.

Friedberg, R. M.: 2000, You can't take it with you? Immigrant assimilation and the portability of human capital, *Journal of Labor Economics* **18**(2), 221–251.

Ghio, D., Bratti, M. and Bignami, S.: 2023, Linguistic barriers to immigrants' labor market integration in Italy, *International Migration Review* **57**(1), 357–394.

Ginsburgh, V. and Weber, S.: 2020, The economics of language, *Journal of Economic Literature* **58**(2), 348–404.

Goldmann, G., Sweetman, A. and Warman, C.: 2015, The portability of new immigrants' human capital: Language, education and occupational matching, *Canadian Public Policy* **41**(Supplement 1), S64–S79.

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.: 2009, Cultural biases in economic exchange?, *The Quarterly Journal of Economics* **124**(3), 1095–1131.

Head, K., Mayer, T. and Ries, J.: 2010, The erosion of colonial trade linkages after independence, *Journal of International Economics* **81**(1), 1–14.

Hutchinson, W. K.: 2005, Linguistic distance as a determinant of bilateral trade, *Southern Economic Journal* **72**(1), 1–15.

Imai, S., Stacey, D. and Warman, C.: 2019, From engineer to taxi driver? Language proficiency and the occupational skills of immigrants, *Canadian Journal of Economics/Revue canadienne d'économique* **52**(3), 914–953.

IPUMS-I: 2020, Integrated public use microdata series, international: Version 7.3 [dataset]. **URL:** *https://doi.org/10.18128/D020.V7.3*

Isphording, I. E.: 2014, Disadvantages of linguistic origin – Evidence from immigrant literacy scores, *Economics Letters* **123**(2), 236–239.

Isphording, I. E. and Otten, S.: 2013, The costs of Babylon – Linguistic distance in applied economics, *Review of International Economics* **21**(2), 354–369.

Isphording, I. E. and Otten, S.: 2014, Linguistic barriers in the destination language acquisition of immigrants, *Journal of Economic Behavior & Organization* **105**, 30–50.

Jasso, G. and Rosenzweig, M.: 1988, How well do U.S. immigrants do? Vintage effects, emigration selectivity, and occupational mobility, *Research in Population Economics* **6**, 229–253.

Karemera, D., Oguledo, V. I. and Davis, B.: 2000, A gravity model analysis of international migration to North America, *Applied Economics* **32**(13), 1745–1755.

Kee, P.: 1995, Native-immigrant wage differentials in the Netherlands: Discrimination?, *Oxford Economic Papers* **47**(2), 302–317.

Keita, S. and Valette, J.: 2019, Natives' attitudes and immigrants' unemployment durations, *Demography* **56**(3), 1023–1050.

Kossoudji, S. A.: 1988, English language ability and the labor market opportunities of Hispanic and East Asian immigrant men, *Journal of Labor Economics* **6**(2), 205–228.

LaLonde, R. J. and Topel, R. H.: 1992, The assimilation of immigrants in the US labor market, *in* G. Borjas and R. Freeman (eds), *Immigration and the Workforce: Economic Consequences for the United States and Source Areas*, University of Chicago Press, Chicago, IL, pp. 67–92.

Lancee, B.: 2021, Ethnic discrimination in hiring: comparing groups across contexts. results from a cross-national field experiment, *Journal of Ethnic and Migration Studies* **47**(6), 1181–1200.

Lohmann, J.: 2011, Do language barriers affect trade?, *Economics Letters* **110**(2), 159–162.

Lubotsky, D.: 2007, Chutes or ladders? A longitudinal analysis of immigrant earnings, *Journal of Political Economy* **115**(5), 820–867.

Mayda, A. M.: 2010, International migration: A panel data analysis of the determinants of bilateral flows, *Journal of Population Economics* **23**(4), 1249–1274.

McManus, W., Gould, W. and Welch, F.: 1983, Earnings of Hispanic men: The role of English language proficiency, *Journal of Labor Economics* **1**(2), 101–130.

McManus, W. S.: 1985, Labor market assimilation of immigrants: The importance of language skills, *Contemporary Economic Policy* **3**(3), 77–89.

Melitz, J.: 2008, Language and foreign trade, *European Economic Review* **52**(4), 667–699.

Melitz, J. and Toubal, F.: 2014, Native language, spoken language, translation and trade, *Journal of International Economics* **93**(2), 351–363.

Mishel, L., Bivens, J., Gould, E. and Shierholz, H.: 2012, *The State of Working America*, 12th edn, Cornell University Press, Ithaca, NY.

Neumark, D.: 2018, Experimental research on labor market discrimination, *Journal of Economic Literature* **56**(3), 799–866.

O'Rourke, K. H. and Sinnott, R.: 2006, The determinants of individual attitudes towards immigration, *European Journal of Political Economy* **22**(4), 838–861.

Ortega, F. and Peri, G.: 2013, The effect of income and immigration policies on international migration, *Migration Studies* **1**(1), 47–74.

Pager, D.: 2007, The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future, *The Annals of the American Academy of Political and Social Science* **609**(1), 104–133.

Pedersen, P. J., Pytlikova, M. and Smith, N.: 2008, Selection and network effects – migration flows into OECD countries 1990–2000, *European Economic Review* **52**(7), 1160–1186.

Peri, G. and Sparber, C.: 2009, Task specialization, immigration, and wages, *American Economic Journal: Applied Economics* **1**(3), 135–69.

Piore, M. J.: 1972, Notes for a theory of labor market stratification. Massachusetts Institute of Technology (MIT), Department of Economics Working Paper No. 95. Available online: `https://dspace.mit.edu/bitstream/handle/1721.1/64001/notesfortheoryof00pior.pdf` (last accessed: 27.07.2024).

Reich, M., Gordon, D. M. and Edwards, R. C.: 1973, A theory of labor market segmentation, *American Economic Review* **63**(2), 359–365.

Rivera-Batiz, F. L.: 1990, English language proficiency and the economic progress of immigrants, *Economics Letters* **34**(3), 295–300.

Rydgren, J. and Ruth, P.: 2013, Contextual explanations of radical right-wing support in Sweden: Socioeconomic marginalization, group threat, and the halo effect, *Ethnic and Racial Studies* **36**(4), 711–728.

Schaafsma, J. and Sweetman, A.: 2001, Immigrant earnings: Age at immigration matters, *Canadian Journal of Economics* **34**(4), 1066–1099.

Sprenger, E.: 2024, What makes us move, what makes us stay: The role of language and culture in intra-EU mobility, *Journal of International Migration and Integration* **25**, 1825–1855.

Strom, S., Piazzalunga, D., Venturini, A. and Villosio, C.: 2018, Wage assimilation of immigrants and internal migrants: the role of linguistic distance, *Regional Studies* **52**(10), 1423–1434.

Swadesh, M.: 1952, Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos, *Proceedings of the American Philosophical Society* **96**(4), 452–463.

Wong, L.: 2023, The effect of linguistic proximity on the labour market outcomes of the asylum population, *Journal of Population Economics* **36**(2), 609–652.

Yao, Y. and van Ours, J. C.: 2015, Language skills and labor market performance of immigrants in the Netherlands, *Labour Economics* **34**, 76–85.

A Summary of language measures used in migration literature

Table A1: Concepts and datasets of linguistic measures used in Economics literature

Dataset	**Coverage**	**Data provided**	**Examples of literature**
World Fact Book[a]	267 world entities (mostly countries)	Lists official and spoken languages in a country (and sometimes the percentage of speakers)	[CL] (Mayda 2010, Ortega and Peri 2013)
Ethnologue: Languages of the World[b]	7,111 known living languages	Linguistic lineage, i.e., the language family	[CL] for (Karemera et al. 2000, Pedersen et al. 2008); [LP] (Desmet et al. 2009, Belot and Hatton 2012, Adsera and Pytlikova 2015, Adserà and Ferrer 2021)
World Atlas of Languages (WALS)[c]	2,679 languages	Linguistic lineage, i.e., the language family	[LP] (Lohmann 2011, Isphording and Otten 2013, Chen 2013)
Dyen percentage cognate matrix of linguistic distances (Dyen et al. 1992)	Indo-European languages	Language-pairwise linguistic proximity information, based on the vocabulary-wise cognation of 200 words from each language (see Swadesh 1952)	[LP] (Belot and Ederveen 2012, Adsera and Pytlikova 2015)
Levenshtein Linguistic Distance Matrix (Brown et al. 2008, Bakker et al. 2009)	(approx.) 3500 languages	Language-pairwise linguistic proximity information, based on the ASJP	[LP] (Isphording and Otten 2014, 2013, Isphording 2014, Adsera and Pytlikova 2015, Strom et al. 2018, Bredtmann et al. 2020, Dale-Olsen and Finseraas 2020, Ghio et al. 2023, Wong 2023)
Linguistic Distance to English (Chiswick and Miller 1998)	43 languages	Information on linguistic proximity to English (based on the difficulty Americans have in learning these languages)	[LP to English] (Chiswick and Miller 2001, 2005, Hutchinson 2005, Isphording and Otten 2013)
CEPII[d]	195 countries	Country-pairwise language information, including common official, spoken and native language, and two measures of linguistic proximity	[CL] (Grogger and Hanson 2011, Beine et al. 2011); [LP] (Melitz and Toubal 2014), (Bar-Haim and Birgier 2024), *our paper*

Notes:

[a] `https://www.cia.gov/library/publications/the-world-factbook/`

[b] `https://www.ethnologue.com/`

[c] `https://wals.info/`

[d] `http://www.cepii.fr/CEPII/en/welcome.asp`, see Melitz (2008), Melitz and Toubal (2014).

CL denotes common language. LP denotes language proximity. Isphording and Otten (2013) provides an intuitive illustration of the methodology used by Bakker et al. (2009). The algorithm that calculates the minimal Levenshtein distance is based on pronunciation and vocabulary of 40 (or 100) words (ASJP code) from each language. See also Ginsburgh and Weber (2020) on a short examination of the different measures, main advantages and shortcomings of the methods, and on more economic literature applying linguistic proximity measures.

B Construction of Common Language Index (CLI)

To examine the relevance of linguistic proximity, we draw on the common language index (CLI) measure conceptualized by Melitz (2008) and Melitz and Toubal (2014). It measures the *ease of communication* between individuals originating from two different countries (in our case, between migrant's origin and destination countries). Basically, it is based on, and adjusts for, the varying relevance of four different measures: Common official language (COL), common native language (CNL), common spoken language (CSL), and linguistic proximity (LP). CNL provides the likelihood that two randomly picked persons, one from each country, share a common mother tongue. Similarly, CSL captures the likelihood that the two persons share any common spoken language (i.e. languages not necessarily their mother tongue, but for instance a global language like English, or other).[21] LP denotes the relative linguistic closeness of two countries' native languages, based on the ASJP index.[22]

Remind that our goal is to have a measure of *ease of communication*, intended to proxy for the role of language in transferring credentials from one country to another. If two countries share a COL, the role of the remainder measures becomes relatively less relevant, as institutional factors likely provide for sufficient transferability of skills. Likewise, if either CNL or CSL is substantially high, LP should be relatively less relevant (Melitz and Toubal (2014) provide some empirical evidence for this argumentation).[23] The CLI reflects these assumptions.

The CLI is bounded by the values of 0 and 1 for the lowest and highest linguistic proximity in our data. For consturcting the index, strictly following Melitz and Toubal (2014), LP and COL are first normalized to each equal one at the sample mean. Then, we normalize $COL + LP$ by dividing it with the highest observed value among the country pairs in our sample. We multiply this series' normalized values by $1 - CSL$. Finally, we construct CLI as the sum of CSL and the adjusted sum of COL and LP (this ensures CLI cannot exceed $1 - CSL$):

$$CLI_{o,d} = (1 - CSL_{o,d}) * ((COL_{o,d} + LP_{o,d})/max_{sample}(COL + LP)) + CSL_{o,d} \quad (4)$$

Note that we deviate from Melitz and Toubal (2014) in that we use the CSL index rather than the CNL index. In table A3, we provide an overview of the top five origin countries per destination country in our sample, together with information on COL, CSL and LP for these country pairs.

[21] Any language considered in the measure must be spoken by at least 4 percent of the respective local populations.

[22] For simplicity, if there is more than one native language in a country, a weighted measure of the two most common native languages of a country were used. In the case of Canada, for instance, English has a weight of 0.70 and French a weight of 0.30 in constructing the measure - indicating both languages' relative domestic relevance in terms of population counts. Other illustrative examples include the United States (weight of 0.85 for English and 0.15 for Spanish), Israel (weight of 0.87 for Hebrew and 0.13 for Russian), or Switzerland (weight of 0.74 for German, 0.26 for French).

[23] Note that LP is denoted 0 *either* when two countries have no linguistic similarity, *or* (following the above argumentation) when CSL is 1 (as in this case linguistic proximity is completely irrelevant).

C Data summary

Table A2: Individual level datasets used in this study

Country	**Dataset**	**Years**
Argentina	Encuesta Permanente de Hogares (*EPH*)	2004-2018
Brazil	IPUMS-published census data	1991, 2000, 2010
Canada	IPUMS-published census data	1981, 1991, 2001, 2011
France	Labor Force Survey (*LFS*)	2003-2012
Germany	Socio-Economic Panel Study (*SOEP*)	1984-2016
Israel	IPUMS-published census data	1983, 1995
Mexico	IPUMS-published census data	1990, 2000, 2010
UK	Labor Force Survey (LFS)	1992-1998, 2000-2007
USA	American Community Survey (*ACS*)	2001-2018

Table A3: Top sample origin countries by destination

Destination	Origin (Top 5)	# of obs.	share in destination migrant sample	Linguistic measures		
				COL	CSL	LP
Argentina	Chile	12,687	0.30	1.00	0.98	1.00
	Paraguay	9,028	0.21	1.00	0.69	0.09
	Bolivia	7,583	0.18	1.00	0.87	0.57
	Peru	3,886	0.09	1.00	0.86	1.00
	Uruguay	2,865	0.07	1.00	0.98	1.00
Brazil	Portugal	7,455	0.23	1.00	1.00	1.00
	Paraguay	2,812	0.09	0.00	0.11	0.08
	Uruguay	2,136	0.07	0.00	0.07	0.42
	Argentina	2,053	0.06	0.00	0.06	0.42
	Italy	1,912	0.06	0.00	0.02	0.37
Canada	United Kingdom	29,141	0.18	1.00	0.85	0.74
	Philippines	14,809	0.09	1.00	0.47	0.03
	Italy	12,672	0.08	0.00	0.30	0.19
	India	11,674	0.07	1.00	0.20	0.00
	USA	11,301	0.07	1.00	0.82	0.66
France	Portugal	5,634	0.17	0.00	0.34	0.29
	Morocco	4,446	0.14	1.00	0.36	0.09
	Algeria	3,969	0.12	1.00	0.57	0.09
	Türkiye	1,318	0.04	0.00	0.07	0.11
	Tunisia	1,310	0.04	1.00	0.64	0.09
Germany	Türkiye	9,518	0.20	0.00	0.15	0.09
	Italy	5,016	0.10	0.00	0.25	0.16
	Poland	4,578	0.09	0.00	0.36	0.09
	Kazakhstan	3,460	0.07	0.00	0.16	0.10
	Greece	3,282	0.07	0.00	0.35	0.09
Israel	Russia	19,143	0.23	0.00	0.12	0.16
	Morocco	13,625	0.16	0.00	0.16	0.24
	Romania	9,448	0.11	0.00	0.15	0.10
	Ukraine	5,972	0.07	0.00	0.08	0.13
	Iraq	5,514	0.07	0.00	0.13	0.24
Mexico	USA	5,962	0.42	0.00	0.20	0.26
	Guatemala	1,875	0.13	1.00	0.85	1.00
	Spain	860	0.06	1.00	0.98	1.00
	Argentina	481	0.03	1.00	0.98	1.00
	El Salvador	462	0.03	1.00	0.99	1.00
UK	India	3,910	0.10	1.00	0.23	0.00
	Ireland	3,755	0.10	1.00	0.97	1.00
	Germany	2,674	0.07	0.00	0.61	0.32
	South Africa	1,667	0.04	1.00	0.29	0.00
	Kenya	1,417	0.04	1.00	0.07	0.05
USA	Mexico	776,056	0.25	0.00	0.20	0.26
	Philippines	172,739	0.06	1.00	0.53	0.05
	India	154,140	0.05	1.00	0.22	0.00
	China	121,734	0.04	0.00	0.01	0.05
	Vietnam	98,229	0.03	0.00	0.00	0.03

Notes: The table lists the top 5 immigrant origin countries in each destination country sample used in our study. It provides the absolute number of observations for each origin country within the destination country sample, their share among all migrants in the respective sample, and the COL, CSL, and LP values for each country-pair. We do not report CLS and CLN here, as these are specific to estimation samples.

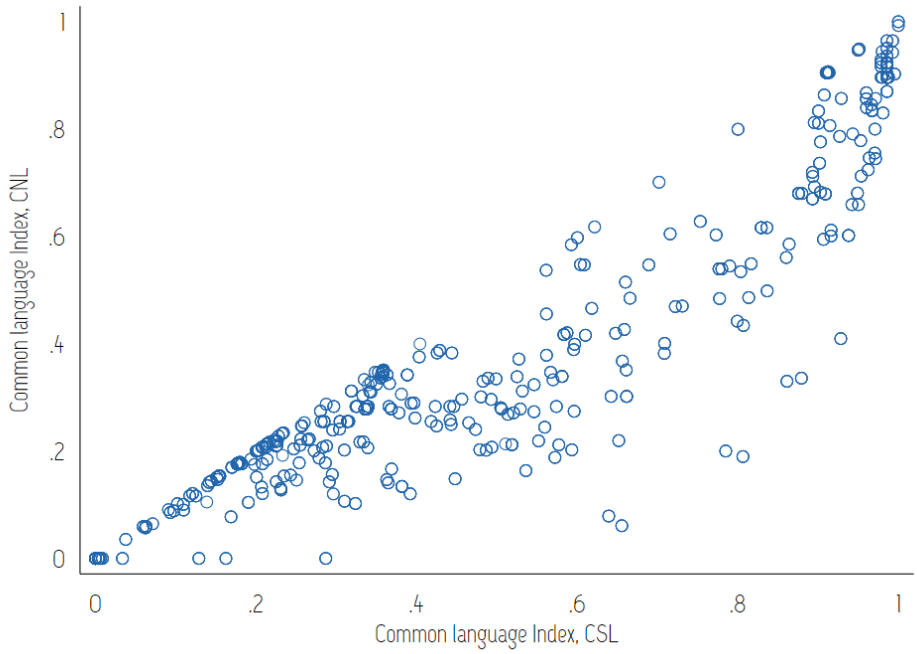
Table A4: ANOVA decomposition of variance in estimated coefficients

Estimate if $immigrant = 1$ **and**	$TE = 1$		$TE = 0$	
	(a)	(b)	(a)	(b)
Residual variance	29.11	13.33	26.39	8.73
Captured variance	50.16	65.93	34.54	52.20
due to destination country	56%	1%	34%	1%
due to origin country	40%	18%	51%	11%
due to destination-by-origin	-	24%	-	34%
due to migrant sample	1%	0%	4%	2%

Notes: specifications denoted by (a) include destination country effects and origin country effects. Specifications denoted by (b) include destination-country effects and origin-country effects as well as destination-by-origin country effects. All specifications include the migrant sample, which has one of the three levels across all specifications: (i) recent migrants; (ii) long-term migrants; (iii) joint estimation for all migrants. Recent migrants are defined as residing shorter than 5 years in a destination country. All components shown are highly significant (bootstrapped standard errors available upon request).

D Sensitivity analysis – estimation for linguistic proximity standardized by common native language

Figure A1: Common language index measures: comparing the standardization by CNL and CSL



Notes: The figure portrays the correlation between common language index measures (CLI), standardized by common native language (CNL), and by common spoken language (CSL). Melitz and Toubal (2014) standardize by common native language (CNL).

Table A5: Table 3: full set of estimates

	(1)	(2) (1) + $CLI_{d,o}$	(3) (2) + Z_o	(4) (3) + $Z_{d,o}$	(5) (4) + $schooling_o$
HE for natives	13.222***	10.974***	10.995***	11.342***	11.520***
	(0.010)	(0.060)	(0.060)	(0.060)	(0.060)
Migrants	-1.029***	-5.539***	-5.406***	-4.153***	-4.248***
	(0.039)	(0.057)	(0.057)	(0.059)	(0.059)
HE for migrants	-3.618***	-2.923***	-3.004***	-2.948***	-2.806***
	(0.042)	(0.085)	(0.085)	(0.085)	(0.086)
Linguistic proximity					
for migrants w/o HE		11.242***	10.951***	8.673***	8.763***
		(0.094)	(0.094)	(0.097)	(0.098)
for migrants w/ HE		0.344***	0.124	0.184	-0.189
		(0.133)	(0.133)	(0.133)	(0.135)
Age	2.068***	1.984***	2.004***	1.968***	1.950***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Age squared	-0.022***	-0.022***	-0.022***	-0.022***	-0.022***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Men	6.387***	6.392***	6.384***	6.384***	6.385***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Years since migration	0.103***	0.185***	0.209***	0.244***	0.211***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Years of schooling					-0.555***
					(0.006)
Constant	12.160***	28.224***	-34.395***	-50.664***	-118.666
	(0.128)	(0.175)	(1.976)	(2.010)	(201.433)
Observations	42,513,808	42,513,808	42,513,808	42,513,808	42,424,420
R-squared	0.259	0.260	0.260	0.260	0.261
Marital status	Yes	Yes	Yes	Yes	Yes
COO FE	Yes	Yes	Yes	Yes	Yes
COD FE	Yes	Yes	Yes	Yes	Yes
Origin X's	No	No	Yes	Yes	Yes
Pair X's	No	No	No	Yes	Yes

Notes: See note under Table 3

Table A6: Robustness check: Table 3, full set of estimates, standardization by CNL

	(1)	(2) (1) + $CLI_{d,o}$	(3) (2) + Z_o	(4) (3) + $Z_{d,o}$	(5) (4) + $schooling_o$
HE for natives	13.222***	11.022***	11.024***	11.189***	11.345***
	(0.010)	(0.051)	(0.051)	(0.052)	(0.052)
Migrants	-1.029***	-3.235***	-3.176***	-2.376***	-2.436***
	(0.039)	(0.051)	(0.051)	(0.052)	(0.052)
HE for migrants	-3.618***	-2.600***	-2.709***	-2.585***	-2.468***
	(0.042)	(0.073)	(0.074)	(0.074)	(0.075)
Linguistic proximity					
for migrants w/o HE		7.381***	7.221***	5.297***	5.295***
		(0.096)	(0.096)	(0.099)	(0.100)
for migrants w/ HE		0.863***	0.632***	0.418***	0.028
		(0.138)	(0.138)	(0.138)	(0.140)
Age	2.068***	1.982***	2.003***	1.976***	1.960***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Age squared	-0.022***	-0.022***	-0.022***	-0.022***	-0.022***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Men	6.387***	6.392***	6.384***	6.385***	6.385***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Years since migration	0.103***	0.185***	0.211***	0.237***	0.203***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
Years of schooling					-0.548***
					(0.006)
Constant	12.160***	24.939***	-39.336***	-53.205***	-122.244
	(0.128)	(0.154)	(1.976)	(2.009)	(201.448)
Observations	42,513,808	42,513,808	42,513,808	42,513,808	42,424,420
R-squared	0.259	0.260	0.260	0.260	0.260
Marital status	Yes	Yes	Yes	Yes	Yes
COO FE	Yes	Yes	Yes	Yes	Yes
COD FE	Yes	Yes	Yes	Yes	Yes
Origin X's	No	No	Yes	Yes	Yes
Pair X's	No	No	No	Yes	Yes

Notes: See note under Table 3

Table A7: Robustness check: Table 4 with common language index standardized by common native language

Specification			$p-value<0.15$	$w=[t-stat]$		$w=[migr_{o,d}]$	
	(1)	(1a)	(2)	(3)	(3a)	(4)	(4a)
TE	0.206***	0.206***	0.232***	0.251***	0.251***	0.216***	0.216***
	(0.021)	(0.023)	(0.022)	(0.016)	(0.024)	(0.020)	(0.026)
Linguistic proximity							
migrants w/o TE	-0.025	-0.025	0.001	-0.052	-0.052	0.003	0.003
	(0.045)	(0.053)	(0.048)	(0.048)	(0.072)	(0.044)	(0.066)
migrants w/ TE	0.108***	0.108***	0.098***	0.103***	0.103***	0.068***	0.068*
	(0.036)	(0.040)	(0.038)	(0.028)	(0.044)	(0.033)	(0.044)
Observations	1,447	1,447	1,260	1,447	1,447	1,447	1,447
R-squared	0.521	0.521	0.576	0.767	0.767	0.552	0.552
OC X's	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DC FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country pair X's	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country pair FE's	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: see note under Table 4. These estimates use measure of linguistic proximity based on common native language.