

Presence of language-learning opportunities abroad and migration to Germany

Early draft (Do not cite!)

Matthias Huber*

University of Jena

Silke Uebelmesser†

University of Jena and CESifo

June 21, 2017

Abstract

This paper analyses the effect of the presence of German language learning opportunities abroad on migration to Germany. We use a unique dataset that provides information on the presence of the Goethe-Institut (GI), an association that promotes German culture and offers language courses and standardized exams, in 81 countries for the period 1965 to 2013. In this multiple origin and single destination framework, we estimate fixed-effects models where we control for multilateral resistance to migration by using the CCE-estimator (Pesaran 2006). We find evidence that the number of language institutes provided by the GI in a country is positively correlated with migration from that country to Germany. We find that the correlation between migration and the number of language institutes is lower for high income countries while differences in linguistic or geographical distance are of no relevance. To establish causality, we show that the probability of opening new institutes is not related to previous migration to Germany in a positive way and we use an instrumental variable approach.

JEL classification: F22, O15, J61.

Keywords: language skills, language learning, international migration, panel data, multilateral resistance.

Funded by the German Science Foundation (DFG, grant number UE 124/2-1)

*University of Jena, Carl-Zeiss-Str. 3, 07743 Jena, matthias.huber@uni-jena.de

†University of Jena, Carl-Zeiss-Str. 3, 07743 Jena, silke.uebelmesser@uni-jena.de

1 Introduction

A large part of the migration literature focuses on migrants' proficiency in the language of the destination country. It has been shown that proficiency improves labour market and integration outcomes. Language skills increase earnings (see e.g. Dustmann and Soest 2001; Chiswick and Miller 1995) and employment probability (Dustmann and Fabbri 2003). At the same time, the probability of intermarriage becomes larger and the likelihood of living in an ethnic enclave decreases (Bleakley and Chin 2010).

Given these benefits of language proficiency in the destination country, potential migrants can be expected to include considerations about languages into their migration decision and their location choice. Indeed, many studies have shown that language is an important determinant of migration flows. For this, measures of linguistic distance are often used to capture the linguistic relationship between the migrants' mother tongue and the language of the destination country. Adserà and Pytliková (2015) and Belot and Hatton (2012) find evidence of a negative effect of linguistic distance on international migration flows based on different measures. Languages which are linguistically more distant from the mother tongue are more difficult to learn and therefore more costly to acquire. However, the concept of linguistic distance neglects actual language acquisition of potential migrants before migration which can alleviate or overcome the negative effects implied by linguistic distance.

We build a random-utility model. Individuals want to maximize expected utility of migration of which the expected wage income net of migration costs is an important component. Acquiring language skills of the host country can increase expected net wage income if the benefits in terms of higher wages abroad exceed the costs of learning. We distinguish children-age and adult-age language learning, as they are different in terms of costs and direction of causality in the context of migration. If language skills are acquired during childhood or adolescence, the decision is more likely determined by factors outside the learner's direct control. These factors may be related to the school system with compulsory foreign language learning and to parents' preferences. It often comes with no or little costs, or they can be regarded as sunk while the benefits of language proficiency acquired during childhood or adolescence might affect the migration decision. Fenoll and Kuehn (2016) use compulsory language learning at school as a measure for language skills beyond linguistic properties and find a positive relationship with migration flows within the European Union.

Adult-age language learning, on the contrary, is more likely to be a decision driven by different motives which can be of personal or economic nature. Economic motives can be related to the local labour market or to migration (intention). A positive migration decision (or intention) might lead to pre-migration language learning. The direction of causality with adult-age language learning can thus be opposite to the direction with children-age language learning. Uebelmesser and Weingarten (2017) analyse determinants of adult-age language learning by using language exams at language institutes worldwide and language course participation at institutes in Germany.

They show that general migration and student migration are indeed important determinants of adult-age language learning.

The aim of this paper is to study the effect of the presence of language learning opportunities abroad on migration to Germany. For this, we use a unique panel dataset for 81 countries and the period 1965 to 2013 containing information about the worldwide presence of Goethe institutes collected from the annual reports of the Goethe-Institut (GI). The GI is an association which constitutes the main actor in promoting German culture and language worldwide. Via its institutes, it offers language courses and standardized language exams as well as information on German culture and society in many different forms, like cultural events and libraries. While the GI is mainly funded by the German government, language courses are self-financed by course fees (Goethe-Institut 2014).¹

We find evidence that the number of language institutes in a country is positively correlated with migration from that country to Germany. By distinguishing between institutes that offer language services and those which do not, we can show that the correlation is indeed driven by language learning opportunities and not by other factors, like the provision of information about German culture and society. The correlation between migration and the number of language institutes is lower for countries which belong to the 40 richest countries in our sample. Furthermore, we find that the results are not driven by EU member countries or by countries which are geographically or linguistically close to Germany. To support a causal interpretation, we follow two routes: We show that the opening of an institute in a country is not positively related to previous migration to Germany from that country and we use an instrumental variable approach.

Our paper is related to the literature about the determinants of bilateral migration flows. While most papers use panel data with multiple destination countries and multiple origin countries (Ortega and Peri 2013; Mayda 2010; Pedersen et al. 2008; Beine et al. 2011), other papers focus on a single destination country and multiple origin countries (Bertoli and Fernández-Huertas Moraga 2013; Bertoli et al. 2016) as we do in our analysis. In a setting with multiple origin countries and one or multiple destination countries, migration decisions are not only influenced by the chosen destination country's attractiveness but also by the attractiveness of other (alternative) destinations. This refers to the concept of multilateral resistance. Ignoring multilateral resistance leads to biased estimates, as the regressors are correlated with the unobserved multilateral resistance in the error term. We follow the approach developed in Bertoli and Fernández-Huertas Moraga (2013) and Bertoli et al. (2016) in order to control for multilateral resistance using the common correlated effects (CCE) estimator (Pesaran 2006). With this approach, migration flows do not only need to depend on characteristics of the origin and destination countries, but can also be affected by characteristics of alternative destinations which relaxes the Independence of Irrelevant Alternatives (IIA) assumption.

¹In this paper we stick to the following convention: when referring to the association of the Goethe-Institut we use the abbreviation "GI". When talking about specific branches of the GI abroad, we refer to them as "institutes".

The remainder of the paper is structured as follows: Section 2 introduces a random utility maximization (RUM) model that describes an individual’s decision to migrate. Section 3 presents the data and provides descriptive statistics. In section 4, the estimation strategy is derived. Section 5 presents our results and robustness checks. In section 6, we provide evidence to support a causal interpretation and section 7 concludes.

2 A random utility maximization model of migration

The micro-foundation of migration choice can be modelled in a RUM model. We follow the approach by Bertoli and Fernández-Huertas Moraga (2013) in order to introduce multilateral resistance to migration into the model. Individual i in origin country j decides to locate in country k out of the set of alternative destination countries D which maximizes his utility

$$U_{ijk} = V_{jk} + \epsilon_{ijk} = w_{jk} - c_{jk} + \epsilon_{ijk} \quad (1)$$

where V_{jk} is the deterministic part of the utility capturing expected income, w_{jk} , which increases individual utility, and costs of migration, c_{jk} , like geographical distance or strict visa regulations, which decrease utility. ϵ_{ijk} is the error term, which will be specific in more detail below. Adult-age language learning likely increases, on the one hand, expected earnings in the destination country by improving migrants’ labour-market outcomes and, on the other hand, the costs of migration which include the costs of language acquisition. Individuals will opt for language learning if their utility is maximized by doing so. Following Bertoli and Fernández-Huertas Moraga (2013), these individual decisions are not explicitly included in the model but captured in the error term. Given that the unit of our analysis of the determinants of migration will be the country level, we will include measures for foreign language learning in the deterministic part of the utility V_{jk} aggregated on the country level.

The resulting choice probabilities for one of the alternatives are determined by the assumptions on the distribution of the error term ϵ_{ijk} . Grogger and Hanson (2011), Mayda (2010) and others adopted the standard logit model, assuming that ϵ_{ijk} follows an independent and identically distributed extreme value type 1 distribution (McFadden 1974). Then, the probability of migrating from country j to k is

$$p_{ijk} = \frac{e^{V_{jk}}}{\sum_{l \in D} e^{V_{jl}}} \quad (2)$$

The assumption on the distribution of the error term implies that the denominator $\sum_{l \in D} e^{V_{jl}}$ does not vary across origin countries. In other words, the attractiveness of destination country k is the same for all individuals across all origin countries j . The ratio of the probabilities of migrating to country k and of not migrating (i.e. staying in the country of origin j) shows this property, known as the Independence of Irrelevant Alternatives (IIA) assumption

$$\frac{p_{ijk}}{p_{ijj}} = \frac{e^{V_{jk}}}{e^{V_{jj}}} \quad (3)$$

The ratio depends only on the relative attractiveness of countries j and k , but not on any alternative destination. In order to relax this restrictive assumption, Bertoli and Fernández-Huertas Moraga (2013) adopted the generalized nested logit model (Wen and Koppelman 2001) which allows for correlation of the error terms across alternative destinations. This framework is suited to introduce multilateral resistance into the model, which results in the following choice probability

$$p_{ijk} = \frac{\sum_m (\alpha_{jkm} e^{V_{jk}})^{\frac{1}{\tau}} (\sum_{l \in b_m} (\alpha_{jlm} e^{V_{jl}})^{\frac{1}{\tau}})^{\tau-1}}{\sum_m (\sum_{l \in b_m} (\alpha_{jlm} e^{V_{jl}})^{\frac{1}{\tau}})^{\tau}} \quad (4)$$

where b_m are nests of destination countries $k \in D$ that are "sharing unobservable sources of attractiveness" (Bertoli and Fernández-Huertas Moraga 2013, p. 81), e.g. 'cultural proximity'. $\alpha_{jkm} \in [0,1]$ indicates the extent to which a destination country k for individuals from origin country j belongs to nest m , where $\sum_m \alpha_{jkm} = 1$. This implies that every destination country can belong to different nests with different proportions, and these proportions can vary by origin country. τ measures the degree of independence within a nest, or, stated differently, τ is inversely related to the correlation of the stochastic components of utility among different pairs of destination countries. Hence, the relative probability of opting for destination k over the probability of not migrating depends not only on characteristics of the origin country and the destination country, but also on characteristics of alternative destinations according to the correlation of the nest structure

$$\frac{p_{ijk}}{p_{ijj}} = \frac{\sum_m (\alpha_{jkm} e^{V_{jk}})^{\frac{1}{\tau}} (\sum_{l \in b_m} (\alpha_{jlm} e^{V_{jl}})^{\frac{1}{\tau}})^{\tau-1}}{\sum_m (\alpha_{jjm} e^{V_{jj}})^{\frac{1}{\tau}} (\sum_{l \in b_m} (\alpha_{jlm} e^{V_{jl}})^{\frac{1}{\tau}})^{\tau-1}} \quad (5)$$

By assuming that the origin country does not share a nest with any destination country, the log odds are given by

$$\ln \left(\frac{p_{ijk}}{p_{ijj}} \right) = \frac{V_{jk}}{\tau} - V_{jj} + r_{jk} \quad (6)$$

The multilateral resistance is represented by

$$r_{jk} = \ln \left(\sum_m (\alpha_{jkm})^{\frac{1}{\tau}} \left(\sum_{l \in b_m} (\alpha_{jlm} e^{V_{jl}})^{\frac{1}{\tau}} \right)^{\tau-1} \right) \quad (7)$$

which is a decreasing function of V_{jl} as long as countries k and l share any nest, and which is

zero otherwise. Hence, an increase in the observed utility of alternative l decreases the log odds of migrating to country k if countries k and l belong to the same nest(s). In Section 4, we depart from equation (6) in order to derive our estimation equation.

3 Data and descriptive statistics

In the following, we describe the data used for the analysis and present descriptive statistics.

3.1 Dependent variables: Immigration flows

As dependent variable, we use yearly immigration flows to Germany over the population size of the origin country. Migration data is provided in the “Wanderungsstatistik” by the German Federal Statistical Office (Destatis). The data documents the number of foreign citizens that move to Germany and register their residence in a given year, where immigrants are categorized according to their citizenship. As this registration is mandatory for all foreign residents staying for more than 2 months (Destatis 2016), this data represents legal immigration to Germany in a comprehensive way. Data on population size comes from the Penn World Table (PWT) 9.0 (Feenstra et al. 2015).

3.2 Independent variable: Language learning opportunities

The data on language learning opportunities is derived from a new dataset comprising information about the presence of the GI and the number of institutes at the country-level. The GI has published annual reports continuously since 1965 in which activities of each institute including statistics of language course and exam participation have been reported. The dataset is constructed from these reports and contains information about the presence of GI on the city-level (see Uebelmesser et al. (2017) for a more detailed description of the dataset).

For our analysis, we aggregate the data on the country-level and construct two variables: a dummy-variable that indicates if at least one institute with language services is present in a given country and year and a variable which contains the number of language institutes in a given country and year. Not every institute offers language services.² For our basic specification we restrict attention to institutes with language services. For robustness checks, we also use information about institutes without language services.

²Also in some cases, the annual reports do not provide information if language services are offered or the numbers are reported jointly with other institutes. This is mainly the case for institutes that are subsidiaries of other main institutes. In these cases we assume that the main institute offers language services and the subsidiaries do not. For more information on the different types of institutes in the dataset, please refer to Uebelmesser et al. (2017).

3.3 Other control variables

As additional variables we include two measures to control for immigration regulations. We construct one dummy which indicates if the country is a member of the European Union in a given year. A second dummy captures if a bilateral agreement on labour recruitment was in place between Germany and the origin country. Germany used these agreements in particular in the 1950s and 1960s to recruit foreign workers as “Gastarbeiter” from South European and North African countries in order to overcome labour shortage.³ Furthermore, we control for the economic conditions in the origin country by including GDP per capita for which we use the variable ‘rgdpe’ and data on population size both from the PWT 9.0 (Feenstra et al. 2015).

3.4 Sample construction

To construct our dataset, we proceed as follows: First, we include all countries for which we have information on GDP and population size for all years between 1965 and 2013. Second, we restrict the sample according to the availability of migration data: We only include countries for which we have migration data in all years but 1990, 2000 and 2001. In these three years, migration data of Destatis includes many missing observations due to changes in the data generation process. We interpolate these missing observations linearly on the basis of the years 1989 and 1991, and 1999 and 2002, respectively. Third, we add the data from our GI dataset about the presence of the GI and the number of institutes per country and year, and assign the value 0 to these variables for countries that are not included in the GI dataset as they never had an institute in the period 1965 to 2013. Finally, we end up with a balanced dataset that includes observations for 81 countries in the period from 1965 to 2013.

3.5 Descriptive statistics

Table 1 provides summary statistics of the variables used in our analysis. Looking more closely at the presence of the GI at the country level in the period 1965 to 2013, we see that there was at least one institute in at least one year in 62 of the 81 countries; in about 50% of our sample (41 countries), the GI was present in all years of the observation period. The worldwide distribution of the countries in our sample is displayed in Figure 1 where countries are grouped according to the number of years in which the GI was present (all years of the observation period, at least one year and no year). The countries in our sample are spread over all continents and so are those with presence of the GI. Note that the (former) Soviet Union and other former socialist countries are not included in our sample. This is due to many newly founded states in the beginning of the 1990s and the lack of GDP and migration data.

Figure 2 presents the number of countries per year in which the GI was present with any type of institute or, respectively, with at least one institute that offered language services. In addition,

³The countries in our sample with which these agreements existed are Italy, Spain, Greece, Turkey, Morocco, Korea, Portugal and Tunisia.

Table 1: Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Emigration to Germany	3969	8057.201	17034.75	3	273690
Migration rate (emigration to Germany/population)	3969	.0001174	.0004487	5.48e-07	.010838
GI present	3969	.928331	.2579716	0	1
GI with language institute present	3969	.9159593	.2774839	0	1
Number of institutes	3969	3.980972	2.952102	0	14
Number of language institutes	3969	3.507703	2.564176	0	9
Number of institutes per 1m inhabitants	3969	.037929	.0696281	0	4.501625
Number of language institutes per 1m inhabitants	3969	.0318764	.0527952	0	2.647548
GDP per capita	3969	10293.64	12596.58	142.3924	91816.98
EU member	3969	.0776906	.2677178	0	1
Bilateral agreement	3969	.0111981	.1052401	0	1
Population in 1m	3969	327.4992	400.6844	.1925593	1279.499

Note: Descriptive statistics are weighted by population size.

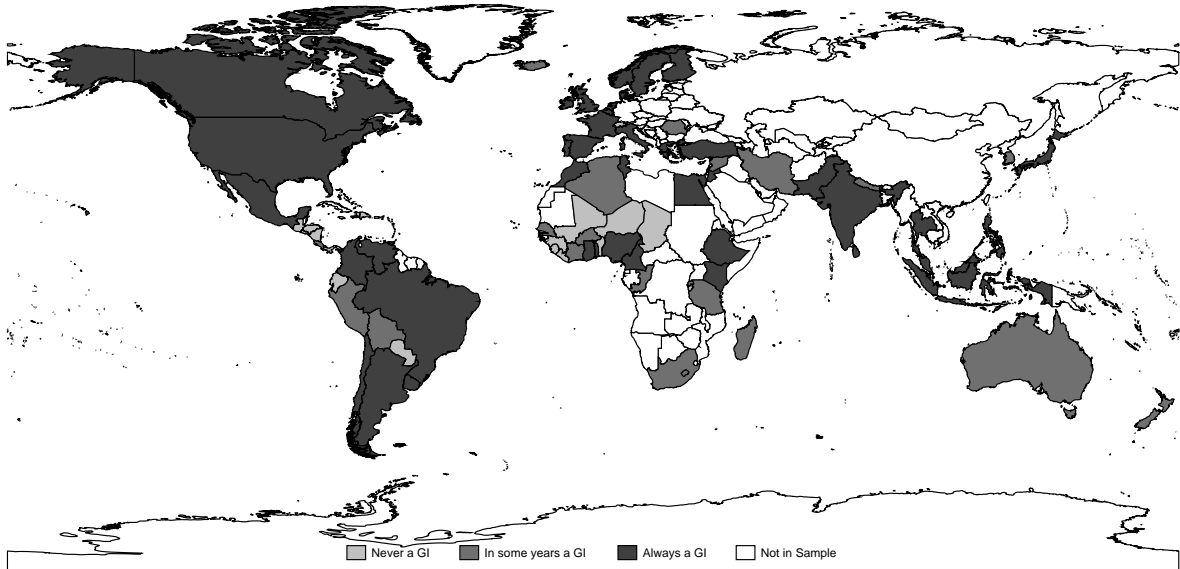


Figure 1: The presence of the GI.

the number of countries without any institute in that year is included. In most of the years, the GI was present in more than 50 out of the 81 countries. Throughout the entire period, the number of countries with institutes which only offered non-language services was negligible.

A different picture emerges when we compare the number of institutes per year with and without language services (Figure 3). While in the entire period 1965 to 2013, the number of institutes which offered language services always exceeded the number of institutes that did not, there was a significant number of institutes of the latter group. Especially in the 1970s, this number amounted to up to 60 institutes, but also afterwards there were around 20 institutes without language services in each year. This number stayed roughly constant also in the 1990s and 2000s, when the number of institutes with language services decreased. The average institute in our sample had 1160 registrations of students for language courses in 1995. This number increased to 1493 in 2005 and to 1939 in 2013.

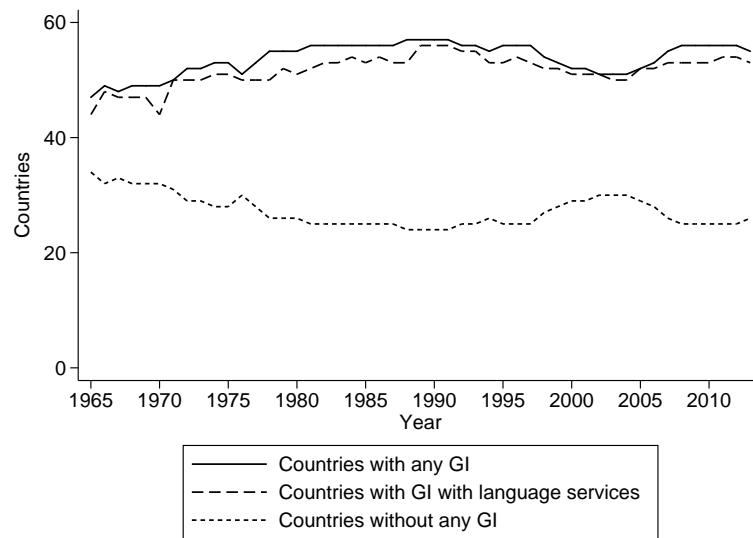


Figure 2: Number of countries with GI (based on 81 countries in the sample).

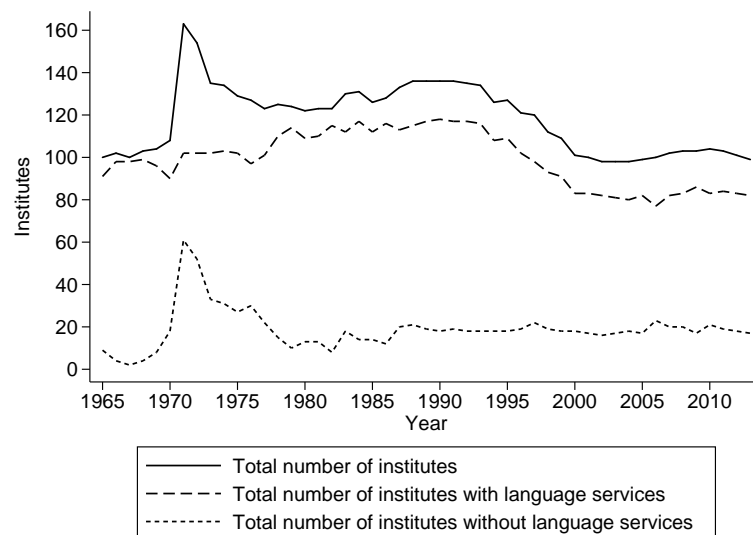


Figure 3: Number of institutes (based on 81 countries in the sample).

4 Estimation strategy

In order to estimate the relationship of language learning opportunities and migration with our panel dataset, we depart from equation (6) and assume that each individual i decides in each period t whether to migrate or not. Taking the average of the individual decisions for each country, we get

$$y_{jkt} = \beta' \left(\frac{x_{jkt}}{\tau} - x_{jjt} \right) + r_{jkt} + \eta_{jkt} \quad (8)$$

where y_{jkt} represents the logarithm of the migration flow from j to k at period t over the number of people that stay in origin country j . x_{jjt} and x_{jkt} , the empirical counterparts of V_{jj} and V_{jk} (cf. equation (1)), represent country-specific characteristics of j and k , respectively, and are orthogonal to the error term η_{jkt} , which is serially uncorrelated and independently and identically distributed over the set of origin-destination dyads.

The resistance term r_{jkt} is serially correlated over time and spatially correlated over origin-destination dyads. Without controlling for the resistance term, the exogeneity assumption might be violated as characteristics of countries might be correlated. Examples are similar movements of GDP or coordinated visa policy, like the Schengen treaty. Bertoli and Fernández-Huertas Moraga (2013) show that the CCE estimator by Pesaran (2006) consistently corrects for multilateral resistance to migration without requiring assumptions about the unobservable nests. For the CCE estimator, cross-sectional averages of all dependent and independent variables interacted with heterogeneous coefficients for all countries have to be included. For our multiple-origin and single-destination setting, the following estimation equation results

$$y_{jt} = \alpha' x_{jt} + \phi_t' d_t + \phi_j' d_j + \lambda_j' \tilde{z}_t + \eta_{jt} \quad (9)$$

with the weighted cross-sectional average defined as

$$\tilde{z}_t = \frac{1}{\sum_j \omega_{jt}} \left(\sum_j \omega_{jt} y_{jt}, \sum_j \omega_{jt} x_{jt} \right)$$

where ω_{jt} gives the weight for country j in t , for which we use population size.

The dependent variable y_{jt} is the logarithm of the migration rate, which is the ratio of the immigration flow from country j to Germany in year t over the population of country j . The vector x_{jt} includes our main variables of interest related to the supply of language services and some further controls. For this, we include a dummy that indicates if at least one language institute is present in the country (*GI with language institute present*) and/or the number of institutes with language services (*Number of language institutes*). To control for the economic condition in the origin country we include the logarithm of GDP/capita (*log GDP per capita*).

A dummy that indicates EU-membership (*EU member*) and a dummy that indicates bilateral agreement on labour recruitment (“Anwerbeabkommen”, *Bilateral agreement*) are included in order to control for immigration restrictions. Furthermore, we control for the log population in the origin country (*log population*). Finally, we add origin dummies d_j to control for all time-invariant characteristics of the origin country and year dummies d_t to control for origin-invariant effects.

5 Results

We estimate equation (9) in several specifications. Our basic specifications in Table 2 employ the GI variables for language institutes. In Tables 3 and 4, robustness checks are presented. As we include country FE in all specifications, the estimated effects only capture within-country variations.

We conduct a test for multilateral resistance. The CCE-test is a F-test on the joint significance of all cross-sectional averages included in the regression. The p-value for that tests is 0.000 in all specifications. Hence, we can conclude that multilateral resistance exists in our setting.

5.1 Basic specifications

For the basic specifications presented in Table 2, we find that the presence of the GI in a country is positively correlated to the migration rate, but only at the 10 % significance level (see column (1)). The relationship between the GI and the migration rate gets stronger and highly significant if language learning opportunities are measured by the number of language institutes (see column (2)): an additional institute that offers language services increases the migration rate by $e^{0.0766} - 1 \approx 8\%$. Column (3) shows when both measures of language learning opportunities are included at the same time, only the number of language institutes are positively correlated. This indicates that the marginally significant coefficient of the presence of GI displayed in column (1) indirectly captures the effect of the number of institutes. Hence, column (2) is our preferred specification for further robustness checks.

The coefficients of log GDP/capita are negative in all specifications.⁴ Improved economic conditions in the origin country reduce the benefits of migration. Furthermore, EU-membership leads to a significant increase in migration to Germany and so does a bilateral agreement on labour recruitment. These two variables capture less restrictive immigration regulations which lower migration costs. The log of population is negatively correlated to the migration rate in all specifications.

⁴The size of the effect is very similar to Clark et al. (2007). They find that a 1 % increase in origin GDP per capita decreases migration to the US by 0.44 %.

Table 2: Estimation results: basic specifications

	(1)	(2)	(3)
DV: log migration rate			
GI with language institute present	0.0822* (0.0444)		0.00929 (0.0490)
Number of language institutes		0.0766*** (0.0113)	0.0705*** (0.0118)
log GDP per capita	-0.400*** (0.0328)	-0.426*** (0.0331)	-0.436*** (0.0331)
EU member	0.604*** (0.121)	0.540*** (0.122)	0.520*** (0.126)
Bilateral agreement	11.98*** (3.048)	15.37*** (2.822)	14.67*** (2.948)
log population	-3.488*** (0.461)	-2.466*** (0.481)	-2.650*** (0.492)
Constant	76.86*** (15.39)	51.70*** (16.66)	55.73*** (16.80)
Observations	3,969	3,969	3,969
Adjusted R-squared	0.976	0.978	0.978
Country FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Countries	81	81	81
Years	1965-2013	1965-2013	1965-2013
CCE-test (p-value)	0.000	0.000	0.000

Observations are weighted by population size of the origin country; results are estimated with the CCE-estimator (Pesaran 2006); the CCE-test is a F-test on the joint significance of the cross-sectional averages of all dependent and independent variables interacted with country dummies. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1;

5.2 Robustness Checks

In Tables 3 and 4, we present robustness checks based on our preferred specification (column (2) in Table 2). First, the positive correlation between the number of language institutes and the migration rate might not measure the effect of language learning opportunities but other aspects that come with the opening of a new institute by the GI. Beyond language services, the institutes provide information about German culture and society. This might reduce uncertainty about life in Germany for potential migrants and therefore increase migration to Germany. The first two columns in Table 3 take this into account. In column (1), we replace the number of language institutes in a given country and year with the number of all institutes, with and without language services. We find that there is no significant effect on the migration rate. In column (2), we split the total number of institutes into two independent variables – the number of language institutes and the number of institutes without language services. We find that only additional language institutes significantly affect the migration rate, whereas institutes without language services do not. The coefficient for the number of language institutes is only slightly smaller than in our preferred specification. From the results of these two robustness checks, we conclude that our

variable indeed measures language learning opportunities and not some other effects of the GI.

Furthermore, we include additionally the first lag of the number of language institutes (column(3)), and the first and second lags (column(4)). Language learning and migration might not take place in the same period as the acquisition of language skills requires some time. Therefore, the effect of an additional institute might be better captured with lags as it might evolve over time. This is indeed what we find: a new institute increases the migration rate to Germany on average by $e^{0.1138} - 1 \approx 12\%$ after one year, as we can see in column (3), and to $e^{0.1341} - 1 \approx 14\%$ after two years (see column (4)).

Finally, it might be the case that only large enough institutes are actually able to influence the migration rate and therefore drive our results. Unfortunately, we cannot measure the actual size of institutes in terms of course participation, because there is no consistent measure available over this long period. In column (5), we therefore relate the number of institutes in a country to the population size. This controls for the possibility that the effect on the migration rate of one more institute in, for example, India and one more institute in Iceland might be different. We find that the coefficient for the number of language institutes per 1m inhabitants remains positive and highly significant.

In our second set of robustness checks, we test if results are driven by different country-groups (see Table 4). To do so, we estimate our preferred specification (column (2) of Table 2) interacting the GI variables with dummies that capture income level, geographic and linguistic distance, as well as EU membership. By ranking the countries according to their mean income for each year, we construct a dummy that indicates if a country belongs to the 40 countries with the highest income level. We also rank countries according to their geographic distance to Germany and construct a dummy for the 40 countries with the smallest distance (Mayer and Zignago 2011). For linguistic distance, the dummy takes the value 1, if the major language spoken in the country is an Indo-European language (Adserà and Pytliková 2015). Table 4 presents the results.

In column (2), we can see that the effect of the number of institutes is significantly smaller for countries with higher income. For all other country-groups the coefficients for the interaction terms are not significantly different from zero (see columns (3) to (5)). As the coefficient for the number of language institutes remains positive and significant in all specifications, we conclude that our results are not driven by specific country-groups.

Table 3: Robustness checks: different specifications of the GI variables

	(1)	(2)	(3)	(4)	(5)
DV: log migration rate					
Number of all institutes	0.0103 (0.00781)				
Number of institutes without language services		0.00441 (0.00846)			
Number of language institutes		0.0585*** (0.0122)	0.0599*** (0.0126)	0.0465*** (0.0132)	
Number of language institutes, lag=1			0.0539*** (0.0123)	0.0399*** (0.0136)	
Number of language institutes, lag=2				0.0477*** (0.0125)	
Number of GI with language services per 1m inhabitants					1.099*** (0.303)
log GDP per capita	-0.440*** (0.0317)	-0.417*** (0.0323)	-0.459*** (0.0340)	-0.500*** (0.0356)	-0.411*** (0.0338)
EU member	0.624*** (0.118)	0.541*** (0.120)	0.520*** (0.124)	0.554*** (0.125)	0.547*** (0.126)
Bilateral agreement	13.92*** (3.596)	5.003 (3.529)	17.52*** (3.265)	24.75*** (4.035)	8.389*** (2.678)
log population	-3.984*** (0.471)	-2.873*** (0.494)	-2.787*** (0.504)	-3.103*** (0.535)	-2.013*** (0.468)
Constant	95.90*** (15.64)	59.87*** (16.65)	56.94*** (17.24)	64.76*** (17.76)	30.07 (25.66)
Observations	3,969	3,969	3,888	3,807	3,969
Adjusted R-squared	0.977	0.977	0.978	0.979	0.977
Country FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Countries	81	81	81	81	81
Years	1965-2013	1965-2013	1966-2013	1967-2013	1965-2013
CCE-test (p-value)	0.000	0.000	0.000	0.000	0.000

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1; observations are weighted by population size of the origin; results are estimated with the CCE-estimator (Pesaran 2006); the CCE-test is a F-test on the joint-significance of the cross-sectional averages of all dependent and independent variables interacted with country dummies.

6 Direction of causality

So far, we have found a positive relationship between the number of language institutes and the migration rate. However, we cannot derive from this the direction of causality, as we cannot preclude that the GI is more likely to open institutes in countries with larger migration to Germany. If this was the case, we would not be able to disentangle the positive correlation of the GI and the migration rate, as estimated in section 5, in the “migration effect” caused by the opening of an institute on the migration rate and the “selection effect” caused by the migration rate on the location decision for a new GI. Furthermore, we would expect the relationship between the number of institutes and the migration rate to be overestimated, i.e. the migration effect would be biased upwards. To study this in more detail, we follow two strategies. First, we estimate the reverse relation in order to see whether there is evidence that previously large migration to Germany determines the opening of new institutes. Second we apply an instrumental variable approach.

Table 4: Robustness checks: interaction effects

Interaction with (DV: log migration rate)	(1) Basic specification	(2) High Income (40 countries)	(3) Geogr. close (40 countries)	(4) Ling. close (53 countries)	(5) EU (14 countries)
Number of language institutes	0.0766*** (0.0113)	0.0976*** (0.0185)	0.0691*** (0.0132)	0.107*** (0.0339)	0.0557*** (0.0121)
Number of language institutes * ...		-0.0512*** (0.0187)	0.0134 (0.0287)	-0.0192 (0.0361)	0.0151 (0.0439)
log GDP per capita	-0.426*** (0.0331)	-0.452*** (0.0347)	-0.466*** (0.0353)	-0.492*** (0.0345)	-0.499*** (0.0363)
EU member	0.540*** (0.122)	0.512*** (0.122)	0.547*** (0.123)	0.457*** (0.129)	0.540*** (0.153)
Bilateral agreement	15.37*** (2.822)	9.266*** (2.936)	15.58*** (2.809)	12.66*** (2.839)	10.85*** (2.942)
log_pop	-2.466*** (0.481)	-2.254*** (0.501)	-2.424*** (0.496)	-2.189*** (0.491)	-2.250*** (0.517)
Constant	51.70*** (16.66)	36.05** (18.14)	50.74 (31.26)	40.82** (17.15)	55.37*** (20.94)
Observations	3,969	3,969	3,969	3,969	3,969
Adjusted R-squared	0.978	0.979	0.978	0.978	0.980
Country FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Countries	81	81	81	81	81
Years	1965-2013	1965-2013	1965-2013	1965-2013	1965-2013
CCE-test (p-value)	0.000	0.000	0.000	0.000	0.000

Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1; observations are weighted by population size of the origin; results are estimated with the CCE-estimator (Pesaran 2006); the CCE-test is a F-test on the joint-significance of the cross-sectional averages of all dependent and independent variables interacted with country dummies.

6.1 Determinants for opening institutes

To find out what drives the GI's decisions where to open new institutes in each year, we estimate the following equation as a linear probability model

$$opening_{jt} = \alpha'x_{jt} + \phi'd + \epsilon_{jt} \quad (10)$$

where $opening_{jt}$ is a dummy indicating the opening of a new institute in country j in a given year t . As explanatory variables x_{jt} we include several factors which potentially determine the opening of a new institute: First, economic factors like the wealth of the country, measured by log GDP per capita, or its economic relation to Germany, measured by German exports (provided by Destatis); second, political factors like the relation with Germany measured by membership dummies for EU, OECD and NATO, and a dummy for bilateral agreements on labour recruitment (see Section 3.3); third, country characteristics like the geographical distance to Germany (Mayer and Zignago 2011) and the population size; fourth, the number of already existing institutes; finally, as our main variable of interest, the stock of migrants from country j in Germany from the Central Register of Foreign Nationals ("Ausländerzentralregister", Destatis) which measures cumulated previous migration to Germany. Table 7 in the Appendix presents summary statistics of the variables used.

We also include a set of dummies d in our estimation. As we want to analyse the GI's decisions in every year, we include year FE in all our specifications. Furthermore, the GI might base its decision on the already realized presence in the respective region. Therefore, we further add regional FE.⁵ In order to capture the regional focus of the GI over time, we also include year-region FE. As we are interested in the between-effect, we prefer not to include country FE. However, as the unobserved heterogeneity of countries might bias our results we also estimate a specification for the within-country effects.

Table 5 presents estimation results of equation (10). Due to data availability of the migrant stock and export data, our sample reduces by two years (1965 and 1966) and by two countries (Belgium and Luxembourg). Columns (1-3) do not include country FE and therefore estimate between-country effects. While column (1) only includes year FE, in columns (2) and (3) further FE are gradually added. The coefficient of the independent variable of interest, the migrant stock, is negative and significant in all three specifications. This indicates that the GI does not open institutes in countries with large migration to Germany. On the contrary, the probability of opening a new institute is higher in countries with lower previous migration to Germany. Column (4) shows the within-country effect of the stock of migrants on the probability of opening a new

⁵Regions are defined according to the organisational structure of the GI used since 2008: Central and Eastern Europe, East Asia, Eastern Europe and Central Asia, North America, North Africa and Middle East, Northwest Europe, South America, South Asia, Southeast Europe, Sub-Saharan Africa, Southwest Europe and Southeast Asia, Australia and New Zealand. Instead of controlling for the organisational level of the GI, we also included continent FE. This did not change the results.

Table 5: Linear probability model for the opening of language institutes

	(1)	(2)	(3)	(4)
DV: Opening				
log GDP per capita	0.00673*** (0.00228)	0.00884** (0.00367)	0.0108*** (0.00384)	-0.00166 (0.00712)
Exports	1.23e-09** (5.62e-10)	1.20e-09** (5.97e-10)	1.57e-09** (7.09e-10)	-0 (1.24e-09)
EU member	-0.0360** (0.0168)	-0.0102 (0.0155)	-0.0106 (0.0160)	-0.0807** (0.0320)
OECD member	0.0169 (0.0122)	0.0555*** (0.0147)	0.0513*** (0.0147)	-0.0165 (0.0245)
NATO member	0.0322* (0.0178)	0.0158 (0.0173)	0.0197 (0.0173)	-0.0553 (0.0615)
Bilateral agreement	0.0468 (0.0359)	0.0644* (0.0365)	0.0537 (0.0389)	0.0560 (0.0384)
Population	0.000102*** (3.09e-05)	9.37e-05*** (3.35e-05)	9.92e-05*** (3.46e-05)	-0.000108 (0.000111)
Geogr. Distance	5.44e-07 (9.18e-07)	-3.93e-06** (1.90e-06)	-3.50e-06* (1.90e-06)	
Number of institutes, lag=1	-0.00447 (0.00330)	-0.00564 (0.00352)	-0.00652* (0.00366)	-0.0468*** (0.0104)
Migrant Stock	-3.85e-08*** (1.08e-08)	-2.62e-08** (1.31e-08)	-3.30e-08** (1.41e-08)	1.88e-08 (3.02e-08)
Constant	-0.0280 (0.0282)	-0.0544 (0.0425)	-0.108*** (0.0377)	0.312** (0.144)
Observations	3,706	3,706	3,706	3,706
Adjusted R-squared	0.046	0.051	0.067	0.124
Country FE	No	No	No	Yes
Year FE	Yes	Yes	Yes	Yes
Region FE	No	Yes	Yes	Yes
Year*Region FE	No	No	Yes	Yes
Countries	79	79	79	79
Years	1967-2013	1967-2013	1967-2013	1967-2013

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

institute. The coefficient of interest becomes positive, but insignificant. Therefore, we conclude that the effect of the GI on the migration rate is not upward biased, because there is no evidence that the GI self-selected into high migration countries.

6.2 Instrumental variable approach

As a second approach to find support for a causal effect of the GI on migration flows to Germany, we employ an instrumental variable strategy. We present results of three (preliminary) instruments. The aim is to identify the effect of the GI on migration flows by only taking into account that variation of the number of institutes that is exogenous to migration. For this, we need an instrument that determines the number of language institutes in a country and, at the same time, is exogenous to migration from that country to Germany. In the following, we describe three instruments and provide evidence that the effect of the number of institutes on migration to Germany can indeed be causally interpreted.

Two instruments use the Correlates of War 2 International Governmental Organizations (IGO) Data Version 2.3 (Pevehouse et al. 2004) which contain information on memberships in IGO for each country and year. An IGO has three characteristics. First, the organization is multilateral, i.e. it has at least three member states. Second, there have to be plenary sessions at least once every ten years. Third, there needs to be a permanent secretary and corresponding headquarters. The dataset includes in total 495 IGOs. Examples are the well known organizations like the EU, NATO, OECD, but the dataset also includes many other less known and more specialized organizations, like the East African High Commission or the Administrative Center for Social Security for Rhine Boatmen. The organizations are often highly specialized and have a regional focus. Although a small number of the organizations deals with migration, the vast majority of IGOs follows goals completely exogenous to migration.

The first instrument which we construct from the IGO data reflects the bilateral political relation between the origin country and Germany. This relation might be important when the GI decides to open or close institutes together with the Federal Foreign Office (FFO, “Auswärtiges Amt”), as main source of funding on the behalf of the German government. Hence, we count the number of joint memberships in IGOs for each origin country and year and use this as an instrument. This instrument is exogenous to migration to Germany because the vast majority of the IGOs does not with with migration. Furthermore, it is not an exclusive bilateral relation as there have to be at least three member states.

The second instrument is a variant of the first one. For this, we count the number of memberships in IGOs in each year and country, independent of a simultaneous German membership. On the one hand, this instrument captures a less strong link between origin countries and Germany as the first instrument such that the exclusion restriction can be expected to hold a fortiori with respect to migration to Germany. On the other hand, the general willingness to cooperate internationally as captured by the number of memberships in IGOs is likely to positively affect the opening of

institutes as this makes it easier for institutes to work properly.

For the third instrument, our strategy is to interact an exogenous time-variant, but country-invariant variable with a variable that varies over countries but not over time. For the latter, time-invariant variable, country characteristics are often used. However, characteristics that explain the number of institutes in a country are often determinants of migration flows as well, like population size. To overcome this problem, we come back to bilateral political relations between the origin country and Germany and construct a variable that reflects membership in NATO and OECD, two organizations that are exogenous to migration, as they focus on different topics. The instrument takes the value 1 if the country is a member of the NATO or the OECD, the value 2 if the country is a member of both and 0 otherwise. As the time-varying variable of our instrument, we use the share of total public expenditure allocated to the FFO. The FFO is the main funding source of the GI and, together with the GI, decides where to open and close institutes (Schneider and Schiller 2000). Therefore, the budget should be correlated with the total number of institutes. However, the absolute size of the budget of the FFO depends on the economic situation of Germany, which itself may affect migration flows to Germany. Hence, in order to sustain exogeneity we relate the budget of the FFO to total public expenditures.

With 2SLS we re-estimate the specification of column (2) in Table 2 with our identifying instruments. Table 7 presents the results. In order to make results comparable, we restricted the time horizon of our analysis to the years 1965 to 2005 in all specifications because of limited data availability of the IGO data. We also re-estimated column (2) in Table 2 for the slightly restricted sample and report it in column (1).

Our estimates confirm the relevance of all three instruments: They are significantly correlated with the number of institutes as indicated by the high F-statistics of the excluded instruments in the first stage. The second-stage results are reported in the lower panel of Table 6. The results show a statistically significant and positive effect of the instrumented variables. The effects with the 2SLS estimations are about 2.5 to 6.5 times larger than in the non-instrumented specification in column (1). Together with the results from section 6.1, this points towards the GI choosing countries for new institutes with previously low migration to Germany. Thus, ignoring to correct for that endogeneity results in a downward bias of the actual effect of language learning opportunities on migration to Germany. Hence, we conclude that the identified effects of the number of language institutes on the migration rate can be causally interpreted. [...]

7 Conclusion

In this paper, we have analysed the effect of the presence of German language learning opportunities abroad on migration to Germany. We find a significant and positive correlation between the number of language institutes of the GI and migration rates to Germany and can show that the channel of the effect of the number of institutes are indeed language learning opportunities. This relationship is stronger for countries with lower income. Furthermore, we analyse the determinants

Table 6: IV estimation results

VARIABLES	(1)	(2)	(3)	(4)
<i>First stage results (DV: Number of language institutes)</i>				
(Budget FFO/ total budget) * OECD/NATO		-47.38*** (9.989)		
Joint Membership in IGO			0.0593*** (0.00604)	
Number of memberships in IGO				0.0317*** (0.00408)
log GDP per capita		0.0264 (0.0562)	0.0383 (0.0554)	0.0540 (0.0558)
EU member		0.152 (0.219)	-0.136 (0.219)	0.0582 (0.218)
Bilateral agreement		-3.492 (5.910)	-3.100 (5.827)	-5.587 (5.863)
log population		-1.256 (0.840)	-0.766 (0.831)	-1.373* (0.833)
<i>Second Stage results (DV: log migration rate)</i>				
Number of language institutes	0.0941*** (0.0129)	0.662*** (0.169)	0.243*** (0.0645)	0.434*** (0.0888)
log GDP per capita	-0.465*** (0.0379)	-0.486*** (0.0453)	-0.470*** (0.0352)	-0.477*** (0.0385)
EU member	0.557*** (0.148)	0.445** (0.178)	0.528*** (0.137)	0.490*** (0.151)
Bilateral agreement	15.87*** (3.989)	18.65*** (4.791)	16.59*** (3.706)	17.53*** (4.062)
log population	-3.446*** (0.567)	-2.591*** (0.717)	-3.223*** (0.533)	-2.935*** (0.589)
Observations	3,321	3,321	3,321	3,321
Countries	81	81	81	81
Years	1965-2005	1965-2005	1965-2005	1965-2005
F-statistic of excluded instrument		22.50	96.67	60.27

All regressions include cross-sectional averages of the dependent and all independent variables, country and year FE. Constant not displayed. Standard errors in parentheses; ***p < 0.01, **p < 0.05, *p < 0.1.

of the opening of new institutes and find that the probability is not significantly positively related to the migration stock in Germany, and therefore to previous migration. The absence of such a selection effect allows us to interpret the effect of the GI on the migration rate to Germany as causal. We further support this causal interpretation with an IV-approach that even increases the effect of language learning opportunities. Thus, we find new evidence that language learning shapes international migration flows beyond linguistic properties, like linguistic distance.

So far, similar effects have been only shown for foreign language learning at school in Europe (Fenoll and Kuehn 2016). Though compulsory children-age language learning leads to more migration, it is hardly within the reach of the policy-maker in the destination country. This is different for adult-age language learning opportunities; in particular, cultural institutes like the GI are often under direct control of the policy-makers. We have shown that by controlling the supply of adult-age language courses the quantity of migration can be affected. The effect on the quality of migration in terms of skill composition and actual language proficiency, however, and its more general effects on the destination country is left for future research.

References

- Adserà, Alícia and Mariola Pytliková (2015). “The Role of Language in Shaping International Migration”. *The Economic Journal* 125(586), F49–F81.
- Beine, Michel, Frédéric Docquier, and Çağlar Özden (2011). “Diasporas”. *Journal of Development Economics*. Symposium on Globalization and Brain Drain 95(1), 30–41.
- Belot, Michèle and Timothy J. Hatton (2012). “Immigrant Selection in the OECD”. *The Scandinavian Journal of Economics* 114(4), 1105–1128.
- Bertoli, Simone, Herbert Brücker, and Jesús Fernández-Huertas Moraga (2016). “The European crisis and migration to Germany”. *Regional Science and Urban Economics* 60, 61–72.
- Bertoli, Simone and Jesús Fernández-Huertas Moraga (2013). “Multilateral resistance to migration”. *Journal of Development Economics*. Migration and Development 102, 79–100.
- Bleakley, Hoyt and Aimee Chin (2010). “Age at Arrival, English Proficiency, and Social Assimilation Among US Immigrants”. *American Economic Journal: Applied Economics* 2(1), 165–192.
- Chiswick, Barry R. and Paul W. Miller (1995). “The Endogeneity between Language and Earnings: International Analyses”. *Journal of Labor Economics* 13(2), 246–288.
- Clark, Ximena, Timothy J Hatton, and Jeffrey G Williamson (2007). “Explaining U.S. Immigration, 1971–1998”. *Review of Economics and Statistics* 89(2), 359–373.
- Destatis (2016). *Qualitätsbericht Wanderungen*.
- Dustmann, Christian and Francesca Fabbri (2003). “Language proficiency and labour market performance of immigrants in the UK”. *The Economic Journal* 113(489), 695–717.
- Dustmann, Christian and Arthur van Soest (2001). “Language Fluency and Earnings: Estimation with Misclassified Language Indicators”. *Review of Economics and Statistics* 83(4), 663–674.
- Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer (2015). “The Next Generation of the Penn World Table”. *American Economic Review* 105(10), 3150–3182.
- Fenoll, Ainhoa Aparicio and Zoë Kuehn (2016). “Does foreign language proficiency foster migration of young individuals within the European Union?” In: *The Economics of Language Policy*. Ed. by Bengt-Arne Wickstroem and Michele Gazzola. Cambridge, MA: MIT Press, 331–355.
- Goethe-Institut (2014). *Jahrbuch 2013/14*.
- Grogger, Jeffrey and Gordon H. Hanson (2011). “Income maximization and the selection and sorting of international migrants”. *Journal of Development Economics* 95(1), 42–57.

- Mayda, Anna Maria (2010). “International migration: a panel data analysis of the determinants of bilateral flows”. *Journal of Population Economics* 23(4), 1249–1274.
- Mayer, Thierry and Soledad Zignago (2011). “Notes on CEPII’s distances measures: The GeoDist database”. *CEPII Working Paper* (2011-25).
- McFadden, Daniel (1974). “Conditional logit analysis of qualitative choice behavior”. In: *Frontiers in Econometrics*. Ed. by Paul Zarembka. New York: Academic Press.
- Ortega, Francesc and Giovanni Peri (2013). “The effect of income and immigration policies on international migration”. *Migration Studies* 1(1), 47–74.
- Pedersen, Peder J., Mariola Pytlikova, and Nina Smith (2008). “Selection and network effects—Migration flows into OECD countries 1990–2000”. *European Economic Review* 52(7), 1160–1186.
- Pesaran, M. Hashem (2006). “Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure”. *Econometrica* 74(4), 967–1012.
- Pevehouse, Jon, Timothy Nordstrom, and Kevin Warnke (2004). “The Correlates of War 2 International Governmental Organizations Data Version 2.0”. *Conflict Management and Peace Science* 21(2), 101–119.
- Schneider, Gerald and Julia Schiller (2000). “Goethe ist nicht überall : eine empirische Analyse der Standortentscheidungen in der Auswärtigen Kulturpolitik”. *Zeitschrift für Internationale Beziehungen* 7(1), 5–32.
- Uebelmesser, Silke, Matthias Huber, and Severin Weingarten (2017). “The German Language Worldwide: A New Dataset on Adult-Age Language Learning”. *University of Jena, mimeo*.
- Uebelmesser, Silke and Severin Weingarten (2017). “A Macro-level Analysis of Adult-age Language Learning”. *University of Jena, mimeo*.
- Wen, Chieh-Hua and Frank S. Koppelman (2001). “The generalized nested logit model”. *Transportation Research Part B: Methodological* 35(7), 627–641.

Appendix

Table 7: Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Opening of new institute	3706	.0226659	.1488562	0	1
GDP per capita	3706	10108.32	11606.77	142.3924	65104.98
Exports to Germany	3706	3696513	1.08e+07	1453	1.02e+08
OECD member	3706	.2490556	.4325244	0	1
NATO member	3706	.1635186	.3698881	0	1
EU member	3706	.1071236	.309312	0	1
Bilateral agreement	3706	.0151106	.1220095	0	1
Distance to Germany	3706	6159.215	3815.971	173.5239	18824.75
Population	3706	39.94605	108.1455	.1989563	1279.499
Number of language institutes	3706	1.245008	1.580373	0	9
Migrant stock	3706	45945.92	187701.5	18	2110223