A New Dataset on Language Course Participation

Early draft (Do not cite!)

Silke Uebelmesser* University of Jena and CESifo

Severin Weingarten† University of Jena

October 27, 2015

Abstract

Language learning can overcome barriers to the integration of migrants into their host societies and may also play an important role in the context of international trade and international relations more generally. However, most survey-based studies on the language skills of individuals, migrants in particular, contain little information about the timing of language acquisition and cannot disentangle effects and causes of language learning. Using a new dataset on adult-age German language learning in Germany and 89 other countries for 1966-2012, we investigate the macro-level drivers of language learning and make two contributions to the literature: First, we distinguish between language learning before and after migration. Second, in our context, the direction of causality from migration and other economic factors to language learning is clearer than in much of the previous literature. Using fixed-effects estimations, we find that German language learning abroad is positively associated with migration from EU, but not from non-EU countries. Language learning in Germany is not strongly associated with current immigration, but with growing migrant stocks more generally. Student migration plays an important role in explaining German language learning abroad and in Germany. Additionally, using random-effects specification that distinguish between within and between effects, we find that countries with large average migration flows to Germany are characterized by more language learning. This relationship weakens as migrant stocks from the respective countries grow, indicating that ethnic enclaves have a negative effect on migrants' incentives to learn German. We also find some evidence of a positive association between trade and language learning in Germany.

JEL classification: F22, J24, J61.

Keywords: language skills, language learning, international migration, migrant networks, migration policies.

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

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

1 Introduction

Language skills play an important role in shaping international migration flows and in determining if and how migrants integrate into their host societies. Based on a new dataset that reports the extent of language course participation at the German Goethe institutes in 89 countries for the period 1966–2013, we examine the determinants and consequences of adult language learning on the macro-level. In this paper we describe the dataset. Additionally, in a first exploratory exercise, we relate language course participation to several key variables, most importantly migration, trade, economic conditions, and linguistic and geographic distance.

The literature on migration choice has long recognized the importance of language skills by controlling for common languages of origin and destination countries (e.g. Mayda 2010; Grogger and Hanson 2011; Belot and Hatton 2012; Ortega and Peri 2013). More recently, Adserà and Pytliková (2015) and Belot and Ederveen (2012) show that linguistic distance, which can be interpreted as the difficulty associated with learning another language, has an important effect on international migration flows.[1] Aparicio Fenoll and Kuehn (forthcoming) find a strong positive effect of school-age language learning on migration between EU countries. Their results show that the study of language learning processes can add value to a literature which has previously focused on linguistic properties. While linguistic properties are beyond the reach of policy makers, language learning is not. It can be part of school curricula, but it can also be encouraged or even made a requirement by the governments of destination countries.

In addition to their importance as a determinant of migration, language skills are also crucial for the integration of migrants into their host societies. They have a positive effect on earnings (Chiswick and Miller 1995; Dustmann and Soest 2001; Dustmann and Fabbri 2003; Bleakley and Chin 2004), employment (Dustmann and Fabbri 2003), and social assimilation (Bleakley and Chin 2010), and they can affect occupational choice (Chiswick and Miller 2007).

Given the importance of language skills, a large number of studies has explored their determinants. Chiswick and Miller (2015, section 4) offer an extensive review of the literature where they group determinants into three categories, which they dub "the three E's": exposure, efficiency and economic incentives. Exposure depends on time since migration, residence in ethnic enclaves and the language spoken by spouse and children. Efficiency variables include age at migration, level of education, linguistic distance, and the

[1] While this paper focuses on international migration, the effect of linguistic barriers is not limited to the factor labor. Lohmann (2011) and Isphording and Otten (2013) show that linguistic distance also has a negative effect on trade.

motive of migration. Two economic incentives for language learning that are addressed in the literature are expected duration of stay and expected gains in earnings from language proficiency.

Since most of these variables vary on the level of the individual migrant, studies on the determinants of migration use censuses or surveys to obtain micro-level data. Typically, these datasets report language skills that respondents possess at the time of the collection of the data. However, the timing of language learning is relevant in its own right. Foreign language acquisition at early ages occurs primarily at school and is determined by the schooling system. For adults on the contrary, the decision to learn a language is more likely to be made in light of a decision to migrate. Migrants who possess language skills at some point in time after their arrival in the host country may have been selected on the basis of pre-existing skills or they may have been motivated to learn the language by their decision to migrate. To the best of our knowledge non of the micro-level studies on the determinants of language skills can distinguish between the incentive-to-learn and the selection effect. However, from the point of view of the policy maker, an understanding of the incentive-to-learn effect is highly relevant, because it allows the targeting of language courses at groups of immigrants who are more likely to lack necessary language skills. While the dataset presented in this paper does not contain micro-level information about individual migrants, it allows for an explicit focus on the motivation effect because it reports on (adult-age) language learning and not on the presence of language skills.

The rest of the paper is structured as follows. Section 2 introduces the new dataset. Section 3 discusses to what extent language learning at the Goethe institutes is a good proxy for language learning in general. Section 4 outlines our hypotheses which are examined using the empirical setup presented in section 5. We discuss our preliminary results in section 6 and conclude in section 7.

2 The Dataset

The Goethe-Institut (GI) is a German association that promotes the study of the German language and culture abroad. Most importantly for our purposes, it maintains institutes in 89 countries, at many of which locals can study the German language and obtain language certificates which are widely recognized. The GI is mainly funded by the German government and through course fees (Goethe-Institut e.V. 2014).

The dataset presented in this paper has two parts: The first part covers language learning abroad. It reports yearly observations of four language-learning-related variables for each of a total of 170 institutes in 89 countries in the period 1966–2012. The four variables are the

number of language exam participants, the average number of students per course term[2], the number of course hours taught, and the number of language teachers employed. The second part of the dataset covers language learning in Germany and reports the number of language course participants for 198 nationalities and the years 1980–2006. Some of the variables are not available for all years due to changes in the reported statistics, but many gaps can potentially be filled on the basis of internal records kept by the institute. Table 1 provides an overview of some descriptive statistics for each variable.

Table 1: Descriptive Statistics GI Dataset[a]

Variable	Interval	Min.	Median	Mean	Max.
Exams	1990–2012	1	110	540	26159
Students	1972–1999[b]	2	427	602	6314
Course Hours	1972–2012	10	4390	6115	38437
Teachers	1978–1999	1	4	6	36
Students in Germany	1980–2006	1	32	150	5206

[a] At the time of submission of this draft of the paper we have not fully digitized all parts of the dataset. Specifically, the years 1966–1971 are still missing and no plausibility checks have been done for the variable 'Teachers'.
[b] While the students variable is no longer reported in the yearbooks after 1999, it is available in internal documents provided by the GI and we plan to extend it to 2013.

Most of the data was digitized from the yearbooks of the GI which have been published continuously since 1966. The yearbooks contain detailed reports on the activities of the GI, as well as key statistics for the entire association and for each institute. The data were typed into CSV files.[3] Country and city names were harmonized and matched to codes used by the UN population division to allow merging with other country-level and city-level datasets.

Figure 1 plots the development of the aggregates of all five variables and the number of institutes that offer language courses over time. The number of institutes is fairly stable, with small increases at the end of the 1970s and 1980s, and a small decrease in the 1990s. Course hours increase steadily throughout the entire period of observation, almost doubling from 500,000 in 1972 to 1 million in 2013. Student numbers at institutes abroad are currently only available until 1999. They are fairly stable, but there are three periods of relatively small temporary increases at the beginning of the 1970, around 1980 and around 1990. Student numbers in Germany follow a similar pattern.

The number of teachers decreases slightly during the 1990s, but this variable is not yet

[2]The length of course terms differs across institutes. The average number of participants per course term can be thought of as the number of students who are currently enrolled in a language course at any point in time during the year.

[3]We are very much indebted to Maik Wehlte for doing a lot of digitization work and to Martin Ahmad, Toni Grimm, and Lars Other, who helped with the digitization of additional data and plausibility checks.

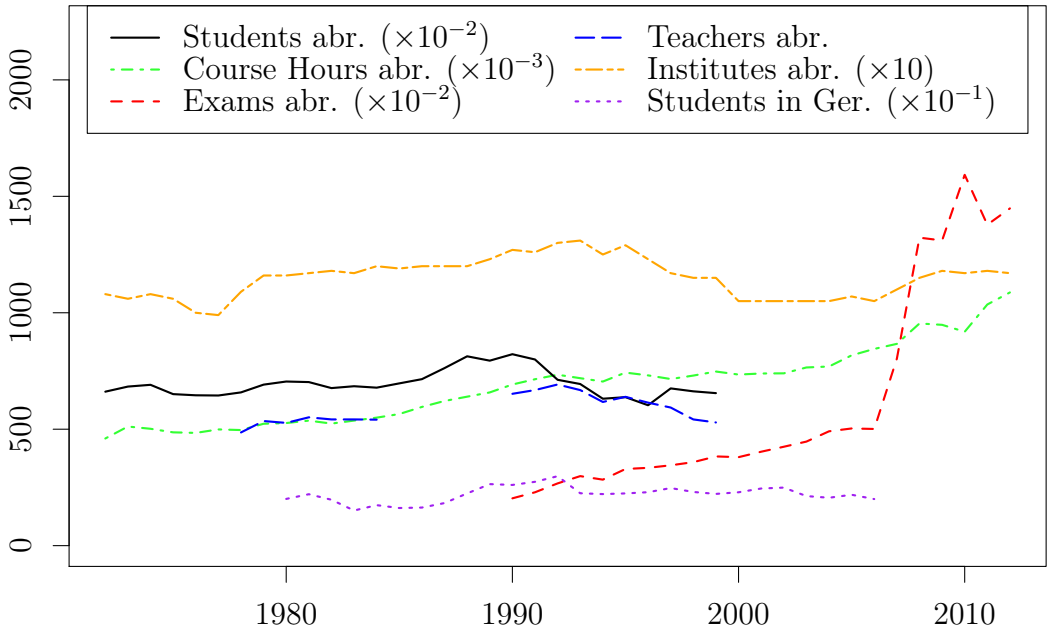

Figure 1: Aggregate development of variables in GI dataset

available for the entire period of observation (see above). Language exam numbers are only available from 1990 onwards and increase steadily until 2006. They then experience a strong surge to over 150,000 in 2010, more than three times the level of 2006.[4]

Figure 2 plots the number of exam participants abroad and student numbers abroad and in Germany for four selected institutes/countries. Two observations can be made regarding the nature of the variance in the data.

First, the average numbers of exam participants and students differ considerably between institutes and countries and these differences cannot be explained by differences in population alone. While the size of the catchment area of an institute is not obvious a priori, neither city population nor country population can explain the difference in exam participant or student numbers for Porto and Amsterdam. For example, Porto has about three times as many language students, but the metropolitan areas of Amsterdam and Porto are roughly the same in terms of population size and Portugal is considerably smaller than the Netherlands. In an analogous comparison, differences in population cannot explain the fivefold difference between average exam participant and student numbers in Ankara and Bangalore. A similar picture emerges from figure 3. While some of the countries with a lot of students are among the most populous in the world (e.g. India, Brazil), others are

[4]This surge could be related to a change in German immigration law. Since 2007, migrants who come to Germany under family reunification provisions (from non-EU countries) have to provide evidence that they possess basic German language skills (equivalent to the A1 level in the Common European Framework of Reference for Languages, CEFR).

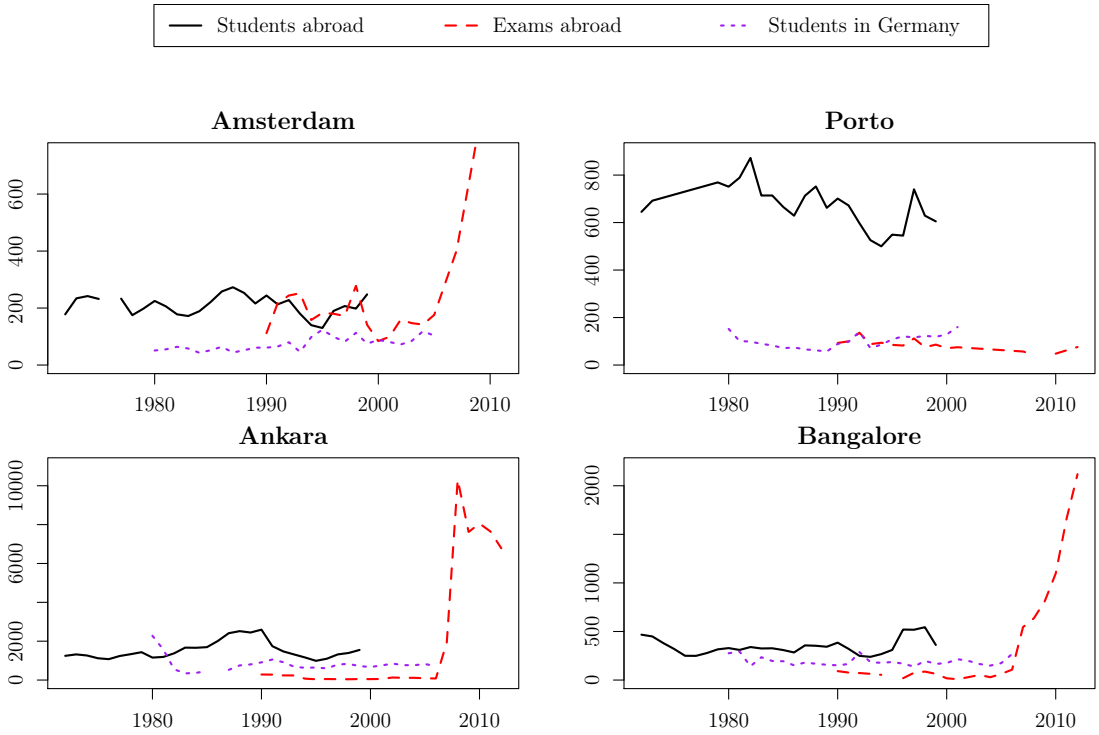

Figure 2: Development of students and exam participants for four selected institutes and countries

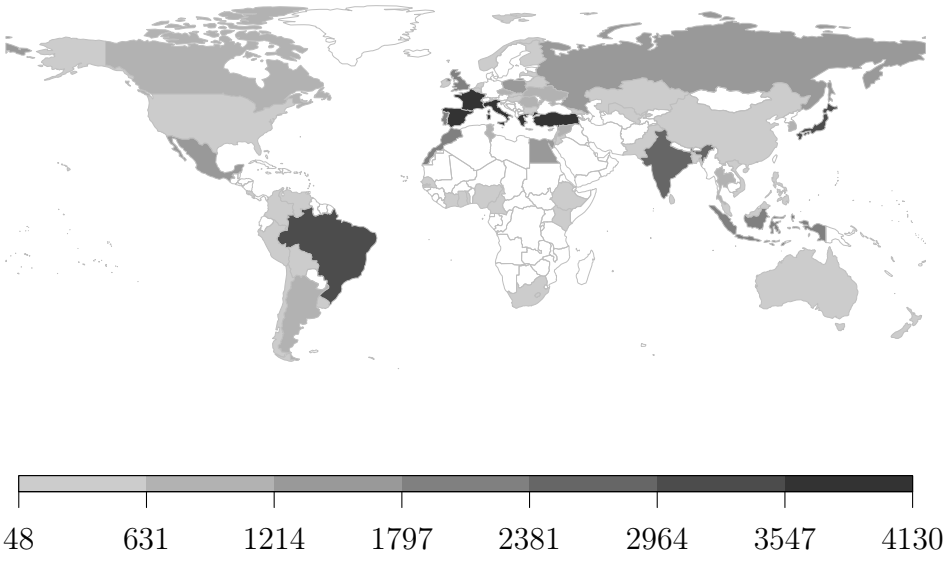

Figure 3: Student numbers at Goethe institutes in 1999

much smaller (e.g. South Korea, Greece).

Second, for each institute/country, all measures vary considerably over time. This variation is not strongly correlated between institutes. While Ankara sees a considerable increase in student numbers in the late 1980s, this increase is much weaker in Amsterdam and Bangalore and barely happens in Porto. Student numbers in Amsterdam and Banaglore increase considerably in the late 1990s, but no similar increase happens in Porto and Ankara. This suggests that student numbers are not only driven by a general trend or changes in Germany, but by country-specific factors. We will attempt to explore these factors in the following.

3 Representativeness of the Goethe Data

While the hypotheses presented in the next section address language learning in general, our data are limited to language learning that occurs at the Goethe institutes. Naturally, there are a large number of other language learning opportunities, including universities, private language schools, and internet platforms. This multitude of other options gives rise to a number of concerns regarding the self-selection of language learners into courses offered by the Goethe institutes that we will address before turning to the hypotheses. Three characteristics on which self-selection may be based are willingness or ability to pay, location, and age:

City	Provider	Course Type	Price / Hour	Currency
Mexico City	GI	Extensive	146.67	MXN
Mexico City	Tecnológico de Monterrey	Extensive	90.91	MXN
Buenos Aires	GI	Extensive A1.1	80	ARS
Buenos Aires	Sprachzentrum Buenos Aires	Extensive A1.1	65	ARS
Rio de Janeiro	GI	Extensive A1	56.21	BRL
Rio de Janeiro	Baukurs	Extensive A1	49.17	BRL
Lissabon	GI	Extensive	5.67	EUR
Lissabon	ilnova	Extensive	6.17	EUR
Ankara	GI	Extensive A1	10.16	TRY
Ankara	Hitit Education Institutions	Extensive A1	11.88	TRY
Tokyo	GI	Intensive	1541.67	JPY
Tokyo	German Office	Intensive	1971.43	JPY

Table 2: Prices of language courses at Goethe institutes and other providers in 2015

Selection on willingness to pay could occur if the prices of courses at the Goethe institutes differed significantly from the costs of other equally suitable learning options. On the one hand, one might suspect the Goethe institutes to be somewhat of a premium provider

of language courses, because they are a semi-official German organization with a long tradition and a good reputation. Such a status would allow them to charge higher prices. On the other hand, one might suspect Goethe courses to be particularily cheap, because the majority of the Goethe institutes' funds comes from the German government.[5] Historical price data on language courses are not available to the best of our knowledge. However, table 2 contains current price data on comparable language courses offered by the Goethe institutes and by other institutes in six cities in different countries.[6] While the data are far from complete or representative, they do not indicate that the Goethe institutes are usually the most expensive provider in the market.

Goethe institutes are usually located in capitals and other major cities. The lack of institutes in rural areas is likely to lead to an underrepresentation of language learners from these areas among participants at the Goethe institutes. However, the bias need not be as large as one would initially expect: Goethe institutes offer both extensive and intensive language courses. Extensive courses are based on weekly lessons and last for several months, but intensive courses are taught en-block. Participants of intensive courses do not necessarily have to live in the vicinity of the respective institute. They may also stay there for the duration of the course only.

The language courses taught by the Goethe institutes are a traditional "offline" form of language learning. At the other end of the spectrum are pure online courses like those offered by "myngle" or "babble". The latter kind of courses may be more attractive to a younger generation of language students, which is more familiar with using the internet in general. While this difference may lead to an overrepresentation of older students among the participants in language courses at the Goethe institute, the advent of online language learning platforms in the late 2000s falls in the very last years of the period of observation from 1982 to 2006 used in our analysis. Consequently, we are confident that the age bias only has a small, if any, effect on our results.

4 Hypotheses

In this section, we present several hypotheses regarding potential determinants of language learning. The first set of hypotheses relates language learning to migration and the second set relates language learning to trade. Generally, the mechanisms described in this section

[5]Several employees of the Goethe institute have stated in conversions with us that their language courses are priced to be self-financing and that government funding is used for other cultural activies. Additionally, the price policies of the individual instistutes take the prices of local competitors into account.

[6]Data comes from the websites of the course providers. The websites of non-Goethe-institute providers were found by searching Google for "language learning" and the name of the respective city in the native language of the respective country.

and the resulting hypotheses apply to all countries, but they may differ in strength for EU and non-EU countries, because EU citizens enjoy unrestricted access to the German labor market. This issue is picked up in section 6 when we discuss the results of our estimations. Similarily, the described mechanisms and resulting hypotheses apply to both language learning abroad and language learning in Germany where no explicit distinction is made.

4.1 Migration

Given the large number of potential benefits of language proficiency and the robust results regarding the effect of proficiency on earnings, migrants should have an incentive to learn the language spoken in their host country. A positive association between immigration and course and exam participation would be in line with the existence of an incentive-to-learn effect, where individuals decide to take language courses to improve their host-country language skills. If our data supports this hypothesis, it will complement the results of previous survey-based studies where the presence of migrant's language skills could either be attributed to an incentive-to-learn or to a selection effect (see section 1).

Hypothesis 1 Immigration from another country to Germany is positively associated with language learning by citizens of that country.

Students from other countries who come to Germany to study at German unversities are a special subgroup of immigrants, but a similar incentive-to-learn rationale applies to them. They may even be more motivated to learn the language than other groups of immigrants, because many of them will enter the job market after completing their degree.

Hypothesis 2 Immigration of students from another country to Germany is positively associated with language learning by citizens of that country.

Migrant Stocks

There are several reasons why not only current migration flows, but also the presence of migrants from a particular country in Germany should increase language learning by those who live in that country or who migrated from that country to Germany: First and most obviously, migrants who arrived in previous years may still be taking language courses in Germany; either because they didn't have the opportunity or motivation to do so earlier or because they are simply continuing their education. Second, a large migrant community in Germany may lead to increased interest in the German language and culture in migrants' home countries. Third, migrant stocks may act as a proxy for short-term migration that are not captured by the residence-based measure of migration flows that we use in our estimations (see table 3).

Hypothesis 3 The presence of a large number of migrants from one country in Germany is positively associated with language learning by citizens of that country.

Migration and Minority Language Concentrations

Minority language concentrations[7] are often considered to improve the ability of migrants to find work in their host country and build social ties to others who speak their native language. As a consequence, speaking the host country's language may be less important for migrants who live in minority language concentration. Several studies find that minority language concentrations are associated with lower levels of language proficiency (Chiswick and Miller 2007; Espenshade and Fu 1997; Lazear 1999; Isphording and Otten 2013). We hypothesise that these results are not exclusively based on the (self-)selection of migrants with worse language skills, but that the negative effect of minority language concentrations extends to the language learning decisions of immigrants and is, thus, at least partially driven by an incentive-to-learn effect. We use the number of citizens of a country of origin who live in Germany as a proxy for the size of the respective minority language concentration.

Hypothesis 4 The presence of a larger number of migrants from another country in Germany weakens the association between migration and language learning.

4.2 Trade

Trade relationships may also be an important driver of language learning decisions. Using different measures of linguistic distance, both Lohmann (2011) and Isphording and Otten (2013) find that linguistic distance has a negative effect on bilateral trade. If language barriers can hinder trade, trade partners should have an incentive to learn each others languages. Therefore, we would expect language learning to be positively related with trade.

Hypothesis 5 Larger trade flows between Germany and another country are associated with more language learning by citizens of that country.

5 Empirical Setup

We are interested in the relationships between language learning on the one hand and several migration-related and trade-related variables on the other hand, where language learning is proxied by two different variables: For language learning abroad, our dependent variable is exam participation. We choose it over the other available variables (course hours

[7] Other authors use the terms "ethnic enclaves" (e.g. Danzer and Yaman 2013) or "migrant networks" (e.g. Bertoli and Fernández-Huertas Moraga 2015).

and students), because it is our broadest measure of language learning. Exam participants may have studied German at one of the Goethe institutes, but they may also have taken a course elsewhere or learned the language on their own. Additionally, only a small number of students take more than one exam per year because exams cover several courses. This reduces the risk of counting the same students more than once per year. For language learning in Germany, the dataset only contains a single variable: the number of course participants per year and nationality.

For the 'abroad' specification, we observe one exam participation number for each institute in each year. This gives rise to a two-level geographical structure, where most explanatory variables are available on the country-level, but where each country may have several institutes. We use institute-level rather than country-level estimations, because they allow us to exploit more of the variance in our explained variable. This approach also avoids the problem of abrupt changes in country-level aggregates of our dependent variable when institutes open and close.

We use two types of estimations: First, OLS regressions with institute/country-fixed effects and year-fixed effects provide estimates that use only variation within institutes/countries and cluster-robust standard errors. Second, random effects regressions specified according to Bell and Jones' (2014) re-formulation of an estimator proposed by Mundlak (1978) allow us to separate the within and between effects of our variables of interest.

Using OLS, we estimate for exam participation abroad

$$P_{ijt} = \alpha + \beta\, x_{jt} + \gamma\, v_{jt} \circ w'_{jt} + \delta\, y_{ijt} + \eta_t\, D_t + \eta_i\, D_i + u_{ijt} \tag{1}$$

and for course participation in Germany

$$P_{jt} = \alpha + \beta\, x_{jt} + \gamma\, v_{jt} \circ w'_{jt} + \eta_t\, D_t + \eta_i\, D_j + u_{jt} \tag{2}$$

where the indices reflect the dimensions across which variables vary: institute i, country j, and time t. P represents exam or course participation and x_{jt} is a vector of our main explanatory and control variables. v_{jt} and w_{jt} are subsets of x_{jt} that are chosen so that their component-wise product gives a set of interactions that are necessary to test the hypotheses outlined in section 4 and to distinguish between effects for EU and non-EU countries. y_{ijt} is a vector of additional, city-level controls and D_i, D_j, and D_t are institute, country, and year dummies, respectively. In equation (1) institute-level dummies capture both country-fixed and institute-fixed effects. α is an intercept and u_{ijt} and u_{jt} are error

terms. In both estimations, the errors are assumed to be clustered on the country level[8].

Using random effects regressions, we estimate for exam participation abroad

$$\begin{aligned} P_{ijt} = {} & \alpha + \beta_W \left(x_{jt} - \bar{x}_j\right) + \beta_B \, \bar{x}_j + \\ & \gamma_{WW} \left(v_{jt} - \bar{v}_j\right) \circ \left(w_{jt} - \bar{w}_j\right)' + \gamma_{WB} \, \bar{v}_j \circ \left(w_{jt} - \bar{w}_j\right)' + \\ & \gamma_{BW} \left(v_{jt} - \bar{v}_j\right) \circ \bar{w}_j' + \gamma_{BB} \, \bar{v}_j \circ \bar{w}_j' + \\ & \delta_W \left(y_{ijt} - \bar{y}_i\right) + \delta_B \, \bar{y}_i + \\ & u_i + u_j + u_t + u_{ijt} \end{aligned} \tag{3}$$

and for course participation in Germany

$$\begin{aligned} P_{jt} = {} & \alpha + \beta_W \left(x_{jt} - \bar{x}_j\right) + \beta_B \, \bar{x}_j + \\ & \gamma_{WW} \left(v_{jt} - \bar{v}_j\right) \circ \left(w_{jt} - \bar{w}_j\right)' + \gamma_{WB} \, \bar{v}_j \circ \left(w_{jt} - \bar{w}_j\right)' + \\ & \gamma_{BW} \left(v_{jt} - \bar{v}_j\right) \circ \bar{w}_j' + \gamma_{BB} \, \bar{v}_j \circ \bar{w}_j' + \\ & u_j + u_t + u_{jt} \end{aligned} \tag{4}$$

where $\bar{x}_j$, $\bar{v}_j$, $\bar{w}_j$, and $\bar{y}_i$ are country-level and institute-level averages and W and B indicate coefficients that are identified by within-country and between-country variation, respectively. This within-between specification also allows us to disentangle our interaction effects into four different terms: Those driven by the within variation of both variables (WW), those driven by the within variation of the first but the between variation of the second variable (WB) and vice versa (BW) and those driven by the between variation of both variables (BB). An example of a mixed within-between interaction is "Immigration (B) $\times$ Migrant Stocks (W)", which can be interpreted as the change in the effect of the country-average of immigration on language learning that is brought about by growing (or shrinking) migrant stocks. u_i, u_j, and u_t are institute-specific, country-specific, and time-specific error terms (random effects). This specification is an extension of Bell and Jones' (2014) equation (12). We add the interaction terms and the additional error term u_i to account for our nested geographical structure of institutes within countries.

All non-dummy variables in our regressions enter in logs so that our coefficients can be interpreted as elasticities. The main advantage of this approach is that it allows variation from countries of different sizes and with completely different magnitudes of language learning, migration, and trade flows to drive the results of our model. An estimation in levels would suffer from considerable heteroskedasticity and the results would necessarily

[8]Cluster-robust standard errors are calculated according to Cameron et al. (2011) using the R package multiwayvcov (version 1.2.2).

be driven by a small number of countries that send a larger number of migrants to Germany or trade with Germany a lot. While immigration from or trade with these countries may be economically more relevant because of its magnitude, our main interest is in identifying the mechanisms that drive language learning more generally and, thus, in using variation from as many countries as possible to identify these effects. Additionally, an estimation in levels would require that we specify to which extent institutes in cities of different sizes are exposed to changes in our country-level explanatory variables. For example, an absolute change in immigration from France should, in absolute terms, have a larger effect on exam participation in Paris than on exam participation in Nancy. Paris is a larger city and the institute there has a larger catchment area. A log-log estimation does away with this concern, because it assumes that both institutes experience the same relative rather than absolute change in exam participation as a result of a relative increase in migration.

Table 3 lists all sources from which we take data for our estimations. We use gross immigration flows rather than net migration, because the latter measure contains return migration and emigration, which should not have an effect on language learning in the home country. We proxy trade flows by dividing total trade revenues with Germany by the GDP of the country in question.

Variable	Source of Data
Lang. exam participants abroad	Digitized GI yearbooks
Lang. course participants in Germany	Digitized GI yearbooks
Population	UN World Population Prospects
City Population	UN World Urbanization Prospects
Migrant stocks in Germany	German Federal Statistical Office
Student migrant stocks in Germany	German Federal Statistical Office
Immigration to Germany	German Federal Statistical Office
Trade flows	German Federal Statistical Office
EU membership	Self-compiled

Table 3: Variables and data sources

6 Preliminary Estimation Results

In this section, we present the results of our fixed-effect and random-effect estimations. Our 'abroad' estimation uses a total of 1358 observations from 84 institutes in 56 countries and covers the time period 1990–2006.[9] The dependent variable is the number of language exams

[9]While data is also available for the years 2007-2010, we omit these years from our estimations, because of a German policy change in 2007 that required family reunification migrants from non-EU countries to present proof of basic German language skills before immigration. This change had a uniquely large effect on the composition of the exam participants at the Goethe institutes that we cannot properly capture

taken at institute i in year t. Our 'Germany' estimations use a total of 1571 observations for 106 nationalities and covers the time period 1992–2006. In these estimations, the dependent variable is the number of language courses taken by individuals with nationality j in year t. In addition to the fixed and random effects described in section 5, all estimations control for EU membership and country population. Additionally, the 'abroad' estimations control for the population of the metropolitan area where the institute is located.

Table 4: Estimation results for course and exam participation abroad and in Germany

	Abroad (FE)	Abroad (RE)	Germany (FE)	Germany (RE)
Immigration (W)	−0.04 (0.06)	0.00 (0.05)	0.02 (0.01)	0.03 (0.02)*
Immigration (B)		0.65 (0.28)**		0.58 (0.18)***
Imm. (W) × EU (W)	0.57 (0.28)**	0.88 (0.34)***	0.11 (0.13)	0.30 (0.20)
Imm. (B) × EU (W)		−0.10 (0.46)		0.35 (0.41)
Imm. (W) × EU (B)		0.63 (0.19)***		0.21 (0.15)
Imm. (B) × EU (B)		0.40 (1.33)		−0.39 (0.81)
Imm. (W) × Mig. Stock (W)	−0.01 (0.03)	0.01 (0.18)	−0.01 (0.01)	−0.02 (0.04)
Imm. (B) × Mig. Stock (W)		−0.17 (0.06)***		−0.08 (0.03)***
Imm. (W) × Mig. Stock (B)		0.01 (0.02)		−0.00 (0.01)
Imm. (B) × Mig. Stock (B)		−0.01 (0.04)		0.02 (0.02)
Student Mig. Stock (W)	0.24 (0.12)**	0.24 (0.05)***	0.07 (0.03)**	0.03 (0.03)
Student Mig. Stock (B)		0.36 (0.17)**		0.40 (0.11)***
SMS (W) × EU (W)	0.16 (0.37)	0.45 (0.54)	0.37 (0.17)**	0.26 (0.42)
SMS (B) × EU (W)		−4.08 (1.26)***		0.21 (0.41)
SMS (W) × EU (B)		−0.10 (0.22)		0.12 (0.26)
SMS (B) × EU (B)		0.94 (0.84)		−0.30 (0.52)
Trade Rev. w/ Ger. (W)	0.28 (0.26)	0.13 (0.10)	−0.02 (0.08)	−0.04 (0.04)
Trade Rev. w/ Ger. (B)		−0.10 (0.23)		0.37 (0.11)***
TR (W) × EU (W)	0.24 (0.30)	0.21 (1.04)	0.28 (0.17)*	0.36 (0.58)
TR (B) × EU (W)		0.96 (0.50)*		−0.22 (0.61)
TR (W) × EU (B)		−0.20 (0.30)		0.46 (0.32)
TR (B) × EU (B)		−0.62 (0.88)		−2.38 (0.44)***
Migrant Stocks (W)	0.38 (0.30)	0.22 (0.10)**	0.28 (0.11)**	0.33 (0.06)***
Migrant Stocks (B)		−0.50 (0.25)**		−0.55 (0.18)***
MS (W) × EU (W)	−0.78 (0.46)*	−3.37 (1.45)**	−0.50 (0.20)**	−1.24 (0.93)
MS (B) × EU (W)		2.56 (1.17)**		−0.66 (0.59)
MS (W) × EU (B)		−0.28 (0.55)		−0.13 (0.58)
MS (B) × EU (B)		−0.02 (0.87)		0.13 (0.71)
EU (W)	−1.04 (0.73)	−1.87 (0.96)*	−0.44 (0.19)**	−0.84 (0.42)**
EU (B)		1.02 (1.83)		0.20 (0.73)
Population (W)	1.21 (1.59)	−1.43 (0.60)**	−0.53 (0.56)	−1.29 (0.24)***
Population (B)		−0.34 (0.16)**		0.40 (0.08)***
Population City (W)	0.69 (1.09)	0.90 (0.40)**		
Population City (B)		0.44 (0.10)***		
R^2	0.73		0.91	
Adj. R^2	0.70		0.90	
Num. obs.	1358	1358	1571	1571
Num. groups: city.de.harm.j		84		
Num. groups: country.iso3		56		106
Num. groups: year		17		15
Variance: city.de.harm.j.(Intercept)		0.27		
Variance: country.iso3.(Intercept)		0.50		0.60
Variance: year.(Intercept)		0.00		0.01
Variance: Residual		0.42		0.26

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 4 reports our main regression results. We begin by discussing results from the

with aggregate migration data.

fixed-effects regressions (columns 1 and 3, within effects) and complement them with results from our random-effects regression (columns 2 and 4, between effects) later.

6.1 Within Effects

With respect to hypothesis 1, we find a relatively strong and significant association between immigration from EU countries and language learning at the Goethe institutes in those countries, but this result does not extend to non-EU countries or to language learning in Germany. Student migration, however, seems to play a more consistent role. The stock of foreign students in Germany is positively associated with language learning abroad and in Germany and the latter association is particularly strong for EU countries. These results support our hypothesis 2 and indicate that students learn German before coming to Germany and, especially if they come from EU countries, after arriving there. Migrant stocks (hypothesis 3) from non-EU countries are positively associated with language learning in Germany, but the coefficient is insignificant for language learning abroad. It is not suprising that migrant stocks play a more important role for language learning in Germany than for language learning abroad, because migrants who are already in Germany will take courses there rather than in their country of origin. In our fixed-effects estimation results we find no support for an interaction between migration flows and stocks (hypothesis 4).

With respect to hypothesis 5, we find a weakly significant association between trade with EU countries and course participation in Germany. Unfortunately, while some of the trade-related coefficients are fairly large, so are the standard errors of these coefficients. This result may indicate that trade-related factors play a role, but that our measure of trade flows (total trade volume divided by GDP) is too unspecific to properly account for these factors. A natural alternative to trade flows would be FDI flows, which capture the establishment of new business-related links between countries. However, data on FDI flows are only available for a relatively small subset of the countries included in our analysis.

6.2 Between Effects and Mixed Interactions

While our fixed-effects approach deliberately ignores all between-country variation, our random-effects approach allows us to consider within and between variation at the same time. Most within coefficients have the same magnitude as in the fixed-effects estimations. However, some smaller differences occur, because the two specifications assume different error structures and because the random-effects estimations disentangle the interaction coefficients from the fixed-effects estimations into three separate coefficients γ_{WW}, γ_{WB},

and γ_{BW} (see equations (3) and (4)).[10] Naturally, the results obtained from between-country variation have to be interpreted particularly carefully, because they are susceptible to bias from omitted time-invariant country-level variables.

With respect to hypotheses 1 and 2, the estimated between coefficients reveal strong significant associations between migration and student migration on the one hand and language learning on the other. Countries with higher average migration and student migration to Germany are characterized by higher levels of language learning. Additionally, the coefficient for the interaction between country-averaged immigration flows and migrant stocks is negative and significantly different from zero in both the 'abroad' and the 'Germany' specification. This result indicates that growing migrant stocks reduce the association between language learning and migration and supports hypothesis 4.

With respect to hypothesis 3, the between effects for migrant stocks are negative in both specifications, highlighting that the positive association between migrant stocks and language learning is driven by the growth in migrant stocks rather than by their average levels. This supports the view that recently arrived migrants or their family members who are preparing to follow them to Germany learn the language, but not the view that migrant stocks lead to a long-term build-up of interest in language learning in their home countries. The latter effect would have to manifest in a positive between effect of migrant stocks.

With respect to hypothesis 5, we find a positive within coefficient for trade in our 'Germany', but not in our 'abroad' specification. This can be seen as another indication that trade-related variables play a role in determining language learning in Germany, but the causal mechanism behind this association requires further investigation.

6.3 Direction of Causality

Above, we provide evidence of a positive association between language learning and several migration-related variables. In this section, we argue that this association is mainly driven by a causal effect of those migration-related variables on language learning.

With respect to language learning in Germany, a causal effect of course participation on our migration-related variables is unlikely, because migrants take the course after arriving in Germany. Theoretically, the availability of post-migration language courses might affect the decision whether and where to migrate. However, since courses that teach migrants the language of the destination country are likely to be available in any potential destination, the potential effect of availability on migration can be neglected in our application.

[10] The variation that identifies the fourth interaction coefficient γ_{BB} from our random-effects specification does not play a role in the fixed-effects models, because it is completely captured by country-fixed effects.

With respect to language learning at institutes outside of Germany, three channels of reverse causality could be particularily relevant: the opening and closing of institutes, the participant recruitment and capacity planning of the institutes, and changes in the individual motivation of participants. First, the opening and closing of institutes may be motivated by the presence of migration and trade flows. This channel of causality would affect the results of our estimations if they "compared" participation in cities in which language courses or exams were offered with zero participation in cities where this was not the case. However, our dataset only includes observations for city-year combinations where language courses were offered and is, thus, not succeptible to the endogenous opening and closing of institutes. Second, the presence of large migration flows from a particular country may motivate institutes in that country to advertise more heavily to "capture" a larger share of the outgoing migrants. While we cannot measure the intensity of advertising of individual institutes, there is no indication that the institutes follow such a strategy. In our conversations with officials at the Goethe institutes, they repeatedly stated that they attempt to adjust to local demand rather than to actively encourage outgoing migrants to participate in courses. Third, the language learning experience at the Goethe institutes may motivate individuals to move to Germany, who initially take the course for non-migration related reasons. While we do not know if this is a good description of the experience of some language learners, the migration choice literature seems to agree that the key determinants of migration decisions are others, income differentials and migration policies in particular (Grogger and Hanson 2011; Ortega and Peri 2013; Bertoli and Fernández-Huertas Moraga 2015).

7 Summary, Conclusions, and Next Steps

In this paper, we use records of the German Goethe-Institut to construct a new dataset on language learning abroad and in Germany. For language learning abroad, the dataset covers 170 Goethe institutes in 89 countries for the years 1966–2012. It contains information on the numbers of students and exams, as well as taught hours and teachers. For language learning in Germany, the dataset reports the number of course participants from each of 198 countries in each year between 1980 and 2006. To the best of our knowledge this is the first large-scale dataset on adult language learning. We use the exams variable from the 'abroad' dataset and the participant number from the 'Germany' dataset for our estimations. Both measures vary considerably between and within institutes and/or countries and this variance is not explained by differences in population alone.

We use the dataset to investigate the determinants of language course and exam participation abroad and in Germany for the period 1990–2006 ('abroad') and for 1992–2006

('Germany'). Our results complement those of studies which use individual-level datasets to investigate the determinants of language skills, but not actual learning decisions, of migrants. We add to the literature by being able to disentangle language learning decisions that are made in the context of migration and other economic factors from the presence of language skills that may have been either the cause or the result of migration decisions. We argue that our results can be interpreted in terms of incentives to learn German.

Using fixed-effects regressions, we find that language learning at the Goethe institutes abroad is strongly associated with immigration from EU countries and with student migration from both EU and non-EU countries. Language learning in Germany is positively associated with student migration from EU and non-EU countries and migrant stocks from non-EU countries, but not with our measure of general migration flows.

We complement these results by running random-effects regressions that disentangle the within and between relationships between language learning and our variables of interest. We find that language learning abroad and in Germany is higher for countries/nationalities with higher average immigration and student immigration to Germany. This holds for both EU and non-EU countries. The positive association between average immigration and language learning weakens considerably as migrant stocks from the respective countries in Germany grow. This result is in line with several micro-level studies which find a negative association between migrants' language skills and their location in ethnic enclaves. We add to this literature, by showing that growing migrant stocks reduce both preparatory and post-migration language learning.

In the area of migration policy, our results allow policy makers to predict the language learning decisions of incoming migrants and to target language-learning programs at groups that would otherwise not invest in language learning. Two important results in this context are, first, that language learning before migration seems to be more common in EU than in non-EU countries, and second, that migrants from countries with larger migrant communities in Germany have a lower incentive to learn German.

We also find some evidence that language learning in Germany is positively associated with trade flows, but further investigation is required to assess the causal mechanisms behind this relationship.

References

Adserà, Alícia and Mariola Pytliková (2015). "The Role of Language in Shaping International Migration". *The Economic Journal* 125(586), F49–F81.

Aparicio Fenoll, Ainhoa and Zoë Kuehn (forthcoming). "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. CESifo Seminar Series.

Bell, Andrew and Kelvyn Jones (2014). "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data". *Political Science Research and Methods* FirstView, 1–21.

Belot, Michèle and Sjef Ederveen (2012). "Cultural barriers in migration between OECD countries". *Journal of Population Economics* 25(3), 1077–1105.

Belot, Michèle and Timothy J. Hatton (2012). "Immigrant Selection in the OECD". *The Scandinavian Journal of Economics* 114(4), 1105–1128.

Bertoli, Simone and Jesús Fernández-Huertas Moraga (2015). "The size of the cliff at the border". *Regional Science and Urban Economics* 51, 1–6.

Bleakley, Hoyt and Aimee Chin (2004). "Language Skills and Earnings: Evidence from Childhood Immigrants". *Review of Economics & Statistics* 86(2), 481–496.

— (2010). "Age at Arrival, English Proficiency, and Social Assimilation among US Immigrants". *American Economic Journal: Applied Economics* 2(1), 165–92.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2011). "Robust Inference With Multiway Clustering". *Journal of Business & Economic Statistics* 29(2), 238–249.

Chiswick, Barry R. and Paul W. Miller (1995). "The Endogeneity between Language and Earnings: International Analyses". *Journal of Labor Economics* 13(2), 246–288.

— (2007). "Computer usage, destination language proficiency and the earnings of natives and immigrants". *Review of Economics of the Household* 5(2), 129–157.

— (2015). "International Migration and the Economics of Language". In: *Handbook of the Economics of International Migration*. Ed. by Barry R. Chiswick and Paul W. Miller. Vol. 1A. Handbooks in Economics. Oxford: North Holland, 211–269.

Danzer, Alexander M. and Firat Yaman (2013). "Do Ethnic Enclaves Impede Immigrants' Integration? Evidence from a Quasi-experimental Social-interaction Approach". *Review of International Economics* 21(2), 311–325.

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 & Statistics* 83(4), 663–674.

Espenshade, Thomas J. and Haishan Fu (1997). "An Analysis of English-Language Proficiency among U.S. Immigrants". *American Sociological Review* 62(2), 288–305.

Goethe-Institut e.V. (2014). *Jahrbuch 2013/2014*.

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.

Isphording, Ingo Eduard and Sebastian Otten (2013). "The Costs of Babylon—Linguistic Distance in Applied Economics". *Review of International Economics* 21(2), 354–369.

Lazear, Edward P. (1999). "Culture and language". *Journal of Political Economy* 107(6), S95.

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

Mayda, Anna (2010). "International Migration: a Panel Data Analysis of the Determinants of Bilateral Flows". *Journal of Population Economics* 23(4), 1249–1274.

Mundlak, Yair (1978). "On the Pooling of Time Series and Cross Section Data". *Econometrica* 46(1), 69–85.

Ortega, Francesc and Giovanni Peri (2013). "The Effect of Income and Immigration Policies on International Migration". *Migration Studies* 1(1), 47–74.