RUHR
ECONOMIC PAPERS

Ingo E. Isphording
Sebastian Otten

Linguistic Distance and the Language Fluency of Immigrants

RUHR UNIVERSITÄT BOCHUM RUB RWI #274

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, email: W.Leininger@wiso.uni-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Christoph M. Schmidt
RWI, Phone: +49 (0) 201/81 49-227, e-mail: christoph.schmidt@rwi-essen.de

Editorial Office

Joachim Schmidt
RWI, Phone: +49 (0) 201/81 49-292, e-mail: joachim.schmidt@rwi-essen.de

Ruhr Economic Papers #274

Responsible Editor: Thomas K. Bauer

All rights reserved. Bochum, Dortmund, Duisburg, Essen, Germany, 2011

ISSN 1864-4872 (online) – ISBN 978-3-86788-319-1

The working papers published in the Series constitute work in progress circulated to stimulate discussion and critical comments. Views expressed represent exclusively the authors' own opinions and do not necessarily reflect those of the editors.

Ruhr Economic Papers #274

Isphording, Ingo E. and Sebastian Otten

Linguistic Distance and the Language Fluency of Immigrants

RUHR UNIVERSITÄT BOCHUM RUB
RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

ISSN 1864-4872 (online)
ISBN 978-3-86788-319-1

Ingo E. Isphording and Sebastian Otten[1]

Linguistic Distance and the Language Fluency of Immigrants

Abstract

We use a newly available measure of linguistic distance developed by the German Max Planck Institute for Evolutionary Anthropology to explain heterogeneity in language skills of immigrants. This measure is based on an automatical algorithm comparing pronunciation and vocabulary of language pairs. Using data from the German Socio-Economic Panel covering the period from 1997 to 2003, the linguistic distance measure is applied within a human capital framework of language acquisition. It is shown that linguistic distance is the most important determinant for host country language acquisition and that it explains a large fraction of language skill heterogeneity between immigrants. By lowering the efficiency and imposing higher costs of language learning, the probability of reporting good language skills is decreasing by increasing linguistic distance.

JEL Classification: F22, J15, J24, J40

Keywords: Linguistic distance; language; immigrants; human capital

August 2011

1 *Ingo E. Insphording, RUB and RUB Research School; Sebastian Otten, RUB and RWI.* *– The authors are grateful to Thomas K. Bauer, John P. Haisken-DeNew, Julia Bredtmann, Carsten Crede, Michael Kind, Jan Kleibrink, and Maren Michaelsen for helpful comments and suggestions. Financial support from the German-Israeli Foundation for Scientific Research and Development (GIF) is gratefully acknowledged. All remaining errors are our own.* *– All correspondence to Ingo Isphording, Chair for Economic Policy: Competition Theory and Policy, Ruhr-Universität Bochum, 44780 Bochum, Germany, E-Mail: ingo.isphording@rub.de.*

1 Introduction

The diversity of languages imposes one of the highest hurdles for migration, international trade and communication. Since the seminal work by Chiswick (1991) and Chiswick and Miller (1995, 1999), the role of language skills for the integration process of immigrants has obtained increasing attention by researchers in the field of migration. Language acts as the medium of everyday and working life, and is therefore a productive trait in itself. Low proficiency may also act as a signal for foreignness, allowing for discrimination and differentiation (Esser 2006). Effects of language fluency on economic outcomes have been analyzed in a multitude of studies, using wages (Dustmann 1994, Chiswick and Miller 2002), employment status (Dustmann and Fabbri 2003), occupational (Chiswick and Miller 2007) and locational choice (Bauer, Epstein and Gang 2005) as outcome variables. All studies conclude that language fluency is a crucial determinant for host country labor market success.

Chiswick and Miller (1995) offer a theoretical framework in which language skills are seen as a typical example of human capital, costly in acquisition, inseparable from the individual, and productive. Language skills are determined by characteristics affecting the efficiency in language acquisition, the exposure to the host country language, and individual economic incentives to learn the host country language. Following this framework, this study specifically focuses on the efficiency of language acquisition. Immigrants might be heterogeneous in their experience of difficulties in learning a foreign host country language, dependent on the similarity of their mother tongue to the host country language. Due to difficulties in measurement, the importance of linguistic distance, hence the dissimilarity between e.g. vocabularies, phonetic inventories, grammar, or script, on host country language fluency has been analyzed only rarely for a small number of English-speaking host countries. For the United States, Chiswick and Miller (1999) use data from the 1990 U.S. Census and include a measure of linguistic distance based on test scores of language classes. The same measure has been used for subsequent analyses using the 1991 Census of Canada (Chiswick and Miller 2001), and the 2000 U.S. Census (Chiswick and Miller 2007), showing a strong negative impact of linguistic distance on the level of language proficiency of immigrants, as it theoretically induces higher costs of learning. So far, the measure developed by Chiswick and Miller (1999) is only applicable for the analysis of the immigration to English speaking countries.

To being able to broaden this literature also to non-English-speaking countries, we propose to use an alternative way of measuring linguistic distance, based on recent developments in linguistic research by the German Max Planck Institute for Evolutionary Anthropology. The so-called Levenshtein distance (Bakker et al. 2009), which provides an easily and transparently computed measure of phonetic dissimilarity, relies on less re-

strictive identification assumptions than previous attempts and can be computed for any pair of host and home country language.

We use this measure in the setting of German language acquisition of immigrants. Determinants of immigrant language skills in Germany have been previously analyzed only by a few studies, starting with early work by Evans (1986). Dustmann (1999) analyzed host country language fluency as a jointly determined outcome along with migration duration, using data from the German Socio-Economic Panel (SOEP). He showed that the investment in language acquisition increases with time of intended stay. Using the same data, Dustmann and van Soest (2001) took into account potential misclassification in self-reported language fluency. Most recently, Danzer and Yaman (2010) analyzed German language fluency as a function of enclave density. However, the impact of country-of-origin characteristics, such like linguistic distance, on the economic integration of immigrants remains unclear and demands for further research (Esser 2006). Here, our new measure may provide new insights, e.g. in explaining the relatively weak economic position of Turks in Germany.

Our results suggest that linguistic distance to German is the strongest predictor of language fluency heterogeneity between immigrants. A higher linguistic distance is strongly related to lower levels of language fluency: Evaluated at the mean, a 1 percent higher linguistic distance between the home country language to German is associated with a decrease in the probability of reporting a *"Very good"* language fluency by 3.6 to 4.6 percent. The effect is stronger in the later stages of language acquisition and is linear in nature across the distribution of linguistic distance.

We briefly discuss our results against the background of current immigration policy addressing the language fluency of immigrants. For the resident immigrant population, the large heterogeneity in language acquisition efficiency, resulting from heterogeneity in linguistic background, requires a further flexibilization of language acquisition support. Points-based immigration schemes taking into account linguistic distance as an additional selection criterion may improve the average efficiency of language acquisition of the immigrant population.

The study proceeds as follows: In section 2 we summarize the theoretical framework by Chiswick and Miller (1995), which we use to analyze the determinants of language fluency, and introduce the Levenshtein distance, a newly available measure of phonetic distance between languages. Section 3 describes the data and the empirical model. In section 4 the results are presented and discussed. Section 5 concludes.

2 Background

2.1 Language Skills as Human Capital

The ability of immigrants to understand, speak, and write the language of the receiving country fulfills all requirements applied to human capital: it is productive as a determinant for wages, it is costly to obtain, either in terms of direct or opportunity costs, and it is a knowledge indivisibly embodied within a person (Chiswick and Miller 2007). Given these characteristics, it may be analyzed in a standard human capital accumulation framework, with some specific determinants, as proposed in Chiswick and Miller (1995). An optimal level of language fluency is determined by the intersection of a demand curve for language fluency (which represents the marginal returns to language acquisition in the labor market) and a supply curve dependent on the direct or opportunity costs of language acquisition, such as direct costs and forgone wages or earnings. Determinants of language fluency may be grouped into three subgroups, labeled by Chiswick and Miller (1995) as the three "E's": exposure, efficiency, and economic incentives. Such a rational choice framework is widely accepted in analyzing language acquisition processes across disciplines, see for example Esser (2006) for an adaption and extension of this model in sociological terms, dividing determinants into motivation, ability, costs, and opportunities.

First, language skills are influenced by the time an immigrant is exposed to the host country language. Immigrants may be exposed to the host country language prior to and after immigration (Chiswick and Miller 1995). Before immigration this may happen if the language of the destination country is a compulsory foreign language at school in the home country of an immigrant. Post-immigration, exposure is determined by neighborhood characteristics such as ethnic composition, family characteristics like number of children, or the working environment.

Immigrants also differ in their *economic incentives* (Chiswick and Miller 1995) for adapting a new language, determined by the expected labor market returns to language fluency. These returns are influenced by other non-language-related human capital components (e.g. occupation-specific knowledge or educational degrees) as well as the expected length of stay in the host country. However, as Dustmann (1999) states, language fluency acquisition and return migration may be jointly determined in a simultaneous decision process.

Finally, Chiswick and Miller (1995) looks at the *efficiency* of language acquisition which determines to what extent each unit of exposure may be transformed into language fluency. Younger people have a higher ability to learn a new language (Newport 1990). Therefore the age at entry in the host country crucially determines efficiency. Literacy and education also affect the ability of adapting a new language.

Most important in the context of this study, the efficiency is influenced by the linguistic distance between first language and acquired language. The further the linguistic distance between two languages, the more difficult it is to learn the host country language, which in turn imposes higher costs of learning.

2.2 Linguistic Distance

Proximity of languages is supposed to be a strong predictor of the decision to adapt a foreign language, as it crucially determines the costs of language acquisition. In linguistics, the distance between languages is a well known research issue. Using their historical development, language trees are developed to order languages into different families. Most prominently, the *Ethnologue* Project (see Lewis 2009) examines all known languages in the world. Unfortunately, although comprehensive, this language tree approach relies on very few increments between languages. As such this approach does not offer the possibility of deriving a continuous measure of linguistic distance.

To rely on such a continuous measure, Chiswick and Miller (1999) use the difficulty of learning languages in standardized language courses (see also Chiswick and Miller 2001, 2005), measured as the average exam score after 24 weeks of receiving lessons in one language. This score-based approach has the advantage of offering a measure of linguistic distance with more parameter values and variation, and of encompassing all dimensions of language differences. However, it might be biased by incentives and motivations to learn a foreign language. Dörnyei and Schmidt (2001) summarizes extrinsic and intrinsic motivations of learning a second language. Extrinsic motivation includes expected utility from being able to communicate in the language. Intrinsic motivation is derived from the fact of learning the language itself, e.g. by boosting the own reputation among friends and peers. These motivations are likely to differ across potential second languages (e.g. it might yield higher economic returns to learn one language instead of another). Although affecting average test scores, these intrinsic and extrinsic incentives do not tell anything about actual linguistic distance. Additionally, as Chiswick and Miller (2001) note, this measure relies on a symmetry assumption, as the fact of learning difficulty for U.S. Americans is assumed to represent the difficulties faced by immigrants in learning English.

The measure of linguistic distance proposed in this study has the advantage of offering a continuous variable that can be used in empirical analyses for any host country language. The measure is based on the *Automatic Similarity Judgement Program* (ASJP) by the German Max Planck Institute for Evolutionary Anthropology.[1] The project aims at developing an automatic procedure to evaluate the phonetic similarity between all of the

[1] Further information can be found at http://www.eva.mpg.de.

world's languages (so far, most languages relevant for migration research are covered).

The project relies on a "lexicostatistical" approach, which uses a core set of vocabulary for each language describing common things and environments, called the *Swadesh list*. Following Bakker et al. (2009), a minimized list of the 40 most stable words, which are shown in Table 1, is used in computing the measure. These words are then expressed in a special phonetic transcription called *ASJP code*, which uses all available characters on a standard QWERTY keyboard to represent all common sounds in human communication.

The words from this 40-word list are automatically judged concerning their similarity leading to a scalar of linguistic distance, the so-called Levenshtein distance. The measure is basically computed as follows: For each word pair, it is evaluated how many additions or subtractions are necessary to transform one word in one language into the same word in another language. For example, the English word *mountain*, expressed in the ASJP code as *maunt3n*, has to be transferred into the German word *Berg (bErk)*. This value is then normalized by the potential maximum distance between both words. The sum of these distances is divided by the number of words that exist in both compared lists and is again normalized by the similarity of phoneme inventories of the language pair. For a more detailed description of the computation, see Bakker et al. (2009).

This measure offers some advantages compared to previous measures of linguistic distance. Isphording and Otten (2011) compare the new measure with Chiswick's approach using data from the U.S. census and find a qualitatively stronger effect of the Levenshtein distance in explaining immigrant's language skills. There are several reasons why the new measure might lead to more precise and efficient results. It is independent of the used data source and is not likely to be biased by economic incentives like the measure used by Chiswick and Miller (1999). Compared to the latter, it offers a much higher variation as it is not restricted to certain parameter values. It is comprehensive (all relevant languages are covered by the ASJP database) and can therefore even be used for rather "exotic" immigrants with few observations that are otherwise typically excluded from the analysis. Also, it can be used for any host country language included in the ASJP database, e.g. for important immigration countries such as the United States, United Kingdom, Canada, Germany, and France.[2] However, because of the comprehensiveness of the database, it may also be used for analyses concerning south-south migration including rather seldom analyzed languages. Maybe most important, the measure is easily and transparently com-

[2]As the SOEP dataset does not offer information on language of birth, we assign languages by country of birth. In multi-lingual countries, languages were assigned as the most prevalent native language (excluding *lingua francas*, i.e. commonly known foreign languages used for trade and communication across different mother tongues), which was identified using a multitude of sources, including factbooks, encyclopedias and Internet resources. A comprehensive index of assigned languages with further explanations is available upon request. This assumption might lead to an attenuation bias of the coefficients of the linguistic distance measure by introducing a measurement error in our linguistic distance measure.

puted. Due to its purely descriptive nature, it is not biased by individual incentives to learn foreign languages.

The computation for all languages included in our dataset for the specific case of Germany results in a right-skewed distribution. The closest languages to German are the Benelux languages (Luxembourgish, 42.12; Dutch, 51.50; Westvlaams, 57.86), and the Scandinavian languages (Norwegian, 64.92; Swedish, 66.56). The furthest languages are Korean (104.30), Arabic (103.72), and Yoruba (spoken in Nigeria, 103.58).

The Levenshtein distance is not without its shortcomings. As Chiswick and Miller (2005) state, languages "differ in vocabulary, grammar, written form, syntax and myriad other characteristics". Clearly, the Levenshtein distance only covers the first of these dimensions. However, it has been shown that the Levenshtein distance is a very good predictor for e.g. genetic distance of languages (Bakker et al. 2009). Correlations in vocabulary result from close historical and evolutionary relationships of languages, and therefore are related to further similarities in additional dimensions.[3] For simplicity, we refer to the Levenshtein distance as linguistic distance throughout our analysis.

3 Data and Empirical Model

3.1 Data

The data used in this study is taken from the German Socio-Economic Panel (SOEP)[4] covering the period between 1997 and 2003. Questions concerning the language fluency of immigrants are included in every second wave. The sample is restricted to individuals who were at least 17 when migrating to Germany and who are not older than 65 at the time of the survey. Therefore, we explicitly exclude individuals who learned German already during their childhood or early adolescence in Germany. After excluding observations containing missing values we use four subsequent cross-sectional samples ranging in size from 1102 observations in 1999, to 1430 observations in 2001.

The sample consists of immigrants from 89 countries. The largest fraction of immigrants has been immigrated from Turkey (24.2 percent in 1997), followed by Italy, Serbia (and Croatia), and Poland. Over time, the sample becomes more diverse in terms of

[3] We focus in this study on the effect of linguistic distance on speaking fluency. Nonetheless, regression results not reported here confirm a similar influence of the linguistic distance on written proficiency.

[4] The SOEP is a panel survey conducted since 1984 and covering roughly 20.000 individuals per wave. For more information see Haisken-DeNew and Frick (2005). The data used in this thesis was extracted using the Add-On package PanelWhiz for Stata. PanelWhiz (http://www.PanelWhiz.eu) was written by John P. Haisken-DeNew (john@PanelWhiz.eu). See Haisken-DeNew and Hahn (2006) for details. The PanelWhiz generated DO file to retrieve the data used here is available upon request. Any data or computational errors in this thesis are the authors'.

country of origin, i.e., the above mentioned sending countries contribute 80.4 percent of the overall sample in 1997, but only 67.2 percent in 2003.

Language proficiency is assessed as an ordered discrete self-reported measure of oral and written German language proficiency ranging from 1 (*"Not at all"*) to 5 (*"Very good"*). The self-assessed ability to speak German *"Good"* is the most frequent answer; only 1.5 percent report not speaking German at all. To avoid estimation problems caused by empty cells, the categories *"Not at all"* and *"Fairly bad"* are joined into one category, leading to the four categories *"Bad"*, *"Not bad"*, *"Good"*, and *"Very good"* used in the empirical analysis. The distribution of language skills changes over time. The relative frequency of reporting *"very good"* language fluency nearly doubles from 10.7 percent in 1997 to 20.5 percent in 2003 (Table 2). German language fluency is negatively correlated to the linguistic distance measure by $r = -0.3$.

The explanatory variables are chosen to represent the slope and shift parameters of the human capital accumulation model discussed in section 2.1, which are the efficiency, exposure, and economic incentives influencing language skill acquisition. Appendix-Table A1 shows descriptive statistics of the variables used in the empirical analysis.[5] The major variable of interest is the linguistic distance between home and host country language as measured by the Levenshtein distance. Years of education is a generated variable taking into account the minimum amount of years needed to reach a certain school degree. Following Dustmann (1994), a dummy is included, coded as one for those who report a *"Good"* or *"Very good"* level of written home country language proficiency to control for the individual level of literacy. We control for whether an immigrant visited a German school after immigration. As the sample is restricted to those who migrated to Germany at age 17 or older, this variable should mostly indicate vocational training schools. Together with a low language distance, entering Germany at a younger age, having higher education, visiting a school in Germany and having a high level of literacy are expected to increase the efficiency of acquiring a new language.

Exposure is expected to be influenced by the time of residence and family characteristics. These are controlled for by the years since migration, dummies for gender, being married and the number of children in the household. The family situation may change the degree of social inclusion and thereby the exposure to the host country language. A priori, it is difficult to hypothesize the sign of this variable. A dummy for neighboring countries of Germany proxies the probability of learning German at school and living in an area with frequent meetings with Germans.

Economic incentives are represented by the self-reported desired time to stay in Germany measured in years. We censor the length of planned duration of stay at age 65

[5] Table A2 in the Appendix provides the detailed description of the variable definitions.

to account only for desired length of stay during the economically active period of life.[6] Further, a dummy for having family abroad controls for return plans that might alter economic incentives.

The geographic distance is included as distance in units of 100 km between Berlin as the capital of Germany and the capital of the country of origin.[7] While controlling for linguistic distance, geographic distance is mostly a proxy for migration costs. Only those immigrants who expect to recover these higher costs by higher labor market incomes will decide to migrate. Potentially, these immigrants are positively selected, e.g. in terms of unobservable motivation. This should lead to a positive coefficient of geographic distance, that could be interpreted as direct evidence for the self-selection of immigrants, as discussed by Borjas (1987). To further control for heterogeneity in country-of-origin characteristics, we include a set of regional dummies.[8]

3.2 Empirical Model

The time-constant nature of our linguistic distance measure as major variable of interest makes it unfeasible to rely on panel data methods. The analysis therefore focuses on the application of standard Ordered Logit models for repeated cross-sections to estimate the determinants of oral language proficiency in Germany.[9] The analysis starts with the linear probability model, assuming cardinality of the dependent variable. This model is also used to compute the contributions in explanatory power of single variables, and acts as a robustness check and benchmark for the following Ordered Logit model.

Both models, the linear probability and the Ordered Logit model, use the fourfold self-reported measure of language proficiency as dependent variable, which is explained by the control variables described in section 3. To account for potential nonlinearities in the effect of linguistic distance on language acquisition, the measure enters the specification in two different ways. In a first set of estimations, absolute values, directly computed as described above, are used.[10] Secondly, a set of four splines constructed as the product

[6] We repeated our estimations without and with lagged values of desire to stay to test for potential endogeneity. The results remained stable.

[7] The geographic distance data are compiled by researchers at Centre d'Etudes Prospectives et d'Informations Internationales (CEPII) and available at http://www.cepii.fr/anglaisgraph/bdd/distances.htm.

[8] We also tried different definitions of regions and specifications including variables for genetic differences. The results regarding linguistic distance remained stable in sign and magnitude.

[9] We also estimated the same specification on a pooled sample including all four waves. The results differed only marginally. Furthermore, all specifications have been estimated by the Generalized Ordered Logit model. The results remain stable. We describe the method and the results in the supplementary Appendix.

[10] The distribution of the measure of linguistic distance is right-skewed, with a majority of languages at the right end of the distribution. To take distributional issues into account, we repeated our estimations with a variable containing percentiles of this distribution instead of absolute values, leading to similar,

of quartile dummies and the absolute values enter the estimation equation to control for a non-linear relationship between linguistic distance and language acquisition. In all specifications the standard errors are allowed to cluster by the home language.

4 Results

Table 3 reports the results of the OLS benchmark estimations for all four waves.[11] Due to the non-cardinal character of the dependent variable, the coefficients can only reasonably be interpreted in sign and compared in magnitude, although the magnitude has no inherent meaning. Therefore, we also report marginal effects of the Ordered Logit model, summarized in Appendix-Tables A3.1 and A3.2, to allow for an interpretation in probabilities of chosing a category.[12] We use the OLS results to determine the contribution of single variables to the explained variance in language skills, expressed by the squared semi-partial correlations following Cohen et al. (2003). Note that these semi-partial correlation constitute a conservative measure. The values do not sum up to the complete R^2. However, they allow for a comparison of explanatory power between different regressors.

The control variables in the OLS specification support the theoretical considerations by Chiswick and Miller (1995). Exposure to the host country language is related to higher reported language skills, indicated by the positively significant coefficients (in 2001 and 2003) of being originated in a neighboring country (pre-immigration exposure), and by the positive coefficients of years since migration (post-immigration exposure). The effect of years since migration diminishes over time. Having a family abroad is negatively correlated to the reported skills, the number of years of desired stay in Germany is positively correlated to the reported language skills, indicating the influence of economic incentives on the language acquisition process.

The results for the control variables in the Ordered Logit model are in line with the OLS results and again indicate an increase of language skills with years since migration (with a decreasing rate). Children are negatively related to the language skills, this relation is stronger than in the OLS results. Having visited a German school is positively related with higher German language skills.

The measure for linguistic distance has a stable and significantly negative effect in the OLS results. Moving up the distribution of linguistic distance decreases the probability of reporting higher categories of language fluency. Contrarily, the geographic distance has a

but less significant results, most likely due to the lower variation in the percentile measure.

[11] All estimation output tables were generated using the Stata routine *estout* by Ben Jann, see Jann (2007).

[12] The underlying coefficients are comparable in sign and significance with those of the OLS regressions and are reported in Table A4 in the supplementary Appendix.

positive effect on language fluency when controlling for linguistic distance, in line with the argument of self-selection discussed in section 3. In terms of explanatory power (measured as squared semi-partial correlations), linguistic distance is the strongest predictor of the specification, apart from the Eastern Europe region dummy. Linguistic distance alone accounts for 10 to 19 percent of the explained variation of the model.

The Ordered Logit results allow for an interpretation in magnitude and reveal a strong negative relationship between linguistic distance and the reported language skills. Increasing the linguistic distance by one unit decreases the probability of reporting the highest category of language skills by 0.3 to 0.6 percentage points. Again, positive marginal effects of geographic distance on the probability to report higher categories of language fluency might indicate a pre-migration self-selection process. For an easier interpretation of the effect of linguistic distance, due to the lack of a natural unit, elasticities and marginal effects multiplied by the interquartile ranges of linguistic distance are reported in Table 4. Increasing linguistic distance by one percent decreases the probability of reporting the highest category by 3.6 to 4.6 percent. Moving up the distribution of linguistic distance from the lower to the upper quartile decreases the probability of reporting "Very Good" language skills by 2.7 to 3.6 percentage points. This interpretation is only meaningful if the effect of linguistic distance is linear across the distribution, which is supported by the inclusion of partially defined splines (as a product of absolute values of linguistic distance and quartile dummies). The respective coefficients and marginal effects are summarized in Table 5. The results indicate no significant or systematic differences of effects across quartiles. This result is stable across categories, models and waves, indicating linearity in the negative effect across the range of linguistic distance.

To demonstrate the importance of our results in a more direct manner, we use the estimates of the Ordered Logit model to predict counterfactual distributions of predicted probabilities to report category *"Very good"* in 2003, changing linguistic distance, but keeping all other covariates constant. The kernel density estimates of these counterfactual distributions are shown in Figure 1.[13] For different counterfactual distributions, the linguistic distance is set to the values of 0 as a benchmark, to German-Dutch (51.50), German-English (72.21), and German-Turkish (99.91). Additionally, the observed distribution without changing the linguistic distance is reported.

Changing the distribution from the observed to the counterfactual "German" distribution moves the probability mass from the left to the right end. About 78 percent of the sample have probabilities of more than 0.8 to report *"Very good"* language fluency. In the observed non-counterfactual distribution, only 4 percent have probabilities of more than 0.8. This result is quite intuitive, as people should most likely report *"Very good"* language

[13]The kernel density estimates of the counterfactual distributions for the years 1997-2001 are presented in Figures A1-A3 in the supplementary Appendix.

fluency when they have no linguistic distance at all to bridge. This result illustrates the importance of linguistic distance as a determinant for language fluency.

5 Conclusion

Linguistic distance, the dissimilarity of languages, supposes to be major determinant of host country language acquisition of immigrants. In this study, we introduced a new measure of linguistic distance to explain differences in speaking fluency among immigrants in Germany. Based on routines developed by the German Max Planck Institute for Evolutionary Anthropology, the measure is computed as a normalized and averaged measure of phonetic similarity between words of different languages.

This measure is easily computed as a continuous variable for any potential language pair and allows for a comprehensive analysis of distance effects in host country language acquisition. Compared to previous attempts to measure linguistic distance, this measure offers advantages in terms of transparency of computation and the necessity of identification assumptions. Due to the comprehensiveness of covered languages, it offers a broad range of applications in economic research.

We applied this measure in the human capital framework introduced by Chiswick and Miller (1995) as a determinant for language acquisition of first generation immigrants in Germany, using data from the German Socio-Economic Panel (SOEP). The linguistic distance captures the effect of higher acquisition costs for those with more distant mother languages. Our main findings can be summarized as follows: (i) robust across all specifications, higher linguistic distance decreases the probability of higher fluency in host country language; (ii) the impact of linguistic distance is more important in later stages of language acquisition; (iii) the negative effect is linear throughout the range of linguistic distance. In our preferred specification, a 1 percent increase in linguistic distance decreases the probability of reporting *"Very good"* language fluency by 3.6 to 4.6 percent. Moving up the distribution of linguistic distance from the lower to the upper quartile decreases the probability of reporting "Very good" language skills by 2.7 to 3.6 percentage points. Spline regressions indicate linearity of this effect across the range of linguistic distance.

Our data also shows some indirect evidence for self-selection. Controlling for linguistic distance, geographic distance as a proxy for migration costs has a positive influence on language fluency. Only those individuals chose to migrate who expect to recover migration costs by higher expected wages, and therefore represent a self-selected group in terms of motivation or unobserved skills.

Our results shed light on a major source of heterogeneity in language acquisition

efficiency, and as such, on the organization and implementation of language acquisition support. Given that increasing language skills among migrants is a commonly accepted political aim, this additional source of heterogeneity has to be taken into account in designing language course systems for the resident migrant population. This is in line with recent claims for further flexibilization of the German integration course system (Bundesministerium des Innern 2006).

Political measures addressing the resident population have to take the distribution of linguistic distance as exogenously given. However, using political measures prior immigration might be able to alter this distribution. The ongoing discussion on the introduction of a points-based immigration scheme, similar to the ones of Australia, Canada, or more recently also in the UK, would allow to take linguistic distance as a proxy for host country language learning efficiency into account, lowering the expected average costs for later language support.

References

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. "Adding typology to lexicostatistics: A combined approach to language classification." *Linguistic Typology*, 13(1): 169–181.

Bauer, Thomas, Gil S. Epstein, and Ira N. Gang. 2005. "Enclaves, language, and the location choice of migrants." *Journal of Population Economics*, 18(4): 649–662.

Borjas, George J. 1987. "Self-Selection and the Earnings of Immigrants." *The American Economic Review*, 77(4): 531–553.

Brant, Rollin. 1990. "Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression." *Biometrics*, 46(4): 1171–1178.

Bundesministerium des Innern. 2006. "Evaluation der Integrationskurse nach dem Zuwanderungsgesetz: Abschlussbericht und Gutachten über Verbesserungspotenziale bei der Umsetzung der Integrationskurse." `http://www.bmi.bund.de/SharedDocs/Downloads/DE/Veroeffentlichungen/evaluation_integrationskurse_de.pdf?__blob=publicationFile`.

Chiswick, Barry R. 1991. "Speaking, Reading, and Earnings among Low-Skilled Immigrants." *Journal of Labor Economics*, 9(2): 149–170.

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

Chiswick, Barry R., and Paul W. Miller. 1999. "English language fluency among immigrants in the United States." In *Research in Labor Economics*. Vol. 17, ed. Solomon W. Polachek, 151–200. Oxford: JAI Press.

Chiswick, Barry R., and Paul W. Miller. 2001. "A Model of Destination-Language Acquisition: Application to Male Immigrants in Canada." *Demography*, 38(3): 391–409.

Chiswick, Barry R., and Paul W. Miller. 2002. "Immigrant earnings: Language skills, linguistic concentrations and the business cycle." *Journal of Population Economics*, 15(1): 31–57.

Chiswick, Barry R., and Paul W. Miller. 2005. "Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages." *Journal of Multilingual and Multicultural Development*, 26(1): 1–11.

Chiswick, Barry R., and Paul W. Miller. 2007. "Modeling Immigrants' Language Skills." Institute for the Study of Labor (IZA) Discussion Paper No. 2974.

Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* 3rd ed. Mahwah, NJ: Routledge Academic.

Danzer, Alexander M., and Firat Yaman. 2010. "Ethnic Concentration and Language Fluency of Immigrants in Germany." Institute for the Study of Labor (IZA) Discussion Paper No. 4742.

Dörnyei, Zoltán, and Richard Schmidt. 2001. *Motivation and second language acquisition.* Vol. 23 of *Technical Report/Second Language Teaching & Curriculum Center* Honolulu, Hawaii: Univ. of Hawaii.

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

Dustmann, Christian. 1999. "Temporary Migration, Human Capital, and Language Fluency of Migrants." *Scandinavian Journal of Economics*, 101(2): 297–314.

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.

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

Esser, Hartmut. 2006. "Migration, Language and Integration: AKI Research Review 4." `http://bibliothek.wz-berlin.de/pdf/2006/iv06-akibilanz4b.pdf`.

Evans, Mariah D. R. 1986. "Sources of Immigrants' Language Proficiency: Australian Results with Comparisons to the Federal Republic of Germany and the United States of America." *European Sociological Review*, 2(3): 226–236.

Greene, William H., and David A. Hensher. 2010. *Modeling Ordered Choices: A Primer.* Cambridge: Cambridge Univ. Press.

Haisken-DeNew, John P., and Joachim R. Frick. 2005. "Desktop Companion to the German Socio-Economic Panel (SOEP): Version 8.0." `http://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.38951.de/dtc.409713.pdf`.

Haisken-DeNew, John P., and Markus Hahn. 2006. "PanelWhiz: A Flexible Modularized Stata Interface for Accessing Large Scale Panel Data Sets." `http://www.panelwhiz.eu/docs/PanelWhiz_Introduction.pdf`.

Isphording, Ingo E., and Sebastian Otten. 2011. "The Costs of Babylon – Linguistic Distance in Applied Economics." mimeo.

Jann, Ben. 2007. "Making regression tables simplified." *Stata Journal*, 7(2): 227–244.

Lewis, Paul M. 2009. *Ethnologue: Languages of the World.* 16th ed. Dallas, Tex: SIL International.

Long, Scott J., and Jeremy Freese. 2006. *Regression models for categorical dependent variables using Stata.* 2nd ed. College Station, Tex: Stata Press.

Newport, Elissa L. 1990. "Maturational constraints on language learning." *Cognitive Science*, 14(1): 11–28.

Peterson, Bercedis, and Frank E. Harrell, Jr. 1990. "Partial Proportional Odds Models for Ordinal Response Variables." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(2): 205–217.

Williams, Richard. 2006. "Generalized ordered logit/partial proportional odds models for ordinal dependent variables." *Stata Journal*, 6(1): 58–82.

Tables

Table 1: 40-Items Swadesh Word List

I	You	We	One
Two	Person	Fish	Dog
Louse	Tree	Leaf	Skin
Blood	Bone	Horn	Ear
Eye	Nose	Tooth	Tongue
Knee	Hand	Breast	Liver
Drink	See	Hear	Die
Come	Sun	Star	Water
Stone	Fire	Path	Mountain
Night	Full	New	Name

Source: Bakker et al. (2009).

Table 2: Distribution of Language Skills across Years

Proficiency	Year				Total
	1997	1999	2001	2003	
Bad	247	164	206	163	780
	18.84	14.88	14.41	13.12	15.34
Not bad	470	383	466	382	1701
	35.85	34.75	32.59	30.76	33.45
Good	454	418	496	443	1811
	34.63	37.93	34.69	35.67	35.61
Very good	140	137	262	254	793
	10.68	12.43	18.32	20.45	15.59
Total	1311	1102	1430	1242	

Notes: – Cells include absolute counts and relative frequencies.

Table 3: Estimation Results Oral Language Fluency – Linear Regression

	1997		1999		2001		2003	
	Coef/StdE	Expl	Coef/StdE	Expl	Coef/StdE	Expl	Coef/StdE	Expl
Female	−0.025	0.123	−0.059	0.544	−0.008	0.014	−0.018	0.052
	(0.059)		(0.062)		(0.060)		(0.055)	
Married	−0.058	0.314	−0.099	0.682	−0.105	1.020	−0.066	0.365
	(0.090)		(0.067)		(0.065)		(0.049)	
Children in the HH. (Ref.= 0)								
One Child	−0.086	0.868	−0.109	0.992	−0.039	0.162	−0.126**	1.375
	(0.055)		(0.082)		(0.096)		(0.039)	
Two Children	−0.119	1.320	−0.146*	1.374	−0.096	0.738	−0.104	0.661
	(0.074)		(0.073)		(0.105)		(0.084)	
Three or more Children	−0.120†	0.911	−0.122	0.621	−0.118	0.782	−0.277***	3.335
	(0.069)		(0.099)		(0.095)		(0.078)	
Years since Migration	0.058***	9.409	0.059**	6.583	0.042**	5.750	0.044**	4.852
	(0.014)		(0.018)		(0.014)		(0.014)	
Years since Migration^2/100	−0.100*	4.724	−0.099*	3.804	−0.083*	4.490	−0.091*	5.008
	(0.037)		(0.039)		(0.038)		(0.035)	
Neighboring Country	0.055	0.187	0.172	1.476	0.228†	3.705	0.269*	4.704
	(0.168)		(0.158)		(0.114)		(0.122)	
Family abroad	−0.066	0.326	−0.155**	1.636	−0.190*	4.833	−0.032	0.117
	(0.061)		(0.056)		(0.086)		(0.048)	
Desired Stay (years)	0.014**	6.204	0.018***	4.997	0.003	0.255	−0.006†	0.492
	(0.004)		(0.003)		(0.004)		(0.003)	
Distance Capitals (100 km)	0.006*	4.031	0.007**	4.565	0.006**	6.776	0.005**	4.252
	(0.003)		(0.003)		(0.002)		(0.002)	
Age at Entry	0.003	0.286	0.011†	1.618	−0.003	0.212	−0.020**	4.593
	(0.008)		(0.006)		(0.005)		(0.006)	
Years of Education	0.115†	1.445	0.187**	3.030	0.220***	4.269	0.158*	2.079
	(0.066)		(0.060)		(0.053)		(0.074)	
Years of Education^2/100	−0.137	0.109	−0.476†	1.056	−0.529*	1.378	−0.172	0.145
	(0.325)		(0.265)		(0.207)		(0.307)	
School attended in Germany	0.470***	5.959	0.579***	8.038	0.301**	2.815	0.350*	3.610
	(0.098)		(0.114)		(0.102)		(0.144)	
Proficiency Home Language	0.253**	6.866	0.324**	7.913	0.193	3.008	0.137	1.461
	(0.084)		(0.095)		(0.138)		(0.095)	
Linguistic Distance	−0.016***	18.333	−0.014***	9.725	−0.013***	19.045	−0.011***	16.937
	(0.003)		(0.004)		(0.002)		(0.002)	
Constant	1.335*		0.519		1.184*		1.883**	
	(0.520)		(0.459)		(0.493)		(0.611)	
Region Dummies	yes		yes		yes		yes	
Adjusted R^2	0.227		0.289		0.257		0.361	
F Statistic	315.28***		196.83***		86.52***		124.54***	
Observations	1311		1102		1430		1242	

*Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Cluster-robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 1 to 4 such that higher values indicate a higher level of oral proficiency. – The explanatory power of each variable is computed as the squared semipartial correlation. The values represent the variables' explained proportion of the total variance.*

Table 4: Elasticities & Interquartile Ranges of Linguistic Distance on Oral Language Fluency – Linear & Ordered Logit Regression

	OLS	Ordered Logit				Ordered Logit Marginal Effects * IQR			
		Bad	Not bad	Good	Very good	Bad	Not bad	Good	Very good
1997	−0.626***	4.019***	1.410***	−2.257***	−4.351***	0.053	0.052	−0.075	−0.029
	(0.115)	(0.918)	(0.427)	(0.443)	(0.936)				
1999	−0.543***	4.545**	2.039**	−2.095***	−4.615**	0.031	0.059	−0.063	−0.027
	(0.139)	(1.469)	(0.737)	(0.607)	(1.443)				
2001	−0.462***	3.974***	1.969***	−1.394***	−3.797***	0.024	0.044	−0.035	−0.033
	(0.073)	(0.691)	(0.407)	(0.223)	(0.620)				
2003	−0.396***	3.901***	2.237***	−1.092***	−3.621***	0.020	0.054	−0.037	−0.036
	(0.071)	(0.777)	(0.481)	(0.234)	(0.698)				

*Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Cluster-robust standard errors are reported in parentheses. – Elasticities and marginal effects are reported at the mean of the covariates vector. – Column 2-6 show the OLS and the Ordered Logit elasticities of linguistic distance. – Column 7-10 show the Ordered Logit marginal effects multiplied by the interquartile ranges of linguistic distance.*

Table 5: Estimation Results Splines Oral Language Fluency – Linear & Ordered Logit Regression

	1997 Coef/StdE	1999 Coef/StdE	2001 Coef/StdE	2003 Coef/StdE
Linear Regression				
Spline 1st Quartile	−0.020***	−0.016**	−0.011***	−0.011***
	(0.004)	(0.005)	(0.002)	(0.002)
Spline 2nd Quartile	−0.019***	−0.015**	−0.010**	−0.011**
	(0.005)	(0.005)	(0.003)	(0.003)
Spline 3rd Quartile	−0.018***	−0.015***	−0.012***	−0.011***
	(0.004)	(0.004)	(0.002)	(0.002)
Spline 4th Quartile	−0.018***	−0.015**	−0.012**	−0.011***
	(0.004)	(0.005)	(0.003)	(0.003)
Ordered Logit Regression				
Spline 1st Quartile	−0.088**	−0.073*	−0.052***	−0.051***
	(0.030)	(0.032)	(0.010)	(0.012)
Spline 2nd Quartile	−0.082**	−0.071*	−0.047***	−0.050***
	(0.029)	(0.031)	(0.012)	(0.012)
Spline 3rd Quartile	−0.078**	−0.068*	−0.050***	−0.049***
	(0.026)	(0.028)	(0.009)	(0.010)
Spline 4th Quartile	−0.078**	−0.071*	−0.049***	−0.048***
	(0.027)	(0.029)	(0.012)	(0.011)
Marginal Effects Ordered Logit Regression	ME/StdE	ME/StdE	ME/StdE	ME/StdE
Bad				
Spline 1st Quartile	0.011**	0.006*	0.005***	0.003***
	(0.004)	(0.003)	(0.001)	(0.001)
Spline 2nd Quartile	0.010**	0.006*	0.004***	0.003***
	(0.003)	(0.003)	(0.001)	(0.001)
Spline 3rd Quartile	0.010**	0.006*	0.004***	0.003***
	(0.003)	(0.002)	(0.001)	(0.001)
Spline 4th Quartile	0.010**	0.006*	0.004***	0.003***
	(0.003)	(0.003)	(0.001)	(0.001)
Not bad				
Spline 1st Quartile	0.011**	0.012*	0.008***	0.009***
	(0.004)	(0.005)	(0.002)	(0.002)
Spline 2nd Quartile	0.010**	0.012*	0.007***	0.009***
	(0.004)	(0.005)	(0.002)	(0.002)
Spline 3rd Quartile	0.010**	0.011*	0.008***	0.009***
	(0.004)	(0.005)	(0.001)	(0.002)
Spline 4th Quartile	0.010**	0.012*	0.008***	0.008***
	(0.004)	(0.005)	(0.002)	(0.002)
Good				
Spline 1st Quartile	−0.016**	−0.013*	−0.007***	−0.006***
	(0.005)	(0.006)	(0.001)	(0.002)
Spline 2nd Quartile	−0.015**	−0.012*	−0.006***	−0.006***
	(0.005)	(0.005)	(0.001)	(0.002)
Spline 3rd Quartile	−0.014**	−0.012*	−0.006***	−0.006***
	(0.005)	(0.005)	(0.001)	(0.001)
Spline 4th Quartile	−0.014**	−0.012*	−0.006***	−0.006***
	(0.005)	(0.005)	(0.001)	(0.001)
Very good				
Spline 1st Quartile	−0.006*	−0.005*	−0.006***	−0.006***
	(0.003)	(0.003)	(0.001)	(0.002)
Spline 2nd Quartile	−0.006*	−0.005*	−0.006**	−0.006***
	(0.002)	(0.003)	(0.002)	(0.002)
Spline 3rd Quartile	−0.005*	−0.005*	−0.006***	−0.006***
	(0.002)	(0.002)	(0.001)	(0.001)
Spline 4th Quartile	−0.005*	−0.005*	−0.006***	−0.006***
	(0.002)	(0.002)	(0.002)	(0.002)

*Notes: – Significant at: *** 0.1% level; ** 1% level; * 5% level; † 10% level. – Cluster-robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 1 to 4 such that higher values indicate a higher level of oral proficiency. – The four splines are constructed as the product of a quartile dummy and the absolute values of Linguistic Distance. – Marginal effects are reported at the mean of the covariates vector.*

Figures

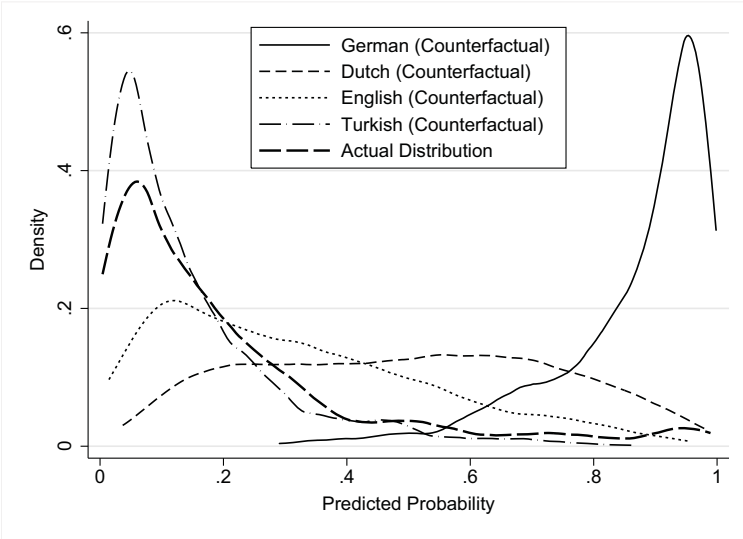

Figure 1: Kernel Density Estimates of Actual and Counterfactual Distributions of Predicted Probabilities to Report Category "Very good" in 2003

Appendix

Table A1: Descriptive Sample Statistics

	1997 Mean/StdD	1999 Mean/StdD	2001 Mean/StdD	2003 Mean/StdD
Oral Proficiency German	2.371	2.479	2.569	2.634
	(0.908)	(0.893)	(0.949)	(0.951)
Female	0.494	0.496	0.515	0.535
	(0.500)	(0.500)	(0.500)	(0.499)
Married	0.869	0.872	0.853	0.849
	(0.338)	(0.334)	(0.354)	(0.359)
Children in the HH.				
No Children	0.475	0.516	0.488	0.530
	(0.500)	(0.500)	(0.500)	(0.499)
One Child	0.225	0.218	0.224	0.219
	(0.418)	(0.413)	(0.417)	(0.414)
Two Children	0.195	0.172	0.179	0.151
	(0.397)	(0.378)	(0.384)	(0.359)
Three or more Children	0.105	0.093	0.108	0.100
	(0.306)	(0.291)	(0.311)	(0.300)
Years since Migration	18.609	19.757	17.615	20.589
	(10.554)	(10.753)	(10.762)	(10.816)
Neighboring Country	0.124	0.124	0.150	0.166
	(0.330)	(0.330)	(0.357)	(0.372)
Family abroad	0.131	0.175	0.316	0.331
	(0.338)	(0.380)	(0.465)	(0.471)
Desired Stay (years)	16.789	16.717	18.616	16.841
	(11.740)	(11.765)	(11.788)	(11.074)
Distance Capitals (100 km)	17.360	17.793	20.022	19.758
	(13.608)	(13.761)	(17.213)	(18.639)
Age at Entry	27.775	27.662	28.192	27.355
	(8.614)	(8.871)	(8.932)	(8.238)
Years of Education	10.022	10.049	10.475	10.750
	(2.430)	(2.427)	(2.400)	(2.596)
School attended in Germany	0.034	0.039	0.041	0.048
	(0.180)	(0.194)	(0.199)	(0.213)
Proficiency Home Language	0.842	0.857	0.882	0.867
	(0.365)	(0.351)	(0.323)	(0.340)
Linguistic Distance	93.562	93.643	92.075	90.934
	(10.378)	(9.573)	(14.242)	(16.994)
Observations	1311	1102	1430	1242

Notes: – Oral Proficiency is measured on a scale of 1 to 4, corresponding to the classifications "Bad", "Not bad", "Good", "Very good".

Table A2: Variables Description

Variable	Description
Oral Proficiency German	Self-reported spoken German proficiency
Female	Dummy = 1 if female
Married	Dummy = 1 if married
Children in the HH.	
No Children	Dummy = 1 if no children in household
One Child	Dummy = 1 if one child in household
Two Children	Dummy = 1 if two children in household
Three or more Children	Dummy = 1 if three or more children in household
Years since Migration	Years since migration to Germany
Neighboring Country	Dummy = 1 if country of origin is a neighboring country of Germany
Family abroad	Dummy = 1 if family lives abroad
Desired Stay (years)	Years desired to stay in Germany
Distance Capitals (100 km)	Geodesic distance between capitals in 100 km
Age at Entry	Age at entry to Germany
Years of Education	Years of education
School attended in Germany	Dummy = 1 if school attended in Germany
Proficiency Home Language	Dummy = 1 if written proficiency in home language is good or very good
Linguistic Distance	Levenshtein distance normalized divided
Region Dummies	
Western Democracies/Japan	Dummy = 1 if country of origin is a western democracy or Japan
Eastern Europe/Soviet Union	Dummy = 1 if country of origin is eastern Europe or the former Soviet Union
Other	Dummy = 1 if country of origin is any other world region

Notes: – Geodesic distances are calculated following the great circle formula, which uses the geographic coordinates of the capital cities for calculating the distance to the capital of Germany. Distance Capitals reports the calculated distance divided by 100.

Table A3.1: Marginal Effects Oral Language Fluency 1997 & 1999 – Ordered Logit Regression

	1997				1999			
	Bad ME/StdE	Not bad ME/StdE	Good ME/StdE	Very good ME/StdE	Bad ME/StdE	Not bad ME/StdE	Good ME/StdE	Very good ME/StdE
Female	0.011	0.010	−0.015	−0.006	0.014	0.026	−0.028	−0.012
	(0.018)	(0.018)	(0.026)	(0.010)	(0.015)	(0.027)	(0.030)	(0.012)
Married	0.016	0.017	−0.023	−0.009	0.021†	0.045	−0.044†	−0.022
	(0.025)	(0.029)	(0.038)	(0.017)	(0.013)	(0.029)	(0.026)	(0.015)
Children in the HH. (Ref.= 0)								
One Child	0.026	0.024	−0.036	−0.013	0.024	0.041	−0.046	−0.018
	(0.018)	(0.016)	(0.024)	(0.010)	(0.024)	(0.036)	(0.044)	(0.016)
Two Children	0.036	0.031	−0.049	−0.018	0.037	0.059†	−0.070†	−0.026†
	(0.028)	(0.020)	(0.034)	(0.013)	(0.025)	(0.032)	(0.042)	(0.015)
Three or more Children	0.031	0.026	−0.042	−0.015	0.023	0.037	−0.043	−0.017
	(0.022)	(0.016)	(0.027)	(0.011)	(0.026)	(0.036)	(0.045)	(0.017)
Years since Migration	−0.018***	−0.017***	0.025***	0.010***	−0.013**	−0.025**	0.027**	0.012**
	(0.004)	(0.005)	(0.006)	(0.003)	(0.004)	(0.008)	(0.009)	(0.004)
Years since Migration^2/100	0.029**	0.029*	−0.042**	−0.016*	0.022*	0.042*	−0.045*	−0.019*
	(0.011)	(0.012)	(0.015)	(0.007)	(0.009)	(0.018)	(0.018)	(0.009)
Neighboring Country	−0.018	−0.020	0.027	0.011	−0.035	−0.080	0.075	0.041
	(0.046)	(0.052)	(0.069)	(0.029)	(0.027)	(0.066)	(0.056)	(0.038)
Family abroad	0.020	0.018	−0.028	−0.010	0.044***	0.068**	−0.082**	−0.030***
	(0.019)	(0.017)	(0.027)	(0.009)	(0.013)	(0.022)	(0.027)	(0.008)
Desired Stay (years)	−0.004***	−0.004**	0.006***	0.002**	−0.004***	−0.007***	0.008***	0.003***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
Distance Capitals (100 km)	−0.002†	−0.002*	0.003†	0.001*	−0.002*	−0.003**	0.003*	0.002**
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.000)
Age at Entry	−0.001	−0.001	0.002	0.001	−0.003*	−0.005†	0.005*	0.002†
	(0.002)	(0.003)	(0.004)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
Years of Education	−0.028	−0.028	0.041	0.016	−0.037**	−0.070*	0.075**	0.032*
	(0.021)	(0.022)	(0.030)	(0.013)	(0.012)	(0.028)	(0.026)	(0.014)
Years of Education^2/100	0.012	0.012	−0.018	−0.007	0.087	0.165	−0.176	−0.076
	(0.107)	(0.106)	(0.153)	(0.060)	(0.058)	(0.123)	(0.123)	(0.058)
School attended in Germany	−0.095***	−0.169***	0.146***	0.118**	−0.077***	−0.249***	0.126***	0.201**
	(0.019)	(0.040)	(0.029)	(0.038)	(0.015)	(0.039)	(0.031)	(0.061)
Proficiency Home Language	−0.090*	−0.058***	0.112**	0.036**	−0.098*	−0.113***	0.160**	0.052***
	(0.036)	(0.016)	(0.038)	(0.011)	(0.040)	(0.028)	(0.052)	(0.014)
Linguistic Distance	0.006***	0.006***	−0.009***	−0.003**	0.005***	0.009**	−0.009***	−0.004*
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)	(0.002)
Region Dummies	yes	yes	yes	yes	yes	yes	yes	yes

Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Cluster-robust standard errors are reported in parentheses. – Marginal effects are reported at the mean of the covariates vector.

Table A3.2: Marginal Effects Oral Language Fluency 2001 & 2003 – Ordered Logit Regression

	2001				2003			
	Bad ME/StdE	Not bad ME/StdE	Good ME/StdE	Very good ME/StdE	Bad ME/StdE	Not bad ME/StdE	Good ME/StdE	Very good ME/StdE
Female	0.005	0.009	−0.007	−0.007	0.006	0.015	−0.011	−0.010
	(0.014)	(0.025)	(0.020)	(0.018)	(0.010)	(0.027)	(0.019)	(0.018)
Married	0.019	0.037	−0.027	−0.029	0.012	0.033	−0.021	−0.023
	(0.012)	(0.026)	(0.018)	(0.020)	(0.007)	(0.022)	(0.013)	(0.017)
Children in the HH. (Ref.= 0)								
One Child	0.012	0.020	−0.017	−0.015	0.024***	0.060***	−0.046***	−0.038***
	(0.019)	(0.033)	(0.029)	(0.024)	(0.007)	(0.016)	(0.013)	(0.011)
Two Children	0.025	0.041	−0.036	−0.030	0.015	0.038	−0.028	−0.024
	(0.024)	(0.036)	(0.034)	(0.026)	(0.015)	(0.032)	(0.025)	(0.022)
Three or more Children	0.031	0.048	−0.045	−0.035	0.059**	0.116***	−0.108***	−0.068***
	(0.023)	(0.033)	(0.033)	(0.023)	(0.022)	(0.027)	(0.032)	(0.018)
Years since Migration	−0.009***	−0.017**	0.013**	0.012**	−0.007***	−0.020**	0.014**	0.013**
	(0.003)	(0.006)	(0.004)	(0.004)	(0.002)	(0.006)	(0.004)	(0.004)
Years since Migration^2/100	0.018**	0.032*	−0.026*	−0.024*	0.015**	0.042**	−0.029**	−0.028**
	(0.007)	(0.015)	(0.011)	(0.011)	(0.005)	(0.014)	(0.010)	(0.010)
Neighboring Country	−0.042*	−0.091*	0.057*	0.077*	−0.037*	−0.118*	0.059**	0.097†
	(0.018)	(0.043)	(0.025)	(0.037)	(0.015)	(0.054)	(0.022)	(0.050)
Family abroad	0.041*	0.066*	−0.059*	−0.049*	0.006	0.017	−0.012	−0.011
	(0.021)	(0.030)	(0.030)	(0.022)	(0.009)	(0.022)	(0.016)	(0.015)
Desired Stay (years)	−0.001	−0.002	0.001	0.001	0.001†	0.003†	−0.002†	−0.002†
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Distance Capitals (100 km)	−0.001**	−0.003**	0.002*	0.002**	−0.001**	−0.002**	0.002*	0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)
Age at Entry	0.000	0.001	−0.001	−0.001	0.003**	0.009**	−0.006*	−0.006***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.002)
Years of Education	−0.042***	−0.075***	0.061***	0.056***	−0.019	−0.052	0.036†	0.035
	(0.011)	(0.018)	(0.014)	(0.016)	(0.013)	(0.032)	(0.022)	(0.024)
Years of Education^2/100	0.100*	0.179*	−0.145*	−0.134*	−0.001	−0.002	0.001	0.001
	(0.042)	(0.075)	(0.058)	(0.060)	(0.055)	(0.148)	(0.103)	(0.100)
School attended in Germany	−0.049***	−0.119**	0.060***	0.109*	−0.042***	−0.149*	0.049**	0.142*
	(0.010)	(0.042)	(0.012)	(0.048)	(0.010)	(0.058)	(0.017)	(0.069)
Proficiency Home Language	−0.048	−0.069	0.067	0.049	−0.026	−0.061	0.049	0.038†
	(0.043)	(0.046)	(0.057)	(0.031)	(0.020)	(0.044)	(0.043)	(0.022)
Linguistic Distance	0.004***	0.008***	−0.006***	−0.006***	0.003***	0.008***	−0.006***	−0.005***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Region Dummies	yes	yes	yes	yes	yes	yes	yes	yes

*Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Cluster-robust standard errors are reported in parentheses. – Marginal effects are reported at the mean of the covariates vector.*

Supplementary Appendix

Table A4: Estimation Results Oral Language Fluency – Ordered Logit Regression

	1997 Coef/StdE	1999 Coef/StdE	2001 Coef/StdE	2003 Coef/StdE
Female	−0.085	−0.158	−0.056	−0.088
	(0.144)	(0.168)	(0.155)	(0.151)
Married	−0.130	−0.265	−0.229	−0.188
	(0.220)	(0.165)	(0.155)	(0.126)
Children in the HH. (Ref.= 0)				
One Child	−0.203	−0.260	−0.128	−0.347***
	(0.140)	(0.241)	(0.211)	(0.091)
Two Children	−0.276	−0.387†	−0.263	−0.216
	(0.194)	(0.227)	(0.241)	(0.193)
Three or more Children	−0.231	−0.241	−0.320	−0.714***
	(0.155)	(0.249)	(0.225)	(0.193)
Years since Migration	0.142***	0.154**	0.104**	0.112***
	(0.034)	(0.049)	(0.034)	(0.034)
Years since Migration^2/100	−0.236**	−0.256*	−0.200*	−0.239**
	(0.087)	(0.104)	(0.087)	(0.081)
Neighboring Country	0.151	0.467	0.558*	0.694*
	(0.392)	(0.385)	(0.259)	(0.323)
Family abroad	−0.154	−0.452**	−0.433*	−0.097
	(0.147)	(0.137)	(0.205)	(0.127)
Desired Stay (years)	0.035***	0.045***	0.010	−0.016†
	(0.010)	(0.009)	(0.010)	(0.008)
Distance Capitals (100 km)	0.015†	0.020**	0.016**	0.013**
	(0.008)	(0.007)	(0.006)	(0.004)
Age at Entry	0.010	0.029*	−0.005	−0.051**
	(0.020)	(0.014)	(0.012)	(0.017)
Years of Education	0.229	0.428**	0.472***	0.297
	(0.170)	(0.157)	(0.113)	(0.191)
Years of Education^2/100	−0.100	−1.008	−1.126*	0.009
	(0.862)	(0.716)	(0.466)	(0.845)
School attended in Germany	1.105***	1.524***	0.726**	0.912*
	(0.260)	(0.312)	(0.251)	(0.379)
Proficiency Home Language	0.626**	0.879**	0.468	0.355
	(0.210)	(0.280)	(0.357)	(0.255)
Linguistic Distance	−0.050***	−0.054**	−0.048***	−0.046***
	(0.011)	(0.017)	(0.008)	(0.009)
Region Dummies	yes	yes	yes	yes
Threshold 1	−0.596	0.921	−0.827	−3.001†
	(1.302)	(1.660)	(1.480)	(1.805)
Threshold 2	1.395	3.160†	1.207	−0.827
	(1.293)	(1.666)	(1.496)	(1.795)
Threshold 3	3.688**	5.584**	3.216*	1.435
	(1.347)	(1.698)	(1.493)	(1.811)
Pseudo-R^2	0.109	0.143	0.123	0.183
Wald χ^2	4327.18***	4286.64***	1871.85***	847.92***
Log-likelihood	-1504.30	-1206.81	-1658.48	-1340.21
Brant χ^2	308.81***	119.74***	138.69***	50.81***
Observations	1311	1102	1430	1242

*Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Cluster-robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 1 to 4 such that higher values indicate a higher level of oral proficiency.*

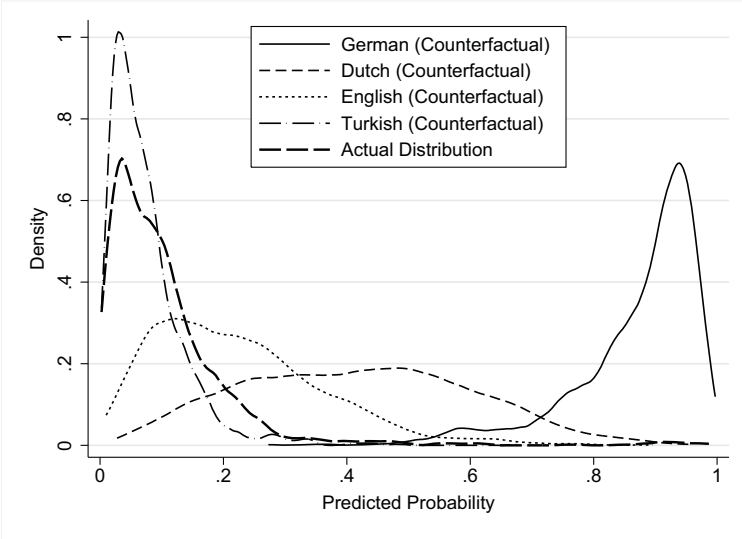

Figure A1: Kernel Density Estimates of Actual and Counterfactual Distributions of Predicted Probabilities to Report Category "Very good" in 1997

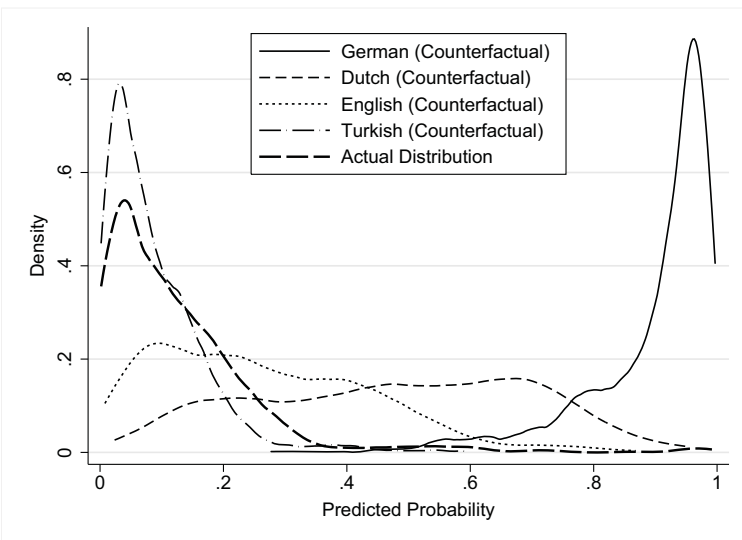

Figure A2: Kernel Density Estimates of Actual and Counterfactual Distributions of Predicted Probabilities to Report Category "Very good" in 1999

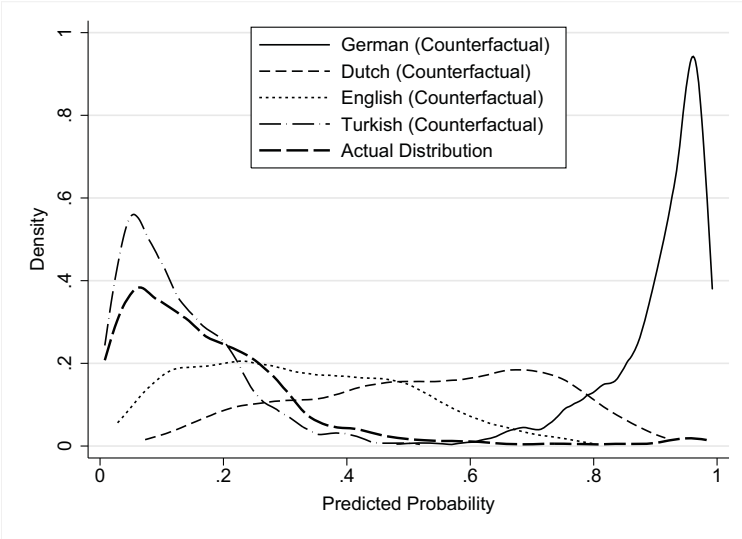

Figure A3: Kernel Density Estimates of Actual and Counterfactual Distributions of Predicted Probabilities to Report Category "Very good" in 2001

Generalized Ordered Logit Model

In this section we explain the Generalized Ordered Logit model – a generalized estimation method for ordinal dependent variables – and describe the estimated results in some detail. Overall, the results of the Generalized Ordered Logit model are similar and comparable to the Ordered Logit results discussed before, therefore we focus in our analysis on the more common Ordered Logit model.

Standard Ordered Logit models impose some strong assumptions, which are likely to be violated in the case of self-reported indicators. Recent developments of ordinal choice models allow for a more appropriate modeling of the cognitive processes underlying the response behavior on self-reported ordinal scales. Standard Ordered Logit models have to assume parallelism of regression lines across the different categories. Hence, it is assumed that the effect of determinants is the same for the step between lower categories (in the beginning of the language acquisition) as it is for latter stages of the language acquisition. Heterogeneous effects would likely violate this Parallel Lines assumption.[1] This assumption, required for identification of the standard Ordered Logit model, states that the slope parameters of the latent variable regression are the same within each category, hence $\beta_1 = \beta_2 = \ldots \beta_J = \beta$. It is unlikely that the Parallel Lines assumption holds in the setting of language acquisition determinants. For example, it is unlikely that units of exposure to the host country language have the same impact in early stages of language acquisition as in later stages, as we should expect decreasing marginal returns.

To take these potentially differing effects into account and to avoid the Parallel Lines assumption, we estimate a Generalized Ordered Logit model as discussed by Williams (2006).[2] It allows for different coefficient vectors across categories, similar to the idea of quantile regressions with differences in coefficients across the distribution of the dependent variable. Thus, apart from imposing less restrictive assumptions, it allows for a more detailed analysis of the effects of determinants of language acquisition.

Instead of estimating one coefficient vector β, $J-1$ coefficient vectors are estimated, which can be interpreted as the effects of covariates on the probability of choosing at least category $j+1$ instead of any lower category. Then, the probability of reporting at least category j is given by

$$P(y_i > j|x_i) = F(\kappa_j + x_i'\beta_j) = \frac{exp(\kappa_j + x_i'\beta_j)}{1 + exp(\kappa_j + x_i'\beta_j)}, \quad j = 1, 2, \ldots, J-1. \tag{1}$$

Hence, the probability of reporting category j is

$$\begin{aligned} P(y_i = 1|x_i) &= 1 - F(\kappa_1 + x_i'\beta_1) && (2)\\ P(y_i = j|x_i) &= F(\kappa_{j-1} + x_i'\beta_{j-1}) - F(\kappa_j + x_i'\beta_j), \quad j = 2, \ldots, J-1 && (3)\\ P(y_i = J|x_i) &= F(\kappa_{J-1} + x_i'\beta_{J-1}) && (4) \end{aligned}$$

with category-specific β_j-vectors. The coefficients are identified in sign and significance

[1] The Parallel Lines assumption is also sometimes called the Proportional Odds or Parallel Regressions assumption.

[2] The Generalized Ordered Logit model was originally developed by Peterson and Harrell (1990). The models presented here were estimated by using the *gologit2*-routine described in Williams (2006).

and are to be interpreted as coefficients of a series of binary logistic regressions. Each one compares categories 1 to $j-1$ with categories j to J. The standard Ordered Logit-case is nested in the Generalized Ordered Logit as the case of $\beta_1 = \beta_2 = \ldots = \beta_J$.

The Brant test (Brant 1990) allows to test if a certain Ordered Logit model violates the Parallel Lines assumption.[3] The test statistics are reported in the Ordered Logit results and all estimated Ordered Logit models are tested positively to violate the Parallel Lines assumption. The Generalized Ordered Logit does not need each coefficient in β to differ across the categories. To rely on a specification as parsimonious as possible, only those coefficients are allowed to differ that are tested to have a categorical-specific influence.

Tables A5, A6.1, and A6.2 show the coefficients and marginal effects obtained by estimating a Generalized Ordered Logit model.[4] The marginal effects are interpreted as effects on the probability of being in exactly one of the four categories. Unlike in the standard Ordered Logit model, due to the category-specific coefficients, marginal effects are allowed to change their signs more than once in the sequence from 1 to J (Greene and Hensher 2010). Again, the inclusion of partially defined splines gives insights in potential non-linear effects of linguistic distance. The according coefficients and marginal effects are summarized in Table A7. The results indicate no systematic differences of effects across quartiles.

[3]The Brant test was generated using the Stata routine *SPost* by Scott Long and Jeremy Freese, see Long and Freese (2006).

[4]The coefficients may be interpreted as coefficients of a series of Logit regressions which analyze the probability of being in a higher than the current category.

Table A5: Estimation Results Oral Language Fluency – Generalized Ordered Logit Regression

	1997			1999			2001			2003		
	Bad Coef/StdE	Not bad Coef/StdE	Good Coef/StdE	Bad Coef/StdE	Not bad Coef/StdE	Good Coef/StdE	Bad Coef/StdE	Not bad Coef/StdE	Good Coef/StdE	Bad Coef/StdE	Not bad Coef/StdE	Good Coef/StdE
Female	−0.091	−0.091	−0.091	−0.158	−0.158	−0.158	−0.065	−0.065	−0.065	−0.093	−0.093	−0.093
	(0.146)	(0.146)	(0.146)	(0.170)	(0.170)	(0.170)	(0.160)	(0.160)	(0.160)	(0.147)	(0.147)	(0.147)
Married	−0.090	−0.090	−0.090	−0.182	−0.182	−0.182	−0.231	−0.231	−0.231	−0.198	−0.198	−0.198
	(0.218)	(0.218)	(0.218)	(0.164)	(0.164)	(0.164)	(0.150)	(0.150)	(0.150)	(0.127)	(0.127)	(0.127)
Children in the HH. (Ref.= 0)												
One Child	−0.209	−0.209	−0.209	0.371	−0.365*	−0.812*	0.154	−0.189	−0.192	−0.343***	−0.343***	−0.343***
	(0.144)	(0.144)	(0.144)	(0.318)	(0.155)	(0.346)	(0.253)	(0.205)	(0.233)	(0.086)	(0.086)	(0.086)
Two Children	0.060	−0.406*	−0.719**	0.372	−0.587**	−0.778***	0.263	−0.392	−0.460†	−0.206	−0.206	−0.206
	(0.144)	(0.205)	(0.234)	(0.322)	(0.199)	(0.207)	(0.179)	(0.264)	(0.241)	(0.193)	(0.193)	(0.193)
Three or more Children	−0.249†	−0.249†	−0.249†	0.309	−0.283	−0.679*	−0.298	−0.298	−0.298	−0.745***	−0.745***	−0.745***
	(0.138)	(0.138)	(0.138)	(0.321)	(0.256)	(0.313)	(0.221)	(0.221)	(0.221)	(0.197)	(0.197)	(0.197)
Years since Migration	0.231***	0.117**	0.063	0.176***	0.129***	0.104**	0.109**	0.109**	0.109**	0.112**	0.112**	0.112**
	(0.028)	(0.038)	(0.056)	(0.047)	(0.037)	(0.034)	(0.033)	(0.033)	(0.033)	(0.035)	(0.035)	(0.035)
Years since Migration^2/100	−0.420***	−0.169†	−0.141	−0.220*	−0.220*	−0.220*	−0.210*	−0.210*	−0.210*	−0.234**	−0.234**	−0.234**
	(0.074)	(0.095)	(0.150)	(0.089)	(0.089)	(0.089)	(0.086)	(0.086)	(0.086)	(0.082)	(0.082)	(0.082)
Neighboring Country	0.023	0.023	0.023	2.200***	0.296	0.277	0.552*	0.552*	0.552*	1.628***	0.691*	0.655†
	(0.338)	(0.338)	(0.338)	(0.456)	(0.271)	(0.544)	(0.254)	(0.254)	(0.254)	(0.369)	(0.303)	(0.350)
Family abroad	−0.111	−0.111	−0.111	−0.469***	−0.469***	−0.469***	−0.421*	−0.421*	−0.421*	−0.086	−0.086	−0.086
	(0.138)	(0.138)	(0.138)	(0.132)	(0.132)	(0.132)	(0.209)	(0.209)	(0.209)	(0.124)	(0.124)	(0.124)
Desired Stay (years)	0.036***	0.036***	0.036***	0.044***	0.044***	0.044***	0.010	0.010	0.010	−0.015†	−0.015†	−0.015†
	(0.010)	(0.010)	(0.010)	(0.007)	(0.007)	(0.007)	(0.010)	(0.010)	(0.010)	(0.008)	(0.008)	(0.008)
Distance Capitals (100 km)	0.013†	0.013†	0.013†	0.017**	0.017**	0.017**	0.018**	0.018**	0.018**	0.014***	0.014***	0.014***
	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)
Age at Entry	0.008	0.008	0.008	0.024†	0.024†	0.024†	−0.005	−0.005	−0.005	−0.076***	−0.044*	−0.040*
	(0.021)	(0.021)	(0.021)	(0.014)	(0.014)	(0.014)	(0.012)	(0.012)	(0.012)	(0.013)	(0.020)	(0.019)
Years of Education	0.332	0.451*	−0.282	0.427**	0.427**	0.427**	0.475***	0.475***	0.475***	0.281	0.281	0.281
	(0.278)	(0.182)	(0.229)	(0.150)	(0.150)	(0.150)	(0.112)	(0.112)	(0.112)	(0.196)	(0.196)	(0.196)
Years of Education^2/100	−0.703	−1.100	2.004*	−1.011	−1.011	−1.011	−1.124*	−1.124*	−1.124*	0.067	0.067	0.067
	(1.309)	(0.940)	(0.896)	(0.672)	(0.672)	(0.672)	(0.467)	(0.467)	(0.467)	(0.852)	(0.852)	(0.852)
School attended in Germany	1.029***	1.029***	1.029***	1.490***	1.490***	1.490***	0.733**	0.733**	0.733**	0.346	1.521**	0.699*
	(0.252)	(0.252)	(0.252)	(0.305)	(0.305)	(0.305)	(0.252)	(0.252)	(0.252)	(0.724)	(0.489)	(0.335)
Proficiency Home Language	1.020***	0.389†	0.499	0.893**	0.893**	0.893**	0.463	0.463	0.463	0.342	0.342	0.342
	(0.201)	(0.221)	(0.326)	(0.301)	(0.301)	(0.301)	(0.364)	(0.364)	(0.364)	(0.249)	(0.249)	(0.249)
Linguistic Distance	0.076**	−0.052***	−0.059***	0.027	−0.050**	−0.054*	0.012	−0.035***	−0.055***	−0.046***	−0.046***	−0.046***
	(0.024)	(0.015)	(0.015)	(0.024)	(0.016)	(0.021)	(0.014)	(0.011)	(0.010)	(0.009)	(0.009)	(0.009)
Constant	−13.548***	−1.909	1.299	−9.940***	−2.845*	−4.154*	−5.279**	−2.724†	−2.897†	3.776*	0.717	−1.603
	(3.541)	(1.474)	(1.949)	(2.286)	(1.416)	(2.029)	(1.710)	(1.630)	(1.585)	(1.835)	(1.909)	(1.791)
Region Dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Pseudo-R^2	0.130			0.166			0.131			0.189		
Wald χ^2	255017.51***			35894.84***			1931.76***			2043.07***		
Log-likelihood	-1470.13			-1175.00			-1643.90			-1331.29		
Observations	1311			1102			1430			1242		

*Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Cluster-robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 1 to 4 such that higher values indicate a higher level of oral proficiency. – "Very good" is the base category.*

Table A6.1: Marginal Effects Oral Language Fluency 1997 & 1999 – Generalized Ordered Logit Regression

	1997				1999			
	Bad ME/StdE	Not bad ME/StdE	Good ME/StdE	Very good ME/StdE	Bad ME/StdE	Not bad ME/StdE	Good ME/StdE	Very good ME/StdE
Female	0.012	0.010	−0.016	−0.006	0.011	0.028	−0.027	−0.012
	(0.020)	(0.017)	(0.026)	(0.010)	(0.013)	(0.030)	(0.031)	(0.012)
Married	0.012	0.010	−0.016	−0.006	0.012	0.033	−0.031	−0.015
	(0.028)	(0.026)	(0.039)	(0.016)	(0.010)	(0.030)	(0.027)	(0.014)
Children in the HH. (Ref.= 0)								
One Child	0.029	0.022	−0.038	−0.013	−0.024	0.115***	−0.039	−0.052**
	(0.020)	(0.016)	(0.026)	(0.010)	(0.018)	(0.029)	(0.040)	(0.018)
Two Children	−0.008	0.106**	−0.058	−0.041**	−0.024	0.169***	−0.096*	−0.049***
	(0.019)	(0.040)	(0.042)	(0.014)	(0.017)	(0.047)	(0.045)	(0.011)
Three or more Children	0.036†	0.025†	−0.045†	−0.015†	−0.020	0.090†	−0.029	−0.042*
	(0.021)	(0.013)	(0.024)	(0.009)	(0.017)	(0.053)	(0.065)	(0.016)
Years since Migration	−0.031***	0.002	0.025**	0.004	−0.013***	−0.020**	0.024***	0.008**
	(0.004)	(0.009)	(0.008)	(0.004)	(0.003)	(0.007)	(0.007)	(0.003)
Years since Migration^2/100	0.057***	−0.015	−0.032	−0.010	0.016**	0.039*	−0.038*	−0.017*
	(0.010)	(0.021)	(0.020)	(0.011)	(0.005)	(0.017)	(0.016)	(0.007)
Neighboring Country	−0.003	−0.003	0.004	0.002	−0.087***	0.014	0.050	0.023
	(0.045)	(0.039)	(0.061)	(0.023)	(0.019)	(0.057)	(0.040)	(0.049)
Family abroad	0.015	0.012	−0.020	−0.007	0.038***	0.078***	−0.084***	−0.032***
	(0.019)	(0.015)	(0.025)	(0.009)	(0.011)	(0.023)	(0.025)	(0.009)
Desired Stay (years)	−0.005***	−0.004**	0.006***	0.002**	−0.003***	−0.008***	0.008***	0.003***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Distance Capitals (100 km)	−0.002†	−0.001*	0.002†	0.001*	−0.001*	−0.003**	0.003*	0.001**
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)
Age at Entry	−0.001	−0.001	0.001	0.001	−0.002†	−0.004	0.004†	0.002
	(0.003)	(0.002)	(0.004)	(0.001)	(0.001)	(0.003)	(0.002)	(0.001)
Years of Education	−0.045	−0.067†	0.130**	−0.019	−0.030**	−0.076**	0.074**	0.033*
	(0.037)	(0.034)	(0.050)	(0.016)	(0.010)	(0.029)	(0.024)	(0.014)
Years of Education^2/100	0.095	0.177	−0.407†	0.135*	0.072	0.181	−0.175	−0.077
	(0.174)	(0.152)	(0.244)	(0.063)	(0.045)	(0.124)	(0.113)	(0.057)
School attended in Germany	−0.099***	−0.149***	0.143***	0.105**	−0.062***	−0.258***	0.123***	0.197***
	(0.020)	(0.038)	(0.026)	(0.035)	(0.010)	(0.045)	(0.025)	(0.055)
Proficiency Home Language	−0.171***	0.077*	0.065	0.029	−0.084*	−0.131***	0.162**	0.053***
	(0.045)	(0.038)	(0.040)	(0.019)	(0.034)	(0.036)	(0.057)	(0.012)
Linguistic Distance	−0.010**	0.023***	−0.009**	−0.004**	−0.002	0.014***	−0.008**	−0.004*
	(0.004)	(0.005)	(0.003)	(0.001)	(0.002)	(0.003)	(0.003)	(0.002)
Region Dummies	yes	yes	yes	yes	yes	yes	yes	yes

Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Cluster-robust standard errors are reported in parentheses. – Marginal effects are reported at the mean of the covariates vector.

Table A6.2: Marginal Effects Oral Language Fluency 2001 & 2003 – Generalized Ordered Logit Regression

	2001				2003			
	Bad ME/StdE	Not bad ME/StdE	Good ME/StdE	Very good ME/StdE	Bad ME/StdE	Not bad ME/StdE	Good ME/StdE	Very good ME/StdE
Female	0.006	0.010	−0.009	−0.007	0.005	0.017	−0.011	−0.011
	(0.016)	(0.024)	(0.022)	(0.018)	(0.009)	(0.027)	(0.018)	(0.017)
Married	0.022	0.035	−0.029	−0.028	0.011†	0.036	−0.022†	−0.025
	(0.014)	(0.023)	(0.018)	(0.019)	(0.007)	(0.023)	(0.013)	(0.017)
Children in the HH. (Ref.= 0)								
One Child	−0.015	0.062*	−0.026	−0.021	0.022***	0.062***	−0.045***	−0.038***
	(0.024)	(0.032)	(0.037)	(0.024)	(0.006)	(0.015)	(0.012)	(0.010)
Two Children	−0.025	0.122*	−0.050	−0.047*	0.013	0.037	−0.027	−0.023
	(0.016)	(0.060)	(0.051)	(0.023)	(0.013)	(0.034)	(0.025)	(0.022)
Three or more Children	0.033	0.042	−0.043	−0.031	0.057**	0.127***	−0.112***	−0.071***
	(0.026)	(0.029)	(0.033)	(0.022)	(0.020)	(0.029)	(0.033)	(0.018)
Years since Migration	−0.011***	−0.016**	0.015**	0.013**	−0.006***	−0.020**	0.013**	0.013**
	(0.003)	(0.005)	(0.005)	(0.004)	(0.002)	(0.006)	(0.004)	(0.004)
Years since Migration^2/100	0.021*	0.031*	−0.028*	−0.024*	0.014**	0.042**	−0.028**	−0.028**
	(0.008)	(0.013)	(0.011)	(0.011)	(0.004)	(0.015)	(0.010)	(0.010)
Neighboring Country	−0.048*	−0.086*	0.061*	0.073*	−0.063***	−0.091†	0.063*	0.091†
	(0.020)	(0.040)	(0.026)	(0.035)	(0.012)	(0.055)	(0.031)	(0.054)
Family abroad	0.045†	0.060*	−0.059†	−0.046*	0.005	0.015	−0.010	−0.010
	(0.024)	(0.029)	(0.031)	(0.022)	(0.007)	(0.022)	(0.015)	(0.015)
Desired Stay (years)	−0.001	−0.001	0.001	0.001	0.001†	0.003†	−0.002†	−0.002†
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Distance Capitals (100 km)	−0.002**	−0.003**	0.002*	0.002***	−0.001**	−0.002**	0.002*	0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)
Age at Entry	0.000	0.001	−0.001	−0.001	0.004***	0.006	−0.006	−0.005*
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.004)	(0.004)	(0.002)
Years of Education	−0.048***	−0.071***	0.064***	0.055***	−0.016	−0.051	0.034	0.033
	(0.010)	(0.018)	(0.015)	(0.015)	(0.011)	(0.035)	(0.022)	(0.024)
Years of Education^2/100	0.113**	0.167*	−0.151*	−0.129*	−0.004	−0.012	0.008	0.008
	(0.044)	(0.073)	(0.061)	(0.058)	(0.050)	(0.154)	(0.103)	(0.101)
School attended in Germany	−0.057***	−0.115**	0.065***	0.107*	−0.018	−0.262***	0.176***	0.103†
	(0.014)	(0.039)	(0.012)	(0.048)	(0.031)	(0.039)	(0.053)	(0.054)
Proficiency Home Language	−0.053	−0.062	0.068	0.047	−0.022	−0.061	0.046	0.037†
	(0.050)	(0.040)	(0.059)	(0.031)	(0.018)	(0.044)	(0.041)	(0.022)
Linguistic Distance	−0.001	0.010***	−0.002	−0.006***	0.003***	0.008***	−0.006***	−0.006***
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Region Dummies	yes	yes	yes	yes	yes	yes	yes	yes

Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Cluster-robust standard errors are reported in parentheses. – Marginal effects are reported at the mean of the covariates vector.

Table A7: Estimation Results Splines Oral Language Fluency – Generalized Ordered Logit Regression

	1997 Coef/StdE	1999 Coef/StdE	2001 Coef/StdE	2003 Coef/StdE
Generalized Ordered Logit Regression				
Bad				
Spline 1st Quartile	−0.039	−0.065**	0.024	0.003
	(0.131)	(0.023)	(0.019)	(0.030)
Spline 2nd Quartile	−0.025	−0.054*	0.032†	0.005
	(0.127)	(0.023)	(0.018)	(0.026)
Spline 3rd Quartile	−0.029	−0.053**	0.020	−0.003
	(0.119)	(0.020)	(0.017)	(0.024)
Spline 4th Quartile	−0.022	−0.054**	0.022	0.003
	(0.117)	(0.021)	(0.018)	(0.024)
Not bad				
Spline 1st Quartile	−0.167***	−0.056*	−0.047**	−0.056***
	(0.041)	(0.023)	(0.017)	(0.016)
Spline 2nd Quartile	−0.158***	−0.054*	−0.040*	−0.052**
	(0.040)	(0.023)	(0.017)	(0.017)
Spline 3rd Quartile	−0.151***	−0.053**	−0.045**	−0.050***
	(0.037)	(0.020)	(0.015)	(0.014)
Spline 4th Quartile	−0.150***	−0.054**	−0.045**	−0.051***
	(0.037)	(0.021)	(0.017)	(0.015)
Good				
Spline 1st Quartile	−0.070***	−0.048*	−0.057***	−0.047***
	(0.019)	(0.023)	(0.011)	(0.011)
Spline 2nd Quartile	−0.075***	−0.054*	−0.059***	−0.050***
	(0.019)	(0.023)	(0.011)	(0.012)
Spline 3rd Quartile	−0.067***	−0.053**	−0.055***	−0.048***
	(0.016)	(0.020)	(0.009)	(0.010)
Spline 4th Quartile	−0.068**	−0.054**	−0.059***	−0.048***
	(0.021)	(0.021)	(0.012)	(0.010)
Marginal Effects Generalized Ordered Logit Regression	ME/StdE	ME/StdE	ME/StdE	ME/StdE
Bad				
Spline 1st Quartile	0.005	0.005**	−0.002	−0.000
	(0.016)	(0.002)	(0.002)	(0.002)
Spline 2nd Quartile	0.003	0.004*	−0.003†	−0.000
	(0.015)	(0.002)	(0.002)	(0.002)
Spline 3rd Quartile	0.004	0.004**	−0.002	0.000
	(0.014)	(0.001)	(0.002)	(0.002)
Spline 4th Quartile	0.003	0.004*	−0.002	−0.000
	(0.014)	(0.001)	(0.002)	(0.002)
Not bad				
Spline 1st Quartile	0.037*	0.009*	0.014***	0.014***
	(0.015)	(0.004)	(0.003)	(0.003)
Spline 2nd Quartile	0.036*	0.010*	0.013***	0.013***
	(0.015)	(0.004)	(0.003)	(0.003)
Spline 3rd Quartile	0.034*	0.010**	0.013***	0.012***
	(0.014)	(0.004)	(0.003)	(0.002)
Spline 4th Quartile	0.035*	0.010*	0.013***	0.012***
	(0.014)	(0.004)	(0.003)	(0.003)
Good				
Spline 1st Quartile	−0.037***	−0.011*	−0.005	−0.008*
	(0.010)	(0.004)	(0.004)	(0.004)
Spline 2nd Quartile	−0.034***	−0.010*	−0.003	−0.007†
	(0.009)	(0.004)	(0.004)	(0.003)
Spline 3rd Quartile	−0.033***	−0.010**	−0.005	−0.006*
	(0.009)	(0.004)	(0.004)	(0.003)
Spline 4th Quartile	−0.033***	−0.010*	−0.004	−0.007*
	(0.009)	(0.004)	(0.004)	(0.003)
Very good				
Spline 1st Quartile	−0.005**	−0.003*	−0.007***	−0.005***
	(0.001)	(0.002)	(0.001)	(0.001)
Spline 2nd Quartile	−0.005**	−0.004*	−0.007***	−0.006***
	(0.002)	(0.002)	(0.002)	(0.002)
Spline 3rd Quartile	−0.004***	−0.004*	−0.006***	−0.006***
	(0.001)	(0.001)	(0.001)	(0.001)
Spline 4th Quartile	−0.004**	−0.004*	−0.007***	−0.006***
	(0.002)	(0.001)	(0.002)	(0.001)

*Notes: – Significant at: *** 0.1% level; ** 1% level; * 5% level; † 10% level. – Cluster-robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 1 to 4 such that higher values indicate a higher level of oral proficiency. – The four splines are constructed as the product of a quartile dummy and the absolute values of Linguistic Distance. – Marginal effects are reported at the mean of the covariates vector.*