DISCUSSION PAPER SERIES

IZA DP No. 8090

Linguistic Barriers in the Destination Language Acquisition of Immigrants

Ingo E. Isphording
Sebastian Otten

April 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Linguistic Barriers in the Destination Language Acquisition of Immigrants

Ingo E. Isphording
IZA

Sebastian Otten
Ruhr University Bochum
and RWI

Discussion Paper No. 8090
April 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 8090
April 2014

ABSTRACT

Linguistic Barriers in the Destination Language Acquisition of Immigrants*

There are various degrees of similarity between the languages of different immigrants and the language of their destination country. This linguistic distance is an obstacle to the acquisition of a language, which leads to large differences in the attainments of the language skills necessary for economic and social integration in the destination country. This study aims at quantifying the influence of linguistic distance on the language acquisition of immigrants in the US and in Germany. Drawing from comparative linguistics, we derive a measure of linguistic distance based on the automatic comparison of pronunciations. We compare this measure with three other linguistic and non-linguistic approaches in explaining self-reported measures of language skills. We show that there is a strong initial disadvantage from the linguistic origin for language acquisition, while the effect on the steepness of assimilation patterns is ambiguous in Germany and the US.

JEL Classification: F22, J15, J24, J40

Keywords: linguistic distance, language skills, immigrants, human capital

Corresponding author:

Ingo E. Isphording
IZA
P.O. Box 7240
53072 Bonn
Germany
E-mail: isphording@iza.org

* An earlier version of this paper circulated as “Linguistic Distance and the Language Fluency of Immigrants”. We are grateful to Thomas K. Bauer, John P. Haisken-DeNew, Julia Bredtmann, Carsten Crede, Michael Kind, Jan Kleibrink, and Maren Michaelsen as well as participants at the annual congress of the European Economic Association 2011, the annual conference of the European Association of Labour Economists 2011, and the International German Socio-Economic Panel User Conference 2012 for helpful comments and suggestions. We also thank the editor William Neilson and an anonymous referee for their comments and suggestions, which greatly improved the manuscript. Financial support from the German-Israeli Foundation for Scientific Research and Development (GIF) is gratefully acknowledged. All remaining errors are our own.

1 Introduction

Already the biblical description of the fall of the Tower of Babel acknowledged the fact that differences and diversity between languages impose major obstacles for human communication. A range of empirical studies have shown that linguistic barriers constitute distinctive hurdles for international factor flows, e.g., in international trade (Lohmann 2011, Isphording and Otten 2013) or international migration flows (Belot and Ederveen 2012, Adsera and Pytlikova 2012). On the individual level, language skills have been analyzed as being a crucial determinant for the economic and social integration of immigrants in their destination country, starting with early work by Carliner (1981) and McManus et al. (1983) and more recently estimating strong wage effects for destination language proficiency (Chiswick and Miller 1995, Dustmann and van Soest 2002, Bleakley and Chin 2004). These wage effects arise from the role of language as a medium of everyday and working life, constituting an important productive trait of individuals (Crystal 2010). Furthermore, low proficiency may also act as a signal of foreignness, facilitating discrimination and differentiation (Esser 2006). Apart from wages, language proficiency is related to further economic outcomes, such as employment status (Dustmann and Fabbri 2003), occupational choice (Chiswick and Miller 2007), and locational choice (Bauer et al. 2005).

Language skills are not randomly distributed: rather, they display the outcome of a systematic human capital investment decision influenced by costs and expected benefits (Chiswick and Miller 1995). This study is concerned with the analysis of a specific cost factor of language acquisition related to the origin of an immigrant. The degree of difficulty in learning a new language depends on the degree of dissimilarity of the mother tongue of immigrants to the language of the destination country. This linguistic distance, denoting differences between vocabularies, phonetic inventories, grammars, scripts, etc., is expected to crucially affect the efficiency of language learning and to raise the costs of human capital investment. In spite of the strong impact of the skills of immigrants in the destination language on their integration process, the literature on the determinants of the acquisition of the language of their destination remains surprisingly scarce. The systematic analysis of the determinants of language proficiency started with the early work by Evans (1986) comparing immigrants in Germany, the US, and Australia. More recently, Chiswick and Miller (1999, 2002, 2005) provide a comprehensive analysis of the language acquisition of immigrants in the US. For Germany, Dustmann (1999) analyzes the language proficiency as a jointly determined outcome along with migration duration. Dustmann and van Soest (2001) takes into account potential misclassification in self-reported language proficiency and Danzer and Yaman (2010) analyze German language proficiency as a function of enclave density. Still, the influence of characteristics related to the country of origin, such as the linguistic distance faced by immigrants, remains an under-researched area (Esser

2006).

The major challenge in analyzing the effect of linguistic barriers on the language acquisition of immigrants is to operationalize the linguistic distance for use in large scale micro data studies. We propose drawing from comparative linguistics and using an innovative linguistically based operationalization of linguistic distance, the so-called normalized and divided Levenshtein distance calculated by the *Automated Similarity Judgment Program* (ASJP). The ASJP approach offers advantages in terms of transparent computation and general applicability. We compare its benefits to those of three other approaches previously used in further applications in the economic literature to measure linguistic distance: (i) The WALS measure, which uses differences in language characteristics, (ii) the TREE measure, which is based on a priori knowledge on language families, and (iii) a measure based on average test scores of native US foreign language students (SCORE). Combining this information on language differences with US and German micro data, we provide a comprehensive analysis of the influence of the linguistic origin on the acquisition of the destination language proficiency. The US and Germany are excellent examples for analyzing the language acquisition of immigrants. Both countries have a long history as significant immigration hubs, receiving immigrants from a large variety of source countries.

The present study contributes to the literature of the determinants of language proficiency in several ways. First, we provide a comprehensive overview of the different methods of deriving a measure of language differences applicable to the analysis of the role of languages in economic behavior. Second, we introduce the ASJP approach as an easily and transparently computed measure of linguistic dissimilarity between languages. Moreover, this new approach to measuring linguistic distance is applicable to any of the world's languages, and offers specific advantages compared to other linguistic and non-linguistic approaches used in the previous literature. We apply the derived methods to explain the language acquisition of immigrants in the US using the American Community Survey (ACS) as a very recent data source. Finally, we contribute to the literature by taking advantage of the general applicability of the linguistically based methods and extend our analysis beyond the case of Anglophone countries using data from the German Socio-Economic Panel (SOEP).

Our results suggest that the linguistic barriers raised by language differences play a crucial role in the determination of the destination-country language proficiency of immigrants. Regardless of the method employed, we estimate large initial disadvantages by linguistic distance for immigrants both in the US and in Germany. In Germany, these initial differences in language skills decrease with a moderate convergence over time. Contrarily, in the US, the initial disadvantages increase over time. The gap between immigrants from different linguistic groups becomes larger with the time of residence. A potential explanation for the opposing results might be found in the higher prevalence of

linguistic enclaves in the US, leading to different long-term incentives for investment in language skill in the US and Germany.

The estimated differences by linguistic origin witness to the great influence of linguistic background on the economic integration of immigrants. This role should be accounted for in the design of integration policy measures. The results allow the identification of potential target groups for policy intervention. Typical measures aiming at increasing the average language proficiency of immigrants have relied on lump sum payments or fixed classroom hours for language classes. Public spending for language acquisition support might be more effective when a priori information about the expected difficulties is taken into account to specifically address target groups prone to insufficient levels.

The remainder of the paper is organized as follows. In Section 2 we provide an overview of the measurement of linguistic differences employed in our analysis. Section 3 describes the data, Section 4 outlines our empirical model. The findings obtained from our empirical analysis are presented and discussed in Section 5, and Section 6 concludes.

2 Measuring Linguistic Distance

The massive increase in migration flows during the last decades have shaped previously homogeneous populations into linguistically and culturally diverse melting pots. Immigrants face very different costs of language acquisition, associated with their linguistic origin. The influence of the first language (L1) on the acquired language (L2) is a common research topic in linguistics: A larger linguistic distance between L1 and L2 is believed to hamper any potential language transfer (the application of knowledge in the mother tongue to second languages) and to make it more difficult to differentiate between different sounds and words. Linguistic studies typically analyze the effect of linguistic distance employing small samples or case studies. An overview and notable exception can be found in the study by Van der Slik (2010).

The effect of linguistic distance on language acquisition can also be interpreted within an economic framework. The acquisition of language skills is an investment in a type of human capital with a high degree of specificity. Analogously to the restricted portability of source-country education (Friedberg 2000), language skills are restricted in their portability across borders. The value of language skills outside a certain country can be very low, and immigrants have to invest in destination language skills as a prerequisite for successful integration. The imperfect portability of source-country language proficiency is a cost factor in the acquisition of the destination language. The linguistic distance indicates this portability of source-country language skills to the destination country. The larger the linguistic distance, the lower is the applicability of source-country language knowledge in the acquisition of the destination language. This leads, ceteris paribus, to greater

difficulties and higher costs in the language acquisition (Chiswick and Miller 1999).

The difficulty in analyzing the relation between linguistic distance and language skills in a large scale micro data setting lies in the operationalization of the concept of linguistic distance. While specialized linguists have dedicated their whole career to studying the difference between two specific languages, our research question requires a simple standardized and continuous measure of differences between a large set of origin and destination languages. We propose to use a measure of linguistic distance relying on the phonetic dissimilarity between languages based on linguistic research by the so-called Automated Similarity Judgment Program (ASJP). This measure, the normalized and divided Levenshtein distance, offers a continuous measure of linguistic differences and is easily computed for any pair of the world's languages. We compare this measure with two linguistic approaches and a test-score based method that have been applied in different settings in the economic literature.

The test score measure

The only work we are aware of that addresses the effect of linguistic distance on language proficiency using large micro data sets are the studies by Chiswick and Miller (1999, 2001, 2005). The construction of that measure is based on average exam scores of US American English native speakers in standardized tertiary education language courses after a fixed amount of class hours. Assuming symmetry in the difficulty of learning languages, the authors state that the difficulty of English native speakers' learning a foreign language resembles the difficulty of speakers of this foreign language in learning English. This symmetry assumption allows using these test scores as a summary statistic for the dissimilarity between languages. The necessary classroom assessments of test scores are provided by Hart-Gonzalez and Lindemann (1993), Chiswick and Miller (1999) report the respective averages by foreign language. For example, US students learning Norwegian reached an average score of 3.0 (the highest potential score). Using this score the linguistic distance for a Norwegian native speaker learning English is defined as the inverse: $LD_{SCORE} = 1/Score = 0.33$. Since Icelandic and Faroese are assumed to be close languages to Norwegian, the same distance is assigned to these languages. Unfortunately, this test-score based measure of linguistic distance is restricted to differences of a finite set of languages from English. An excerpt of the scores and resulting distances provided by Chiswick and Miller (1999) can be found in Table 1.

The approach, especially the underlying symmetry assumption, has been widely disputed in the linguistic literature (see, e.g., Van der Slik 2010). A further disadvantage of such a test-score based approach is a potential bias by incentives and motivations to learn a foreign language that cannot be separated from the effect of differences between languages. These incentives can include different economic prospects from learning a language

(differences in the applicability in the labor market), or the prestige from learning new, difficult or "hip" languages. These potential biases might lead to rather counter-intuitive assessments, such as the similarly low distance between Swahili and English or Dutch and English.

Linguistic approaches: The TREE and the WALS measure

Comparative linguistics, a branch of linguistics that is concerned with the analysis of family ties and similarities within language families, provides alternatives to the test-score based method. To retrace the historical development of languages, language trees have been developed to arrange languages into different families. These language trees depict the "genealogical" relations between languages and allow of tracing back the development of languages to likely extinct common ancestors. Most prominently, the *Ethnologue* Project (see, Lewis 2009) aims at evaluating the family relations between all known languages in the world. Using this information about the family relations between languages, it is possible to derive a measure of the linguistic distance between languages by counting the number of branches between the languages. While offering a convenient and continuous measure of linguistic distance, although with a comparably low number of increments, the resulting measure is build on strong and arbitrary assumptions of cardinality along the language tree and makes it difficult to include isolated languages (such as Korean or Basque) in the analysis. Two recent studies apply this approach to measure the effect of linguistic distance in a macroeconomic framework. Desmet et al. (2009) use a measure based on steps through the branches of a language tree to assess the effect of linguistic diversity on redistribution. Adsera and Pytlikova (2012) use a language tree approach to analyze the role of linguistic barriers in migration flows. Using the *Ethnologue* information, they define a language proximity index that takes on the value of 0 for languages without any family language relation, and 1 for being the same language. Between these extreme values, the language proximity indicator takes on values of 0.1, 0.25, 0.45 and 0.7 for sharing up to four levels of family relations. As both approaches by Desmet et al. (2009) and Adsera and Pytlikova (2012) rely on more or less arbitrarily chosen assumptions on cardinality and functional form, we employ the one by Adsera and Pytlikova (2012) due to its straightforward computation. Figure 1 illustrates a subset of a language tree to outline its computation. Since Portuguese and Spanish share the first four common branches: Indo-European, Italic, Romance, and Italo-Western, this is coded as a linguistic proximity of 0.7. English and German only share three branches: Indo-European, Germanic, and West. Therefore, the approach leads to a proximity indicator for this language pair of 0.45. The linguistic distance is again defined as the inverse of this proximity indicator: $LD_{TREE} = 1/Proximity$.

Apart from *Ethnologue*, a second information source about languages is the *World Atlas*

of Language Structure (WALS). The WALS offers an online database of the structural properties of languages, such as the phonological, grammatical and lexical features of more than 2,500 different languages. The 144 different characteristics include, for example, different cases, word order or syntax. Specific grammatical features from WALS have been used recently to analyze the relation between language structure and economic behavior, such as the encoding of present and future savings behavior (Chen 2013) or gender systems and female political participation (Santacreu-Vasut et al. 2013). Panel C of Table 2 lists some examples of English and German WALS features. Both languages share a low consonant-vowel ratio, but while English possesses a vowel nasalization, German does not. Using the full information on all features offered by WALS, Lohmann (2011) derives an index of linguistic dissimilarity between 0 and 1 by counting and averaging shared characteristics between languages to explain international trade flows. While conveniently summarizing linguistic differences in a number of different dimensions, the approach relies on the more or less arbitrary assumption of the equal importance of each linguistic feature. More importantly, the WALS database suffers from highly unbalanced data, since not every WALS characteristic is assessed for every language. This leads to the fact that the distance between some languages relies on a very small subsets of the commonly assessed WALS features, which potentially generates a large measurement error in the variable. To reduce this measurement error (with the trade-off of losing observations), we only include distances between languages that are based on at least 20 out of the 144 available characteristics.

The Automatic Similarity Judgement Program

The main focus of our analysis is the application of a new and innovative way of measuring linguistic distance, the so-called *Automatic Similarity Judgement Program* developed by the German Max Planck Institute for Evolutionary Anthropology.[1] This project aims at developing an automatic procedure to evaluate the phonetic similarity between all of the world's languages and offers a convenient way of deriving a continuous measure of linguistic differences that is purely descriptive in nature. As such, it might be used to derive language trees (which is its original purpose) but does not rely on any prior expert opinion on language families, as does the TREE approach. The basic idea behind the ASJP is the automatic comparison of the pronunciation of words across languages. A more similar pronunciation proxies the number of *cognates*, word pairs between languages with common ancestors, which then again indicates a closer relation between the languages. Petroni and Serva (2010) and Brown et al. (2008) demonstrate that the language relations predicted by the ASJP coincide closely with expert opinions on language relations taking into account any available language characteristics, despite the fact that it is only based

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

on simple comparisons of word lists.

To implement this "lexicostatistical" approach, the ASJP uses a core set of vocabulary for each language, describing common things and environments, called the *Swadesh word list* (Swadesh 1952). The Swadesh list consists of words which are deductively chosen according to their availability in as many languages as possible, so that synonyms for these words exist in almost any potential language. Panel A of Table 2 lists the words used, which comprise parts of the human body, environmental descriptions, and basic words of human communication such as classifiers or personal pronouns. To focus on the pronunciation instead of the written word, these words are transcribed into a phonetic script, the *ASJP code*. The ASJP code uses all available characters on a standard QWERTY keyboard to represent sounds of human communication. For example, the English word *mountain* is transcribed in the ASJP code as *maunt3n*, while its German counterpart *Berg* is transcribed as *bErk*. The English word *you* is transcribed as *yu*, the respective German *du* is the same in the ASJP code, *du*.

In the following, we go through the algorithm that leads to the continuous measure of language dissimilarity. In the first step, all word pairs from the transcribed 40-word list are compared with regard to their similarity in pronunciation. For each word pair, the minimum distance between the transcribed phonetic strings is measured as the Levenshtein distance, a measure of distance between string variables. The Levenshtein distance counts how many additions and/or subtractions are necessary to transform the string of the pronunciation of a word in language A into the string of the pronunciation of the respective word in Language B. For example, to transform the English *yu* into the very similar German *du*, only the first sound has to be changed. Whereas for the very dissimilar words *mountain* transcribed as *maunt3n* and *Berg (bErk)*, all of the seven sounds of *maunt3n* have to be changed or removed. This first step results in a word-by-word absolute distance $D(\alpha_i, \beta_i)$ between item i of two languages α and β.[2] Examples for the transcription and determination of the word-by-word minimum distance are listed in Panel B of Table 2.

Taking a simple average across all M word pairs α_i and β_i, $i = 1, .., N$ results in the normalized Levenshtein distance (LDN):

$$LDN(\alpha, \beta) = \frac{1}{M} \sum_i D(\alpha_i, \beta_i). \tag{1}$$

This simple normalized Levenshtein distance might indicate a closeness between languages if languages shared the same set of commonly used sounds in communication. These potential similarities in phonetic inventories (the sum of speech sounds used in a particular language) between two compared languages do not conclusively hint at a genealogical relation between the languages, but might rather produce a similarity by chance. To filter

[2]We draw in our notations from Petroni and Serva (2010).

out similarities by common phonetic inventories, a global average distance $\Gamma(\alpha, \beta)$ between all non-related items of the languages α and β is defined by comparing each word of the first language with all non-related words from the second language. This distance takes into account the overall similarity in phonetic inventories irrespective of the meanings of the words:

$$\Gamma(\alpha, \beta) = \frac{1}{M(M-1)} \sum_{i \neq j} D(\alpha_i, \beta_j). \tag{2}$$

The final normalized and divided linguistic distance is then defined as the quotient between the normalized linguistic distance and the global distance between α and β:

$$LDND(\alpha, \beta) = \frac{LDN(\alpha, \beta)}{\Gamma(\alpha, \beta)}. \tag{3}$$

The resulting continuous measure can be broadly interpreted as a percentage measure of dissimilarity between languages, with lower numbers indicating a closer relation. In a few cases, the resulting numbers are bigger than 100%, indicating a dissimilarity that exceeds a potentially incidental similarity between languages that would be expected due to similar phonetic inventories. The ASJP algorithm allows including or excluding loan words from different languages, e.g., the predominance of former Latin words in many of the European languages. While it makes sense to exclude these loan words in the analysis of the long-term development of languages, we include these loan words in our analysis, as they lead to certain similarities of languages that might ease the later language transfer in the acquisition process.[3]

The normalized and divided Levenshtein distance offers some advantages compared to previous measures of linguistic distance, which lead to more precise and efficient results in economic and social science applications. First, the measure is easily and transparently computed and is purely descriptive in nature, as such it does not rely on any a priori expert information on language relations. Second, due to this purely descriptive nature, it is not likely to be biased by economic incentives. Third, it offers a high variation as it is not restricted to certain parameter values. Lastly, it is comprehensive (all relevant languages are covered by the ASJP database) and can be used for any destination-country language included in the ASJP database. Therefore, it not only allows the analysis of important immigration countries such as the US, the United Kingdom, Canada, Germany, and France, but also permits the analysis of immigrants from rather "exotic" countries with typically few observations that are otherwise excluded from datasets. The comprehensiveness of the database further allows analyses concerning South-South migration, including rather seldom analyzed languages. This is a major advantage compared to the test-score based

[3]The necessary software to compute the distance matrix is available at http://www.eva.mpg.de. The complete distance matrix used in our analysis is available upon request.

approach of Chiswick and Miller (1999), which is restricted to distances from English.

Identification issues

We rely in our estimations on four measures of linguistic differences between the destination- and source-country language that differ in their ranges of availability and in the restrictiveness of their necessary assumptions. The test-score based measure (SCORE) is compelling with its encompassing nature, but relies on a strong symmetry assumption and is potentially biased by differences in incentives to learn a specific language. Most importantly, it is restricted to distances from English. In our US sample, it is available for up to 56 different source-country languages, when expert opinions on close language relations are taken into account to maximize the scope of the measure (Chiswick and Miller 1999). Compared to this test-score measure, linguistically based approaches offer a more general framework to asses the distance between languages. The tree approach (TREE) derives a measure of distance by counting the number of shared branches in language trees, relying on prior knowledge of language family relations. It is based on strong assumptions on functional form and cardinality. Due to the completeness of the language family classifications by *Ethnologue*, this approach is available for the distances of 85 languages toward English and 83 languages toward German. Using external databases on language characteristics and pronunciation, the WALS and the ASJP approach offer ways to assess the differences between languages in a more descriptive manner. Neither approach relies on a priori expert knowledge of language families. However, the WALS approach has to make assumptions on cardinality. The data restrictions of the WALS database reduce the number of available languages to 67 languages in the case of the US sample and 68 languages in the case of the German sample. The ASJP database does not suffer from these restrictions, offering sufficient information for almost any language in the samples. We can rely on information for 85 languages in the US sample, and 83 languages in the German sample, providing the same applicability as that of the TREE approach. Because of its general applicability and descriptive nature, we argue that the ASJP approach, based on simple comparisons of pronunciations of word lists, offers the most appropriate way to measure linguistic distance and is superior for the application at hand. Although the ASJP approach includes much broader information on source-country languages, for the sake of comparability we restrict our estimations to immigrants from those source countries for which we have common information using all four approaches.

Table 3 summarizes the three closest and the three most distant languages from English and German according to the four different measures of linguistic distance. Consistent across the different measures, the closest languages consist of members of the Germanic language family. Some advantages and disadvantages of the measures employed are already apparent in this table. Due to the low number of increments within the measurement

scale, both the TREE and the SCORE approach show only a small variation between the closest and furthest languages. Therefore, a range of languages shares the closest and the most distant position, respectively. In contrast, the ASJP and the WALS measure offer a high variation in the data. The comprehensiveness of the ASJP database allows including more remote languages, such as the Caribbean Creole languages, in the analysis, which are not covered by the other approaches.

Regardless of the approach employed, the identification of an effect of linguistic barriers on the language acquisition might be biased by a correlation between linguistic distance and unobservable further cultural differences in habits and behavior (Chen et al. 1995). These unobserved cultural differences might hamper the identification of language barriers in terms of an omitted variable bias. To address this identification issue, we additionally control for the geographic distance between the destination and the immigrant's source country. Moreover, we use a measure of genetic differences between populations as a proxy for cultural differences. Spolaore and Wacziarg (2009) combine the frequencies of gene manifestations in populations sampled by Cavalli-Sforza et al. (1994) and the ethnicity composition of countries compiled by Alesina et al. (2003) to derive a measure of the average genetic distance between countries. The change in genes, the emergence of new alleles, happens randomly at an almost constant rate. This constant rate of change over time makes it a reasonable proxy for the time populations spent separated, making the genetic distance an "excellent summary statistic capturing divergence in the whole set of implicit beliefs, customs, habits, biases, conventions, etc. that are transmitted across generations—biologically and/or culturally—with high persistence." (Spolaore and Wacziarg 2009, p. 471). Including this measure of genetic distance as a proxy for cultural distance and assuming a reasonable correlation between the measured genetic differences and any unobservable cultural differences should allow the identification of the isolated effect of linguistic distance in the estimations.[4]

The linguistic, genetic, and geographic distance are, due to their parallel emergence over time, likely to be highly correlated. High pairwise correlations could lead to difficulties in the identification of single effects, but the pairwise rank correlations in Table 4 are far from perfect. The linguistic distance measures are highly correlated among each other, increasing our confidence in these measures. The correlation between linguistic and geographic and especially between linguistic and genetic distance is distinctively lower.[5]

[4]The data on genetic differences was originally gathered by Cavalli-Sforza et al. (1994) for 42 subpopulations. Spolaore and Wacziarg (2009) extended this data to genetic differences between 180 countries by weighting it using data on the composition of ethnicities of countries compiled by Alesina et al. (2003). It is stressed again at this point that the measure of genetic distance focuses solely on genetic distance based on neutral change, not caused by evolutionary pressure, and therefore does not explain differences in language acquisition due to superior skills or ability.

[5]Due to the lag of normally distributed measures, we report the rank correlations instead of the Pearson correlation coefficients.

3 Data

To assess how the different approaches to measuring linguistic dissimilarities fare in explaining the differences in the language acquisition of immigrants, two sources of individual data are employed for the estimations. Large scale micro data from the American Community Survey (ACS) offers a comprehensive representative sample of the American immigrant population. Furthermore, using a dataset from an English speaking destination country allows us to compare the linguistically based approaches with the test-score measure by Chiswick and Miller (1999) due to its restriction to distances toward English. To take advantage of the comprehensive nature of the linguistically based measures of linguistic distance, we further use data from the German Socio-Economic Panel (SOEP) to analyze the influence of the linguistic origin on the language acquisition in a non English-speaking country. Besides this new application, the SOEP offers the benefit of a broader range of individual characteristics that are unobservable in census-like data such as the ACS.

The ACS data is taken from the 2006–2010 Public Use File and used as a pooled cross section. The dataset includes a self-reported measure of language skills which indicates English proficiency on a four point scale ranging from "Not at all/Bad" to "Very Well," which constitutes our dependent variable. To focus on the potential workforce, the sample is restricted to immigrants between 17 and 65 years of age. As we want to concentrate our analysis on immigrants who acquire a destination language as an additional language, we restrict the sample to immigrants arriving at an age of 17 or older, and who originate from a non-English speaking country. After excluding observations with missing information, the pooled sample consists of 514,874 observations. A disadvantage of using the ACS is that it only offers scarce background information. As explanatory variables in our model, we use information on the time of residence, the age at arrival, individual education, sex, and marital status.[6] We also include indicators of the source countries' geopolitical world region and the year of observation to control for region- and time-fixed effects.[7]

To bring the analysis beyond the case of English-speaking destination countries, we use the German SOEP as a long-run panel which is an excellent data source for immigration- and integration-specific research, due to its over-sampling of immigrants and a migration-specific background questionnaire.[8] The sample used in this study covers the period

[6]We recode the information on highest degree to compute years of schooling using a modified version of the definition proposed by Jaeger (1997) adapted to the categories of the ACS. Specifically, we recode: No schooling completed = 0, Nursery school to grade 4 = 4, Grade 5 or grade 6 = 6, Grade 7 or grade 8 = 8, Grade 9 = 9, Grade 10 = 10, Grade 11 = 11, Grade 12, no diploma = 12, High school graduate = 12, Some college, but less than one year = 13, One or more years of college, no degree = 13, Associate's degree = 14, Bachelor's degree = 16, Master's degree = 18, Professional school degree = 18, Doctoral degree = 18.

[7]The geopolitical regions are defined following the MAR project, see http://www.cidcm.umd.edu/mar.

[8]The SOEP is a panel survey conducted since 1984 covering more than 20,000 individuals per wave.

between 1997 and 2010. Until 2007, questions concerning the language proficiency of immigrants were included in every second wave, and on an annual basis after 2007. Analogously to the ACS sample, we restrict the SOEP sample to immigrants between 17 and 65 years of age who were at least 17 years of age when migrating to Germany, and who were born in a non-German speaking country. Furthermore, we exclude Ethnic Germans and asylum seekers from the sample. After excluding observations with missing values, we end up with a sample of 5,803 person-year observations which we use in a pooled cross-section.

Similarly to the ACS, the SOEP offers information on self-reported German (oral) proficiency. The self-reported measure of language proficiency is fivefold, but because of the small number of individuals indicating the category "Not at all," we recode this information to derive an analog fourfold ordinal measure ranging from "Not at all/Bad" to "Very Well."

The survey character of the SOEP offers a broader range of information about the individual characteristics shaping the language acquisition process. The factors influencing the language acquisition of immigrants can be divided into three groups: the exposure to the destination-country language, the efficiency of their learning ability, and the economic incentives of learning the new language (Chiswick and Miller 1995). Our main variable of interest—the linguistic distance—affects the efficiency in acquiring the new language, decreasing the potential of any lexical transfer or portability of their proficiency in the source-country language. The efficiency of learning a new language is further controlled for by individual years of education, an indicator of good proficiency in the source-country language (as a proxy for literacy) and the age at entry, related to neurobiological research demonstrating a decreased efficiency for older arrivers (Newport 2002).

We model the effect of exposure to the destination-country language by including five variables in our estimation model. The simple 'learning by doing' effect is captured by a function of the years since migration. Moreover, we account for family composition characteristics captured by the number of children, marital status, and the German nationality of the spouse. The relation of these factors to the language acquisition process is ambiguous, because they lead to a social exclusion or inclusion of immigrants. Finally, an indicator for neighboring countries of Germany serves as a proxy for the probability of pre-migration exposure to the German language.

The economic incentives for learning a new language are primarily influenced by the expected length of stay, shaping the time horizon of the expected benefits. An indicator variable for having family ties abroad captures potential return plans that might alter

For more information, see Haisken-DeNew and Frick (2005). The data used in our analysis 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.

the economic incentives to invest in the destination language. Our estimation model also includes an indicator for immigrant's sex and controls for the source country's geopolitical world region and the year of observation using region- and time-fixed effects.

We augment both individual datasets—the ACS and the SOEP—with a set of aggregated country characteristics. These characteristics capture aspects of the relation between the immigrant's country of birth and the country of residence that might be correlated with the linguistic distance. First, we include the share of immigrants from the migrant's source country among the destination country's population to capture potential network and enclave effects. The data on bilateral migrant stocks are taken from United Nations (2012). Ethnic enclaves may reduce the incentives for immigrants to acquire destination-country specific abilities such as proficiency in the official language. Although the share of immigrants of the same source country is only a raw proxy for the immigrant's neighborhood, it might still provide some insights into the role of networks and enclaves in the acquisition of foreign language skills. Second, we control for the geographic distance, which serves as a proxy for the individual costs of migration. The geographic distance is defined as the geodesic distance between the capitals of the source and the destination country in 100 kilometers.[9] Lastly, we include a measure of the genetic distance between the source and the destination country as discussed in Section 2, which serves as a proxy for cultural differences.[10]

As neither of our micro data sources (the ACS and the SOEP) offer information on the mother tongue of an immigrant, the linguistic distance is assigned by the predominant language of the country of birth. In multi-lingual countries, languages are 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 is identified using a multitude of sources, including factbooks, encyclopedias, and Internet resources.[11] To allow easier comparison between the differently defined measures, we standardize each measure to have a mean of zero and a standard deviation of one.

4 Method

This data setup, the ACS and SOEP micro data combined with the measures of linguistic distance, allows us to estimate the language proficiency L as a function of the linguistic distance and the control variables, both on an aggregated and on the individual level.

[9] 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.

[10] Descriptive statistics for the ACS and the SOEP sample are presented in Table A1 in the Appendix. Table A2 in the Appendix provides a description of the variables used in our estimations.

[11] A comprehensive index of assigned languages with further explanations is available upon request.

To get a first glimpse into the relationship between linguistic barriers and the language acquisition of migrant groups, we start with estimations on the aggregated level. In these estimations, we explain the average language proficiency by source country and year of observation. As dependent variable, we use predictions from a first stage explaining the individual language proficiency L_{it} by a fully interacted set of source-country (c_j^S) and time indicators (T_k) and a set of individual characteristics X_{it} (gender, marital status, years since migration and age at entry):

$$L_{it} = \beta_0 + X'_{it}\beta + \sum_{j=1}^{J}\sum_{k=1}^{K}\gamma c_j^S T_k + \varepsilon_{it}. \quad (4)$$

From this first stage, we derive averages of the predicted language proficiency by source country and year of observation ($\widehat{L_{jt}}$). In the second step, we then explain these predicted values by the respective linguistic distance (LD_j) and a set of aggregated source-country and country-pair characteristics (Z_{jt}):

$$\widehat{L_{jt}} = \delta_0 + \delta_1 LD_j + Z'_{jt}\eta + \varepsilon_{jt}. \quad (5)$$

Although this specification on an aggregated country-of-origin level provides some first insights in the relation between linguistic barriers and the language acquisition, it ignores further available information on individual migration experience and potential interactions between the linguistic barriers and individual characteristics. Therefore, in a second step we change to the individual level and model the destination language proficiency as:

$$L_{it} = \beta_0 + \beta_1 LD_i + \beta_2 YSM_{it} + X'_{it}\gamma + \varepsilon_{it}. \quad (6)$$

Here, LD depicts the linguistic distance between the source- and destination-country languages, YSM represents the years since migration, and X is a vector including the control variables. In the following, we refer to the model depicted by Equation 6 as Model 1.[12]

In Model 1, β_1 represents an average effect of linguistic origin for all immigrants. However, it is likely that the linguistic distance not only imposes an initial barrier to language acquisition, but also affects the steepness of the language acquisition. Two different profiles are imaginable. On the one hand, recent immigrants with a distant linguistic background might have higher incentives to invest in language skills than

[12]For the sake of brevity, we present here and in the following only the linear notation of our estimation models.

linguistically close immigrants due to decreasing returns to invested effort. This would lead to a convergence over time. On the other hand, the hurdles imposed by language barriers can discourage investments and might lead to flatter acquisition profiles for distant immigrants. This would then lead to a divergence for immigrants from different linguistic origins, leaving linguistically distant immigrants worse off.

To address this potential convergence or divergence, we allow the disadvantage by the linguistic distance to vary with the years since migration. We include an interaction of both variables $LD \times YSM$ in Equation 7. We will refer to this specification as Model 2:

$$L_{it} = \beta_0 + \beta_1 LD_i + \beta_2 YSM_{it} + \beta_3 LD_i \times YSM_{it} + X'_{it}\gamma + \varepsilon_{it}. \quad (7)$$

In Model 2 the main effect indicated by β_1 shows the effect of linguistic distance on language ability at the time of immigration and β_3 depicts the change in the steepness of the assimilation profile. A convergence in skill levels over time should be represented in a positive coefficient β_3, indicating a catching up to immigrants with a lower linguistic distance. A negative β_3 would imply a divergence. Linguistically more distant immigrants would then face flatter assimilation profiles than immigrants with a lower linguistic distance.

We start our analysis by estimating our models using Ordinary Least Squares (OLS), separately for the four measures of linguistic distance in the US case and three measures in the German case. To interpret the OLS results using the ordinal language proficiency variable quantitatively, we have to impose strong cardinality assumptions. To take into account this ordinal character of the dependent variable and to derive quantitatively interpretable results, we repeat the estimations using Ordered Logit regressions and use graphical representations to interpret the interaction between linguistic distance and years since migration. Throughout all specifications in our analysis, we use (cluster)-robust standard errors to correct for possible heteroskedasticity in the data.

5 Results

A first descriptive look at the relation between language proficiency and the different measures of linguistic distance is provided in Table 5. The distribution of language skills in the US and Germany is quite different. While in the US about 37% of all immigrants report a "Very Well" proficiency, in Germany only 15% report the highest category. The expected negative relation between linguistic distance and language proficiency does not appear in the unconditional means reported in Table 5 in the US sample, ASJP, WALS and SCORE even suggest a marginally positive relation. In the German sample, the relation between linguistic distance and language proficiency is distinctively negative on

the descriptive level: across all three available measures, we observe a decrease in the linguistic distance as the language skills increases, with the lowest average distance in the "Very Well" category. However, it remains to be seen how potentially correlated individual characteristics change this first descriptive picture.

The results of the estimations of equation 5 on the aggregated source country level are summarized in Table 6.[13] We find a strong negative relationship between linguistic barriers and the average language proficiency. This result is robust, and can be observed both in the US and the German data, but differs distinctively in magnitude across different measures of linguistic distance. Assuming cardinality in our dependent variable, the coefficient of linguistic distance measured by the ASJP approach indicates that an increase of the linguistic distance by one standard deviation (roughly the difference in the distance to English between German and Romanian) decreases the average language proficiency by 0.17 points on the 0–3 scale in the US sample and 0.19 points in the German sample. Using the WALS or TREE approach shows a decrease by only 0.11 points in the US sample, while the TREE approach indicates a decrease of 0.2 points in the German sample, comparable to the ASJP sample. Concerning the control variables, migrant stocks are negatively related to the average language proficiency, hinting at potential negative influences of ethnolinguistic enclaves, see also Dustmann and Fabbri (2003), Chiswick and Miller (2002) and Cutler et al. (2008). We further find positive relationships between geographic and genetic distance (as a proxy for cultural differences) which we interpret as indirect evidence for selection on unobservable motivation and ability, while the positive coefficients of GDP per capita capture potential differences in pre-migration language exposure and education.

The results provided in Table 7 bring the analysis to the individual level. Table 7 summarizes the results of the OLS estimations for the ACS sample, separately for the different measures of linguistic distance. As already seen in the aggregated results, the estimations of the effect of the linguistic distance remain very volatile to the choice of employed measure. This highlights the importance of applying different available measures, rather than relying on only one approach, to get a comprehensive insight into the relation of linguistic barriers and the language acquisition. The results of Model 1 are summarized in Panel A. Across all different methods, the effect is highly significant and negative. Similar to the aggregated results, the ASJP approach indicates the strongest influence of the linguistic origin on the language acquisition: an increase by one standard deviation is related to a lower language proficiency by 0.24 points on the 0–3 scale, while estimations using the TREE, the WALS and the SCORE approach indicate a decrease by 0.10 to 0.12 points.

The coefficients in Panel A represent an average effect of linguistic distance for immi-

[13] We generated all estimation output tables using the Stata routine *estout* by Ben Jann (see, Jann 2007).

grants sharing a common linguistic background. To analyze whether this disadvantage is increasing or decreasing over time of residence, Model 2 includes an interaction term between the years since migration and the linguistic distance. The respective results are included in Panel B. In this specification, the main effect of linguistic distance is to be interpreted as an initial disadvantage at the time of immigration. Compared to Model 1, this initial disadvantage is smaller than the average difference by linguistic origin in Panel A. This results from a negative interaction between linguistic distance and the years since migration. Although we observe an overall positive language assimilation over time, the language assimilation profile becomes flatter with increased linguistic distance. Linguistically distant immigrants not only experience a higher initial disadvantage in their language acquisition, but also seem to experience a slower acquisition of English as destination language. After immigration, the initial gap between the immigrants from close and from distant linguistic origins increases over time. This pattern is robust across all four different models, while again the effect is strongest for the ASJP approach.[14]

To drop the cardinality assumption and to take the ordinal character of the self-reported language proficiency into account, we estimate both models using Ordered Logit regressions instead of OLS. Table 8 provides the marginal effects of the linguistic distance on the probability of reporting specific categories of language proficiency in Model 1.[15] Increasing the linguistic distance quantified by the ASJP approach by one standard deviation decreases the probability of reporting "Very Well" language skills in English by about 20 percentage points. Due to the non-linear Ordered Logit model and the inclusion of an interaction term, the marginal effects of linguistic distance in Model 2 are best interpreted in a graphical manner. Figure 2 depicts predicted probabilities of reporting the highest category of language proficiency by different levels of linguistic distance over the time of residence. Linguistically close immigrants in the 1^{st} percentile of the distance distribution face a initially steeper assimilation profile, linguistically distant migrants are outpaced.

While this pattern sheds some light on the effect of the heterogeneity in linguistic origin on the language acquisition of immigrants in the US, the large differences in immigration policy regimes and differences in selection patterns make it difficult to generalize the results to other countries. Previous analyses using the SCORE approach have been restricted to English-speaking destination countries (Chiswick and Miller 1999). However, English, as a lingua franca in international trade, the Internet, and communication technology, might enjoy very different incentives for being learned, compared to languages which lack this worldwide predominance. Against this restriction, a major advantage of the linguistically

[14]Regarding the influence of the control variables, Model 1 and Model 2 do not differ much, neither does the influence of the control variables vary with the measure of linguistic distance applied. For the sake of brevity, we do not further discuss the influence of the control variables. The respective coefficients can be found in Table A5 in the Appendix.

[15]The underlying coefficients and the marginal effects of Model 1 and 2 of the Ordered Logit regressions are presented in Tables A6–A8 in the Appendix.

based methods of language differences is the general applicability to any pair of languages. Taking advantage of this general applicability, we are able to extend the analysis beyond immigration to English-speaking countries. Specifically, we turn to Germany, one of the most important non English-speaking destinations for international migration.

The German SOEP sample allows a similar analysis as that of the ACS data, but with a richer set of control variables including the number of children, literacy, family ties abroad, and having a native spouse.[16] Table 9 lists the respective OLS results of Model 1 and Model 2.[17] Again, we find a negative effect of the linguistic distance between mother tongue and destination language on the language acquisition process, which differs strongly by the employed approach. To derive a quantitative interpretation, we again turn to results of an Ordered Logit model in Table 8. The marginal effects of Model 1 show a negative effect of linguistic distance on reporting "Very Well" German proficiency by 1.9 to 4.4 percentage points, which is moderate compared to the US results.[18]

However, the results for Model 2 draw a very different picture for the German SOEP sample compared to the US results. The interaction term between the linguistic distance and the years since migration in Model 2 turns out to have a positive sign but is insignificant across all different estimations in the German case (see Table 9, Panel B). This slight convergence is more distinctive in the Ordered Logit results, which are illustrated in Figure 3 in terms of predicted probabilities of reporting "Very Well" proficiency. Immigrants from a more distant linguistic origin therefore face a steeper assimilation profile than immigrants with a close linguistic background. Instead of observing a divergence by linguistic origin, we find a convergence in language skills. Over the time of residence, the gap between the linguistically close and distant immigrants closes, linguistically distant immigrants are able to catch up.

We might speculate about the driving factors of the difference between the divergence and convergence patterns in the US and in Germany. English and German are very closely related Germanic languages. This raises doubts that the differences are simply driven by purely linguistic reasons, such as that one language possesses particularly strong obstacles, e.g., by very special grammatical features, that would lead to the observed divergence. A more economically based potential explanation are differences in unobservable characteristics by different selection patterns of immigrants in the US and Germany. A perceived higher difficulty of German compared to English could lead to a self-selection of immigrants with superior skills in the acquisition of foreign languages into Germany. If this selection pattern were stronger for linguistically more distant immigrants

[16]Following Dustmann (1999), literacy is assumed for individuals reporting being able to write in their mother tongue.

[17]The coefficients, omitted in Table 9, are included in Table A9 in the Appendix.

[18]The underlying coefficients and the marginal effects of Model 1 and 2 of the Ordered Logit regressions are presented in Tables A10–A12 in the Appendix.

(where the initial returns to the language acquisition would be higher), the observed competing patterns of divergence in the US and convergence in Germany could arise. However, as we control in both samples for individual education, which is expected to be correlated with unobserved ability, we should at least partially capture such a selection process.

A second, in our opinion more plausible, explanation might be related to enclave effects in language acquisition. A range of studies have addressed the potentially discouraging effects of linguistical enclaves on investments in language skills (e.g., Chiswick and Miller 2002, Dustmann and Fabbri 2003, Cutler et al. 2008). Living in a linguistic enclave reduces the need for and potential advantages of learning the destination language, as immigrants can communicate in daily life in their mother tongue. Danzer and Yaman (2010) argue that the probability of moving into a neighborhood dominated by speakers of their own mother tongue is positively related to the own learning costs. The initial learning costs are strongly related to the linguistic distance between the mother tongue and the destination language, making it more likely for linguistically distant immigrants to move into segregated neighborhoods. Neighborhood segregation needs time to take place: due to its longer migration history, the ethnic segregation within cities is much more pronounced in the US than in Germany with its comparably short migration history. Therefore, the observed differences in assimilation patterns are potentially driven by the larger prevalence of linguistic enclaves in the US (e.g., the famous Chinatowns and Little Italy's in US cities).

In order to test the robustness of our results, we use different subsamples of our datasets. In doing so, we split the sample: (i) by gender, (ii) between low-skilled and high-skilled immigrants, and (iii) by excluding the majority immigrant groups, i.e., Mexican immigrants in the US and Turkish immigrants in Germany, from our regressions. A summary of these sensitivity checks is provided in Tables A3 and A4 in the Appendix. The underlying pattern of initial disadvantage and divergence in the US is stable across all subsamples. Linguistic barriers seem to play a larger role in the case of low-skilled immigrants (having a high school degree or less) than for high-skilled immigrants. The results for Germany are less robust, likely due to the low number of observations in the SOEP data. The negative main effect of linguistic distance remains robust across all different subsamples. The interaction term between linguistic distance and the years since migration becomes positively significant for high-skilled immigrants, who seem to drive the observed convergence in Germany. However, the convergence profile becomes insignificant when we split the sample by gender, by skill-level, and when Turks are excluded.

To summarize, our results highlight the importance of linguistic origin as a factor of typically unobservable heterogeneity in the integration process of immigrants. The initial disadvantages only marginally disappear over time of residence, but linguistic barriers

remain even after a long period of stay. Given the large impact of language proficiency on labor market outcomes (Chiswick and Miller 1995, Dustmann and van Soest 2002, Bleakley and Chin 2004), it is likely that these differences are transferred into labor market disadvantages. Disadvantages in the language acquisition process prevent the social integration of immigrants by reducing their ability to communicate with natives. In addition, imperfect language skills can act as a signal for foreignness, opening the way to discriminatory behavior of employers and decreasing the productiveness of individuals, leading to lower employment probabilities and wages.

Against the background of immigration policy design, our results hint at a way to identify target groups for supportive integration policy measures. Immigrants obviously differ strongly in their costs of language acquisition, dependent on their linguistic background. This heterogeneity is seldom accounted for in the design of integration policies. Policies aiming at the support of immigrant language acquisition, as currently practiced in Germany with the "Integrationskurse" system ("integration classes"), often include a lump sum payment for public language classes. This lump sum payment, restricting class hours irrespective of the actual need for support, is likely to lead to a inefficient spending of public money. In a class system that does not distinguish language students by their actual need for support, linguistically close immigrants are provided too many class hours, while linguistically distant immigrants might be outpaced. A means-tested voucher taking into account the expected costs by linguistic origin might lead to a more efficient spending of public money than a lump sum policy measure.

6 Conclusion

International labor migration is a worldwide and steadily growing phenomenon. According to UN estimates, in 2010 roughly 215 million individuals lived in a country different from their country of birth (World Bank 2011). On a first glimpse, this is a massive number but it still accounts for only around 3% of the world's population, a surprisingly low number given the large differentials in economic conditions. While technological progress in transportation and communication have led to a significant decrease in the initial costs of migration, cultural and linguistical borders continue to play an important role for international migration flows (Belot and Ederveen 2012, Adsera and Pytlikova 2012). In this study, we provide a in-depth analysis and quantification of the linguistic barriers in destination language acquisition in Germany and the US.

For immigrants, proficiency in the destination-country language leads to substantial economic returns (Dustmann and van Soest 2002, Bleakley and Chin 2004). However, large fractions of the immigrant population possess only insufficient levels of proficiency in the destination language. While the investment in language capital has been thoroughly

analyzed in human capital frameworks (Chiswick and Miller 1995), our knowledge of the influence of typically unobservable heterogeneity in the linguistic origin of immigrants remains limited. The linguistic distance between languages is a concept that is difficult to operationalize for its implementation in empirical models. In this study, we demonstrate four different methods providing continuous measures of linguistic differences and compare their specific advantages and shortcomings. More specifically, we draw from linguistic research and propose using a measure of linguistic distance based on comparisons of pronunciation between word lists. This method, referred to as the ASJP approach, offers a convenient way to derive a continuous measure of linguistic differences. Given its purely descriptive measurement and general applicability to any potential pair of languages, it provides an advantageous measure for the application at hand. We compare its performance with further linguistic approaches using information about language relations (TREE measure) and language characteristics (WALS measure) and a measure based on average test scores (SCORE) by Chiswick and Miller (1999).

All four measures of language differences are applied to the analysis of the destination language acquisition of immigrants in the US using data of the American Community Survey (ACS). To take advantage of the general applicability of the linguistically based methods beyond the analysis of English-speaking destination countries, we extend the analysis to German microdata from the German Socio-Economic Panel (SOEP). In both scenarios, we use the different measures of linguistic distance to explain differences in self-reported measures of immigrant's destination language proficiency.

Our results indicate that the linguistic distance, the dissimilarity between the origin and destination languages, has a distinctively negative average effect on the language acquisition of immigrants. Immigrants with a distant linguistic origin face higher costs in the language acquisition than immigrants with a closer linguistic background. Furthermore, we analyze differences in the slope of the language assimilation curve that can be attributed to differences in the linguistic origin. We find different assimilation patterns for the US and Germany. In Germany, immigrants with a more distant source-country language display a steeper language assimilation profile. Initial disadvantages are reduced over time, leading to a convergence in average proficiency for immigrants from different linguistic origins. For the US, we estimate the opposite picture of diverging profiles. Gaps in the proficiency of linguistically close and distant immigrants tend to increase over time of residence. We interpret this difference in assimilation patterns as a potential outcome of stronger enclave effects in the US. This crucial difference highlights the importance of extending the analysis beyond the case of Anglophone countries.

The initial disadvantages and differences in assimilation patterns attributable to linguistic distance are able to explain a large fraction of the explained variation in the destination language proficiency. This highlights the importance of linguistic differences

for the analysis of the skill acquisition of immigrants, as an influencing factor that was previously part of the "black box" of culture in the economic literature (see, Epstein and Gang 2010). This additionally explained variation might play an important role in the design of integration policy measures. Lump sum payments for language classes might turn out to be inefficient in the presence of a high heterogeneity in the actual need for language acquisition support and compared to means-tested vouchers taking into account the expected costs of language acquisition.

References

Adsera, A. and Pytlikova, M. (2012). The role of language in shaping international migration. CReAM Discussion Paper No. 06/12.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. and Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth* 8: 155–194.

Bauer, T., Epstein, G. S. and Gang, I. N. (2005). Enclaves, language, and the location choice of migrants. *Journal of Population Economics* 18: 649–662.

Belot, M. and Ederveen, S. (2012). Cultural barriers in migration between OECD countries. *Journal of Population Economics* 25: 1077–1105.

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

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

Carliner, G. (1981). Wage differences by language group and the market for language skills in Canada. *Journal of Human Resources* 16: 384–399.

Cavalli-Sforza, L. L., Menozzi, P. and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press.

Chen, J., Sokal, R. R. and Ruhlen, M. (1995). Worldwide analysis of genetic and linguistic relationships of human populations. *Human Biology* 67: 595–612.

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

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

Chiswick, B. R. and Miller, P. W. (1999). English language fluency among immigrants in the United States. In Polachek, S. W. (ed.), *Research in Labor Economics*. Oxford, UK: JAI Press, *17*, 151–200.

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

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

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

Chiswick, B. R. and Miller, P. W. (2007). Modeling Immigrants' Language Skills. IZA Discussion Paper No. 2974.

Crystal, D. (2010). *The Cambridge Encyclopedia of Language*. Cambridge, UK: Cambridge University Press, 3rd ed.

Cutler, D. M., Glaeser, E. L. and Vigdor, J. L. (2008). When are ghettos bad? Lessons from immigrant segregation in the United States. *Journal of Urban Economics* 63: 759–774.

Danzer, A. M. and Yaman, F. (2010). Ethnic Concentration and Language Fluency of Immigrants in Germany. IZA Discussion Paper No. 4742.

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

Dustmann, C. (1999). Temporary migration, human capital, and language fluency of migrants. *Scandinavian Journal of Economics* 101: 297–314.

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

Dustmann, C. and van Soest, A. (2001). Language fluency and earnings: Estimation with misclassified language indicators. *Review of Economics and Statistics* 83: 663–674.

Dustmann, C. and van Soest, A. (2002). Language and the earnings of immigrants. *Industrial and Labor Relations Review* 55: 473–492.

Epstein, G. S. and Gang, I. N. (2010). Migration and culture. In Epstein, G. S. and Gang, I. N. (eds), *Frontiers of Economics and Globalization: Migration and Culture*. Bingley, UK: Emerald Group Publishing Limited, *8*, 1–21.

Esser, H. (2006). Migration, language and integration: AKI research review 4. `http://bibliothek.wz-berlin.de/pdf/2006/iv06-akibilanz4b.pdf`.

Evans, M. 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: 226–236.

Friedberg, R. M. (2000). You can't take it with you? Immigrant assimilation and the portability of human capital. *Journal of Labor Economics* 18: 221–51.

Haisken-DeNew, J. P. and Frick, J. R. (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, J. P. and Hahn, M. (2006). PanelWhiz: A Flexible Modularized Stata Interface for Accessing Large Scale Panel Data Sets. `http://www.panelwhiz.eu/docs/PanelWhiz_Introduction.pdf`.

Hart-Gonzalez, L. and Lindemann, S. (1993). Expected Achievement in Speaking Proficiency. Foreign Service Institute, US Department of State.

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

Jaeger, D. A. (1997). Reconciling the old and new Census Bureau education questions: Recommendations for researchers. *Journal of Business & Economic Statistics* 15: 300–309.

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

Lewis, P. M. (2009). *Ethnologue: Languages of the World*. Dallas, TX: SIL International, 16th ed.

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

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

Newport, E. L. (2002). Critical periods in language development. In Nadel, L. (ed.), *Encyclopedia of Cognitive Science*. London, UK: Macmillan Publishers Ltd./Nature Publishing Group, 737–740.

Petroni, F. and Serva, M. (2010). Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications* 389: 2280–2283.

Santacreu-Vasut, E., Shoham, A. and Gay, V. (2013). Do female/male distinctions in language matter? Evidence from gender political quotas. *Applied Economics Letters* 20: 495–498.

Spolaore, E. and Wacziarg, R. (2009). The diffusion of development. *The Quarterly Journal of Economics* 124: 469–529.

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

United Nations (2012). Trends in International Migrant Stock: Migrants by Destination and Origin. United Nations database, Department of Economic and Social Affairs.

Van der Slik, F. W. P. (2010). Acquisition of Dutch as a second language. *Studies in Second Language Acquisition* 32: 401–432.

World Bank (2011). *Migration and Remittances Factbook 2011*. Washington, DC: World Bank Publications, 2nd ed.

Figures and Tables

Table 1: Average Test Scores of US Language Students

Average Test Scores	Linguistic Distance	Languages (Examples)
1.00	1.00	Japanese, Korean, Laotian
1.25	0.80	Cantonese, Hakka, Mien
1.50	0.67	Arabic, Syriac, Vietnamese
1.75	0.57	Bengali, Greek, Nepali
2.00	0.50	Finnish, Serbo-Croatian, Turkish
2.25	0.44	Danish, Spanish, Yiddish
2.50	0.40	French, Italian, Portuguese
2.75	0.36	Bantu, Dutch, Swahili
3.00	0.33	Afrikaans, Norwegian, Swedish

Notes: Average test scores of American students learning foreign languages. Numbers provided by Hart-Gonzalez and Lindemann (1993), reproduced from Chiswick and Miller (1999), Appendix B.

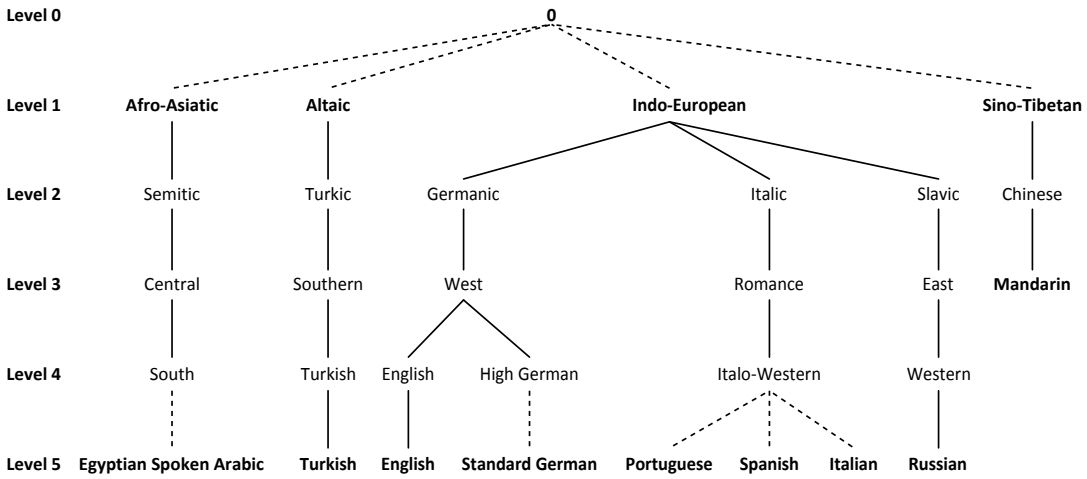

Figure 1: Language Relations in the TREE Approach

Table 2: 40-Items Swadesh Word List with Computational Examples

A. 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

B. ASJP Computation

Word	**English**	**German**	**Distance**
fish	*fiS*	fiS	0
you	*yu*	du	1
hand	*hEnd*	hant	2
tree	*tri*	baum	4
mountain	*maunt3n*	bErk	7

C. Examples for WALS Features

Feature	**English**	**German**
Consonant-vowel ratio	Low	Low
Vowel Nasalization	Present	Absent
Number of cases	2	4

Notes: Panel A displays the 40-item Swadesh sublist used in the computation of the ASJP approach. – Panel B: Examples for the computation of the linguistic distance between English and German. – Panel C: Examples for differences in WALS features between English and German.

Table 3: Closest and Furthest Languages to English and German

	ASJP Approach				WALS Approach			
	Closest		Furthest		Closest		Furthest	
	Language	Distance	Language	Distance	Language	Distance	Language	Distance
Distance to English								
	St. Vincent Creole	41.34	Somali (Somalia)	103.03	German	0.302	Burmese (Burma)	0.627
	Jamaican Creole	41.95	Twi Asante (Ghana)	103.43	French	0.338	Guarani (Paraguay)	0.632
	Krio (Sierra Leone)	47.44	Vietnamese	104.06	Russian	0.348	Somali (Somalia)	0.667
Distance to German								
	Luxembourgish	42.12	Korean	104.30	French	0.283	Vietnamese	0.681
	Bernese German	48.34	Palestinian Arabic	103.72	English	0.302	Guarani (Paraguay)	0.685
	Dutch	51.50	Yoruba (Nigeria)	103.58	Russian	0.310	Zulu (South Africa)	0.688
	TREE Approach*				**SCORE Approach***			
	Closest		Furthest		Closest		Furthest	
	Language	Distance	Language	Distance	Language	Distance	Language	Distance
Distance to English								
	Westvlaams (Belgium)	0.55	Amoy Minnan Chinese	1.00	Norwegian Bokmal	0.33	Palestinian Arabic	0.67
	Dutch	0.55	Palestinian Arabic	1.00	Swedish	0.33	Korean	1.00
	Bernese German	0.55	Vietnamese	1.00	Romanian	0.33	Japanese	1.00
Distance to German								
	Luxembourgish	0.30	Korean	1.00	–	–	–	–
	English	0.55	Palestinian Arabic	1.00	–	–	–	–
	Westvlaams (Belgium)	0.55	Yoruba (Nigeria)	1.00	–	–	–	–

*Notes: – The table shows the three closest and furthest languages toward English and German according to different measures of linguistic distance. – Only languages spoken within samples are listed. – Geographic origin of language in parentheses. – *Due to the small number of increments, a range of languages shares the "top" positions.*

Table 4: Rank Correlations among Linguistic, Geographic, and Genetic Distance Measures

	Linguistic Distance ASJP	Linguistic Distance WALS	Linguistic Distance TREE	Linguistic Distance SCORE	Geographic Distance (100 km)	Genetic Distance
ACS Sample						
Linguistic distance ASJP	1					
Linguistic distance WALS	0.84	1				
Linguistic distance TREE	0.84	0.83	1			
Linguistic distance SCORE	0.81	0.87	0.73	1		
Geographic distance (in 100 km)	0.79	0.72	0.66	0.72	1	
Genetic distance	0.48	0.54	0.70	0.35	0.25	1
SOEP Sample						
Linguistic distance ASJP	1					
Linguistic distance WALS	0.75	1				
Linguistic distance TREE	0.88	0.85	1			
Geographic distance (in 100 km)	0.61	0.47	0.61		1	
Genetic distance	0.69	0.81	0.80		0.64	1

Notes: – Spearman Rank correlations reported. – Number of observations in the ACS sample: 514,874. – Number of observations in the SOEP sample: 5,803.

Table 5: Distribution of Language Skills across Samples

Proficiency	Observations	Mean Linguistic Distance			
		ASJP	WALS	TREE	SCORE
ACS Sample					
Bad	58,523 0.11	93.79	0.40	0.91	0.46
Not bad	128,384 0.25	94.71	0.42	0.92	0.49
Well	139,223 0.27	95.62	0.45	0.93	0.52
Very well	188,744 0.37	95.20	0.46	0.92	0.52
Total	514,874				
SOEP Sample					
Bad	1,057 0.18	96.21	0.50	0.95	–
Not bad	1,918 0.33	95.14	0.48	0.94	–
Well	1,946 0.34	92.53	0.45	0.90	–
Very well	882 0.15	88.01	0.39	0.84	–
Total	5,803				

Notes: – Column 2 shows the absolute number of observations and the relative frequencies for each category.

Table 6: Language Ability and Linguistic Distance – Aggregated Results

	ACS Sample				SOEP Sample		
	ASJP Coef/StdE	WALS Coef/StdE	TREE Coef/StdE	SCORE Coef/StdE	ASJP Coef/StdE	WALS Coef/StdE	TREE Coef/StdE
Linguistic distance	-0.173^{***}	-0.105^{***}	-0.105^{***}	-0.128^{***}	-0.188^{***}	$-0.094^{\dagger}$	-0.195^{***}
	(0.022)	(0.014)	(0.013)	(0.012)	(0.031)	(0.048)	(0.032)
Migrant stock (% of population)	-0.207^{***}	-0.204^{***}	-0.218^{***}	-0.213^{***}	-0.341^{***}	-0.329^{***}	-0.341^{***}
	(0.020)	(0.021)	(0.022)	(0.019)	(0.044)	(0.045)	(0.045)
Geographic distance (in 100 km)	0.006^{***}	0.006^{***}	0.005^{***}	0.005^{***}	-0.001	-0.000	-0.003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
Genetic distance	0.003	-0.003	0.004	-0.007	0.054^{***}	0.039^{**}	0.063^{***}
	(0.006)	(0.006)	(0.006)	(0.005)	(0.014)	(0.014)	(0.015)
ln GDP per capita (in USD)	0.141^{***}	0.136^{***}	0.149^{***}	0.175^{***}	0.120^{**}	0.135^{**}	0.101^{*}
	(0.016)	(0.016)	(0.015)	(0.015)	(0.045)	(0.048)	(0.047)
Constant	0.622^{**}	0.759^{***}	0.562^{**}	0.497^{*}	0.716	0.723	$0.929^{\dagger}$
	(0.195)	(0.198)	(0.193)	(0.197)	(0.496)	(0.525)	(0.503)
Region dummies	yes	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	yes	yes	yes
Adjusted R^2	0.609	0.579	0.602	0.600	0.244	0.202	0.241
F Statistic	103.0	75.6	79.1	75.4	19.4	13.3	19.0
Observations	395	395	395	395	423	423	423

Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Robust standard errors are reported in parentheses. – Dependent variable: predicted average language skills by source country and year. – Level of observation is the source country, destination-country language ability evaluated as source-country averages of the predicted language proficiency.

Table 7: OLS Results of Linguistic Distance – ACS Sample

	ASJP	WALS	TREE	SCORE
	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE
Panel A: Model 1				
Linguistic distance	-0.241^{***}	-0.103^{***}	-0.107^{***}	-0.124^{***}
	(0.004)	(0.003)	(0.002)	(0.002)
Adjusted R^2	0.406	0.404	0.405	0.409
F Statistic	16,051.0	15,351.8	16,102.4	15,706.5
Observations	514,874	514,874	514,874	514,874
Panel B: Model 2				
Linguistic distance	-0.180^{***}	-0.038^{***}	-0.086^{***}	-0.075^{***}
	(0.004)	(0.004)	(0.002)	(0.003)
Years since migration	0.015^{***}	0.014^{***}	0.014^{***}	0.014^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
LD × YSM	-0.004^{***}	-0.005^{***}	-0.001^{***}	-0.003^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Adjusted R^2	0.407	0.405	0.405	0.410
F Statistic	14,965.9	14,489.8	15,076.1	14,881.2
Observations	514,874	514,874	514,874	514,874

Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 0 to 3 such that higher values indicate a higher level of language proficiency.

Table 8: Ordered Logit Marginal Effects of Linguistic Distance – Model 1 ACS & SOEP Sample

	Bad	Not bad	Well	Very well
	ME/StdE	ME/StdE	ME/StdE	ME/StdE
ACS Sample				
Linguistic distance ASJP	0.063^{***}	0.175^{***}	-0.034^{***}	-0.204^{***}
	(0.001)	(0.003)	(0.001)	(0.003)
Linguistic distance WALS	0.021^{***}	0.057^{***}	-0.012^{***}	-0.066^{***}
	(0.001)	(0.002)	(0.000)	(0.002)
Linguistic distance TREE	0.032^{***}	0.090^{***}	-0.018^{***}	-0.105^{***}
	(0.001)	(0.002)	(0.000)	(0.002)
Linguistic distance SCORE	0.021^{***}	0.058^{***}	-0.012^{***}	-0.067^{***}
	(0.000)	(0.001)	(0.000)	(0.001)
SOEP Sample				
Linguistic distance ASJP	0.017^{*}	0.035^{*}	-0.033^{*}	-0.019^{*}
	(0.008)	(0.016)	(0.015)	(0.009)
Linguistic distance WALS	0.038^{***}	0.080^{***}	-0.074^{***}	-0.044^{***}
	(0.011)	(0.024)	(0.022)	(0.013)
Linguistic distance TREE	0.027^{**}	0.056^{**}	-0.052^{**}	-0.031^{**}
	(0.009)	(0.020)	(0.018)	(0.011)

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 0 to 3 such that higher values indicate a higher level of language proficiency. – Marginal effects are reported at the mean of the covariates vector.

Table 9: OLS Results of Linguistic Distance – SOEP Sample

	ASJP	WALS	TREE
	Coef/StdE	Coef/StdE	Coef/StdE
Panel A: Model 1			
Linguistic distance	-0.079^{**}	-0.160^{**}	-0.117^{**}
	(0.030)	(0.051)	(0.037)
Adjusted R^2	0.373	0.375	0.376
F Statistic	34.55	35.80	34.59
Observations	5,803	5,803	5,803
Panel B: Model 2			
Linguistic distance	-0.121^{*}	-0.187^{**}	-0.182^{***}
	(0.050)	(0.059)	(0.054)
Years since migration	0.017^{***}	0.017^{***}	0.017^{***}
	(0.004)	(0.004)	(0.004)
LD × YSM	0.002	0.002	0.003
	(0.002)	(0.002)	(0.002)
Adjusted R^2	0.373	0.375	0.377
F Statistic	34.41	34.69	34.59
Observations	5,803	5,803	5,803

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 0 to 3 such that higher values indicate a higher level of language proficiency.

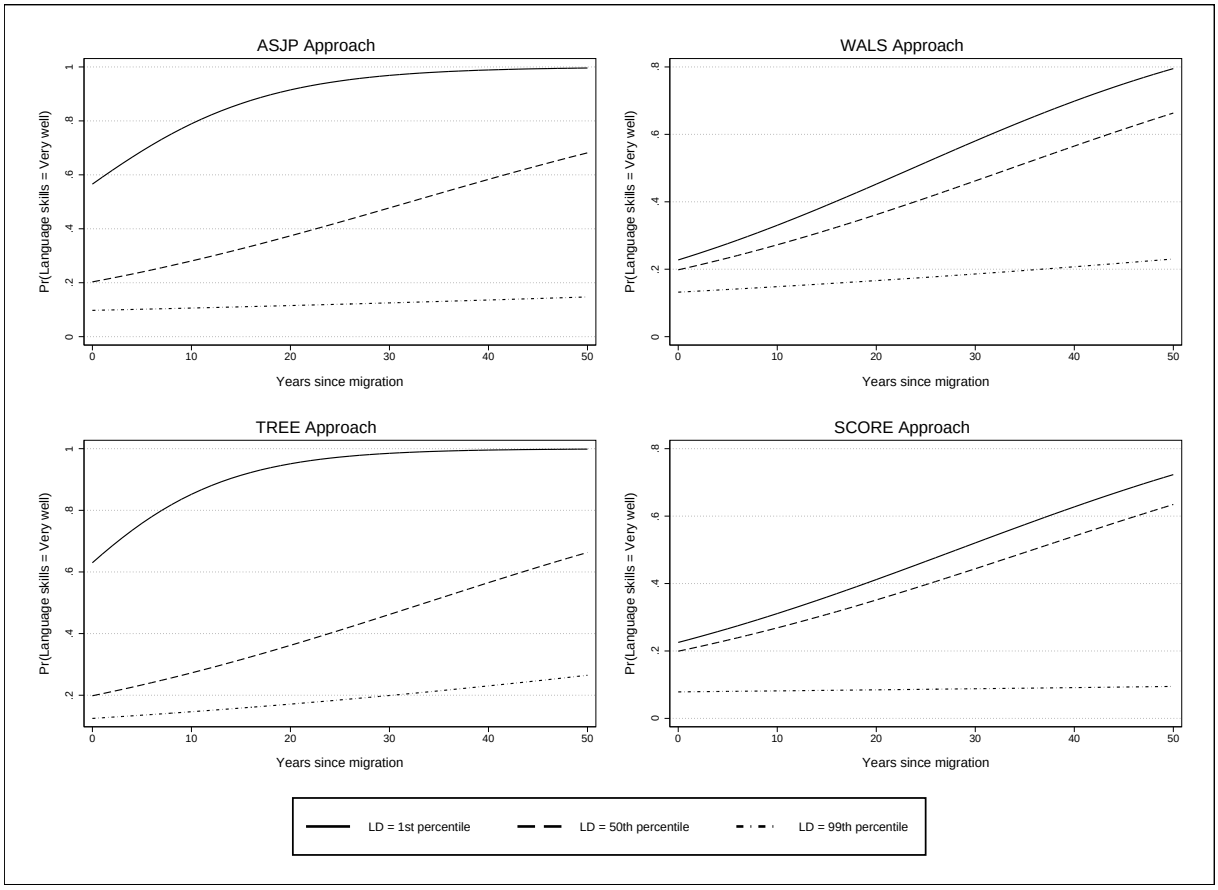

Figure 2: Predicted Language Assimilation Profiles for the ACS Sample

Notes: – The predicted assimilation profiles are based on Ordered Logit regressions of Model 2.

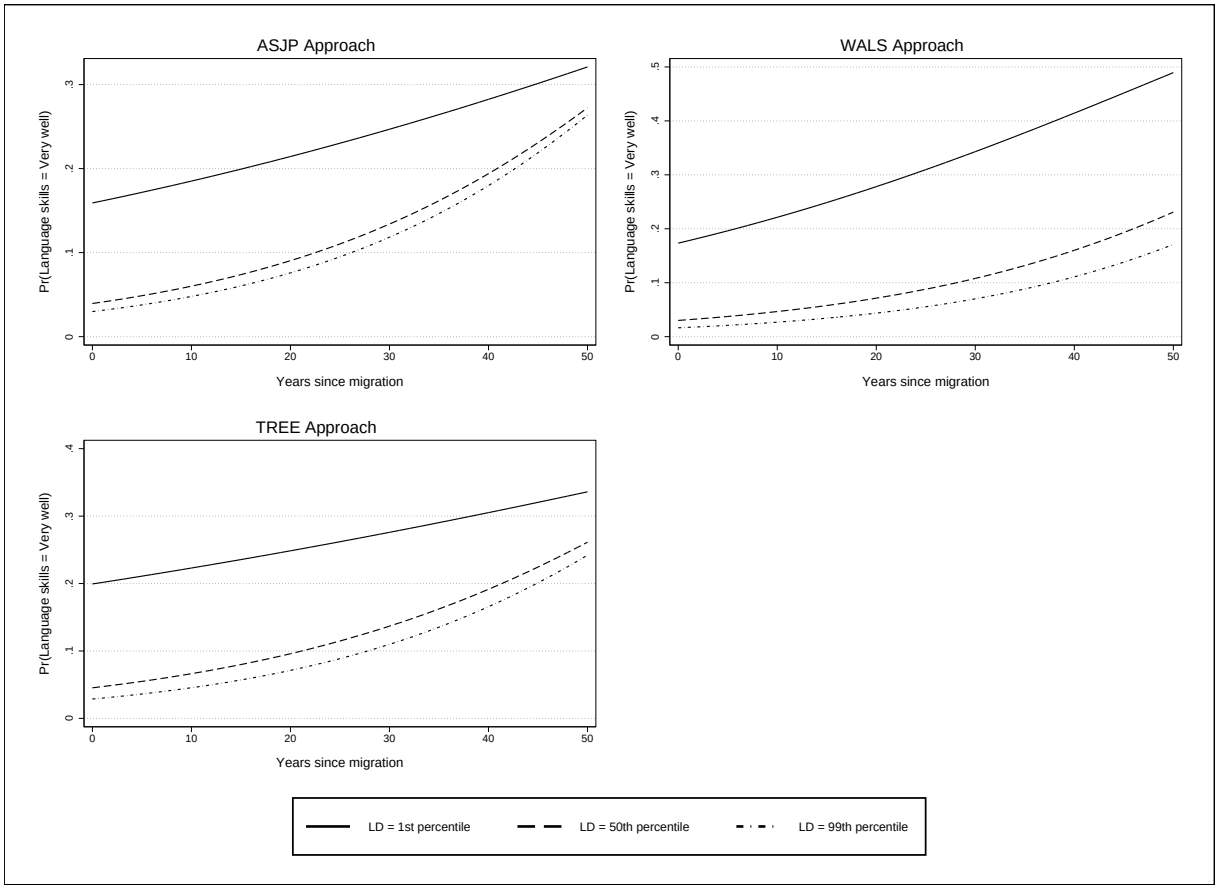

Figure 3: Predicted Language Assimilation Profiles for the SOEP Sample

Notes: – The predicted assimilation profiles are based on Ordered Logit regressions of Model 2.

Appendix

Table A1: Descriptive Statistics – ACS & SOEP Sample

	ACS Sample		SOEP Sample	
	Mean	StdD	Mean	StdD
English proficiency	1.890	1.029	–	–
Oral German proficiency	–	–	1.457	0.957
Linguistic distance ASJP	95.030	5.118	93.376	8.437
Linguistic distance WALS	0.439	0.090	0.458	0.107
Linguistic distance TREE	0.921	0.065	0.915	0.101
Linguistic distance SCORE	0.507	0.121	–	–
Years since migration	14.804	9.980	23.688	11.268
Age at entry	27.323	8.398	25.331	6.885
Years of education	12.159	4.459	10.192	2.555
Female	0.437	0.496	0.534	0.499
Married	0.630	0.483	0.850	0.357
Children in the HH.				
No children	–	–	0.564	0.496
One child	–	–	0.207	0.405
Two children	–	–	0.148	0.355
Three or more children	–	–	0.081	0.273
Native German partner/spouse	–	–	0.218	0.413
Family abroad	–	–	0.263	0.440
Proficiency home language	–	–	0.876	0.330
Desired stay (years)	–	–	14.722	11.266
Neighboring country	–	–	0.079	0.270
Migrant stock (% of population)	1.485	1.682	1.621	1.264
Geographic distance (in 100 km)	68.951	43.260	19.434	17.727
Genetic distance	8.813	2.686	3.734	2.923
ln GDP per capita (in USD)[a]	8.294	1.055	9.029	0.970
Observations	514,874		5,803	

Notes: – Unweighted means and standard deviations reported. – English and German proficiency is defined on a scale of 0 to 3, corresponding to the classifications "Bad", "Not bad", "Well", "Very well". – ln GDP per capita is based on 510,220 observations in the ACS sample and 5,801 observations in the SOEP sample.

Table A2: Variables Description – ACS & SOEP Sample

Variable	Description
English proficiency	Self-reported English proficiency
German proficiency	Self-reported oral German proficiency
Linguistic distance ASJP	Levenshtein distance normalized divided
Linguistic distance WALS	Measure based on structural language features by Lohmann (2011)
Linguistic distance TREE	Language tree measure based on Adsera and Pytlikova (2012)
Linguistic distance SCORE	Test-score measure by Chiswick and Miller (1999)
Years since migration	Years of residence in destination country
Age at entry	Age at entry into destination country
Years of education	Years of education
Female	Dummy = 1 if female
Married	Dummy = 1 if married
Children in the HH.	
No children	Dummy = 1 if no children live in the household
One child	Dummy = 1 if one child lives in the household
Two children	Dummy = 1 if two children live in the household
Three or more children	Dummy = 1 if three or more children live in the household
Native German partner/spouse	Dummy = 1 if partner/spouse is native German
Family abroad	Dummy = 1 if family lives abroad
Proficiency home language	Dummy = 1 if written proficiency in mother tongue is well or very well
Desired stay (years)	Years desired to stay in Germany
Neighboring country	Dummy = 1 if country of origin is a neighboring country of Germany
Migrant stock (% of population)	Migrant stock by source country as percentage of the total population
Geographic distance (in 100 km)	Geodesic distance between capitals in 100 km
Genetic distance	Weighted F_{ST} genetic distance, divided by 100 (Spolaore and Wacziarg 2009)
ln GDP per capita (in USD)	Logarithm of GDP per capita in constant 2000 US dollars
Region dummies	6 region dummies, indicating the source countries' geopolitical world region. The geopolitical regions are defined following the MAR project as: 1) Western democracies and Japan, 2) Eastern Europe and the former Soviet Union, 3) Asia, 4) North Africa and the Middle East, 5) Sub-Saharan Africa, and 6) Latin America and the Caribbean.
Year fixed effects	Time dummy variables for each sample year

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 the US and Germany, respectively. Geographic distance reports the calculated distance divided by 100.

Table A3: Robustness Checks: OLS Results of Linguistic Distance – Model 2 ACS Sample

	ASJP Coef/StdE	WALS Coef/StdE	TREE Coef/StdE	SCORE Coef/StdE
Male Subsample				
Linguistic distance	-0.165^{***}	-0.056^{***}	-0.081^{***}	-0.051^{***}
	(0.006)	(0.005)	(0.003)	(0.004)
LD × YSM	-0.007^{***}	-0.006^{***}	-0.003^{***}	-0.006^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Years since migration	0.015^{***}	0.014^{***}	0.014^{***}	0.014^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Adjusted R^2	0.394	0.393	0.392	0.396
F Statistic	8,976.37	8,850.42	9,019.55	8,889.42
Observations	290,127	290,127	290,127	290,127
Female Subsample				
Linguistic distance	-0.199^{***}	-0.026^{***}	-0.093^{***}	-0.100^{***}
	(0.006)	(0.006)	(0.003)	(0.004)
LD × YSM	-0.002^{***}	-0.004^{***}	-0.000^{***}	-0.001^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Years since migration	0.013^{***}	0.013^{***}	0.013^{***}	0.013^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Adjusted R^2	0.429	0.427	0.428	0.433
F Statistic	6,976.10	6,664.67	7,023.96	6,985.29
Observations	224,747	224,747	224,747	224,747
Low-skilled Subsample				
Linguistic distance	-0.381^{***}	-0.059^{***}	-0.178^{***}	-0.097^{***}
	(0.011)	(0.010)	(0.006)	(0.007)
LD × YSM	-0.001	-0.002^{***}	0.000	-0.001^{*}
	(0.000)	(0.000)	(0.000)	(0.000)
Years since migration	0.020^{***}	0.019^{***}	0.020^{***}	0.019^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Adjusted R^2	0.208	0.203	0.206	0.204
F Statistic	3,149.23	2,813.24	3,116.19	2,848.01
Observations	268,271	268,271	268,271	268,271
High-skilled Subsample				
Linguistic distance	-0.142^{***}	-0.049^{***}	-0.072^{***}	-0.096^{***}
	(0.004)	(0.004)	(0.002)	(0.003)
LD × YSM	-0.002^{***}	-0.003^{***}	-0.001^{***}	-0.002^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Years since migration	0.008^{***}	0.008^{***}	0.007^{***}	0.007^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Adjusted R^2	0.238	0.236	0.236	0.252
F Statistic	2,302.87	2,201.65	2,330.84	2,436.70
Observations	246,603	246,603	246,603	246,603
Subsample excluding Mexican immigrants				
Linguistic distance	-0.205^{***}	-0.057^{***}	-0.089^{***}	-0.095^{***}
	(0.004)	(0.004)	(0.002)	(0.003)
LD × YSM	-0.002^{***}	-0.003^{***}	-0.000^{***}	-0.002^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Years since migration	0.011^{***}	0.011^{***}	0.010^{***}	0.010^{***}
	(0.000)	(0.000)	(0.000)	(0.000)
Adjusted R^2	0.403	0.399	0.400	0.406
F Statistic	7,676.43	7,256.76	7,620.57	7,450.67
Observations	344,686	344,686	344,686	344,686

Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 0 to 3 such that higher values indicate a higher level of language proficiency.

Table A4: Robustness Checks: OLS Results of Linguistic Distance – Model 2 SOEP Sample

	ASJP Coef/StdE	WALS Coef/StdE	TREE Coef/StdE
Male Subsample			
Linguistic distance	-0.115	-0.223^{*}	-0.185^{*}
	(0.085)	(0.096)	(0.083)
LD × YSM	0.002	0.002	0.003
	(0.004)	(0.004)	(0.004)
Years since migration	0.016^{**}	0.015^{**}	0.015^{**}
	(0.005)	(0.005)	(0.005)
Adjusted R^2	0.302	0.307	0.308
F Statistic	13.55	14.18	13.32
Observations	2,703	2,703	2,703
Female Subsample			
Linguistic distance	-0.085	$-0.139^{\dagger}$	$-0.114^{\dagger}$
	(0.058)	(0.077)	(0.066)
LD × YSM	0.000	-0.000	0.000
	(0.002)	(0.003)	(0.002)
Years since migration	0.020^{***}	0.021^{***}	0.020^{***}
	(0.005)	(0.005)	(0.005)
Adjusted R^2	0.448	0.449	0.449
F Statistic	30.74	31.52	31.00
Observations	3,100	3,100	3,100
Low-skilled Subsample			
Linguistic distance	-0.101	-0.178^{**}	-0.158^{*}
	(0.062)	(0.062)	(0.070)
LD × YSM	-0.000	-0.001	0.000
	(0.002)	(0.003)	(0.002)
Years since migration	0.021^{***}	0.022^{***}	0.020^{***}
	(0.004)	(0.004)	(0.004)
Adjusted R^2	0.334	0.337	0.338
F Statistic	24.68	25.13	26.07
Observations	5,067	5,067	5,067
High-skilled Subsample			
Linguistic distance	-0.208^{**}	-0.150	-0.286^{***}
	(0.064)	(0.142)	(0.069)
LD × YSM	0.011^{**}	$0.008^{\dagger}$	0.014^{***}
	(0.004)	(0.005)	(0.004)
Years since migration	0.007	0.005	0.011
	(0.007)	(0.008)	(0.007)
Adjusted R^2	0.294	0.279	0.313
F Statistic	5.11	4.94	5.83
Observations	736	736	736
Subsample excluding Turkish immigrants			
Linguistic distance	-0.076	$-0.131^{\dagger}$	-0.136^{*}
	(0.053)	(0.069)	(0.061)
LD × YSM	-0.000	-0.001	0.001
	(0.002)	(0.002)	(0.002)
Years since migration	0.014^{**}	0.015^{**}	0.014^{**}
	(0.005)	(0.005)	(0.005)
Adjusted R^2	0.281	0.281	0.285
F Statistic	17.46	17.75	17.51
Observations	4,032	4,032	4,032

*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 0 to 3 such that higher values indicate a higher level of language proficiency.*

Supplementary Appendix

Table A5: OLS Results – Model 1 & 2 ACS Sample

	ASJP		WALS		TREE		SCORE	
	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE
Linguistic distance	−0.241***	−0.180***	−0.103***	−0.038***	−0.107***	−0.086***	−0.124***	−0.075***
	(0.004)	(0.004)	(0.003)	(0.004)	(0.002)	(0.002)	(0.002)	(0.003)
Years since migration	0.014***	0.015***	0.014***	0.014***	0.014***	0.014***	0.014***	0.014***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
LD × YSM	–	−0.004***	–	−0.005***	–	−0.001***	–	−0.003***
		(0.000)		(0.000)		(0.000)		(0.000)
Age at entry	−0.018***	−0.018***	−0.018***	−0.018***	−0.018***	−0.018***	−0.018***	−0.018***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	0.096***	0.096***	0.097***	0.097***	0.097***	0.097***	0.097***	0.097***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	−0.026***	−0.028***	−0.024***	−0.026***	−0.027***	−0.027***	−0.028***	−0.028***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Married	0.106***	0.106***	0.105***	0.102***	0.107***	0.107***	0.108***	0.106***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Migrant stock (% of population)	−0.037***	−0.037***	−0.036***	−0.036***	−0.039***	−0.039***	−0.040***	−0.040***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Geographic distance (in 100 km)	0.011***	0.011***	0.011***	0.011***	0.010***	0.010***	0.008***	0.008***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Genetic distance	−0.024***	−0.023***	−0.026***	−0.025***	−0.023***	−0.023***	−0.031***	−0.031***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.516***	0.504***	0.686***	0.655***	0.595***	0.595***	1.014***	1.008***
	(0.016)	(0.016)	(0.017)	(0.017)	(0.015)	(0.015)	(0.013)	(0.013)
Region dummies	yes	yes	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	yes	yes	yes	yes
Adjusted R^2	0.406	0.407	0.404	0.405	0.405	0.405	0.409	0.410
F Statistic	16,051.0	14,965.9	15,351.8	14,489.8	16,102.4	15,076.1	15,706.5	14,881.2
Observations	514,874	514,874	514,874	514,874	514,874	514,874	514,874	514,874

*Notes: – Significant at: *** 0.1% level; ** 1% level; * 5% level; † 10% level. – Robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 0 to 3 such that higher values indicate a higher level of language proficiency.*

Table A6: Ordered Logit Results – Model 1 & 2 ACS Sample

	ASJP		WALS		TREE		SCORE	
	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE
Linguistic distance	−1.035***	−0.719***	−0.339***	−0.198***	−0.533***	−0.374***	−0.344***	−0.234***
	(0.017)	(0.017)	(0.009)	(0.010)	(0.011)	(0.011)	(0.004)	(0.006)
Years since migration	0.036***	0.039***	0.035***	0.034***	0.036***	0.035***	0.035***	0.035***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
LD × YSM	–	−0.028***	–	−0.011***	–	−0.016***	–	−0.008***
		(0.001)		(0.000)		(0.001)		(0.000)
Age at entry	−0.044***	−0.045***	−0.044***	−0.045***	−0.044***	−0.045***	−0.045***	−0.045***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	0.227***	0.227***	0.228***	0.228***	0.228***	0.228***	0.231***	0.231***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Female	−0.074***	−0.078***	−0.070***	−0.073***	−0.073***	−0.076***	−0.080***	−0.079***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Married	0.237***	0.231***	0.234***	0.227***	0.240***	0.233***	0.245***	0.240***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Migrant stock (% of population)	−0.072***	−0.074***	−0.066***	−0.067***	−0.079***	−0.081***	−0.078***	−0.079***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Geographic distance (in 100 km)	0.028***	0.029***	0.026***	0.027***	0.024***	0.025***	0.018***	0.018***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Genetic distance	−0.047***	−0.042***	−0.068***	−0.064***	−0.028***	−0.021***	−0.087***	−0.087***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
Region dummies	yes	yes	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	yes	yes	yes	yes
Threshold 1	0.516***	0.602***	−0.061	−0.003	0.398***	0.453***	−1.200***	−1.186***
	(0.041)	(0.041)	(0.043)	(0.044)	(0.040)	(0.041)	(0.038)	(0.038)
Threshold 2	2.604***	2.700***	2.023***	2.092***	2.484***	2.548***	0.890***	0.909***
	(0.041)	(0.041)	(0.044)	(0.044)	(0.041)	(0.041)	(0.038)	(0.038)
Threshold 3	4.184***	4.283***	3.593***	3.664***	4.059***	4.126***	2.476***	2.496***
	(0.041)	(0.042)	(0.044)	(0.044)	(0.041)	(0.042)	(0.039)	(0.039)
Pseudo-R^2	0.194	0.195	0.190	0.191	0.193	0.194	0.194	0.195
Wald χ^2	126,505.3	125,775.2	129,153.7	128,591.9	125,523.4	124,729.3	129,363.1	129,271.9
Log-likelihood	-13,780,358	-13,758,860	-13,841,570	-13,824,642	-13,795,170	-13,778,354	-13,771,129	-13,761,387
Observations	514,874	514,874	514,874	514,874	514,874	514,874	514,874	514,874

*Notes: – Significant at: *** 0.1% level; ** 1% level; * 5% level; † 10% level. – Robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 0 to 3 such that higher values indicate a higher level of language proficiency.*

Table A7: Ordered Logit Marginal Effects – Model 1 ACS Sample

	ASJP				WALS			
	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE
Linguistic distance	0.063^{***}	0.175^{***}	-0.034^{***}	-0.204^{***}	0.021^{***}	0.057^{***}	-0.012^{***}	-0.066^{***}
	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.002)	(0.000)	(0.002)
Years since migration	-0.002^{***}	-0.006^{***}	0.001^{***}	0.007^{***}	-0.002^{***}	-0.006^{***}	0.001^{***}	0.007^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age at entry	0.003^{***}	0.008^{***}	-0.001^{***}	-0.009^{***}	0.003^{***}	0.008^{***}	-0.002^{***}	-0.009^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	-0.014^{***}	-0.038^{***}	0.008^{***}	0.045^{***}	-0.014^{***}	-0.039^{***}	0.008^{***}	0.045^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	0.004^{***}	0.012^{***}	-0.002^{***}	-0.014^{***}	0.004^{***}	0.012^{***}	-0.002^{***}	-0.014^{***}
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Married	-0.014^{***}	-0.040^{***}	0.008^{***}	0.047^{***}	-0.014^{***}	-0.040^{***}	0.008^{***}	0.046^{***}
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Migrant stock (% of population)	0.004^{***}	0.012^{***}	-0.002^{***}	-0.014^{***}	0.004^{***}	0.011^{***}	-0.002^{***}	-0.013^{***}
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
Geographic distance (in 100 km)	-0.002^{***}	-0.005^{***}	0.001^{***}	0.005^{***}	-0.002^{***}	-0.004^{***}	0.001^{***}	0.005^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Genetic distance	0.003^{***}	0.008^{***}	-0.002^{***}	-0.009^{***}	0.004^{***}	0.011^{***}	-0.002^{***}	-0.013^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
LD × YSM	no	no	no	no	no	no	no	no
Region dummies	yes	yes	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	yes	yes	yes	yes

	TREE				SCORE			
	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE
Linguistic distance	0.032^{***}	0.090^{***}	-0.018^{***}	-0.105^{***}	0.021^{***}	0.058^{***}	-0.012^{***}	-0.067^{***}
	(0.001)	(0.002)	(0.000)	(0.002)	(0.000)	(0.001)	(0.000)	(0.001)
Years since migration	-0.002^{***}	-0.006^{***}	0.001^{***}	0.007^{***}	-0.002^{***}	-0.006^{***}	0.001^{***}	0.007^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age at entry	0.003^{***}	0.008^{***}	-0.001^{***}	-0.009^{***}	0.003^{***}	0.008^{***}	-0.002^{***}	-0.009^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	-0.014^{***}	-0.039^{***}	0.008^{***}	0.045^{***}	-0.014^{***}	-0.039^{***}	0.008^{***}	0.045^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	0.004^{***}	0.012^{***}	-0.002^{***}	-0.014^{***}	0.005^{***}	0.014^{***}	-0.003^{***}	-0.016^{***}
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Married	-0.015^{***}	-0.041^{***}	0.008^{***}	0.047^{***}	-0.015^{***}	-0.042^{***}	0.009^{***}	0.048^{***}
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Migrant stock (% of population)	0.005^{***}	0.013^{***}	-0.003^{***}	-0.016^{***}	0.005^{***}	0.013^{***}	-0.003^{***}	-0.015^{***}
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
Geographic distance (in 100 km)	-0.001^{***}	-0.004^{***}	0.001^{***}	0.005^{***}	-0.001^{***}	-0.003^{***}	0.001^{***}	0.004^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Genetic distance	0.002^{***}	0.005^{***}	-0.001^{***}	-0.006^{***}	0.005^{***}	0.015^{***}	-0.003^{***}	-0.017^{***}
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
LD × YSM	no	no	no	no	no	no	no	no
Region dummies	yes	yes	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	yes	yes	yes	yes

Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 0 to 3 such that higher values indicate a higher level of language proficiency. – Marginal effects are reported at the mean of the covariates vector.

Table A8: Ordered Logit Marginal Effects – Model 2 ACS Sample

	ASJP				WALS			
	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE
Linguistic distance	0.066^{***}	0.187^{***}	-0.036^{***}	-0.217^{***}	0.022^{***}	0.060^{***}	-0.012^{***}	-0.069^{***}
	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.002)	(0.000)	(0.002)
Years since migration	-0.002^{***}	-0.006^{***}	0.001^{***}	0.007^{***}	-0.002^{***}	-0.006^{***}	0.001^{***}	0.007^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age at entry	0.003^{***}	0.008^{***}	-0.001^{***}	-0.009^{***}	0.003^{***}	0.008^{***}	-0.002^{***}	-0.009^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	-0.014^{***}	-0.039^{***}	0.007^{***}	0.045^{***}	-0.014^{***}	-0.039^{***}	0.008^{***}	0.045^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	0.005^{***}	0.013^{***}	-0.003^{***}	-0.015^{***}	0.004^{***}	0.012^{***}	-0.003^{***}	-0.014^{***}
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Married	-0.014^{***}	-0.039^{***}	0.007^{***}	0.045^{***}	-0.014^{***}	-0.039^{***}	0.008^{***}	0.045^{***}
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Migrant stock (% of population)	0.004^{***}	0.013^{***}	-0.002^{***}	-0.015^{***}	0.004^{***}	0.011^{***}	-0.002^{***}	-0.013^{***}
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
Geographic distance (in 100 km)	-0.002^{***}	-0.005^{***}	0.001^{***}	0.006^{***}	-0.002^{***}	-0.005^{***}	0.001^{***}	0.005^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Genetic distance	0.002^{***}	0.007^{***}	-0.001^{***}	-0.008^{***}	0.004^{***}	0.011^{***}	-0.002^{***}	-0.013^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
LD × YSM	yes	yes	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	yes	yes	yes	yes

	TREE				SCORE			
	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE
Linguistic distance	0.035^{***}	0.100^{***}	-0.019^{***}	-0.116^{***}	0.021^{***}	0.058^{***}	-0.012^{***}	-0.066^{***}
	(0.001)	(0.002)	(0.000)	(0.002)	(0.000)	(0.001)	(0.000)	(0.001)
Years since migration	-0.002^{***}	-0.006^{***}	0.001^{***}	0.007^{***}	-0.002^{***}	-0.006^{***}	0.001^{***}	0.007^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Age at entry	0.003^{***}	0.008^{***}	-0.001^{***}	-0.009^{***}	0.003^{***}	0.008^{***}	-0.002^{***}	-0.009^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of education	-0.014^{***}	-0.039^{***}	0.007^{***}	0.045^{***}	-0.014^{***}	-0.039^{***}	0.008^{***}	0.045^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female	0.005^{***}	0.013^{***}	-0.002^{***}	-0.015^{***}	0.005^{***}	0.013^{***}	-0.003^{***}	-0.015^{***}
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Married	-0.014^{***}	-0.040^{***}	0.007^{***}	0.046^{***}	-0.015^{***}	-0.041^{***}	0.008^{***}	0.047^{***}
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)
Migrant stock (% of population)	0.005^{***}	0.014^{***}	-0.003^{***}	-0.016^{***}	0.005^{***}	0.013^{***}	-0.003^{***}	-0.015^{***}
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)
Geographic distance (in 100 km)	-0.001^{***}	-0.004^{***}	0.001^{***}	0.005^{***}	-0.001^{***}	-0.003^{***}	0.001^{***}	0.004^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Genetic distance	0.001^{***}	0.004^{***}	-0.001^{***}	-0.004^{***}	0.005^{***}	0.015^{***}	-0.003^{***}	-0.017^{***}
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
LD × YSM	yes	yes	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	yes	yes	yes	yes

Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Robust standard errors are reported in parentheses. – The dependent variable is defined on a scale of 0 to 3 such that higher values indicate a higher level of language proficiency. – Marginal effects are reported at the mean of the covariates vector.

Table A9: OLS Results – Model 1 & 2 SOEP Sample

	ASJP		WALS		TREE	
	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE
Linguistic distance	-0.079^{**}	-0.121^{*}	-0.160^{**}	-0.187^{**}	-0.117^{**}	-0.182^{***}
	(0.030)	(0.050)	(0.051)	(0.059)	(0.037)	(0.054)
Years since migration	0.017^{***}	0.017^{***}	0.017^{***}	0.017^{***}	0.016^{***}	0.017^{***}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
LD × YSM	–	0.002	–	0.002	–	0.003
		(0.002)		(0.002)		(0.002)
Age at entry	-0.018^{***}	-0.018^{***}	-0.020^{***}	-0.020^{***}	-0.018^{***}	-0.019^{***}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Years of education	0.103^{***}	0.103^{***}	0.102^{***}	0.101^{***}	0.100^{***}	0.099^{***}
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Female	$-0.081^{\dagger}$	$-0.080^{\dagger}$	$-0.076^{\dagger}$	$-0.076^{\dagger}$	$-0.081^{\dagger}$	$-0.079^{\dagger}$
	(0.047)	(0.047)	(0.046)	(0.046)	(0.047)	(0.047)
Married	-0.179^{**}	-0.178^{**}	-0.198^{***}	-0.196^{***}	-0.184^{**}	-0.182^{**}
	(0.059)	(0.059)	(0.057)	(0.057)	(0.058)	(0.059)
Children in the HH. (Ref.= 0)						
One child	−0.035	−0.035	−0.036	−0.040	−0.029	−0.030
	(0.056)	(0.055)	(0.055)	(0.055)	(0.055)	(0.055)
Two children	−0.036	−0.037	−0.030	−0.031	−0.038	−0.039
	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)
Three or more children	$-0.157^{\dagger}$	$-0.154^{\dagger}$	-0.173^{*}	-0.171^{*}	-0.160^{*}	$-0.156^{\dagger}$
	(0.080)	(0.080)	(0.080)	(0.080)	(0.081)	(0.081)
Native German partner/spouse	0.297^{***}	0.299^{***}	0.312^{***}	0.314^{***}	0.286^{***}	0.293^{***}
	(0.069)	(0.069)	(0.066)	(0.067)	(0.068)	(0.068)
Family abroad	−0.073	−0.070	−0.051	−0.049	−0.075	−0.073
	(0.064)	(0.064)	(0.061)	(0.061)	(0.063)	(0.063)
Proficiency home language	0.279^{***}	0.279^{***}	0.294^{***}	0.293^{***}	0.277^{***}	0.278^{***}
	(0.059)	(0.059)	(0.058)	(0.058)	(0.058)	(0.059)
Desired stay (years)	0.004	0.005	0.003	0.003	0.004	0.004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Migrant stock (% of population)	-0.208^{***}	-0.210^{***}	-0.155^{**}	-0.158^{***}	-0.202^{***}	-0.209^{***}
	(0.053)	(0.052)	(0.048)	(0.047)	(0.050)	(0.049)
Neighboring country	0.286^{**}	0.283^{**}	0.238^{*}	0.237^{*}	0.266^{*}	0.253^{*}
	(0.107)	(0.107)	(0.108)	(0.108)	(0.107)	(0.107)
Geographic distance (in 100 km)	0.002	0.002	0.002	0.002	−0.000	−0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Genetic distance	0.041^{*}	0.041^{**}	0.042^{**}	0.043^{**}	0.048^{**}	0.050^{**}
	(0.016)	(0.016)	(0.015)	(0.015)	(0.016)	(0.016)
Constant	0.352	0.357	0.309	0.337	0.432	0.444
	(0.286)	(0.284)	(0.273)	(0.276)	(0.275)	(0.273)
Region dummies	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	yes	yes
Adjusted R^2	0.373	0.373	0.375	0.375	0.376	0.377
F Statistic	34.55	34.41	35.80	34.69	34.59	34.59
Observations	5,803	5,803	5,803	5,803	5,803	5,803

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 0 to 3 such that higher values indicate a higher level of language proficiency.

Table A10: Ordered Logit Results – Model 1 & 2 SOEP Sample

	ASJP		WALS		TREE	
	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE
Linguistic distance	-0.210^{*}	-0.331^{*}	-0.478^{***}	-0.546^{***}	-0.335^{**}	-0.525^{**}
	(0.093)	(0.165)	(0.139)	(0.162)	(0.118)	(0.174)
Years since migration	0.043^{***}	0.043^{***}	0.043^{***}	0.043^{***}	0.040^{***}	0.041^{***}
	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)	(0.010)
LD × YSM	–	0.006	–	0.004	–	0.008
		(0.007)		(0.006)		(0.007)
Age at entry	-0.047^{***}	-0.048^{***}	-0.056^{***}	-0.056^{***}	-0.050^{***}	-0.051^{***}
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Years of education	0.271^{***}	0.270^{***}	0.267^{***}	0.266^{***}	0.264^{***}	0.262^{***}
	(0.029)	(0.029)	(0.028)	(0.028)	(0.029)	(0.028)
Female	−0.197	−0.197	−0.179	−0.180	−0.199	−0.197
	(0.124)	(0.124)	(0.121)	(0.120)	(0.124)	(0.124)
Married	-0.443^{**}	-0.444^{**}	-0.495^{***}	-0.490^{***}	-0.454^{**}	-0.454^{**}
	(0.154)	(0.153)	(0.146)	(0.147)	(0.151)	(0.151)
Children in the HH. (Ref.= 0)						
One child	−0.096	−0.098	−0.095	−0.104	−0.085	−0.089
	(0.149)	(0.149)	(0.147)	(0.146)	(0.147)	(0.147)
Two children	−0.125	−0.128	−0.109	−0.112	−0.133	−0.137
	(0.173)	(0.173)	(0.174)	(0.174)	(0.173)	(0.173)
Three or more children	$-0.422^{\dagger}$	$-0.413^{\dagger}$	-0.471^{*}	-0.467^{*}	-0.432^{*}	$-0.420^{\dagger}$
	(0.217)	(0.217)	(0.218)	(0.218)	(0.218)	(0.218)
Native German partner/spouse	0.765^{***}	0.773^{***}	0.807^{***}	0.815^{***}	0.735^{***}	0.753^{***}
	(0.189)	(0.189)	(0.183)	(0.184)	(0.189)	(0.187)
Family abroad	−0.180	−0.174	−0.120	−0.117	−0.191	−0.190
	(0.171)	(0.172)	(0.163)	(0.163)	(0.171)	(0.172)
Proficiency home language	0.730^{***}	0.731^{***}	0.780^{***}	0.775^{***}	0.730^{***}	0.734^{***}
	(0.160)	(0.160)	(0.159)	(0.159)	(0.159)	(0.160)
Desired stay (years)	0.010	0.011	0.005	0.005	0.010	0.010
	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)
Migrant stock (% of population)	-0.513^{***}	-0.518^{***}	-0.371^{**}	-0.379^{**}	-0.503^{***}	-0.522^{***}
	(0.151)	(0.149)	(0.131)	(0.128)	(0.141)	(0.136)
Neighboring country	0.750^{*}	0.743^{*}	0.601^{*}	0.600^{*}	0.685^{*}	0.647^{*}
	(0.302)	(0.301)	(0.300)	(0.300)	(0.305)	(0.300)
Geographic distance (in 100 km)	0.009	0.009	0.008	0.007	0.002	0.001
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Genetic distance	0.102^{*}	0.103^{*}	0.110^{**}	0.111^{**}	0.124^{**}	0.129^{**}
	(0.046)	(0.046)	(0.041)	(0.041)	(0.047)	(0.046)
Region dummies	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	yes	yes
Threshold 1	0.763	0.745	0.880	0.791	0.562	0.502
	(0.794)	(0.791)	(0.737)	(0.753)	(0.761)	(0.760)
Threshold 2	2.893^{***}	2.879^{***}	3.018^{***}	2.933^{***}	2.693^{***}	2.640^{***}
	(0.804)	(0.800)	(0.746)	(0.762)	(0.770)	(0.769)
Threshold 3	5.270^{***}	5.254^{***}	5.402^{***}	5.315^{***}	5.088^{***}	5.034^{***}
	(0.806)	(0.803)	(0.753)	(0.767)	(0.775)	(0.775)
Pseudo-R^2	0.180	0.180	0.182	0.182	0.182	0.183
Wald χ^2	572.3	584.4	577.0	576.5	569.7	584.5
Log-likelihood	-22,897,234	-22,887,498	-22,830,347	-22,824,832	-22,833,466	-22,812,425
Observations	5,803	5,803	5,803	5,803	5,803	5,803

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 0 to 3 such that higher values indicate a higher level of language proficiency.

Table A11: Ordered Logit Marginal Effects – Model 1 SOEP Sample

	ASJP				WALS				TREE			
	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE
Linguistic distance	0.017^{*}	0.035^{*}	-0.033^{*}	-0.019^{*}	0.038^{***}	0.080^{***}	-0.074^{***}	-0.044^{***}	0.027^{**}	0.056^{**}	-0.052^{**}	-0.031^{**}
	(0.008)	(0.016)	(0.015)	(0.009)	(0.011)	(0.024)	(0.022)	(0.013)	(0.009)	(0.020)	(0.018)	(0.011)
Years since migration	-0.003^{***}	-0.007^{***}	0.007^{***}	0.004^{***}	-0.003^{***}	-0.007^{***}	0.007^{***}	0.004^{***}	-0.003^{***}	-0.007^{***}	0.006^{***}	0.004^{***}
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
Age at entry	0.004^{***}	0.008^{***}	-0.007^{***}	-0.004^{***}	0.004^{***}	0.009^{***}	-0.009^{***}	-0.005^{***}	0.004^{***}	0.008^{***}	-0.008^{***}	-0.005^{***}
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
Years of education	-0.022^{***}	-0.045^{***}	0.042^{***}	0.025^{***}	-0.021^{***}	-0.045^{***}	0.042^{***}	0.025^{***}	-0.021^{***}	-0.044^{***}	0.041^{***}	0.024^{***}
	(0.003)	(0.005)	(0.005)	(0.003)	(0.002)	(0.005)	(0.005)	(0.003)	(0.002)	(0.005)	(0.005)	(0.003)
Female	0.016	0.033	−0.030	−0.018	0.014	0.030	−0.028	−0.016	0.016	0.033	−0.031	−0.018
	(0.010)	(0.021)	(0.019)	(0.012)	(0.010)	(0.020)	(0.019)	(0.011)	(0.010)	(0.021)	(0.019)	(0.011)
Married	0.035^{**}	0.074^{**}	-0.069^{**}	-0.041^{**}	0.039^{***}	0.083^{***}	-0.077^{***}	-0.046^{***}	0.036^{**}	0.076^{**}	-0.071^{**}	-0.042^{**}
	(0.013)	(0.026)	(0.024)	(0.014)	(0.012)	(0.025)	(0.023)	(0.014)	(0.012)	(0.025)	(0.024)	(0.014)
Children in the HH. (Ref.= 0)												
One child	0.008	0.016	−0.015	−0.009	0.008	0.016	−0.015	−0.009	0.007	0.014	−0.013	−0.008
	(0.012)	(0.025)	(0.023)	(0.014)	(0.012)	(0.025)	(0.023)	(0.014)	(0.012)	(0.025)	(0.023)	(0.013)
Two children	0.010	0.021	−0.019	−0.012	0.009	0.018	−0.017	−0.010	0.011	0.022	−0.021	−0.012
	(0.014)	(0.029)	(0.027)	(0.016)	(0.014)	(0.029)	(0.027)	(0.016)	(0.014)	(0.029)	(0.027)	(0.016)
Three or more children	$0.034^{\dagger}$	$0.071^{\dagger}$	$-0.065^{\dagger}$	$-0.039^{\dagger}$	0.037^{*}	0.079^{*}	-0.073^{*}	-0.043^{*}	0.034^{*}	0.072^{*}	-0.067^{*}	-0.040^{*}
	(0.017)	(0.037)	(0.034)	(0.020)	(0.017)	(0.037)	(0.034)	(0.020)	(0.017)	(0.037)	(0.034)	(0.020)
Native German partner/spouse	-0.061^{***}	-0.128^{***}	0.119^{***}	0.071^{***}	-0.064^{***}	-0.135^{***}	0.125^{***}	0.074^{***}	-0.058^{***}	-0.123^{***}	0.114^{***}	0.067^{***}
	(0.015)	(0.032)	(0.029)	(0.019)	(0.015)	(0.031)	(0.029)	(0.018)	(0.015)	(0.032)	(0.029)	(0.018)
Family abroad	0.014	0.030	−0.028	−0.017	0.010	0.020	−0.019	−0.011	0.015	0.032	−0.030	−0.018
	(0.014)	(0.029)	(0.026)	(0.016)	(0.013)	(0.027)	(0.025)	(0.015)	(0.013)	(0.029)	(0.026)	(0.016)
Proficiency home language	-0.058^{***}	-0.122^{***}	0.113^{***}	0.067^{***}	-0.062^{***}	-0.131^{***}	0.121^{***}	0.072^{***}	-0.058^{***}	-0.122^{***}	0.113^{***}	0.067^{***}
	(0.013)	(0.027)	(0.026)	(0.015)	(0.013)	(0.027)	(0.026)	(0.015)	(0.013)	(0.027)	(0.026)	(0.015)
Desired stay (years)	−0.001	−0.002	0.002	0.001	−0.000	−0.001	0.001	0.000	−0.001	−0.002	0.001	0.001
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Migrant stock (% of population)	0.041^{***}	0.086^{***}	-0.080^{***}	-0.047^{***}	0.029^{**}	0.062^{**}	-0.058^{**}	-0.034^{**}	0.040^{***}	0.084^{***}	-0.078^{***}	-0.046^{***}
	(0.012)	(0.025)	(0.023)	(0.014)	(0.011)	(0.022)	(0.020)	(0.012)	(0.011)	(0.024)	(0.022)	(0.013)
Neighboring country	-0.060^{*}	-0.125^{*}	0.116^{*}	0.069^{*}	-0.048^{*}	-0.101^{*}	0.093^{*}	0.055^{*}	-0.054^{*}	-0.115^{*}	0.106^{*}	0.063^{*}
	(0.024)	(0.051)	(0.048)	(0.028)	(0.024)	(0.051)	(0.047)	(0.027)	(0.024)	(0.052)	(0.048)	(0.028)
Geographic distance (in 100 km)	−0.001	−0.002	0.001	0.001	−0.001	−0.001	0.001	0.001	−0.000	−0.000	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Genetic distance	-0.008^{*}	-0.017^{*}	0.016^{*}	0.009^{*}	-0.009^{**}	-0.018^{**}	0.017^{**}	0.010^{**}	-0.010^{**}	-0.021^{**}	0.019^{**}	0.011^{**}
	(0.004)	(0.008)	(0.007)	(0.004)	(0.003)	(0.007)	(0.006)	(0.004)	(0.004)	(0.008)	(0.007)	(0.004)
LD × YSM	no	no	no	no	no	no	no	no	no	no	no	no
Region dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	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. – The dependent variable is defined on a scale of 0 to 3 such that higher values indicate a higher level of language proficiency. – Marginal effects are reported at the mean of the covariates vector.

Table A12: Ordered Logit Marginal Effects – Model 2 SOEP Sample

	ASJP				WALS				TREE			
	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE	Bad ME/StdE	Not bad ME/StdE	Well ME/StdE	Very well ME/StdE
Linguistic distance	0.016^{*}	0.034^{*}	-0.032^{*}	-0.019^{*}	0.036^{***}	0.076^{***}	-0.070^{***}	-0.041^{***}	0.027^{**}	0.057^{**}	-0.053^{**}	-0.031^{**}
	(0.008)	(0.016)	(0.015)	(0.009)	(0.012)	(0.026)	(0.023)	(0.014)	(0.009)	(0.020)	(0.018)	(0.011)
Years since migration	-0.003^{***}	-0.007^{***}	0.007^{***}	0.004^{***}	-0.003^{***}	-0.007^{***}	0.007^{***}	0.004^{***}	-0.003^{***}	-0.007^{***}	0.006^{***}	0.004^{***}
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
Age at entry	0.004^{***}	0.008^{***}	-0.007^{***}	-0.004^{***}	0.004^{***}	0.009^{***}	-0.009^{***}	-0.005^{***}	0.004^{***}	0.009^{***}	-0.008^{***}	-0.005^{***}
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
Years of education	-0.022^{***}	-0.045^{***}	0.042^{***}	0.025^{***}	-0.021^{***}	-0.045^{***}	0.041^{***}	0.024^{***}	-0.021^{***}	-0.044^{***}	0.041^{***}	0.024^{***}
	(0.002)	(0.005)	(0.005)	(0.003)	(0.002)	(0.005)	(0.005)	(0.003)	(0.002)	(0.005)	(0.005)	(0.003)
Female	0.016	0.033	−0.031	−0.018	0.014	0.030	−0.028	−0.017	0.016	0.033	−0.031	−0.018
	(0.010)	(0.021)	(0.019)	(0.012)	(0.010)	(0.020)	(0.019)	(0.011)	(0.010)	(0.021)	(0.019)	(0.011)
Married	0.035^{**}	0.074^{**}	-0.069^{**}	-0.041^{**}	0.039^{***}	0.082^{***}	-0.076^{***}	-0.045^{***}	0.036^{**}	0.076^{**}	-0.071^{**}	-0.042^{**}
	(0.013)	(0.026)	(0.024)	(0.015)	(0.012)	(0.025)	(0.023)	(0.014)	(0.012)	(0.025)	(0.024)	(0.014)
Children in the HH. (Ref.= 0)												
One child	0.008	0.016	−0.015	−0.009	0.008	0.017	−0.016	−0.010	0.007	0.015	−0.014	−0.008
	(0.012)	(0.025)	(0.023)	(0.014)	(0.012)	(0.025)	(0.023)	(0.013)	(0.012)	(0.025)	(0.023)	(0.013)
Two children	0.010	0.021	−0.020	−0.012	0.009	0.019	−0.017	−0.010	0.011	0.023	−0.021	−0.013
	(0.014)	(0.029)	(0.027)	(0.016)	(0.014)	(0.029)	(0.027)	(0.016)	(0.014)	(0.029)	(0.027)	(0.016)
Three or more children	$0.033^{\dagger}$	$0.069^{\dagger}$	$-0.064^{\dagger}$	$-0.038^{\dagger}$	0.037^{*}	0.078^{*}	-0.073^{*}	-0.043^{*}	$0.033^{\dagger}$	$0.071^{\dagger}$	$-0.065^{\dagger}$	$-0.038^{\dagger}$
	(0.017)	(0.037)	(0.033)	(0.020)	(0.017)	(0.037)	(0.034)	(0.020)	(0.017)	(0.037)	(0.034)	(0.020)
Native German partner/spouse	-0.062^{***}	-0.129^{***}	0.119^{***}	0.071^{***}	-0.065^{***}	-0.137^{***}	0.127^{***}	0.075^{***}	-0.060^{***}	-0.126^{***}	0.117^{***}	0.069^{***}
	(0.015)	(0.032)	(0.029)	(0.019)	(0.015)	(0.031)	(0.029)	(0.018)	(0.015)	(0.032)	(0.029)	(0.018)
Family abroad	0.014	0.029	−0.027	−0.016	0.009	0.020	−0.018	−0.011	0.015	0.032	−0.030	−0.017
	(0.014)	(0.029)	(0.026)	(0.016)	(0.013)	(0.027)	(0.025)	(0.015)	(0.013)	(0.029)	(0.026)	(0.016)
Proficiency home language	-0.058^{***}	-0.122^{***}	0.113^{***}	0.068^{***}	-0.062^{***}	-0.130^{***}	0.121^{***}	0.071^{***}	-0.058^{***}	-0.123^{***}	0.114^{***}	0.067^{***}
	(0.013)	(0.027)	(0.026)	(0.015)	(0.013)	(0.027)	(0.026)	(0.015)	(0.013)	(0.027)	(0.026)	(0.015)
Desired stay (years)	−0.001	−0.002	0.002	0.001	−0.000	−0.001	0.001	0.000	−0.001	−0.002	0.002	0.001
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Migrant stock (% of population)	0.041^{***}	0.087^{***}	-0.080^{***}	-0.048^{***}	0.030^{**}	0.064^{**}	-0.059^{**}	-0.035^{**}	0.041^{***}	0.088^{***}	-0.081^{***}	-0.048^{***}
	(0.012)	(0.025)	(0.023)	(0.014)	(0.010)	(0.022)	(0.020)	(0.012)	(0.011)	(0.023)	(0.022)	(0.013)
Neighboring country	-0.059^{*}	-0.124^{*}	0.115^{*}	0.069^{*}	-0.048^{*}	-0.101^{*}	0.093^{*}	0.055^{*}	-0.051^{*}	-0.109^{*}	0.101^{*}	0.059^{*}
	(0.024)	(0.051)	(0.047)	(0.028)	(0.024)	(0.051)	(0.047)	(0.027)	(0.024)	(0.051)	(0.047)	(0.027)
Geographic distance (in 100 km)	−0.001	−0.001	0.001	0.001	−0.001	−0.001	0.001	0.001	−0.000	−0.000	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
Genetic distance	-0.008^{*}	-0.017^{*}	0.016^{*}	0.010^{*}	-0.009^{**}	-0.019^{**}	0.017^{**}	0.010^{**}	-0.010^{**}	-0.022^{**}	0.020^{**}	0.012^{**}
	(0.004)	(0.008)	(0.007)	(0.004)	(0.003)	(0.007)	(0.006)	(0.004)	(0.004)	(0.008)	(0.007)	(0.004)
LD × YSM	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Region dummies	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Year fixed effects	yes	yes	yes	yes	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. – The dependent variable is defined on a scale of 0 to 3 such that higher values indicate a higher level of language proficiency. – Marginal effects are reported at the mean of the covariates vector.