Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Kathleen Donovan April 14, 2015

Globalization, Education, and Linguistics: Determinants of English Proficiency in EU Countries

by

Kathleen Donovan

Jasminka Ninkovic
Adviser

Department of Economics

Jasminka Ninkovic
Adviser

Hugo Mialon

Committee Member

Hiram Maxim

Committee Member

April 14, 2015

Globalization, Education, and Linguistics: Determinants of English Proficiency in EU Countries

By

Kathleen Donovan

Jasminka Ninkovic
Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Economics

2015

Abstract

Globalization, Education, and Linguistics: Determinants of English Proficiency in EU Countries

By Kathleen Donovan

English's role as a global lingua franca in business and politics has made improving English skills a primary concern for many countries throughout the world, particularly in the European Union. The existing research and literature focusing on the countrywide determinants of English proficiency is quite limited. The literature also leaves room for improvement in terms of methodology and variables used. This paper uses a panel data approach to consider both time-series and cross-country variation in English proficiency in the EU; by lagging potentially endogenous variables when possible, this project attempts to find a causal relationship between globalization indicators, education variables, linguistic variables, and English proficiency.

The results show strong evidence to suggest that government expenditures on education, similarity of the native language to English, and the economic importance of the native language all have explanatory power over the differences in English proficiency between EU countries. For the globalization variables, although strong correlation with English proficiency exists, not much evidence to suggest a causal relationship is found. However, there is preliminary evidence to suggest that one globalization variable, the strength of the country's tourism industry, might have explanatory power over variation in English proficiency within countries.

In the future, this project could be improved upon when there are more Eurobarometer survey data, thus increasing the sample size. The methodology could also be improved upon with the use of instrumental variables to better account for endogeneity bias.

Globalization, Education, and Linguistics: Determinants of English Proficiency in EU Countries

By

Kathleen Donovan

Jasminka Ninkovic
Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Economics

2015

Acknowledgements

I would like to thank my adviser, Dr. Ninkovic, for her valuable insight and encouragement. I am also grateful to her for, in the beginning of my college career, helping to spark my interest in Economics and particularly in issues dealing with globalization and international economics.

I also would like to thank my committee members Dr. Maxim and Dr. Mialon for offering me their unique perspectives and ideas, as well as constructive criticism, throughout the research and writing process.

Table of Contents

List of figures

List of tables

I) Introduction

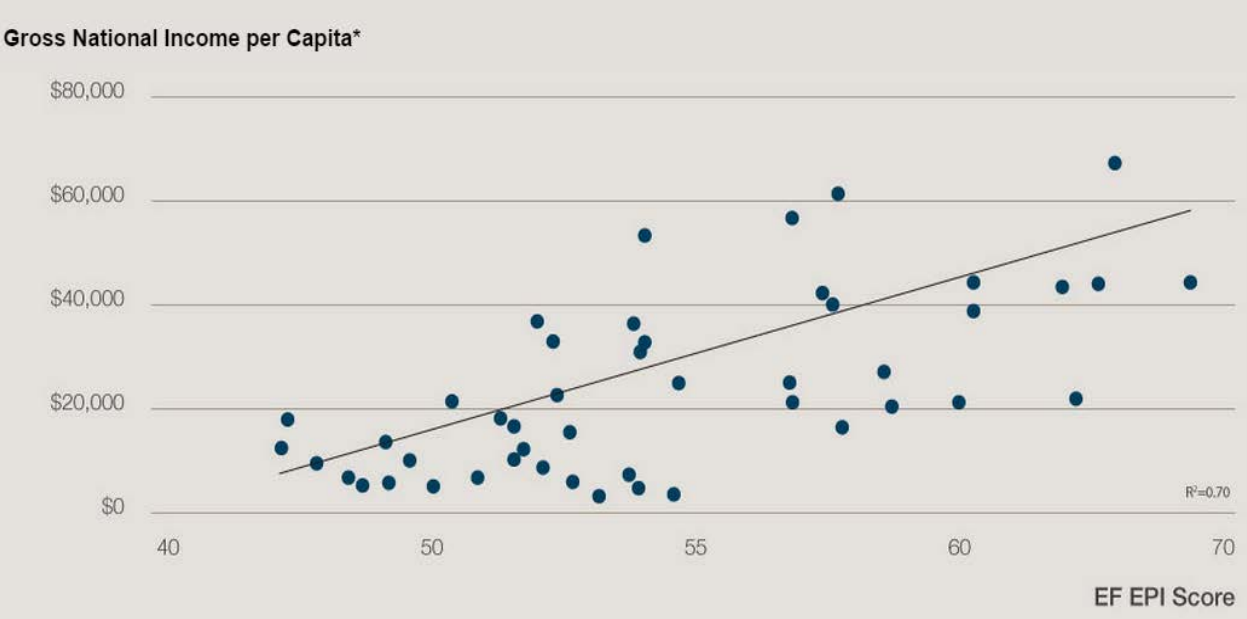
As the modern global economy becomes more interconnected, with most countries' economies depending heavily on international trade and foreign investment, it is no surprise that a single global *lingua franca*—a language used for communication between two peoples who share no other common language—is emerging as the language of international business and political relations. Currently, this language is English, which has at least 900 million native and non-native speakers, making it the second-most widely spoken language in the world[1] (*Ethnologue: Languages of the World,* Summer Institute of Linguistics 2013). Many international organizations use English for their global communications. There are over 50 countries and over 20 "sovereign and non-sovereign entities" in the world that have English as an official language (World Factbook), and some of these countries have some of the world's largest economies. Indeed, when one adds up the GDPs of all of these countries (plus the United States, which has no official language but clearly speaks English), one finds that their combined GDP consists of 36.3% of the entire global economy. This means that *over a third* of the world economy is "dominated by English" (Hjorth-Andersen 2006), and this is not including the countless business interactions conducted in English between non-native English speaking parties.

Consequently, since there are so many English-speaking people in the world, and the English-speaking countries are together an extremely significant economic force, learning English as a foreign language should theoretically be correlated with monetary benefits for both individuals (income) and entire countries (GDP). This relationship between English skills and monetary benefits has been empirically examined by a significant amount of economic research. Several studies have shown that, even in non-English speaking countries, proficiency in English is correlated with high returns to the individual in terms of income, job opportunities, and

[1] Note: This number is disputed, as it is very difficult to estimate. Other sources, such as the British Council and *Encyclopaedia Britannica* estimate higher numbers, from 1 billion to 1.5 billion English speakers (Hammond, 2012)

promotions (see Education First, Fidrmuc, and Azam et al.). Some studies have also shown that countries with higher levels of English proficiency are more globally competitive, having a higher GNI per capita (see Figure 1), and attracting more foreign investment and fostering international trade relations (see *The Economist,* Fidrmuc, and Education First). However, in none of these studies is the direction of causality clear—current research still has not reached a consensus as to whether English proficiency causes all these benefits, or if these benefits (income, GDP, trade, FDI) are what incentivize people to learn English. Also, a surprisingly small amount of research has been dedicated to finding what the country-level determinants of a non-English-speaking country's English proficiency are. Therefore, the purpose of this project is to do just that: to attempt to determine what country-level factors have an influence on a nation's overall English proficiency.

Figure 1: GNI/capita correlates with English proficiency (Education First, 2012)



Better English and higher income go hand in hand

The interaction between English proficiency and gross national income per capita is a virtuous cycle, with improving English skills driving up salaries, which in turn give governments and individuals more money to invest in English training. The relationship also applies anecdotally on a micro level, where improved English skills allow individuals to apply for better jobs and raise their standards of living.

*Source: United Nations, GNI per capita PPP($), 2012

Why the EU?

Although the rather small amount of literature that has examined the determinants of English proficiency considered countries all around the world (prior literature will be discussed in the following section), this project will focus on one region: the European Union. Focusing on one region reduces the number of possible confounding variables and therefore will give these results more validity, even if the results are admittedly only valid for the EU. The EU offers compelling research possibilities because the importance of having a single common language is likely especially significant for Europe compared to the rest of the world. This is due to the EU being an economic union with the largest combined GDP in the world, but with cooperation dependent on compromises between very culturally and linguistically diverse countries. The EU, realizing the importance of mutual intelligibility across nations, has implemented an official language policy, which is that students must learn two foreign languages in school (Hjorth-Andersen 2006). The European Union has made several official statements acknowledging the importance of learning foreign languages to foster economic and political cooperation (European Commission, "Language Policy" 2014), and it is also their view that "multilingualism... is an important element in Europe's competitiveness" (European Commission, "Language Policy" 2014). While English is technically not required to be one of those two foreign languages that students learn, in 2012, *96.7%* of all upper secondary-level pupils were studying English*[2] (Eurostat, news release 2014). This push for EU citizens to learn foreign languages combined with the popularity of English as one of the languages learned has caused a universal upward trend in English proficiency over time, in almost all EU countries (see graphs in Appendix 2). This paper examines the determinants of this trend.

[2] This number is excluding the United Kingdom and Ireland, where the vast majority of people speak English natively, so citizens of these countries would learn other languages as their foreign languages. For the purposes of this paper, any time the EU is mentioned, the UK and Ireland are excluded.

An additional observation that makes the EU a particularly interesting region to examine in this project is that, despite this centralized policy to increase foreign language skills across all countries of the EU, and the fact that nearly 100% of students learn English at some point, there is a huge disparity in how well people in different EU countries actually learn English and maintain that knowledge. This is evident from a variety of sources: According to the European Commission's 2012 special Eurobarometer survey, the proportion of people in EU countries who claim to speak English "well enough...to have a conversation" ranges from 20% in Hungary and 22% in Spain, to 86% in Sweden and 90% in the Netherlands (see Figure 3 in Appendix 3) (Eurobarometer 77.1, number 386, 2012). Tests administered by Education First as well as TOEFL (Test of English as a Foreign Language) show similarly wide ranges in English proficiency within the EU in 2012 (Education First, "Regional Spotlight Europe" 2013; TOEFL, "Test and Score Data Summary" 2013). So, this paper also aims to examine what causes these cross-country differences in English proficiency in the EU.

II) Literature review

As previously mentioned, there is some research showing economic returns to English skills for individuals and for countries, as well as research quantifying the global economic importance of the English language. (Education First; Azam et al.; Fidrmuc; Hjorth-Andersen). However, the literature on the country-level determinants of a non-English-speaking country's overall English proficiency is quite small, excluding the brief consideration that perhaps the correlation between Trade, Investment, and GDP is due to those things incentivizing English-learning rather than vice versa (see Fidrmuc, Education First, and *The Economist*). To date, there appear to be only two papers that address this precise topic: "Linguistic And Nonlinguistic Factors Determining Proficiency Of English As A Foreign Language: A Cross-Country Analysis" (Kim and Lee, 2010) and "Economic, statistical, and linguistic factors affecting success on the test of English as a foreign language (TOEFL)" (Snow 1998). Snow's paper regresses countries' average 1996 TOEFL scores on

GNP/capita, Export share of GNP, Percentage taking the TOEFL (PERPOP), and Relatedness of a TOEFL candidate's native language to English (IHRE). IHRE is an integer from 1-7 that ranks language families (i.e. Germanic, Romantic, Slavic, non-Indo-European) from most to least similar to English. The purpose of PERPOP is to adjust for some countries' TOEFL scores being very positively skewed, due to only a very small percentage of their population taking the exam. This assumes that only people who feel reasonably confident in their English abilities will pay to take the exam. Snow finds that all four of these variables are statistically significant in explaining variation in countries' average TOEFL scores at the 5% significance level, and they all have the expected coefficients (positive for all except negative for PERPOP).

Kim and Lee's paper acknowledges Snow's findings and expands upon them, including many more variables and looking at two different years, producing much more nuanced results. They still use country average TOEFL scores as the proxy for English proficiency. They look at 1997/1998 and 2004/2005, performing a cross-sectional OLS analysis for each year. Kim and Lee's independent variables are as follows, where necessary descriptions of the variables are in brackets [ ], and the variable names are in parentheses ( ):

- Linguistic factors: Historical affinity to English Value, or HAV [similar to Snow's IHRE index; an index denoting how similar a country's official language is to English] (*AFFINITY*), Word order [a binary variable indicating whether or not the country's official language has the same semantic word order as English] (*WordOrder*), and Linguistic fractionalization [a variable to reflect if a large amount of different languages are spoken within the same country; "the probability that two randomly selected individuals from a population belong to different language groups"] (*LingFrac*)
- Non-linguistic factors: GDP/capita, Expected years of schooling (*SCHOOLING*), Export share of GDP (*EXPORT*), Number of inbound international travelers (*INBOUND*), Number of Internet users (*INTERNET*), the KOF Globalization Index (*GLOBAL*) [to be used in place of *EXPORT*, *INBOUND*, and *INTERNET*--the "openness" indicators--in some regression models], Colonial experience by English-speaking countries (*ExColony*), and the Percentage of population taking the TOEFL (*PerPop*).

Kim and Lee found the linguistic factors to be statistically significant at the 5% level for both years. The HAV Index ("*AFFINITY*") is collinear with WordOrder, so they do not use them in regressions together. For the non-linguistic factors, they found that EXPORT and ExColony were not statistically significant in either year. More variables were significant in 1997 than in 2004. For 1997, the non-linguistic factors that Kim and Lee found to be statistically significant at the 5% level were SCHOOLING, INBOUND, INTERNET, and PerPop. The KOF Globalization Index is significant when the three "openness" indicators are not included. For 2004, the KOF Index remains significant at the 5% level when it is used, but many other variables become insignificant. Only SCHOOLING remains significant at the 5% level. Kim and Lee acknowledge that "the reason why the openness measures have a weaker relation with proficiency in English requires further study," but they explain that the reduced significance of INTERNET on TOEFL scores likely has to do with the fact that, in 1997, over 75% of webpages were in English, but by 2004 that number had fallen to 50% (Kim and Lee, 2010). They also find that GDP/capita is very highly correlated with SCHOOLING, INTERNET, and INBOUND. As a consequence, for the 1998 results, when GDP is used in regressions with those variables, it is not significant. It only becomes significant when those variables are omitted. In 2005, GDP is significant when used in regressions with those three variables, but it is not robust and when significant it has the opposite sign as expected. When those three variables are omitted, it becomes significant with the expected positive sign. Overall, both the Snow and Kim and Lee studies find strong evidence to suggest that linguistic factors are important in explaining variation in TOEFL scores across countries, regardless of the year. Kim and Lee also find that the duration of compulsory education is significant regardless of year. However, the openness indicators--INTERNET, EXPORT, and INBOUND--have a weaker relationship with English proficiency in 2005 than in 1998.

III) Data and Hypotheses

Dependent variable

Perhaps the biggest key difference between this project and the literature is the different method of measuring the dependent variable, English proficiency. Snow as well as Kim and Lee use TOEFL scores as the approximant for countries' proficiency in English. However, there are many problems with this metric as a proxy for English proficiency. The TOEFL is a standardized exam. A major purpose for taking it is because it is an entrance requirement for many colleges and universities in English-speaking countries. Universities use it and similar English exams to gauge potential international students' English abilities. In addition, the TOEFL costs money to take: between $160 and $250 (ETS, 2014). Because there is a significant cost to taking the test, and because it is taken typically by people who wish to go to university in an English-speaking nation, there is a clear self-selection bias associated with the TOEFL. Only people who already feel relatively confident in their English abilities will take the test: People who are very poor in their English skills wouldn't be considering university in an English-speaking nation and would be deterred by the cost of the exam. Granted, the previous researchers knew very well this bias associated with using the TOEFL to estimate English proficiency. This was the reasoning behind including the PerPop variable, which adjusts slightly for the bias, but does not completely eliminate it. Also, for Snow and Kim and Lee, there was no better alternative to the TOEFL, because they were interested in so many countries all over the world, and the TOEFL is the most widely-used English exam in the world. This project, however, is focusing only on one region, Europe. For the European Union, there is a more representative measure of countries' English proficiencies: selected Eurobarometer surveys that asked respondents about language proficiency.

The Eurobarometers are surveys conducted by the European Commission that ask citizens in each of the EU countries about various social and political issues. The Eurobarometers 34.0, 41.0, 50.0, 52.0, 55.1, 64.3, and 77.1--conducted in the years 1990, 1994, 1998, 1999, 2001, 2006, and 2012 respectively--queried survey respondents about language proficiency, asking them what languages they spoke "well enough in order to be able to have a conversation," excluding their mothertongue. (Eurobarometer 64.3, 77.1). The exact wording of the question varies only slightly from year to year. The European Commission has released the original raw datasets for all of these years. In these datasets, whether or not a person mentioned *English* as one of the languages they speak at a conversational level is its own binary variable, so from that it is possible to calculate the proportion of survey respondents for each country who speak English. However, one can also simply use the published summary statistics in the surveys' codebooks and press releases for their findings, which have already done that calculation and state the proportion of people in each EU country who speak English, for each of these years (see again Figure 3 in Appendix 3: the proportions from the 2012 survey). These calculated proportions are what serve as the dependent variable in this study.

The Eurobarometer surveys are preferable over TOEFL scores, because they yield much more representative estimates for the EU countries' English proficiencies: Whereas national average TOEFL scores carry a self-selection bias toward people who speak English relatively well, the Eurobarometers are much more likely to capture a broader cross-section of the population without self-selection bias. The surveys ask about a huge number of different topics; the questions about language proficiency are not the main purpose of the surveys. Also, the European Commission invites EU citizens to participate in Eurobarometer surveys through a multi-stage, stratified random sampling procedure to ensure a representative random sample of the population. The researchers stratify "by the distribution of the … population in terms of metropolitan, urban and rural areas, i.e. proportional to the population size … and to the population density" to ensure

"total coverage of the country." These strata are designated "primary sampling units (PSU)," and it is ensured that "each of the administrative regions in every country" are represented by at least one PSU. The researchers then randomly select "a cluster of addresses" from each PSU. From those addresses, one single member of each household is randomly selected to participate in the Eurobarometer survey, which is a face-to-face in-home interview, conducted in the participant's native language. ("Sampling and Fieldwork," GESIS Eurobarometer Data Service, 2013). Following data collection, to further ensure a representative sample, statistical weights are applied to each individual observation to account for any under- or over-represented demographic groups in each country, after comparing the demographic makeup of the sample with the true demographics of the universe (the population). Universe demographic information comes from "National Research Institutes and/or by EUROSTAT" ("Weighting overview: Standard & Special EB," GESIS, 2014). In most countries, around 1,000 people complete the survey, but some smaller countries such as Luxembourg have smaller samples of around 500 ("Countries, regions, population coverage" GESIS, 2015). Thus, with large, random, representative samples, one may assume that the Eurobarometers present unbiased estimates of the true proportions of people in each EU country who speak English at each time period. The next best alternative data source on English proficiency is Education First's English Proficiency Index (EPI). Education First obtains this number for many countries in the world through a free online test. This has some of the same problems as using the TOEFL does, but not as severely, since Education First's test is free and not used for entrance to universities. However, the Eurobarometers are still better for the purposes of this project. The EPI does not have data for before 2007 while the Eurobarometers have data from 1990-2012, and this project partly concerns itself with countries' changes in English proficiency over time.

As comprehensive and representative as the Eurobarometer surveys are, the data for the purposes of this project still has significant limitations. It is important to note that data on English proficiency for all the current EU countries are not available for every year. Since the

Eurobarometers have always only surveyed people in EU member states, there is only data on countries' English proficiency for those years *following* their addition to the EU. The biggest restriction, therefore, is on the 10 countries that joined the EU in 2004 (Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, and Slovenia). For those countries there is only data on English proficiency for two years, 2006 and 2012. Because of this, the dataset is unbalanced. Despite the limited information on some of the countries, the Eurobarometer surveys are still the best and most representative source of information on English proficiency over time in the EU.

An interesting question was to see if the Eurobarometers would yield similar results as the TOEFL scores when using similar methodology as Kim and Lee. As shall be explained in the following section, the independent variables and empirical strategy used in this project is quite different from what the literature has done. But as a supplement to this project, one may also roughly approximate the literature's methodology and use the literature's independent variables, but using the percentages given by the Eurobarometers as the dependent variable instead of TOEFL scores. This approach yields some results that are similar to the literature's findings, and some results that are quite different. Because this analysis is supplemental and is not part of the primary project, the results table and interpretation may be found in Appendix 4, rather than in the main body of the paper.

Independent variables

The independent variables in this project differ in many respects from those used in the literature, but there are some similarities. It is hypothesized that the following factors have explanatory power on an EU country's English proficiency:

- Quality of general education
    - Government expenditures on education as a percentage of GDP
- Emphasis on foreign-language (FL) or English-language education
    - Percentage of students learning English in earlier stages of education (primary and lower secondary)
- Economic globalization
    - The Openness Index: Total value of imports plus exports, divided by GDP
    - Foreign Direct Investment, inward and outward
    - Amount of tourism
- Social globalization
    - Internet users per 100 people
- How similar a country's official language is to English
- The relative economic importance of a country's official language

Brief descriptions of the variables that make up these factors, the variable names, and their data sources are in the following table. Henceforth, variables will be referred to as their variable name, or as a non-ambiguous abbreviation of that name, e.g. "Expend" for "Expend_AllLevels." For summary statistics of these variables, see Table 2 in Appendix 3.

Table 1: Descriptions of variables

Variable name	*Variable description and code, if applicable*	*Data source*
PercEng *(Dependent variable)*	**Percentage of population who speaks English "well enough to have a conversation"**	**Eurobarometer surveys**
Expend_AllLevels	"Total public expenditure as a % of GDP. All levels [SE.XPD.TOTL.P]"	World Bank, Education Statistics
PercLearnEng_Primary	"L03_16 - Pupils learning English at ISCED level 1 (aka 'primary' education) as a percentage of the total pupils at this level"	Eurostat
PercLearnEng_LowerSec	"L03_1 - Pupils learning English at ISCED level 2 (aka 'middle' or 'lower secondary' education) as a percentage of the total pupils at this level"	Eurostat
Internet	"Internet users (per 100 people) [IT.NET.USER.P2]"	World Bank, Development Indicators
Trade_PercGDP	"Trade (% of GDP) [NE.TRD.GNFS.ZS]" - also known as the Openness Index	World Bank, Development Indicators
FDI_InStock	"Inward FDI stocks in % of GDP (tec00105)" - total value as percentage of GDP of FDI stocks held by the reporting country, from all other countries in the world	Eurostat
FDI_OutStock	"Outward FDI stocks in % of GDP (tec00106)" – total value as % of GDP of FDI stocks from the reporting country, held by all other countries in the world	Eurostat
TouristsperPerson	"International tourism, number of arrivals [ST.INT.ARVL]", divided by population, "Population, total [SP.POP.TOTL]"	World Bank, Development Indicators
KOF	KOF Globalization Index – to be used in place of other globalization variables in some regression models, as literature has previously done	ETH: Swiss Federal Institute of Technology Zürich
IHRE_Index	The "Index of Historical Relatedness to English" – a number from 1-7 indicating how closely related a language is to English, historically	Snow
RelativeImpLang_Bin	Binary variable: 1 = this country speaks a relatively "economically important" language according to Andersen's 2006 analysis; 0 = this country does not speak an economically important language	Hjorth-Andersen, original variable

The reason why it is hypothesized that English proficiency is determined by the quality of education and the amount that education focuses on foreign language (FL) learning is because education is the primary method "*how"* people learn an FL, including English, since the EU requires students to learn two FLs.

For this project's measure of the overall quality of education, I choose to use government expenditures on education over Kim and Lee's "expected number of years" because in this data specific to the EU, the variable "Duration of compulsory education" has relatively little sample variation, compared to the variable Expend_AllLevels which has more variance. Since EU countries tend to be at a relatively similar level of development, the duration of compulsory education has not changed much in recent times, nor does it differ much between countries. Government expenditures also may have a closer connection with the *quality* of education than the number of years in school, which is more of a measure of *quantity* of education, which is how Kim & Lee interpret it. Using a measure for *quality* of education is more appropriate than a measure of *quantity,* because in general, better-funded and higher-quality school systems produce more successful students. Another choice to note is that I have chosen to use expenditures on education for *all levels* rather than using the two, more specific measures of expenditures for the *primary* and *secondary* levels. This is because I am working with a relatively small sample, so one variable that roughly combines two variables and captures both of their effects is preferable to using both variables.

Kim and Lee included no education variable that is specific to English, simply because they had no data for this, since they analyzed so many different countries. However, for only the EU, there is data for this: Eurostat's percentage of students learning English at the Primary and Lower Secondary levels of education. I choose to include these variables to avoid omitted variable bias--the number of students learning English should certainly have explanatory power for a country's English proficiency, and it may be correlated with education expenditures. This is because, in almost all countries, the number of students studying English at the "upper secondary" (ISCED level 3) level of education is very close to 100% for all countries for all time periods for which data are available (Eurostat, variable L03_2). If a country's students are learning English at this time *in addition to* at the primary or lower secondary levels, there may be more education spending.

Next, economic globalization indicators may have explanatory power over English proficiency because English is used so often as a lingua franca in international business and political interactions. Therefore, countries which are more involved in the global economy are likely to have more jobs within their own borders which require or encourage English skills. This is similar to Kim and Lee's justification for including a measure of international trade: "Residents of a country with a relatively high level of international trade are expected to have more opportunities to be exposed to English as well as greater incentive to learn English as a second language, as good English skills would give them enhanced opportunities to get higher paying jobs." (Kim & Lee, pg 2351). Economic globalization factors can be thus thought of as a driving force for "*why"* people learn English.

I use the Openness Index, or total trade as a percentage of GDP, instead of Kim and Lee's "EXPORTS" because the Openness Index is a broader measure of economic globalization than only considering "export share of GDP." Imports are an important a part of economic globalization—in particular, how large the influence of international trade is on domestic activities--and thus should have explanatory power over English proficiency for the same reason as exports. Additionally, imports are likely to be correlated with exports, so excluding imports would overstate the effect of exports through omitted variable bias. Also, when Kim and Lee use the KOF Globalization Index in place of their other "openness indicators" in certain regressions, this Index is actually using the Openness Index, not simply exports ("Variables and Weights," ETH). To explore the logic more specifically behind the relationship between English proficiency and the Trade aspect of economic globalization, when a business in a country chooses to export goods to a foreign country, these interactions may be conducted in English. Similarly, if this country accepts imports from a foreign exporting country, these interactions are also likely to be conducted in English.

Kim and Lee did not include any measures of Foreign Direct Investment (FDI), but FDI is quite an important part of economic globalization and may cause omitted variable bias if not

included. As mentioned in the introduction of this paper, previous studies found a correlation between English proficiency and the amount of foreign investment entering a country (see Fidrmuc and Education First). The specific logic behind the relationship with FDI and English proficiency is similar to that for Trade: Any international business interactions will likely be conducted in English. Other sources consider FDI to be a part of economic globalization: Eurostat groups their FDI variables with International Trade variables together as "economic globalization indicators" (Eurostat), and the KOF Globalization Index, which again is used in Kim and Lee's paper, actually weights FDI *more* heavily than it does Trade ("Variables and Weights," ETH). Therefore, if any measure of international trade is included in my regressions, I cannot omit FDI, because the two have similar justifications for their relationship with English, and they are equally important measures of economic globalization. There are two variables for FDI in this project: total *out*ward stocks and total *in*ward stocks of FDI, as percentages of GDP. Outward stocks is the total amount of FDI stocks held by all other countries in the world, from the reporting country. Inward stocks is essentially the inverse: total amount of FDI stocks held by the reporting country, from all other countries in the world. Both values are represented as percentages of the country's GDP to control for the size of the economy.

Another one of Kim and Lee's globalization or "openness" indicators is their variable "INBOUND," which is defined as the total "number of inbound international travelers." The corresponding variable in this project, TouristsperPerson, is almost the same, except I choose to divide the number of incoming tourists by the total population of the country, in order to control for the size of the country and therefore make cross-country comparisons more meaningful. Simply using the number of inbound travelers does not tell us much, since the same amount of tourists make a larger impact on the economy of a smaller country than they do on a larger country. The logic behind this variable's relationship with English proficiency specifically is as follows: The larger the ratio of tourists to citizens, the more important the tourism industry will be to the

country's economy, thus creating more jobs in the tourism and hospitality industries which require knowledge of many foreign languages, including English. An exception to note is that this variable may be a poor predictor of English skills in one country in particular, Austria, where the largest source country for tourists is Germany, which shares a common official language with Austria. This project ignores the possibility of Austria therefore potentially being an outlier, but future research may choose to omit Austria in certain regressions or find a way to take into account the native language or country of origin of tourists.

The KOF Globalization Index is used in the literature as a single variable to replace all "openness indicators" in certain regressions. The expectation is that KOF captures all the effects of all my globalization indicators. It is useful in this project in particular because it increases the sample size and the number of observations per panel group, and it decreases the number of variables.

Social globalization represents less formal methods of *"how"* people learn English, which is why it is expected to have an effect on English proficiency. English is unique in that, along with being used often in political and business interactions, it is also heard throughout the world via pop culture, in television, music, and movies. Additionally, the Internet is primarily in English, although this number has been shrinking with time as Internet access becomes more and more common throughout the world. In 1998 over 75% of webpages were in English (Ho, 2003), but in 2013 55% of all webpages were in English (w3techs.com). Regardless, this is still a majority, and no other language comes close to having this strong of a presence online. Thus, I hypothesize that many social globalization factors, including Internet, and the popularity of English-language television, films, and music, affect how much a person is exposed to English outside of school, serving as another method *"how"* a person learns English.

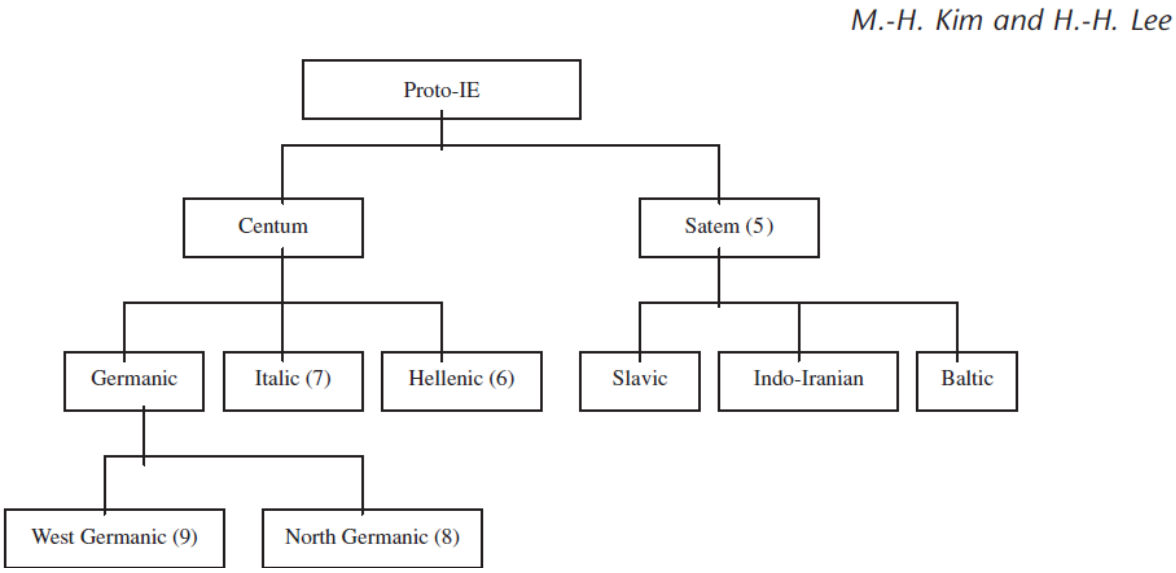
The only measure of social globalization for which data are available is the number of Internet users in the country. This project uses "Internet" for the same reasons as Kim and Lee use

it: to proxy people's everyday casual exposure to English. I also use the exact same measure of Internet users as they did: the World Bank's "number of Internet users per 100 people."

The first linguistic factor is how similar a country's official language(s) is/are to English. This is hypothesized to have an effect on English proficiency, because the more similar a foreign language is to one's native language, the less time may be required to become proficient in that language. If a particular language can be learned in a relatively short time, people may be both more motivated to learn it and more likely to continue learning it until proficiency is reached. Both of the most relevant papers in the literature, Kim and Lee's and Snow's, include linguistic variables that indicate how similar languages are to English.

Linguistic similarity to English shall be represented by Snow's IHRE (Index of Historical Relatedness to English). Values of the IHRE go from 1-7, with higher numbers indicating higher "historical relatedness" to English. Historical relatedness is directly related to how similar the modern language is to English: The farther back in history the nearest common ancestor between a language and English is, the less similar that language will be to English today. Consequently, people in countries which have higher IHRE values most likely speak a native language that is more similar to English, making it easier (in terms of a time commitment) for them to learn English. The values of the IHRE come from historical linguistic research which, by "observing systematic correspondences of sounds between significant numbers of words with related meanings in two or more different languages," traces linguistic development over time and determines common ancestors of languages (Cardona et al., 1970; Lehman, 1974). This development can be visualized as a "tree" of Indo-European (IE) languages, with language families making up "branches" of the tree. The full IE language tree is quite complicated and goes beyond the scope of this paper. Here is a simplified IE language tree Kim and Lee include in their paper to explain their HAV Index, but it is also just as useful for understanding the concept of "historical relatedness" behind the IHRE:

Figure 2: Simplified Indo-European (IE) language tree (Kim and Lee, 2010)



The following list explains the meanings of the seven values of the IHRE and which countries in this project are associated with each value:

1. Non-Indo-European: Turkey, Finland, and Hungary
2. Other Indo-European languages: Greece and Cyprus
3. Slavic: Bulgaria, Croatia, Czech Republic, Estonia, Latvia, Lithuania, Poland, Slovak Republic, and Slovenia
4. Romance: Belgium, France, Italy, Portugal, Romania, and Spain
5. Scandinavian (aka North Germanic): Denmark and Sweden
6. German: Germany, Austria, and Luxembourg
7. Dutch[3]: The Netherlands

I choose the IHRE over Kim and Lee's HAV Index because the IHRE is easier to interpret, and the limitations of the IHRE are not as problematic for this dataset as they were for Kim and Lee's. The primary advantage of the HAV Index is that there is a gap in the index from 1-4 to account for the extremely substantial difference between non-Indo-European languages and Indo-European (IE) languages, compared to the differences between IE languages. The IHRE is simply 1-7, with no gaps. This can be slightly problematic, because it assumes that the linguistic difference between, for

[3] German and Dutch are both part of the West Germanic language family. For this reason, in Kim and Lee's HAV Index they are assigned the same number. However, the West Germanic family is further divided into many subfamilies (see Figure 4 in Appendix 3), and with this greater specificity Dutch is technically more closely historically related to English. Snow finds this distinction important and therefore assigns German and Dutch different values in his IHRE Index.

example, Chinese and Greek (1 and 2 in the IHRE), is the same as the difference between Swedish and German (5 and 6 in the IHRE). It is advantageous that the HAV Index takes care of this limitation. However, this advantage is also a disadvantage in that it is no longer valid to interpret the HAV Index's coefficient in regressions in terms of marginal increases in the HAV Index's value. Using the IHRE therefore allows one to interpret the Index more meaningfully. The disadvantage of the IHRE does not pose much of a problem for this sample, because only three out of the 26 countries in my dataset speak non-IE languages. Also, two of these countries (Hungary and Turkey) together only contribute three observations of PercEng (out of 109 total observations). Therefore, using the IHRE is not as problematic for this analysis as it was for Kim and Lee's, which had many non-IE languages in the sample.

In my sample, the IHRE Index is sufficient for proxying linguistic similarity to English without having to include other linguistic variables. Kim and Lee also include a binary variable, WordOrder, denoting whether or not the language has the same syntactical structure as English. But in this sample which has only EU countries, there is only one country (Turkey) which speaks a language that does not share English's SOV (Subject-Object-Verb) sentence structure. Kim and Lee also use a binary variable called "ExColony" which indicates if the country was once a colony of the British empire or not. But again, in this project this variable would only be relevant for one country, Cyprus. So, in the interest of not having too many variables, both WordOrder and ExColony are omitted.

The second linguistic factor refers to how "economically important" a country's official language is. Why this concept could have explanatory power over English proficiency requires some explanation, as it has not been explored in the literature. One may assume that a driving economic force in "*why"* people learn English is that it gives them higher earnings power and opportunities for career advancement. If this is the case, then one would be less motivated to learn English if their own native language also granted plenty of opportunities by being relatively useful

in international economic and political activities. *The Economist* addresses this idea in an article about English skills around the world, saying, "the larger the number of speakers of a country's main language, the worse that country tends to be at English." For example, this could be "why Spain was the worst performer [in English abilities] in western Europe...Spanish's role as an international language in a big region [Latin America] dampens incentives to learn English" (*The Economist,* 2011). So, I hypothesize that the relative importance of a country's official language affects how motivated the people of that country are to learn English.

The binary variable *RelativeImpLang_Bin* somewhat crudely approximates the relative "economic importance" of a country's official language. Figuring out how to measure this concept is rather difficult. Some researchers suggest using the number of worldwide speakers a language has (*The Economist,* Hjorth-Andersen). But in his paper titled, "The Relative Economic Importance the European Languages," Hjorth-Andersen with the University of Copenhagen asserts that "the number of people speaking a language may often be quite irrelevant from an economic point of view," and points out that "in economics, we are used to measuring the importance of different countries by their GNP rather than the number of inhabitants" (Hjorth-Andersen). Thus, he suggests another approach for measuring the economic importance of languages: adding up the GNPs of all countries that speak a certain language and considering that number to be roughly the size of the "economy" of that language. He also divides that number by the total GNP of the world to determine the percentage of the global economy that that language "dominates." Additionally, he repeats this approach with GDP in Purchasing Power Parity to better allow comparisons across all countries in the world. With the PPP approach, Hjorth-Andersen finds that, in 2006, English was by far the most economically important language in Europe and in the world, with its countries' GDPs adding up to 36.3% of the global economy. Next, Spanish was roughly tied in economic importance with German and French, with each of those languages "dominating" around 5.5% of the global

economy. Next was Italian and Portuguese, controlling 3% each, and last was Dutch, with about 2% (Hjorth-Andersen, 2006).

The next challenge was determining how to use Hjorth-Andersen's 2006 findings in my research, which spans the years 1990-2012. His hierarchy of which European languages are the most important relative to each other is only valid for 2006, since that is when he did the calculations. While it is probably safe to assume that English has remained the most economically important language since 1990, one cannot assume that the hierarchy of German/French/Spanish, followed by Italian/Portuguese, followed by Dutch has remained the same. There was not enough time over the course of this project to repeat Hjorth-Andersen's methodology for every other year for which I have data on English proficiency, so it was decided to create a variable that would roughly extend Andersen's 2006 findings to all time periods in this sample. This is the binary variable *RelativeImpLang_Bin.* For this variable, I assign a country the value of 1 (for all years -- the variable is time-invariant) if at least one of its official languages is one of these six most economically important European languages, according to Hjorth-Andersen's 2006 analysis. All other countries receive a value of 0. With this method, those countries with a value of 1 can be thought of as speaking languages which are economically important in the world, affording them career opportunities which less economically important languages do not offer. One can assume that while the order of which European languages were more important than others may have changed since 1990, the fact that these six languages are "relatively important" in the world probably has not.

It is hypothesized that the signs of the coefficients for all these variables in this analysis will be positive, except for *RelativeImpLang.* This coefficient should be negative, since if people's native language provides them with ample opportunities, then there is less motivation to learn English for professional reasons.

The literature includes GDP/capita to control for the wealth of economies, but I omit such a measure because the literature finds that it may not be necessary. As mentioned in the Literature Review section, in their 1997/1998 regressions, Kim and Lee find that GDP/capita is only significant when other variables are omitted. They explain this by pointing out the strong correlation between GDP/capita and SCHOOLING, INTERNET, and INBOUND (their education variable and some globalization variables) (Kim and Lee, 2010). The collinearity implies that the wealth of the country is already controlled for in those variables, and therefore a measure of GDP is not necessary.

IV) Empirical Strategy

Because I wish to look at differences in English proficiency both *within* countries over time (time series), as well as *between* countries at a given time period (cross-sectional), this project lends itself quite well to a panel or longitudinal data approach.

Fixed-(within) effects (time-series analysis)

The fixed-effects (fe) model for panel data is quite useful in examining the upward trend in English proficiency over time that one may observe across almost all EU countries (see graphs in Appendix 2). This is because the fe model explains variation over time, within countries. One may assume that time-invariant, country-specific effects are important for this project and need to be controlled for, which implies that a fixed effects model is preferable over a random effects model. An example of a country-specific fixed effect could be linguistic factors. I verify this assumption statistically by performing the Hausman test comparing a fixed-effects regression model with a random-effects model, in which I reject the null hypothesis at P=.05 . [4] The fe model controls for time-invariant effects in part of the error term, which means in this model one does not have to

[4] Hausman test: Ho: Country-specific errors (θi) are not correlated with regressors (use random effects). Ha: Country-specific errors (θi) are correlated with regressors, i.e. the disturbances are not identically distributed over the panels (use fixed effects).

explicitly account for the linguistic variables. This is because linguistic factors, such as how similar a country's official language is to English, do not change with time. My binary variable approximating which countries speak "economically important" languages, *RelativeImpLang_Bin,* is also time-invariant and so is also omitted from the fixed effects models. Thus, the form of the fixed effects regression model is as follows:

$$PercEng_{it} = \alpha + \beta_1 Expend_{it} + \beta_2 PercLearn\_Primary_{it} + \beta_3 PercLearn\_LowerSec_{it} + \beta_4 Internet_{it} + \beta_5 Trade_{it} + \beta_6 FDI\_InStock_{it} + \beta_7 FDI\_OutStock_{it} + \beta_8 TouristsperPerson_{it} + year + \theta_i + u_{it}$$

for *i* units, *i*=1,...,26, because there are 26 countries in this dataset, measured at times t=1,...,T_i. T_i is the total number of years for which a country has observations for PercEng. T_i is used instead of T because this dataset is unbalanced. T_i ranges from 2-7. α is the constant, u_{it} is the unexplained error term, and Θ_i represents the country-specific time-invariant fixed effects, or in other words, those errors that are correlated with the regressors. *PercEng* represents the proportion of people in country *i* in year *t* who speak English, as given by the Eurobarometer surveys. All independent variables are those that were described in the "Data" section. The linguistic variables are omitted, because they are time-invariant. *Internet, Trade, FDI_InStock, FDI_OutStock,* and *Tourists,* are all globalization indicators and so shall be replaced with the KOF Globalization Index in some regressions, to see if both methods yield similar results, as the literature has done. Finally, dummy variables for years are included (e.g. *yr29*==1 if *year*==2012) to control for any country-invariant time effects that the model does not account for.

It must be now acknowledged that the globalization indicators are likely to be endogenous. As previously mentioned, researchers have found a correlation between English proficiency and some measures of economic globalization such as amount of trade and foreign investment, "though

it's not clear which factor causes which" (R.L.G., *The Economist*). In my analysis, I am attempting to find a causal link from globalization to English proficiency: the more economically globalized a country becomes, the more jobs require or encourage English proficiency; and with the more everyday exposure the people have to English-language media through Internet, television, and films, the better they become at English. However, it could very well be the other way around: When many people speak English well, it becomes easier for companies to find qualified employees to allow them to do business internationally and become more economically globalized. Also, if many people speak English, English-language media including the Internet becomes more popular. It could also be a case of simultaneous causation in which both sides of the equation influence the other. The literature has not addressed this endogeneity problem. One way to account for this bias is by lagging the endogenous variables. Lagged variables may actually be a more accurate way of modeling the relationship between these variables and English proficiency, since a population's overall proficiency in any language would not respond immediately to changes in its determinants, as people need time to learn the language. A country's level of globalization wouldn't be related to its current English proficiency, but rather English proficiency several years later as more people learn the language in response to increased globalization. Lagging the endogenous variables is an imperfect solution, since it causes serial correlation bias and cannot be used in the between-effects model discussed in the following section of this paper. Ideally to solve an endogeneity problem one would use instrumental variables and 2SLS, but I have not found valid instruments for every endogenous variable. The variables that will be lagged are *Internet, Trade, FDI_InStock, FDI_OutStock,* and *Tourists,* as well as *KOF* when it is used. For the individual globalization indicators, the length of the lag that has been chosen is three years. For *KOF,* the lag is one year. Lag length was determined by various statistical tests, which is discussed in detail in Appendix 1.

Other than the endogeneity problem, the primary limitation of the fixed effects model is that its results can only be meaningfully interpreted as explaining variation in English proficiency *within*

countries over time. In order to see how the variables affect differences in English proficiency *between* countries, one must turn to another approach in addition to the fixed effects.

Between -effects (cross-sectional analysis)

The between-effects model allows one to consider differences in English proficiency between countries, differences which can be quite large, as mentioned in the introduction. In contrast with the fixed effects model which is purely time series, the between effects model is purely cross-sectional. This makes this approach using both fixed- and between-effects preferable over random effects, because it allows me to see what affects differences *between* countries in contrast with *within* countries. I cannot make this distinction with random-effects, which is a weighted average of the fixed- and between-effects results. Analyzing both therefore gives me greater specificity.

The between-effects model is also known as "regression on group means" because of the way it is mathematically calculated. This regression equation removes time effects by regressing the group mean y on the group mean x's. Of course, this also implies that, unlike in the fixed effects model, time-invariant effects are not already controlled for, so I must now include the linguistic variables. So, the form of the between effects regression model is as follows:

$$\begin{aligned}\overline{PercEng_i} \;=\; & \alpha + \beta_1\overline{Expend_i} + \beta_2\overline{PercLearn\_Primary_i} \\ & + \beta_3\overline{PercLearn\_LowerSec_i} + \beta_4\overline{Internet_i} + \beta_5\overline{Trade_i} \\ & + \beta_6\overline{FDI\_InStock_i} + \beta_7\overline{FDI\_OutStock_i} + \beta_8\overline{TouristsperPerson_i} \\ & + \beta_9 IHRE\_Index_i + \beta_{10} RelativeImpLang\_Bin_i + \theta_i + \bar{u}_i\end{aligned}$$

again for for *i* units, *i*=1,...,26. There is no *t* or dummy variables for years in this model since the model regresses group mean *PercEng* on country mean independent variables taken from each country's observations from all years. A limitation of the between effects model is it does not allow me to mitigate the endogeneity bias by lagging variables.

V) Results and Discussion

Fixed-effects (within country, time series) – Regression Tables 1 and 2

For analyzing variation in English proficiency over time within countries, the results are mostly inconclusive. In the models without lagged variables, the only variables to be statistically significant at all are TouristsperPerson (at 5% level), and Internet and FDI_OutStock (at 10% level). The statistically significant variables all have their expected positive signs. These results imply that for every unit increase in the number of tourists per citizens, English proficiency rises by 17.24%. This seems like a very strong effect, but recall the summary statistics for TouristsperPerson: the highest value is 2.8, and 90% of the values are below 2. Therefore, an increase of 1 is relatively large. The coefficient on Internet implies that for every single increase in the percentage of people who are Internet users in the country, English proficiency increases by .25%. The coefficient on FDI_OutStock implies that, if total FDI stocks held by foreign countries from the respective country increase by 1% of the respective country's GDP, PercEng increases by .24%. Trade, PercLearnEng_Primary, FDI_InStock, and KOF all have negative coefficients which is the opposite of what was expected, but none of them are statistically significant.

The education variables cause some sampling bias by restricting the sample to later years. No data for the two PercLearn variables are available before 1998, and many countries report data on Expend rather irregularly and/or did not start reporting it until later years, creating many gaps in the data. Therefore, I am interested in seeing how the results change when I remove the education variables, because doing so greatly increases the sample size and the number of observations per country, although it may cause omitted variable bias. However, when I remove the education statistics from the model, the three statistically significant globalization variables (Internet, FDI_OutStock, and TouristsperPerson) actually become *in*significant. Many coefficients also change fairly drastically when a group of variables is removed: Compare the coefficients on

PercLearnEng_LowerSec and Tourists between Models 1, 2, and 3 for the most extreme examples of changing coefficients . This could be a sign that these results may not be very robust. That the globalization variables lose significance when the education variables are omitted, and that the coefficients are volatile, is somewhat difficult to interpret. It could be a result of the fact that certain globalization variables and certain education variables are moderately correlated. Upon examining the correlation matrix for all variables in the model, the following pairs of variables have a correlation of at least .5: Expend and FDI_OutStock; PercLearnEng_LowerSec and Trade; and PercLearnEng_LowerSec and FDI_InStock (see correlation matrix in Appendix 3). This is not high enough to designate a collinearity problem, but it could be high enough to suggest that perhaps the variables interact in such a way that removing some affects the others. In other words, the education variables and the globalization variables may be jointly statistically significant. When I perform a Wald test for joint statistical significance on *PercLearnEng_LowerSec* (the education variable in the first model with the highest t-statistic), *FDI_OutStock, Internet,* and *Tourists* (the three statistically significant variables in model 1), I reject the null hypothesis at the 5% level.[5] However, another possibility for explaining why the globalization variables lose significance when the education variables are removed, is that the model without the education variables is simply more accurate in estimating the true coefficients of the globalization variables, due to having a 46% larger sample size than the model with all variables. If this is the case, then the globalization variables have less explanatory power over PercEng than the model with all variables (Model 1) would imply.

When I lag the globalization variables to decrease endogeneity bias (Regression Table 2), in Model 1, TouristsperPerson remains as significant as it was before and has a similar coefficient. This implies that Tourists is a fairly robust variable that may have explanatory power over English

[5] In a Wald test for joint statistical significance, Ho: $X_1 = 0$ AND $X_2 = 0$ AND $X_3 = 0$; alternatively written: $X_1 = X_2 = X_3 = 0$. Ha: Ho is false; at least one variable's coefficient is not equal to 0.

proficiency, rather than the causality being the other way around. However, one still cannot be completely certain of the direction of causality, due to serial correlation bias. Direction of causality is even less certain for the other globalization variables. When it is lagged, FDI_OutStock loses its significance. Internet also loses its significance in Model 1, but it actually gains significance in the model without the education variables (Model 3), in which it is significant at the 5% level. Again, this is hard to interpret. It could be that the higher significance is a more true representation of L3.Internet's explanatory power over PercEng, because of the larger sample size. Alternatively, it could also be a case of omitted variable bias due to removing the education variables. Internet could be correlated with Expend, although there is not a particularly strong statistical correlation (corr = .3). Regardless of the low statistical correlation, Expend could affect Internet, because if a government spends more on education, school systems are more likely to have better resources to offer supplemental classes that teach children how to use computers and the Internet. Overall, in the fixed-effects model, for all globalization variables except Tourists and perhaps Internet, after lagging the variables I cannot definitively say that the direction of causality is what I expect it to be.

An interesting observation arises when one looks at the year dummy variables in both the lagged and non-lagged models, comparing their significance when different groups of variables are included. In the model with all variables (Model 1) and the model with no education variables (Model 3), almost all year dummies are insignificant. However, in the model in which the globalization variables are removed and replaced with KOF (Model 2), three out of the four included year dummies become significant at the 1% level. Finally, when all variables are removed except for KOF (Model 4), all six of the year dummies are significant. This seems to imply that, although not many variables are statistically significant on their own, they may be jointly significant, in that the year dummies pick up significance when they are removed. In other words, together at least some of these variables do explain some variation in English proficiency over time, because unexplained time effects are not significant when they are included, but they are significant

when the variables are omitted. With this interpretation, the globalization variables appear to have more to do with time variation than the education variables do. I come to this conclusion by comparing the time dummies in Model 2 and Model 3. When education variables are included but globalization variables are not, time effects are significant. Conversely, when globalization variables are included but education variables are not, time effects are not significant. This implies that it is the removal of the globalization variables, not the education variables, that causes time effects to pick up significance. Further support for this claim comes from testing for joint statistical significance with a Wald test. I find that in both the lagged and non-lagged models, the globalization variables together are jointly significant at around the 5% level, while the education variables are quite far from being jointly significant.

Overall it seems that in analyzing time-series, within-country variation in PercEng, the only variables that on their own may have explanatory power over rising English proficiency with time are TouristsperPerson and maybe Internet. FDI_OutStock shows correlation, but because it is never significant when it is lagged, we cannot claim that it causes English proficiency rather than the other way around. Although not many variables are significant on their own, the globalization variables may be jointly significant in explaining English proficiency, but the education variables are not.

Regression Table 1: Fixed effects (within-country) regression results

Dependent var: PercEng	Model 1: All vars	Model 2: Replace Global with KOF	Model 3: No Educ vars	Model 4: only KOF
Expend_AllLevels	0.136 (0.33)	0.057 (0.11)	--	--
PercLearnEng_Primary	-0.028 (0.38)	0.001 (0.02)	--	--
PercLearnEng_LowerSec	0.252 (1.39)	-0.108 (0.57)	--	--
Internet	0.251 (1.85)*	--	0.107 (1.37)	--
FDI_OutStock	0.243 (1.75)*	--	0.010 (0.23)	--
FDI_InStock	-0.133 (0.89)	--	-0.089 (1.47)	--
TouristsperPerson	17.236 (2.75)**	--	2.365 (0.97)	--
Trade_PercGDP	-0.077 (0.58)	--	0.051 (0.91)	--
KOF	--	-0.327 (0.66)	--	-0.169 (1.13)
yr11 (1994)[6]	--	--	--	5.643 (2.64)**
yr15 (1998)	--	-10.712 (2.90)***	-5.344 (1.14)	9.587 (3.73)***
yr16 (1999)	-3.499 (1.55)	-11.278 (3.51)***	-4.573 (1.09)	11.074 (4.14)***
yr18 (2001)	-8.766 (2.49)**	-10.571 (3.57)***	-6.051 (1.75)*	10.706 (3.72)***
yr23 (2006)	-8.641 (1.34)	-1.161 (0.59)	-0.468 (0.30)	20.473 (7.17)***
yr29 (2012)	-14.728 (1.59)	--	--	21.235 (7.25)***
_cons	4.480 (0.23)	82.810 (1.69)	40.908 (6.20)***	44.530 (4.37)***
R^2	0.84	0.57	0.81	0.75
Adjusted R2	0.54	0.13	0.67	0.64
N	54	62	79	109

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

*Regression output given by the statistical software *Stata*, version 13

[6] When a year dummy variable is reported as missing, it is because it must be omitted from that regression due to collinearity.

Regression Table 2: Fixed effects (within-country) regression results, lagged globalization variables

Dependent var: PercEng	Model 1: All vars	Model 2: Replace Global with KOF	Model 3: No Educ vars	Model 4: only KOF
Expend_AllLevels	0.184 (0.32)	0.054 (0.11)	--	--
PercLearnEng_Primary	0.036 (0.57)	-0.009 (0.11)	--	--
PercLearnEng_LowerSec	-0.091 (0.55)	-0.083 (0.44)	--	--
L3.Internet	0.148 (1.61)	--	0.155 (2.43)**	--
L3.FDI_OutStock	0.011 (0.09)	--	-0.043 (1.20)	--
L3.FDI_InStock	-0.038 (0.47)	--	-0.022 (0.38)	--
L3.TouristsperPerson	16.554 (2.50)**	--	1.522 (0.51)	--
L3.Trade_PercGDP	0.199 (1.44)	--	0.070 (1.21)	--
L1.KOF	--	-0.149 (0.31)	--	-0.025 (0.11)
yr15 (1998)	5.945 (0.82)	-10.749 (2.82)***	-1.521 (0.35)	3.267 (1.86)*
yr16 (1999)	5.177 (0.74)	-11.224 (3.45)***	-0.797 (0.19)	4.634 (2.47)**
yr18 (2001)	0.995 (0.18)	-10.586 (3.55)***	-3.291 (0.90)	4.043 (1.82)*
yr23 (2006)	2.369 (0.80)	-1.139 (0.57)	1.131 (0.63)	13.783 (6.42)***
yr29 (2012)	--	--	--	14.479 (6.20)***
_cons	6.821 (0.23)	66.463 (1.43)	37.017 (5.63)***	39.151 (2.33)**
R^2	0.80	0.57	0.78	0.70
Adjusted R2	0.45	0.12	0.61	0.56
N	56	62	78	99

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Between-effects (between countries, cross-sectional) - Regression Table 3

When analyzing differences in English proficiency between countries, regardless of time, I find much more significant results, and most variables have the expected coefficients. I shall only interpret coefficients for the model with all variables, because in the models that omit variables, coefficients are sometimes higher (in absolute value), possibly due to omitted variable bias.

Of the education variables, only Expend is significant. It is significant at the 1% level in both regressions in which it appears. It is positive, as expected, and it implies that for every extra percentage point of the country's GDP that the government spends on education, the percentage of people who speak English increases by 1.5%.

Of the globalization variables, there are many significant results, but there are also some surprises regarding the signs of coefficients. First, in the regression with all variables, all globalization variables are significant. Internet and Trade are significant at the 5% level, and FDI_OutStock, FDI_InStock, and Tourists are significant at the 1% level. In the model without the education variables, Internet and Trade lose their significance completely, but the other three variables remain significant. Most coefficients have the expected signs: The coefficient on FDI_OutStock is positive, and it says that if total FDI stocks held by foreign countries from the respective country increase by 1% of the respective country's GDP, PercEng increases by .69%. The coefficient on TouristsperPerson is also positive, and implies that an increase of 1 in the Tourists/Citizens ratio causes PercEng to increase by 8.83%. This is a considerably weaker effect than this variable has in the fixed-effects regressions. The coefficient on Trade also has the expected positive sign, and it says that if total trade (exports + imports) increases by one percentage point of GDP, then PercEng increases by .17%.

The coefficient on Internet is negative, which is highly unexpected. This coefficient implies that for every increase in the percentage of people who are Internet users in the country, English

proficiency *decreases* by .37%. In considering how to interpret this result, I look to the literature. Kim and Lee found that Internet is not significant in their 2005 regressions, whereas it is in the 1998 regressions. They explain this by pointing out the fact that a smaller proportion of webpages were in English in 2005 than was in 1998. Perhaps my sample is even more affected by this observation than their sample is. Because so many countries in my sample are only observed for PercEng in 2006 and 2012, in total the observations in my sample are skewed toward more recent times, when English has a less overarching presence online than it used to. And also, since my sample consists of only EU countries which are more developed than the average country in the world, EU citizens may be more likely to be able to browse the Internet in their native tongue than many other people in the world. To go further, then, more Internet users in an EU-country could imply that more of the Internet is in that country's language, meaning they actually have *less* exposure to English on the Internet than people in a country with relatively few Internet users. Conversely, people in a country with fewer Internet users would then more likely to browse English-speaking websites since fewer people who speak their native language are online. This directly contradicts the initial reasoning for including Internet as a variable: that more Internet use correlates with more English exposure. This assumption is still valid, but with this sample and this model, that effect appears to be empirically weaker than the opposite effect.

The coefficient for FDI_InStock is also negative, which is also unexpected. However, we justify this finding by reasoning that FDI_InStock actually might have less reason to be positively correlated with PercEng than FDI_OutStock does. In the following explanation, the term "home country" refers to the reporting country for FDI_OutStock and InStock. So for OutStock, it is the country who is investing abroad, and for InStock, it is the country who is receiving investment *from* abroad. FDI_OutStock implies initiative on the part of the "home" country. Businesses choose to and put the effort and resources into expanding their operations to other countries. Since they may be expanding to multiple countries, English works well as a lingua franca to communicate with any

country in which they may be investing. FDI_InStock, however, describes businesses in *other* countries initiating contact with the "home" country to invest in it. Communications between the "home" and "original" country may often be conducted in English, but businesses rarely expand to foreign countries without having at least a few people in charge of the expansion who speak the language of the foreign country. They also will hire local employees in the home country. For example, consider the Switzerland-based multinational corporation Nestle's investment in Hungary. Communications between Nestle's global headquarters in Vevey, Switzerland and the Hungarian headquarters in Budapest would likely often be in English, but Nestle factories in Hungary would employ Hungarian workers and managers, and communication there would be in Hungarian. Each of these factories contributes to Hungary's value of FDI_InStock, but they do not have anything to do with people speaking English. So, because for InStock, the home country has more of a passive role in foreign direct investment, essentially letting other countries come to it rather than actively seeking foreign involvement, the true coefficient for InStock may indeed not be positively correlated with English proficiency like OutStock's may be.

One should note that the results in the between-effects approach for the globalization variables do not imply causation. We cannot mitigate the endogeneity bias since we cannot lag variables in these models. Regardless, comparing these results with the non-lagged fixed effects regressions does imply that in this sample, the globalization variables are correlated with more between-country variation than within-country variation. However, I cannot conclusively say that the globalization variables *cause* between-country differences in English proficiency, because it still could be the other way around, that English proficiency affects changes in these variables.

The linguistic variables are very robust, have high coefficients, and have the expected signs. Both the IHRE_Index and RelativeImpLang remain strongly significant at the 1% level in every model. The coefficient on IHRE says that each "point" on the Index of Historical Relatedness to

English contributes to a 6.13% increase in the percentage of people who speak English. In other words, a country should have 6.13% higher PercEng than a country with a 1 lower value of the IHRE. For example, a German-speaking country should have a 24.54% higher PercEng value than a Greek-speaking country, all other things equal.[7] RelativeImpLang, the binary variable denoting if a country's official language can be considered to be relatively economically important, has the expected negative coefficient. This means that EU countries that speak Spanish, French, German, Italian, Portuguese, or Dutch (the most economically important languages in the EU besides English, according to Hjorth-Andersen's 2006 analysis) have a 37.98% lower PercEng value than other EU countries, all else equal. To reiterate the reasoning for this variable, this is because those people already speak a relatively important language that grant them sufficiently lucrative career opportunities without requiring them to learn English.

[7] IHRE value for German: 6; IHRE value for Greek: 2. (6 * 6.13) - (2 * 6.13) = 24.54.

Regression Table 3: Between-country effects (regression on group means) results

Dependent var: PercEng	Model 1: All vars	Model 2: Replace Global with KOF	Model 3: No Educ vars	Model 4: only KOF
Expend_AllLevels	1.543 (3.71)***	2.410 (3.08)***	--	--
PercLearnEng_Primary	0.016 (0.19)	-0.133 (0.84)	--	--
PercLearnEng_LowerSec	0.182 (1.29)	0.309 (1.28)	--	--
Internet	-0.366 (2.98)**	--	0.018 (0.12)	--
FDI_OutStock	0.685 (8.06)***	--	0.742 (7.33)***	--
FDI_InStock	-0.355 (3.58)***	--	-0.353 (2.84)**	--
TouristsperPerson	8.830 (3.17)***	--	13.185 (4.58)***	--
Trade_PercGDP	0.168 (2.41)**	--	-0.025 (0.36)	--
IHRE_Index	6.129 (4.48)***	8.947 (3.77)***	6.941 (4.71)***	8.544 (3.45)***
RelativeImpLang_Bin	-37.980 (7.81)***	-24.355 (3.16)***	-39.282 (7.41)***	-25.560 (3.22)***
KOF	--	0.738 (1.41)	--	1.208 (2.68)**
_cons	-29.584 (1.59)	-116.363 (2.91)***	18.575 (2.68)**	-75.106 (2.17)**
R^2	0.96	0.73	0.88	0.52
Adjusted R2	0.93	0.64	0.84	0.45
N	54	62	79	109

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Remarks on both models

Insignificant results do not necessarily mean uninteresting results. An interesting observation across all models in both fixed-effects and between-effects is that the KOF Globalization Index is almost never significant. It is not even significant when it is the only variable in the fixed-effects regression model, besides the year dummies (Model 4 in Reg Tables 1 and 2). The only regression in which it is significant is in Model 4 of the between-effects models (Reg Table 3), in which it is the only variable besides the linguistic variables. But in that model, omitted variable bias is likely causing KOF to pick up the significance of both Expend as well as the globalization variables. However, it is not the case that globalization as a whole has no correlation with English proficiency. I did find certain globalization variables to be significant on their own (particularly in between-effects), they are all always jointly statistically significant, and in the fixed-effects models they have an effect on the significance of time effects. The KOF Index is supposed to be an approximation of globalization as a whole, and it takes into account all of my globalization variables and more. Therefore, it is surprising for KOF to be almost always insignificant. It is also highly unexpected because the literature found KOF to be a very robust, significant variable in explaining variation in TOEFL scores worldwide.

Consequently, the question arises of why the KOF Index seems to be better at explaining variation in TOEFL scores between countries worldwide than it is in explaining variation in the percentages of people who speak English in EU countries. It could be the case that with the way that the KOF Index is calculated, it is not as good at accounting for differences in globalization within or between *EU countries* as it is for the rest of the world. Several of the variables used in the calculation of the KOF Index may in fact not be very relevant for the EU, because they do not vary much within or between EU countries, which are all relatively developed and globalized by worldwide standards. The following variables that partially make up the KOF index may not be very

relevant for the EU: “number of televisions per 1,000 people,” “trade in newspapers as a percentage of GDP,” “Number of McDonalds restaurants,” “number of embassies in country,” “membership in international organizations,” and “participation in U.N. security council missions.” (“Variables and Weights” ETH Zurich, 2014). If these variables do not vary much between and/or within EU countries, this would then mean that the KOF Index is not accounting for aspects of globalization relevant for comparing EU countries.

A promising result across both models is that the constant term is always statistically insignificant in the models with all variables. It tends to become significant when one removes variables, and also its coefficient tends to strengthen. This says that overall, the variables I have chosen are together relatively good at explaining variation in PercEng. If the constant term were significant when all variables were included, then that would tell us that we are not controlling for something extremely important.

Another value one can comment on is the R2 value. When one interprets the R2 value in Model 1, the model with all variables, it is best to interpret the Adjusted R2. R2 increases with the number of variables, which makes it quite misleading when one has many variables. Adjusted R2 fixes this problem. In both fixed-effects models, R2 is almost twice as high as Adjusted R2 in Model 1, so it is very important to only interpret the more statistically valid value. In the between-effects Model 1, the values do not differ very much, probably because the between-effects model does not use year dummies, and so it has fewer variables. Adjusted R2 suggests that in Model 1 for non-lagged fixed-effects, the model explains around 54% of the variation in PercEng over time. In the lagged Model 1, the model explains around 45% of the variation in PercEng. In between-effects, Model 1 has a very high Adjusted R2, saying that the model explains 93% of the variation in PercEng between countries.

VI) Conclusion

The European Commission places great importance on the learning of foreign languages in the European Union, publicly stating the need for multilingualism and instituting an EU-wide "two foreign languages" education policy (European Commission, "Language Policy" 2014). English is the most popular language to learn in the EU, with 96% of students learning English as of 2012. This is due to English's role as a global lingua franca in international business and political interactions. English-language pop culture, such as music and films, is also very popular and could also be related to the widespread study of English in Europe. Despite English's importance in the world and in the EU in particular, very little research on the country-level determinants of English proficiency exists. This project contributes to this small field by conducting a panel data analysis of English proficiency in the EU using the Eurobarometers' reported percentages of people in EU countries who speak English, between the years 1990 and 2012.

Since I am interested both in the trend across the EU of increasing English proficiency with time, as well as the disparities in English proficiency between countries, I decided to separate these effects using a fixed-effects (time-series) and between-effects (cross-sectional) approach, rather than using random effects which is an average of these two effects. After analysing the within-country variation and the between-country variation separately, the only variable to be robust and significant in *both* models is TouristsperPerson. This means that how important a country's tourism industry is must be a very important factor both in explaining the rise in English proficiency over time, and the ongoing differences in PercEng between EU countries. Another interesting remark for both models is that the KOF Index, which is significant and robust in the literature, is overall not very good at explaining variation in PercEng within or between EU countries. This could be because the KOF Index focuses on certain ways of measuring globalization that do not vary much between or within EU countries, which are relatively developed and globalized compared to the world

average. Finally, when we include all variables, the constant term is insignificant and has a weak coefficient relative to models in which variables are omitted. This observation supports the strength of the variables overall.

Other than Tourists, the only individual variables to have any significance in the within-effects regressions are Internet and FDI_OutStock. But because both lose their significance in Model 1 of the lagged fixed-effects models, they may have only been significant due to endogeneity bias. In fixed-effects, the globalization variables together do show some joint statistical significance, even though they are not significant on their own. They also cause time effects to become significant when they are removed, implying that they together do explain some time-series variation.

In the between-effects results, *all* of the variables are statistically significant, except for the two PercLearnEng variables and KOF (except when it is almost the only variable in the regression). However, when I remove the education variables, Internet and Trade lose significance entirely. Thus, for between-effects, the most robust variables are Expend, FDI_OutStock, FDI_InStock, Tourists, IHRE, and RelativeImpLang. Internet and FDI_InStock have negative coefficients, which is the opposite of what was expected, and possible reasoning for this finding has been proposed.

A possible reason why the globalization variables are so much more significant in the between-effects results compared to the fixed-effects results, is that this project does not have a method for accounting for endogeneity bias in between-effects. Because I do not have appropriate instrumental variables, I lag variables to reduce endogeneity bias, but this is not possible to do in between-effects. Still, this does not explain why the between-effects results are so much more significant than the *non-lagged* fixed effects results. For this, I propose another explanation:

Another reason why the fixed effects results are less significant than the between effects results, could be because there are many inconsistencies in the within-country trends in English proficiency as assessed by the Eurobarometers. In the trend graphs (Appendix 2) there are often

unexpected jumps and dips (dips at 2000 or jumps at 2006), and for a couple of countries it could be argued that they do not show a steady trend at all (e.g. Belgium, Portugal, Luxembourg). Also, the small amount of observations of PercEng per country creates a very serious limitation for this project. Some countries have seven observations, but many have only two. The countries with only two observations are the primarily Eastern European countries that joined the EU in 2004. Attempting to analyze within-country variation using a fixed-effects model is quite difficult when about a third of the countries in the dataset offer only two observations of English proficiency. These two obstacles primarily affect the time-series analysis, which could explain why the fixed effects regression results were more inconclusive than the between-effects results.

In conclusion, this paper offers strong evidence to suggest that two linguistic variables--how similar a country's language is to English and how economically important a country's language is--have explanatory power over the differences between EU countries' English proficiencies. The results also provide evidence that government expenditures on education, representing the quality of education, is a significant driving force in how well people learn English in EU countries, compared to other EU countries. As for the globalization variables, this project suggests that the strength of a country's tourism industry--represented by the number of tourists per citizen--has explanatory power over English proficiency, both within- and between-countries. The other globalization variables--number of Internet users, outward FDI stocks, inward FDI stocks, and the Openness Index--do exhibit correlation with differences in English proficiency only between-countries, but there is not convincing evidence to suggest that this correlation is causation. For within-countries, these globalization variables only exhibit weak correlation individually. Still, their exclusion does cause the time effects as well as the constant term to become statistically significant. This implies that they are at least jointly significant, which is supported by a statistical test.

There are many ways that this project could be improved upon. Since the primary weakness of my research is the small number of observations per country, this project could be much stronger if repeated in the future when there is more data from the Eurobarometers, assuming the European Commission continues to survey EU citizens on language proficiency. Also, given the occasionally irregular trend graphs (Appendix 2), statistical tests for linearity could be conducted in future research to examine whether or not it was correct to assume that rising English proficiency is in fact a linear trend. Another severe obstacle to this area of research is the difficulty in overcoming the endogeneity bias of the globalization variables. It will be very interesting if a future researcher can come up with an appropriate instrumental variable for each globalization variable. This project could also be better if there were a more nuanced way of evaluating the economic importance of languages besides the dummy variable used in this project. In the future, someone could repeat Hjorth-Andersen's methodology for different years and actually use the hierarchy of which European languages are relatively more important than others for multiple years as an index variable. One could then see if it matters whether someone speaks, for example, German vs Italian, or if all the "economically important" languages roughly equally demotivate their speakers to learn English, as the binary variable suggests. Additionally, future research may wish to further examine why certain results in this project are unexpected, such as the negative coefficients on Internet and FDI_InStock and the insignificance of the KOF Index.

Finally, there may be many other variables that affect English proficiency in EU countries that have not been accounted for in this project. Future research may focus on, for example, other forms of social globalization such as consumption of foreign music, films, and television programs. Several other education factors could also be examined, given that government expenditures on education may not be the best all-encompassing variable for education quality. There is a distinct possibility that public spending on education experiences diminishing returns to the quality of education—for example, the U.S. has the highest per-pupil educational expenditures in the world,

but student performance lags behind many other developed countries (University of Southern California, USA Rossier Online). Therefore it may be important to look at other factors such as pedagogy, teacher education, and the effectiveness of foreign-language-focused educational practices such as bilingual schooling and Content and Language Integrated Learning (CLIL). For education factors, it may also be more effective to research English proficiency at the local- or even school-level rather than country-level, to account for the many differences in educational quality that may exist at those levels.

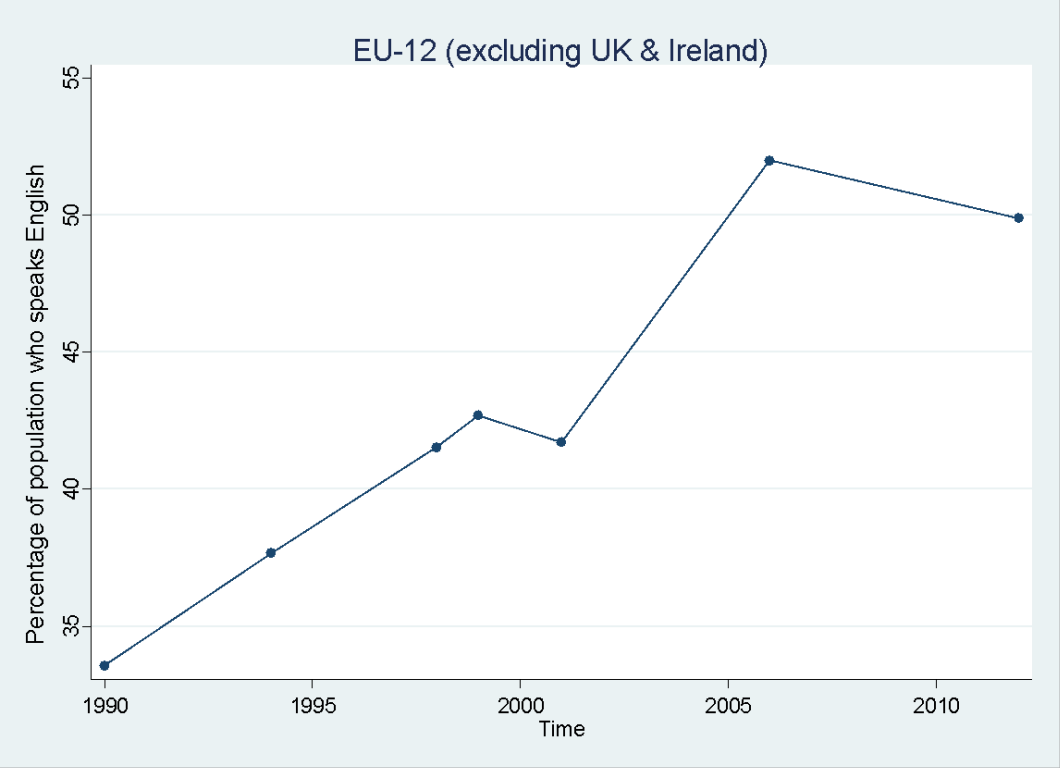
VII) Appendices

Appendix 1 – further explanation of choosing lag lengths

One may test for optimal lag length by running likelihood ratio (LR) tests, and computing four information criteria: "the final prediction error (FPE), Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (SBIC), and the Hannan and Quinn information criterion (HQIC)" (StataCorp., Base Reference Manual). Each metric, each using different methods of calculation, yields its own "optimal lag." In this project, four out of the five metrics always yield the same optimal lag to use, so for each country I choose the lag given by those four metrics. Unfortunately, the tests only work with time-series data, not panel data, so I may only test one country at a time. This limitation does not pose a problem for this analysis, however, because the overwhelming majority of countries give the same results, allowing me to choose the optimal lag length for the globalization variables (3 years) and for *KOF* (1 year). I ran the tests for each individual country in the dataset and counted how many countries had optimal lags of 1, 2, and 3 years. The average optimal lag for KOF was 1.3 (1 for twenty countries; 2 for four countries, 3 for two countries), so I round down to 1. The average optimal lag for the other globalization variables was 2.7 (1 for one country, 2 for four countries; 3 for twenty-one countries), so I round up to 3.

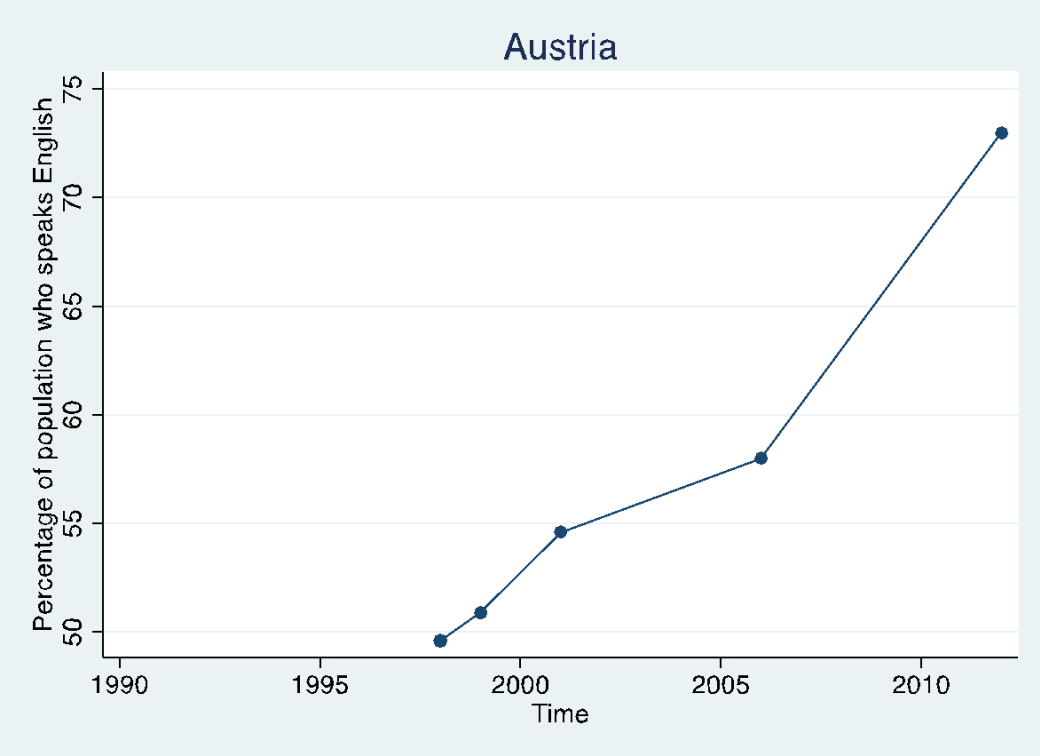
Appendix 2 – trend graphs of English proficiency throughout the EU and in individual countries, as measured by the Eurobarometers from 1990 – 2012

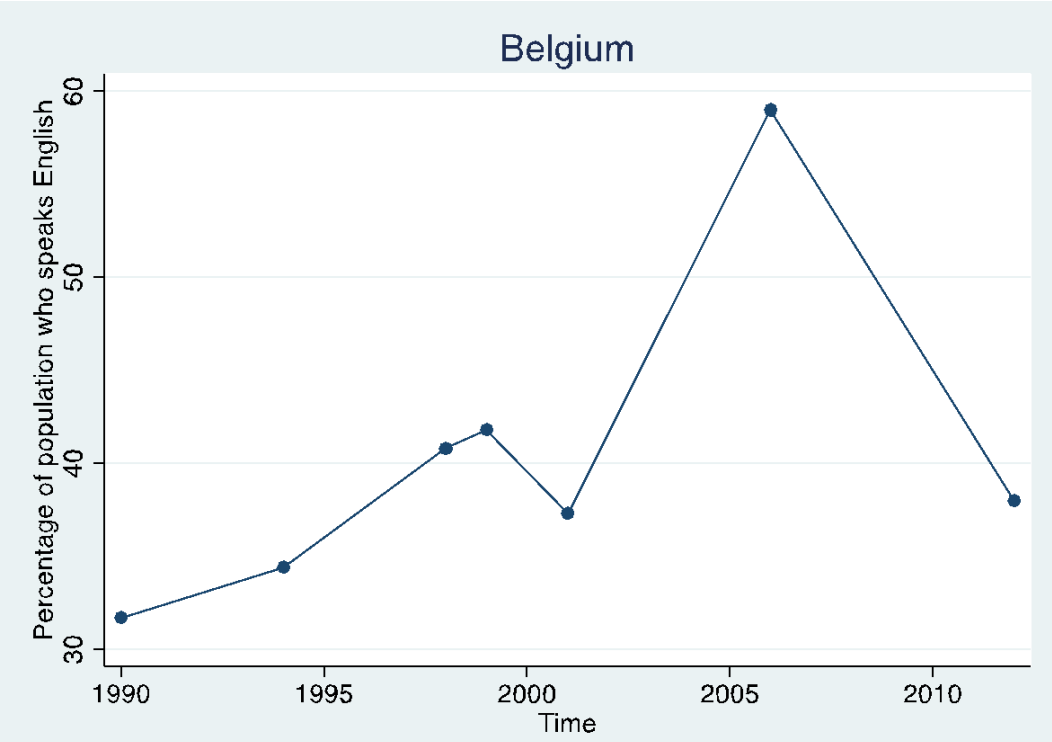
To allow for meaningful comparisons over time, this graph is restricted to the 10 countries in the sample that were members of the EU, and thus have been surveyed by the Eurobarometers, since 1990.

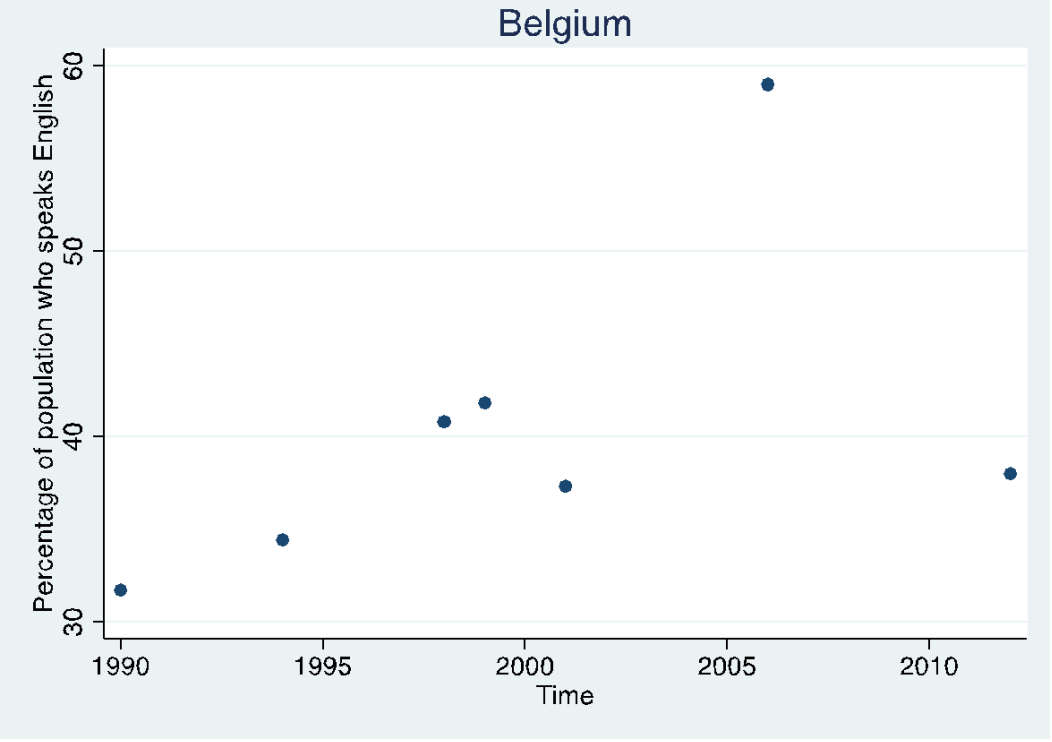


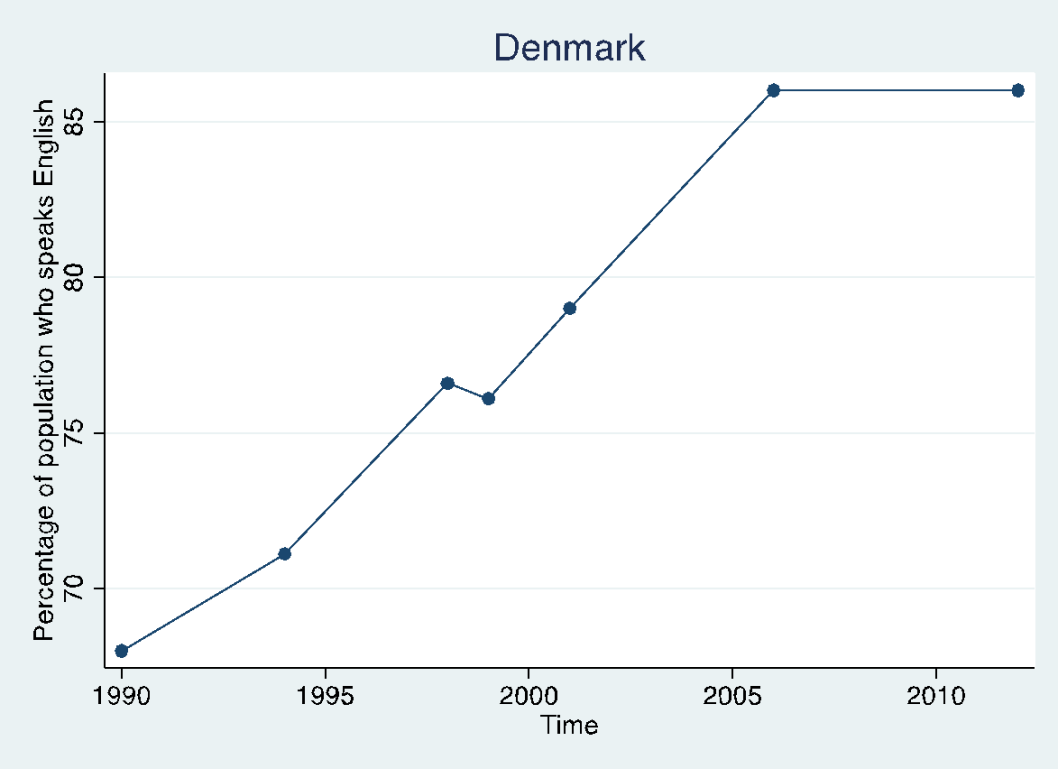
The following graphs only show countries which have been surveyed by the Eurobarometers on English proficiency at least five times. The countries which have only been surveyed twice (newer EU members, surveyed in 2006 and 2012) also demonstrate upward trends in English proficiency, but it is less meaningful to visually examine a graph with only two datapoints.

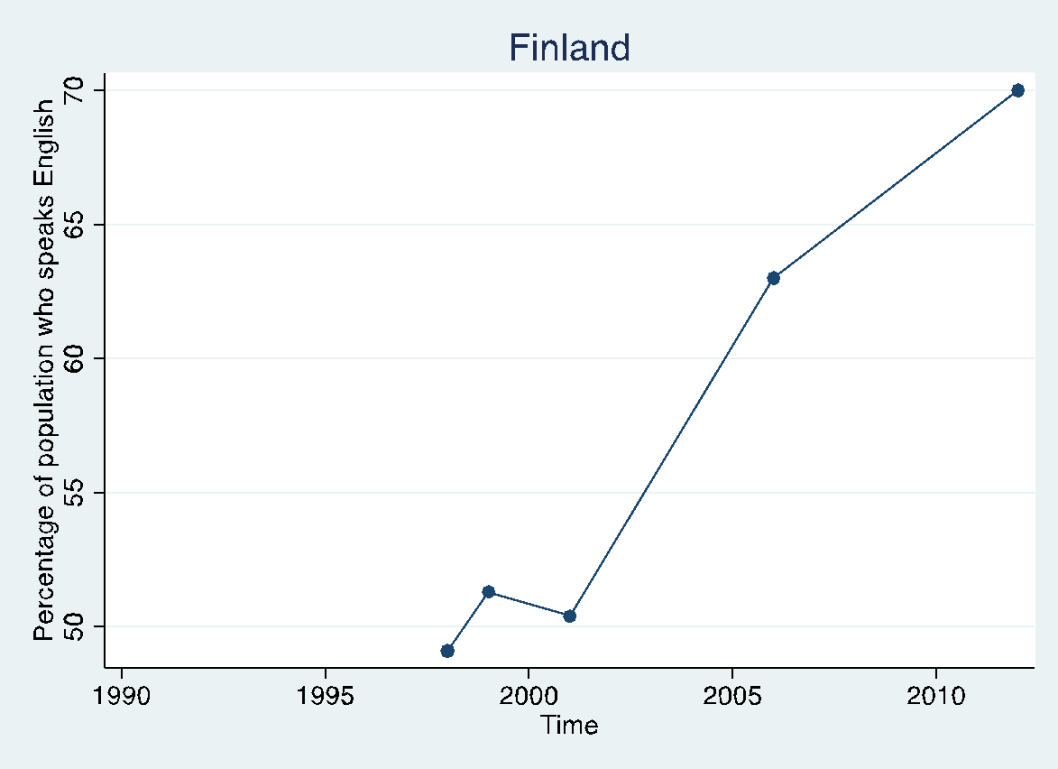
When a country's trend is inconsistent to the point that it may be questionable whether the country truly demonstrates a linear upward trend in English proficiency, a scatterplot without connecting lines is also shown.

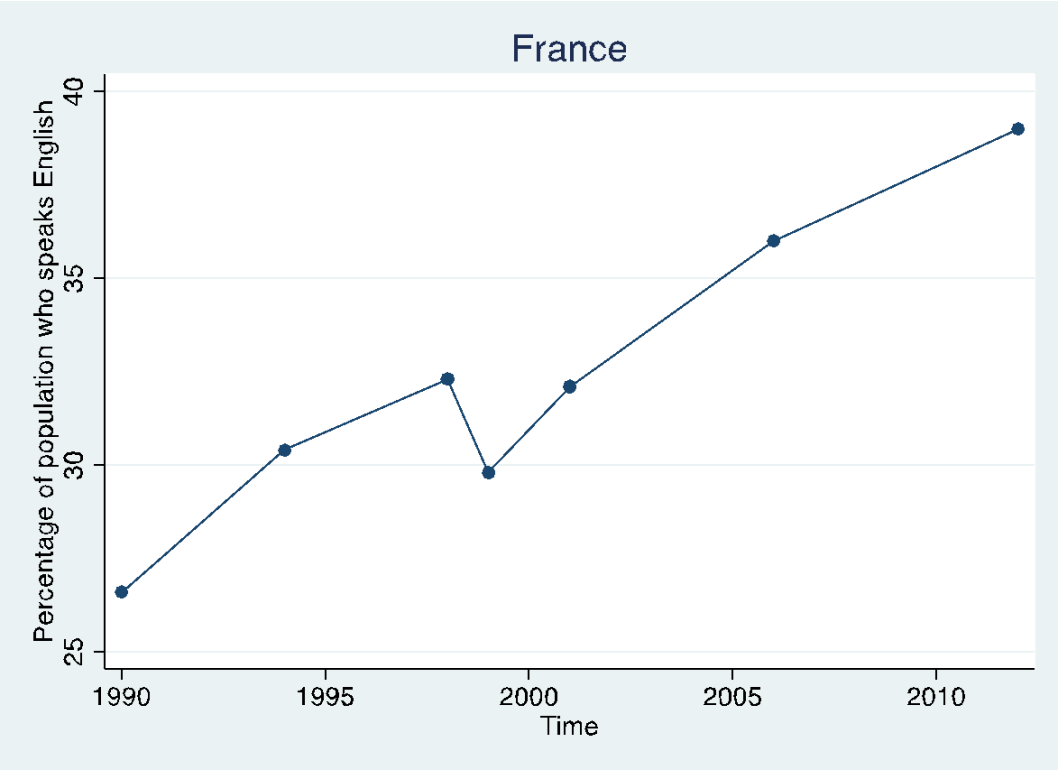



Belgium
Percentage of population who speaks English
30
40
50
60
1990
1995
2000
2005
2010
Time

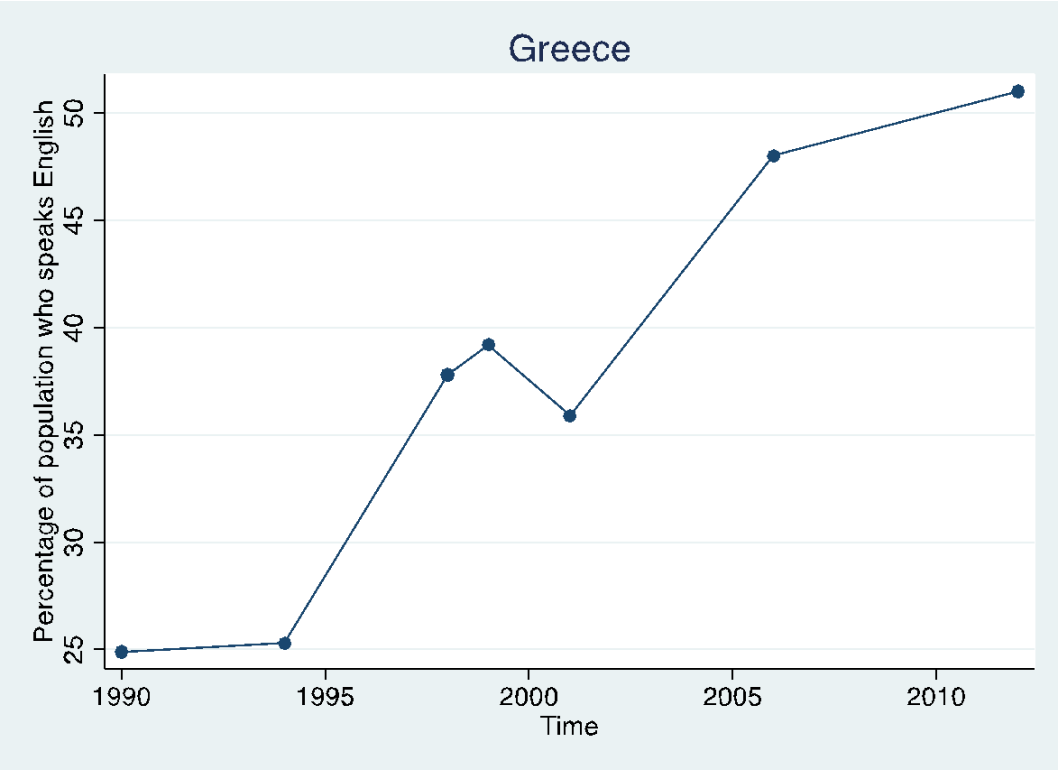

Belgium
Percentage of population who speaks English
30
40
50
60
1990
1995
2000
2005
2010
Time

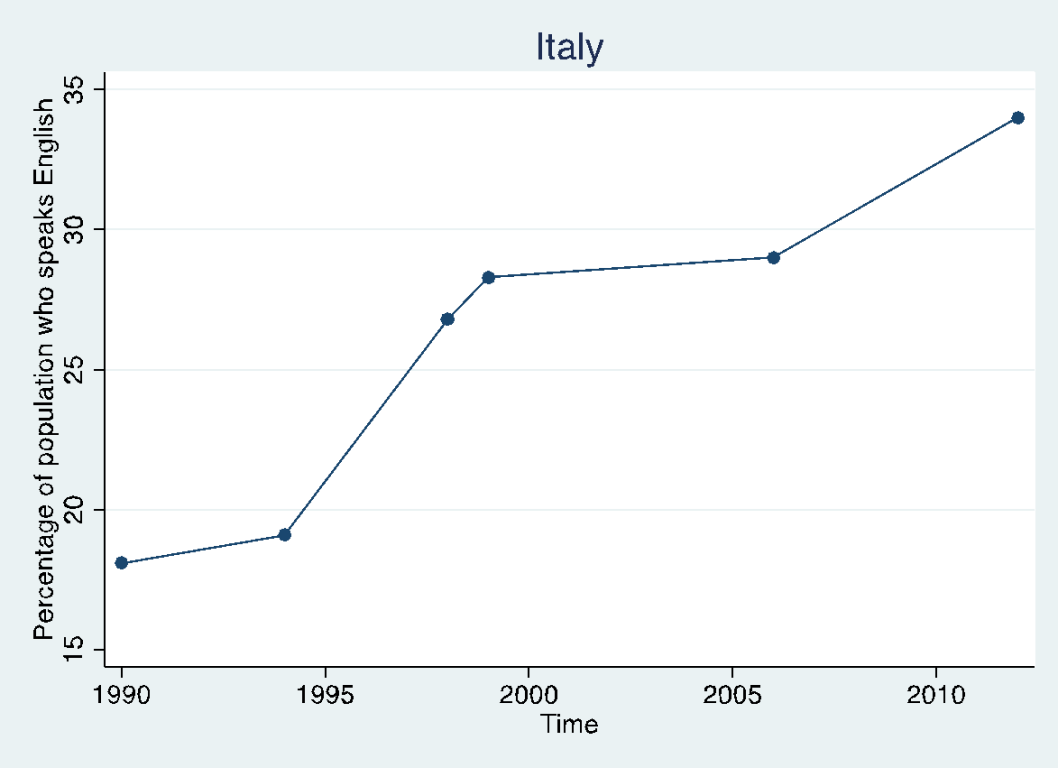

Denmark
Percentage of population who speaks English
70
75
80
85
1990
1995
2000
2005
2010
Time

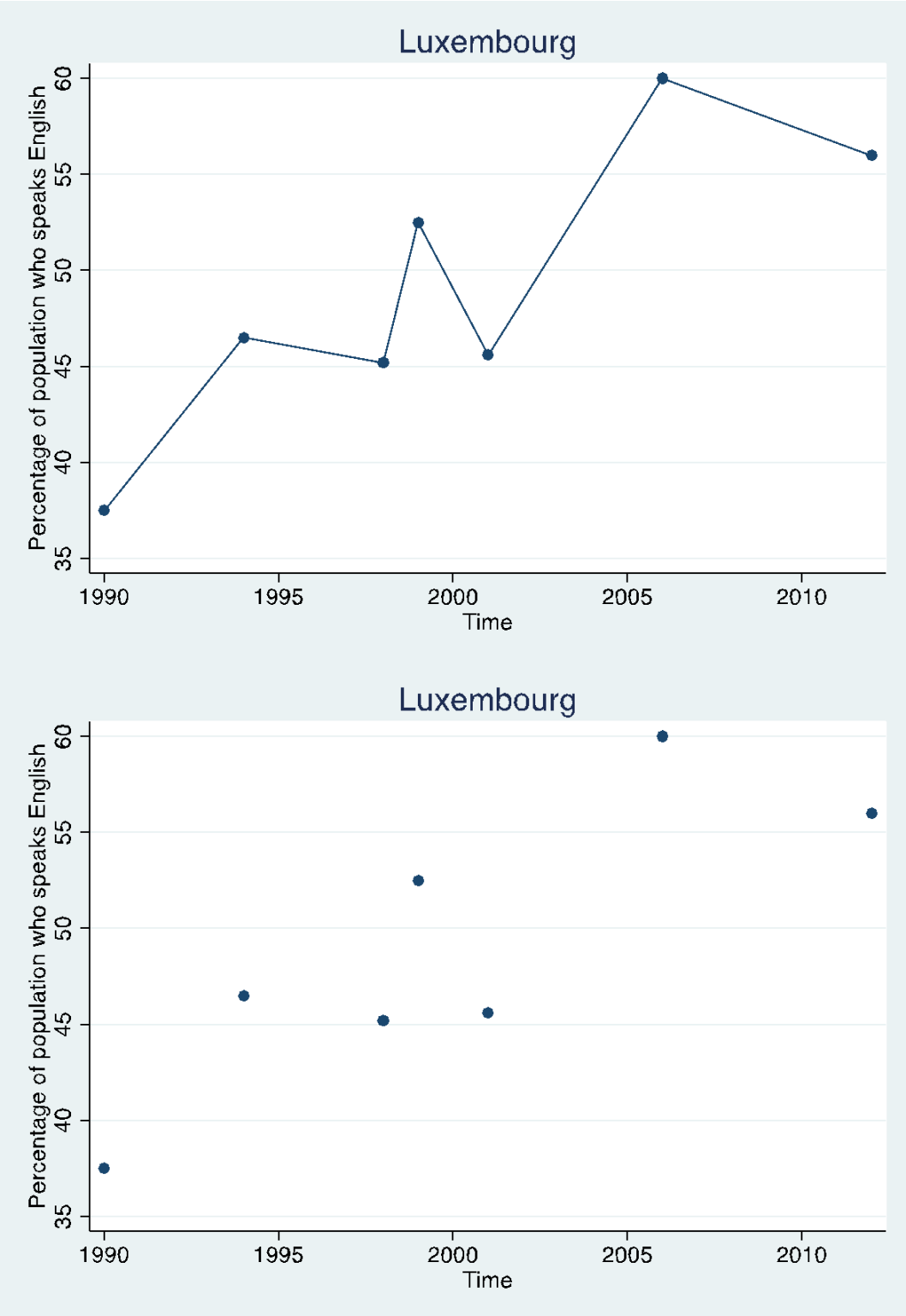

Finland
Percentage of population who speaks English
50
55
60
65
70
1990
1995
2000
2005
2010
Time

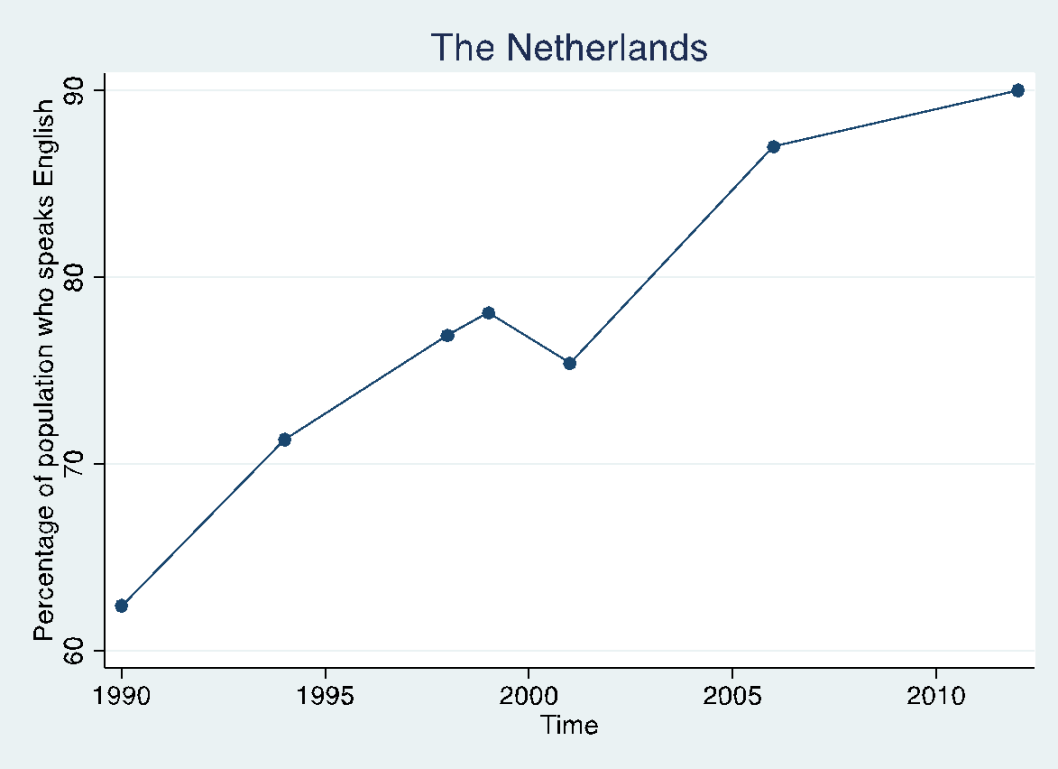

France
Percentage of population who speaks English
25
30
35
40
1990
1995
2000
2005
2010
Time

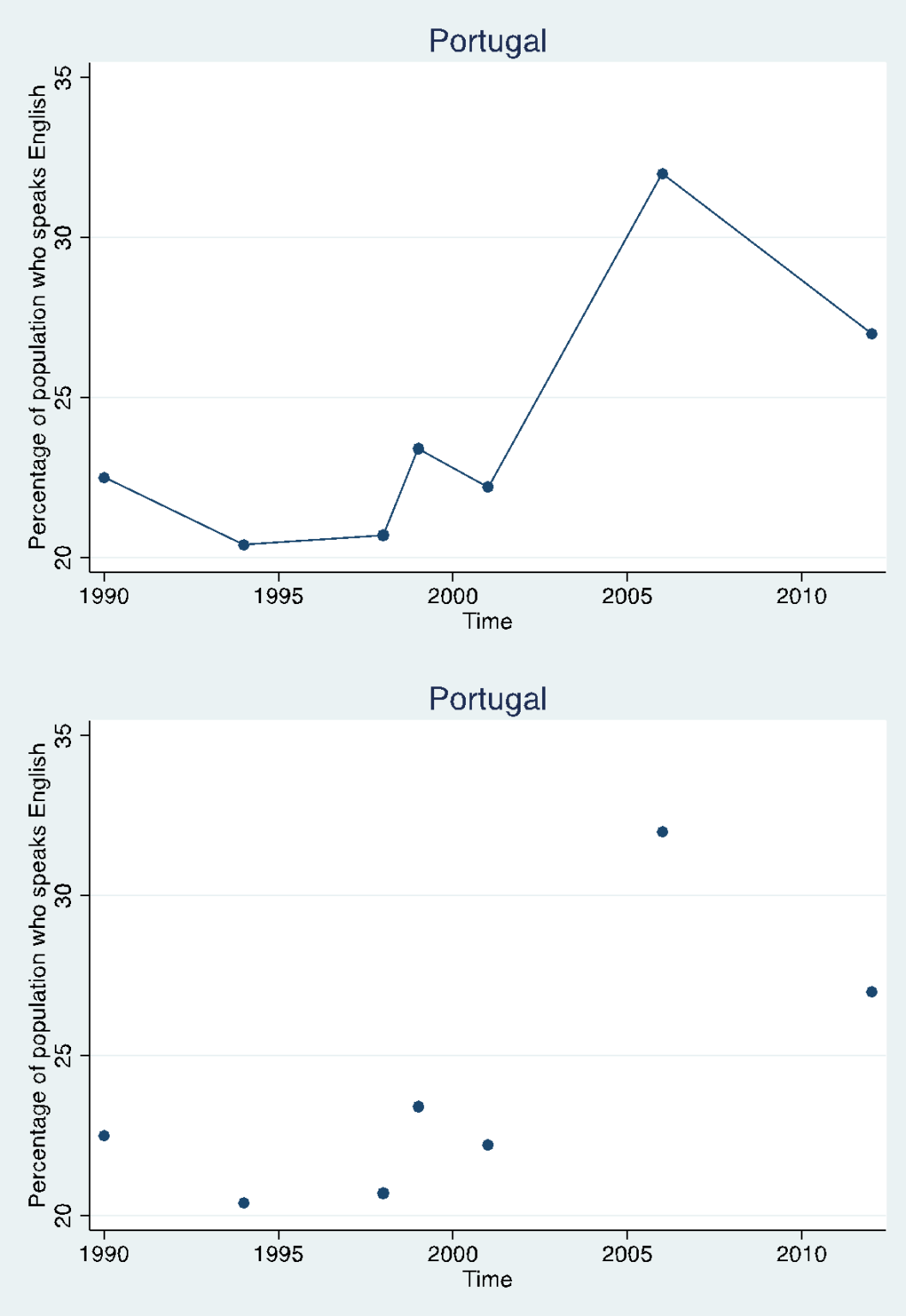

Germany
Percentage of population who speaks English
35
40
45
50
55
1990
1995
2000
2005
2010
Time

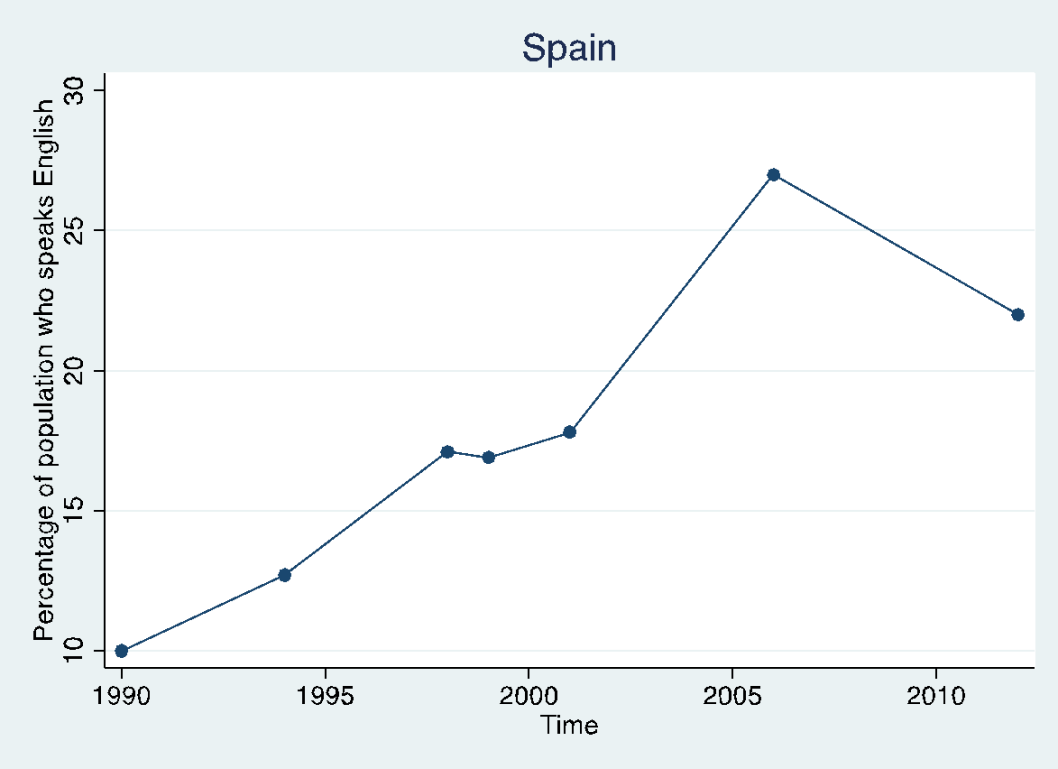

Greece
Percentage of population who speaks English
25
30
35
40
45
50
1990
1995
2000
2005
2010
Time

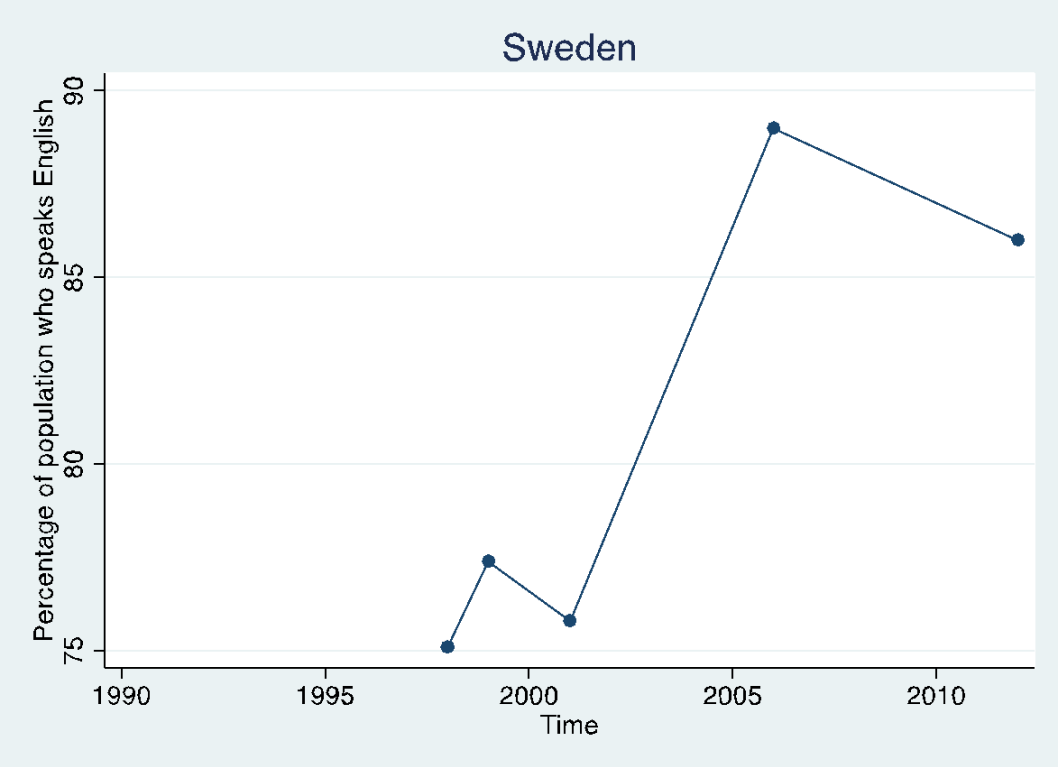

Italy
Percentage of population who speaks English
15
20
25
30
35
1990
1995
2000
2005
2010
Time


Luxembourg
Percentage of population who speaks English
35
40
45
50
55
60
1990
1995
2000
2005
2010
Time
Luxembourg
Percentage of population who speaks English
35
40
45
50
55
60
1990
1995
2000
2005
2010
Time


The Netherlands
Percentage of population who speaks English
60
70
80
90
1990
1995
2000
2005
2010
Time


Portugal
Percentage of population who speaks English
20
25
30
35
1990
1995
2000
2005
2010
Time
Portugal
Percentage of population who speaks English
20
25
30
35
1990
1995
2000
2005
2010
Time


Spain
Percentage of population who speaks English
10
15
20
25
30
1990
1995
2000
2005
2010
Time


Sweden
Percentage of population who speaks English
75
80
85
90
1990
1995
2000
2005
2010
Time

Appendix 3 – other supplemental graphs, tables, figures

Figure 3 –percentages of people in EU countries who speak various foreign languages in 2012

D48T Languages that you speak well enough in order to be able to have a conversation - TOTAL
THREE MOST WIDELY KNOWN LANGUAGES (% per country)

EU27	
English	38%
French	12%
German	11%

BE	
English	38%
French	45%
German	22%

BG	
English	25%
Russian	23%
German	8%

CZ	
English	27%
Slovakian	16%
German	15%

DK	
English	86%
German	47%
Swedish	13%

DE	
English	56%
French	14%
German	10%

EE	
Russian	56%
English	50%
Finnish	21%

IE	
Irish/Gaellic	22%
French	17%
English	6%

EL	
English	51%
French	9%
German	5%

ES	
English	22%
Spanish	16%
Catalan	11%

FR	
English	39%
Spanish	13%
German	6%

IT	
English	34%
French	16%
Spanish	11%

CY	
English	73%
French	7%
Greek	5%

LV	
Russian	67%
English	46%
Latvian	24%

LT	
Russian	80%
English	38%
German	14%

LU	
French	80%
German	69%
English	56%

HU	
English	20%
German	18%
French	3%

MT	
English	89%
Italian	56%
French	11%

NL	
English	90%
German	71%
French	29%

AT	
English	73%
French	11%
Italian	9%

PL	
English	33%
German	19%
Russian	18%

PT	
English	27%
French	15%
Spanish	10%

RO	
English	31%
French	17%
Italian	7%

SI	
Croatian	61%
English	59%
German	42%

SK	
Czech	47%
English	26%
German	22%

FI	
English	70%
Swedish	44%
German	18%

SE	
English	86%
German	26%
French	9%

UK	
French	19%
English	10%
German	6%

Source: Special Eurobarometer 386, Wave EB77.1 “EUROPEANS AND THEIR LANGUAGES” June 2012

Table 2: Summary statistics

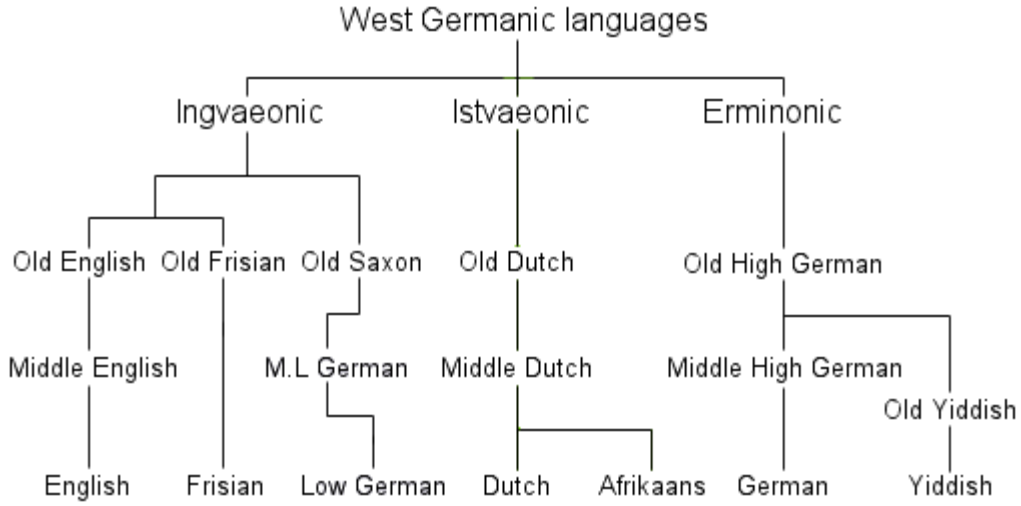
Variable name	*Obs*	*Mean*	*Std. Dev.*	*Min*	*Max*	*Years available*
PercEng *(Dependent variable)*	109	44.659	20.741	10	90	1990, '94, '98, '99, 2001, '06 '12
Expend_AllLevels	429	23.511	4.899	9.98	36.993	all
PercLearnEng_Primary	319	53.458	27.018	0	100	1998-2012
PercLearnEng_LowerSec	323	84.446	18.402	31.3	100	1998-2012
Internet	571	31.026	29.407	0	93.18	all
Trade_PercGDP	587	96.829	48.674	30.48	333.53	all
FDI_InStock	382	42.269	32.094	4.901	201	1994-2012
FDI_OutStock	372	28.788	35.485	-.4	244	1994-2012
TouristsperPerson	440	.91359	.63602	.1107	2.8649	1994-2012
KOF	589	74.925	12.909	34.05	92.5	all
IHRE_Index	754	3.5769	1.5495	1	7	time-invariant
RelativeImpLang_Bin	754	.34615	.47605	0	1	time-invariant

*"all" years means that the variable is available for all the years the dependent variable is

(obs=54)

	PercEng	Expend~s	PercLe..	PercLe~c	Internet	Trade_~P	FDI_In~k	FDI_Ou~k	T~tspe~n	IHRE_I~x	Relati~n	KOF
PercEng	1.0000											
Expend_All~s	0.5915	1.0000										
PercLea~mary	0.2138	0.1483	1.0000									
PercLearnE~c	0.3448	0.1270	0.5981	1.0000								
Internet	0.5436	0.4052	0.2413	0.0600	1.0000							
Trade_Perc~P	0.1692	-0.0583	-0.3571	-0.5519	0.4248	1.0000						
FDI_InStock	0.1365	0.0472	-0.2628	-0.5092	0.5180	0.8452	1.0000					
FDI_OutStock	0.6117	0.4988	0.0570	0.0660	0.6088	0.1621	0.3336	1.0000				
Touristspe~n	0.1592	0.1612	0.2318	0.0789	0.0482	0.2186	0.1363	0.0069	1.0000			
IHRE_Index	0.2751	0.2032	0.1484	0.0767	0.1033	0.0992	0.0930	0.3782	0.2563	1.0000		
RelativeIm~n	-0.3079	0.0995	0.0281	-0.0605	-0.2328	-0.2434	-0.1763	0.2491	0.2619	0.5529	1.0000	
KOF	0.4569	0.5671	0.2012	0.0798	0.3858	0.0044	-0.0053	0.6350	0.1840	0.3749	0.3180	1.0000

Figure 4: West Germanic language tree



Appendix 4 – Regressions for the purpose of comparison with the literature

An interesting question was to see if the Eurobarometers would yield similar results as the TOEFL scores when using similar methodology as Kim and Lee, since this project's primary methodology and independent variables are quite different from those used in the literature. Kim and Lee looked at 2004/2005 TOEFL scores. The closest approximation to Kim and Lee's reported 2004/2005 TOEFL scores is Eurobarometer data on English proficiency from 2006. OLS regressions for only the year 2006 of PercEng on Kim and Lee's variables that are readily available and relevant for the EU yield results that are in some ways similar, and some ways very different, compared to Kim and Lee's results (see Regression Table 4 below). The KOF Index is insignificant, unlike in the literature. Internet is significant as is was in the literature, but it has a negative coefficient instead of the expected positive. These unexpected results are also found in the primary analysis of this paper, and so the explanation for them here is the same as the proposed explanation in the "Results and Discussion" section. Next, tourismarrivals is also significant as it was in the literature, but has an unexpected negative coefficient. As explained in the "Independent Variables" section, tourismarrivals is actually more of a measure of the size of the country, rather than the relative prevalence of international travellers. Country size could indeed be negatively correlated with English proficiency through country size being negatively related to globalization, since a large country has access to more resources and thus has less need for foreign involvement. GDP is not significant when used with all other variables, but it gains significance as the other variables (education and openness indicators) are removed in Models 4 and 5, which corresponds with Kim and Lee's findings. The fact that duration of education is not significant, whereas SCHOOLING was highly significant in Kim & Lee's analysis, can be explained by the fact that the current sample only consists of EU countries, which all have rather similar durations of compulsory education. ExColony and WordOrder are included in only in one regression because they are only relevant for a couple of observations in this sample (see reasoning for omitting them in the main project in the section

"Independent Variables"). It is unclear why the HAV index is not significant in this analysis, whereas it was for Kim and Lee. Overall, it seems that in this supplemental analysis, results that contradict the literature may be due to a variable behaving differently for the EU than for the rest of the world, due to this analysis having a small sample of only 26 observations, or due to some other unexplained reason.

Regression Table 4: Regression Table 4: Results using literature's methodology for purposes of comparison (OLS regressions, only year 2006, using their variables)

Dep. variable: PercEng	Model 1: all vars	Model 2: remove ExColony and WordOrder	Model 3: replace openness indicators with KOF	Model 4: replace openness indicators with KOF, remove Educ	Model 5: remove all vars that are correlated with GDP
GDP_capita	0.000 (1.70)	0.001 (2.36)**	0.001 (2.08)*	0.001 (2.18)**	0.001 (2.95)***
Educ_Duration[8]	0.134 (0.07)	1.190 (0.52)	1.099 (0.35)	--	--
Trade_PercGDP	-0.138 (2.65)**	-0.177 (2.97)***	--	--	--
tourismarrivals[9]	-0.000 (2.84)**	-0.000 (3.41)***	--	--	--
Internet	0.699 (3.63)***	0.482 (2.37)**	--	--	--
HAV_Index	1.395 (1.51)	1.215 (1.21)	0.730 (0.50)	0.823 (0.58)	0.941 (0.67)
ExColony	36.282 (3.50)***	--	--	--	--
WordOrder	-3.903 (0.32)	--	--	--	--
KOF	--	--	0.468 (0.75)	0.561 (1.01)	--
_cons	15.750 (1.00)	18.026 (1.05)	-19.175 (0.44)	-17.038 (0.40)	24.909 (3.25)***
R^2	0.87	0.78	0.44	0.43	0.41
Adjusted R2	0.81	0.71	0.33	0.36	0.35
N	26	26	26	26	26

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

[8] Educ_Duration is very similar to Kim and Lee's "SCHOOLING:" Number of years of compulsory education.

[9] Tourismarrivals is the same as Kim & Lee's "INBOUND:" Total number of tourism arrivals in that year.

VIII) References

Azam, Mehtabul; Chin, Aimee; Prakash, Nishith. "The Returns to English-Language Skills in India." *Economic Development and Cultural Change* Vol. 61, No. 2 (January 2013) (pp. 335-367)

Central Intelligence Agency. Official language; "Field Listing - Languages". The World Factbook

Educational Testing Service (ETS). "About the TOEFL IBT® Test." *TOEFL IBT: About the Test*. N.p., 2014. Web. <https://www.ets.org/toefl/ibt/about>.

Educational Testing Service (ETS). *Test and Score Data Summary for TOEFL IBT® Tests, January 2013– December 2013*. N.p., 2014. Web.

Ethnologue*: Languages of the World*. "English." 2014. Web. <http://www.ethnologue.com/language/eng>.

European Commission. *Europeans and Their Languages, Special Eurobarometer 386*. N.p., June 2012. Web.

European Commission. *Eurobarometers 34.0, 41.0, 50.0, 52.0, 55.1, 64.3, and 77.1.* Raw datasets accessed through http://zacat.gesis.org/webview/ZACAT, Leibnitz Institute for the Social Sciences. Web. (1990, 1994, 1998, 1999, 2001, 2006, 2012).

European Commission. *EUROSTAT*. Last updated 11 Dec. 2014.

Eurostat. Eurostat Press Office. *News Release: European Day of Languages*. N.p., 25 Sept. 2014. Web.

Eurydice: The Information Network on Education in Europe. *Key Data on Teaching Languages at School in Europe*. N.p., 2005. Web.

Eurydice: The Information Network on Education in Europe. *Key Data on Teaching Languages at School in Europe*. N.p., 2012. Web.

Fidrmuc, Jan and Jarko. "Languages and trade." *Economics and Finance Working Paper Series,* number 09-14. Department of Economics and Finance, Brunel University West London. (2009)

GESIS. "Countries, regions, population coverage." GESIS Eurobarometer Data Service, 09 Apr. 2015. Web. <http://www.gesis.org/eurobarometer-data-service/survey-series/standard-special-eb/sampling-and-fieldwork/>.

GESIS. "Sampling and Fieldwork." GESIS Eurobarometer Data Service, 03 Dec. 2013. Web. <http://www.gesis.org/eurobarometer-data-service/survey-series/standard-special-eb/sampling-and-fieldwork/>.

GESIS. "Weighting overview: Standard & Special EB." GESIS Eurobarometer Data Service, 12 Nov. 2014. Web. <http://www.gesis.org/eurobarometer-data-service/survey-series/standard-special-eb/sampling-and-fieldwork/>.

Hammond, Alex. "The most widely spoken languages." 2 Jan. 2012. <http://blog.esl-languages.com/blog/esl/most-spoken-languages-world/>

Hjorth-Andersen, Christian. "The Relative Importance of the European Languages" Working paper number 06-23. Department of Economics, University of Copenhagen, Cph. (2006)

Johnson Language Blog. "Who Speaks English?" *The Economist*. The Economist Newspaper, 05 Apr. 2011. Web. <http://www.economist.com/blogs/johnson/2011/04/english>.

Kim, Myung-Hee, and Hyun-Hoon Lee. "Linguistic And Nonlinguistic Factors Determining Proficiency Of English As A Foreign Language: A Cross-Country Analysis." Applied Economics 42.16-18 (2010): 2347-2364.

McManus, Walter S. “Labor Market Effects of Language Enclaves: Hispanic Men in the United States.” *The Journal of Human Resources* Vol. 25, No. 2 (Spring, 1990) (pp. 228-252)

Swiss Federal Institute of Technology. *KOF Globalization Index*. Last updated: 15 April 2014. Web. http://globalization.kof.ethz.ch/

Snow, M. S. (1998) “Economic, statistical, and linguistic factors affecting success on the test of English as a foreign language (TOEFL)”, Information Economics and Policy, 10, 159–72.

StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.

StataCorp. 2013. Stata 13 Base Reference Manual. College Station, TX: Stata Press.

The World Bank. *World Development Indicators*. Last updated: 16 Dec. 2014.

The World Bank. *Education Statistics*. Last updated: 29 Sep. 2014.