Returns to Language Skills in the US Labor Market[*]

Natalie Cook[†]

May 10, 2021

Abstract

Figuring out which language gives you the biggest wage increase in the US holding as much other stuff constant as possible. Using linear regression plus lots of fancy data science techniques. This is really for fun and I don't think there are any serious implications of the findings. Maybe on a personal level people could choose which language to learn based on what will make them the most money.

[*] Acknowledgements here, if any.

[†] Department of Economics, University of Oklahoma. E-mail address: natalie.c.cook-1@ou.edu

1 Introduction

Returns to language are often cited colloquially as highly significant in a increasingly interconnected global economy. There is a large body of research which focuses on the positive economic effects of language acquisition in migrants. The ability of a migrant adult or child to pick up the language of their new country is often a key factor in their successful assimilation and subsequent positive outcomes. There is also much research which seeks to study the returns to bilingualism more generally. However, many researchers (myself included) encounter limitations because of the data that exists. There is very little data which specifically captures native English speakers who have acquired skill in an additional language.

There are many policy issues which can be informed by this type of research. Examples include foreign language requirements in education, and official language distinctions at a federal level. Additionally, individuals face a personal decision regarding their investment in language acquisition. The human capital model is often used to shed light on these decisions and capture the effects of language within an economic framework.

This analysis seeks to explore the impact of language skills on earnings for a specific subset of the population within the US labor market. Although the original intent was to capture those individuals whose native language is English and have developed skills in an additional foreign language, the reality of the available data inhibits this. In actuality those individuals who are studied likely speak a non-English language as their mother tongue, and have acquired English as their second language. This limitation caused by the data is common among the research.

This analysis is divided into the following sections. There is a literature review which presents the existing research regarding the topic. This is followed by a presentation of the data employed in this analysis, and the methods used on that data. Finally there is a section which details the results of this analysis, and a conclusion which presents final thoughts.

2 Literature Review

Albert Breton was among the first to employ the human capital model as a means of understand the economic benefits of language skills. His paper is a fascinating mix of micro and macro economic theory which seeks to apply the tools of economic theory to the complex world of language and culture. He described it as something that could be invested in using . He distinguished between physical capital such as machinery and knowledge or human capital based on their economic status. He argues that while physical capital can be sold, human capital can only be rented. He argues that the benefit must be calculated as a difference between the investment required to achieve the language and the return that is realized as result of the language. He points out that the return on investment for a mother tongue is very high. This is a unique way of thinking about basic literacy. He concedes that return on investment for additional languages is much more difficult to qualtify. (Can)

An analysis conducted by Isphording and Sinning used American Community Survey data from 2010 to study returns to language skills. Their focus was on immigrant populations, and they took special care to differentiate between the experiences of migrants and their children. They use an instrumental variable to address endogeneity which looks at differences in language acquisition for immigrants from English speaking and non English speaking countries. Their analysis found significant returns to language skills for both adult and child migrants. For the children, they found that education was very important to language acquisition and had an effect on wages. They use duration of residence quite a bit to add nuance to their analysis and better understand the relationship between language skills and earnings. Despite the common background differences between adult and child migrants, they did not find a significant difference between their returns to language skills. (Ingo Isphording)

Williams' study regarding returns to language skills utilizes the European Community Household Panel 1994-1999 to study individuals from 14 countries in Western Europe. This research stands out from the rest because its scope includes multiple countries. The researchers prioritize actual language use in a job setting over reported proficiency which is also unique. They found large variance across countries regarding the proportion of individuals that reported using a second language at work. This ranged from 6 percent in the UK to 78 percent Luxembourg. Notably, the most common foreign language that workers reported using at their jobs was English, which was followed by French. The rate of second language usage was highest in the business sector for most countries. Using a fixed effects model they found more modest returns to language than their original OLS model which they attribute to productivity differentials. They find that returns are generally around 3 to 5 percent. (Wes)

Lopez conducted a foundational study in 1999 using the National Adult Literacy Survey of 1992 from the US Department of Education. He uses his findings to argue against making English the official language of the US or over-emphasizing English proficiency. He found that bilingual individuals earned slightly more than those who spoke only English, given a set of controls. He encounters a common issue with the data: individuals who originally speak English and learn an addition language are excluded because language questions are only posed to language minority individuals. Therefore, although he uses the term bilingual in his analysis, he is really referring to individuals belonging to a language minority who have maintained their native language and then developed English skills. His most significant result was a difference between those whose only language is non-English, and those who speak English in addition to their native language. He also found that the degree of proficiency was more significant that the particular language an individual was proficient in. He calls for additional researched focused on understanding the outcomes of native English speakers who learn an additional language, as opposed to language minority individuals who learn English. (Lop)

More recent literature provides an answer to this call.Researchers at the Federal Reserve Bank of Philadelphia investigated the economic returns associated with speaking a foreign language. Their empirical analysis found that the hourly earnings of individuals who spoke a foreign language were more than 2 percent higher than the earnings of those who did not. Their research is highly relevant to this analysis because they seek to understand the returns to speaking a second language for college graduates who are native speakers of English. This research stands out from the rest, because the vast majority of the literature focuses on returns to learning English for people who have migrated to the US. They overcame one of the core challenges of this inquiry which is finding a data set which captures native English speakers who also have learned a second language and includes additional information on the individual including academic performance. The data set that they used is called the ”Baccalaureate and Beyond Longitudinal Study” from the National Center for Education Statistics for the years 1992, 1993, 1994, and 1997 which included 9274 individuals. (Saiz and Zoido (2002))

Their analysis included extensive work to control for as many additional factors as possible. They use control variables to mitigate selection bias, and address ability bias using SAT scores, GPA, parental education, and quality of the college attended. They control for the major chosen by the student which may serve as an indicator of career preferences. They use the longitudinal dimension of the data set to account for constant unobserved individual characteristics by comparing earnings growth for individuals between 1993 and 1997. Finally, they employ instrumental variables to address the problem of selection by earnings. These instruments are high school foreign language requirements in the state where the individual attended high school, and college second language requirements. These inclusion of these instrumental variables in particular led to increased estimates of the returns to a second language. The estimates with the instrumental variables had high standard errors, but calculated returns as being between 14 and 30 percent. (Saiz and Zoido (2002))

3 Data

The primary data source for this research is the 2018 American Community Survey from the US Census Bureau. An extract was creating using IPUMS in order to pull the relevant variables. The dependent variable for this analysis is income. The Wage and Salary variable measures the individual's total pre-tax wage and salary income for the previous year. This was chosen to most accurately capture explicit and conventional earnings.

Language is the primary variable of interest. This is measured in a notable way. For the ACS language records the language that the respondent spoke at home, particularly if a language other than English was spoken. In practice, those individuals that are recorded as speaking another language in addition to English speak that language as their mother tongue, and have learned English as a second language. There are 94 unique languages recorded by this variable.

Speaks English is a categorical variable which was condensed into a binary for this analysis in order to subset the data to include only those individuals who spoke English plus an additional language. Employment status is also a categorical variable that indicates whether the individual is a part of the workforce. This was used to filter the data to only include working individuals.

Industry is a variable which records the industry in which the individual performed their occupation. There are many hundred unique industries which can be condensed into greater sectors. This variable serves as a control and is also involved as an interaction effect with language in one of the models that follow.

A number of common control variables were included in the model. Age is continuous and is measured conventionally in years. Race is a categorical variable which is comprised of 9 categories where white is set as the reference group. Sex is a binary categorical variable which is measured conventionally with male as the base group. Place of work: state was chosen as a control for regional differences. It is a categorical variable which records the state in which the individual works.

Now I will talk about the survey in general a little bit and the go into depth on the variables I will use and how they are measured. This is important for interpretation later on.

I used IPUMS to make an extract with some variables, then ”ipumsr” to load that extract into R with the codebook. Then I started manipulating stuff. I made a logwage variable

4 Empirical Methods

The primary empirical models for the analysis are shown in the following equations:

$$LogIncome = \beta_1 Language + \beta_{2j} Industry_j + \beta_3 Location + \beta_4 Age + \beta_5 Sex + \beta_7 Race \tag{1}$$

$$LogIncome = \beta_{1i}Language_i + \beta_{2j}Industry_j + \beta_3 Location + \beta_4 Age + \beta_5 Sex + \beta_6 Race \qquad (2)$$

$$LogIncome = \beta_{1ii}Language_i * Industry_j + \beta_3 Location + \beta_4 Age + \beta_5 Sex + \beta_6 Race \qquad (3)$$

The dependent variable remains the same across the three models. It is a log transformation of the income variable which is continuous. The parameter of interest is largely the same as well. It the coefficient associated with the language variable. The three models also share identical control variables for location, age, sex, and race. Location, sex, and race are factor or categorical variables. Age is numeric. The models are evaluated using Ordinary Least Squares Regression.

In the first model, language is a binary which measures whether or not an individual speaks any additional language besides English. In this first model, the industry variable is a vector which is comprised of many industries grouped into sectors. The first model is designed to calculate whether speaking any additional language is correlated with either a positive or negative effect on log income.

The second model displays language as a vector, similar to the industry vector which is also included. Instead of simply measuring whether or not an individual speaks an additional language, this records which language in particular the individual speaks. In practice the model is comprised of a long list of dummy variables which are coded with 0's and 1's to indicate which language is spoken. This second model is designed to display the income effects of each language. Both the sign and magnitude of the effects will be interesting to study.

Finally, the third model includes both the language and industry vectors, but creates an interaction term by multiplying them together. This third model is designed to calculate the effect of working in a given industry while speaking a certain language for each of the possible industry-language combinations. In theory the sign and magnitude of these coefficients could indicate which industry-language combinations are correlated with the greatest income effects.

5 Research Findings

The main results are reported in Table 3. This is where I try to interpret my results. I'll comment on whether the model overall is valid/significant. I'm hoping that it is. I'll talk about heteroskedasticity and robust standard error and other things that I definitely remember that make a model legit. Then I'll interpret the coefficients that get calculated for each variable. Maybe not each variable since some of them are supposed to be controls.

6 Conclusion

If you want to pick a language to learn that will increase your salary - please do an actually nuanced assessment of your situation. If you're gonna go into national security or intelligence of course you should learn Russian or Korean or Arabic or Mandarin. If you need something that is broadly applicable you should learn Spanish. This is dumb and not actually useful and you shouldn't use it as advice because I can almost guarantee that whatever it tells you to do will actually be wrong. It's a fun exercise though so there's that. Don't actually take this conclusion seriously, this is for my learning and entertainment purposes.

Table 1: Results

	Dependent variable:
	INCWAGE
Observations	1,415,468
R^2	0.192
Adjusted R^2	0.192
Residual Std. Error	60,158.470 (df = 1415068)
F Statistic	841.972*** (df = 399; 1415068)
Note:	*p<0.1; **p<0.05; ***p<0.01

References

???? Tech. rep.

???? "Does Speaking a Second Language Affect Labor Market Outcomes? Evidence from the National Adult Literacy Survey of 1992." .

???? "Economic Approaches to Language and Bilingualism. New Canadian Perspectives." .

Ingo Isphording, Mathias Sinning. ???? "The Returns to Language Skills in the US Labor Market." .

Saiz, Albert and Elena Zoido. 2002. "The returns to speaking a second language." Tech. rep.

Figures and Tables

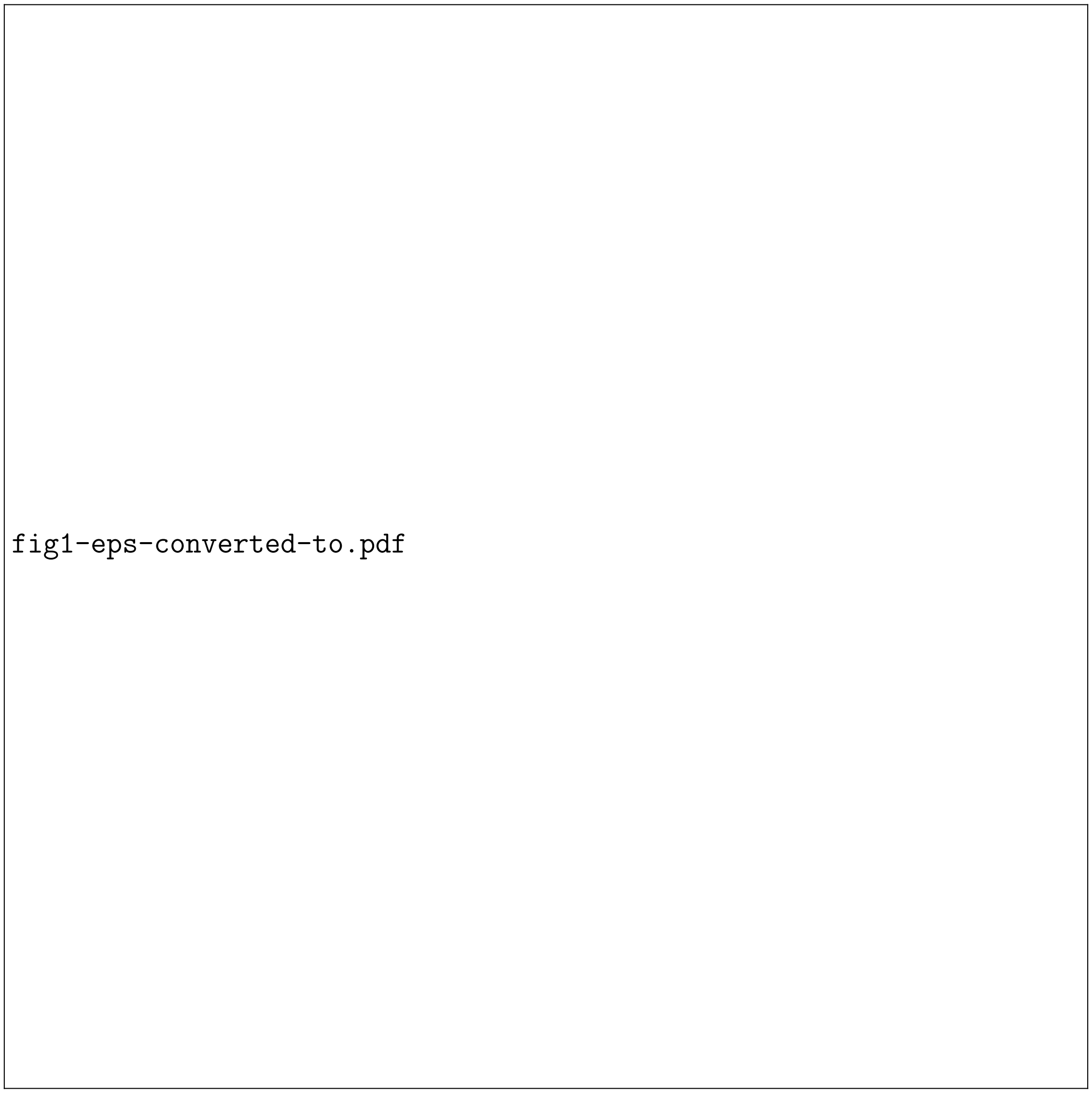

Figure 1: Figure caption goes here

Table 2: Summary Statistics of Variables of Interest

Panel A: Summary Statistics for Variables of Interest

	Mean	Std. Dev.	Min	Max
Outcome variable 1	4.127	1.709	0.000	8.516
Outcome variable 2	1.293	0.648	0.000	0.216
Policy variable	0.685	0.464	0.000	1.000
Control variable 1	0.451	0.497	0.000	1.000
Control variable 2	0.322	0.467	0.000	1.000

Panel B: Sample Means of Outcome Variables for Subgroups

	Group 1	Group 2	Group 3	Group 4
Outcome variable 1	1.782	2.181	3.749	4.127
Outcome variable 2	0.824	0.971	1.215	1.693
N	25,796	75,879	37,157	33,839

Notes: Put any notes about the table here. Sample size for all variables in Panel A is $N = 172,671$.

Table 3: Empirical estimates of parameter of interest

	Few Controls	Many Controls
Variable of interest	-1.977***	-0.536**
	(0.219)	(0.214)
Individual characteristics	✓	✓
Firm characteristics		✓
Location dummies		✓
N	172,671	172,671

Notes: Table notes here. Standard errors in parentheses. ***Significantly different from zero at the 1% level; **Significantly different from zero at the 5% level.