Data Analysis 2 Term Project

Gyongyver Kamenar (2103380)

Introduction

This paper analyze the differneces in English correctness between English learners of several different native languages.

HERE COMES THE MOTIVATION WHY THIS IS A MEANINGFUL PROJECT AND WHAT IS THE MAIN GOAL!

Data

The Massachusetts data are ... Further information is available (here)[https://www.rdocumentation.org/packages/AER/versions/1.2-9/topics/MASchools].

ECT.

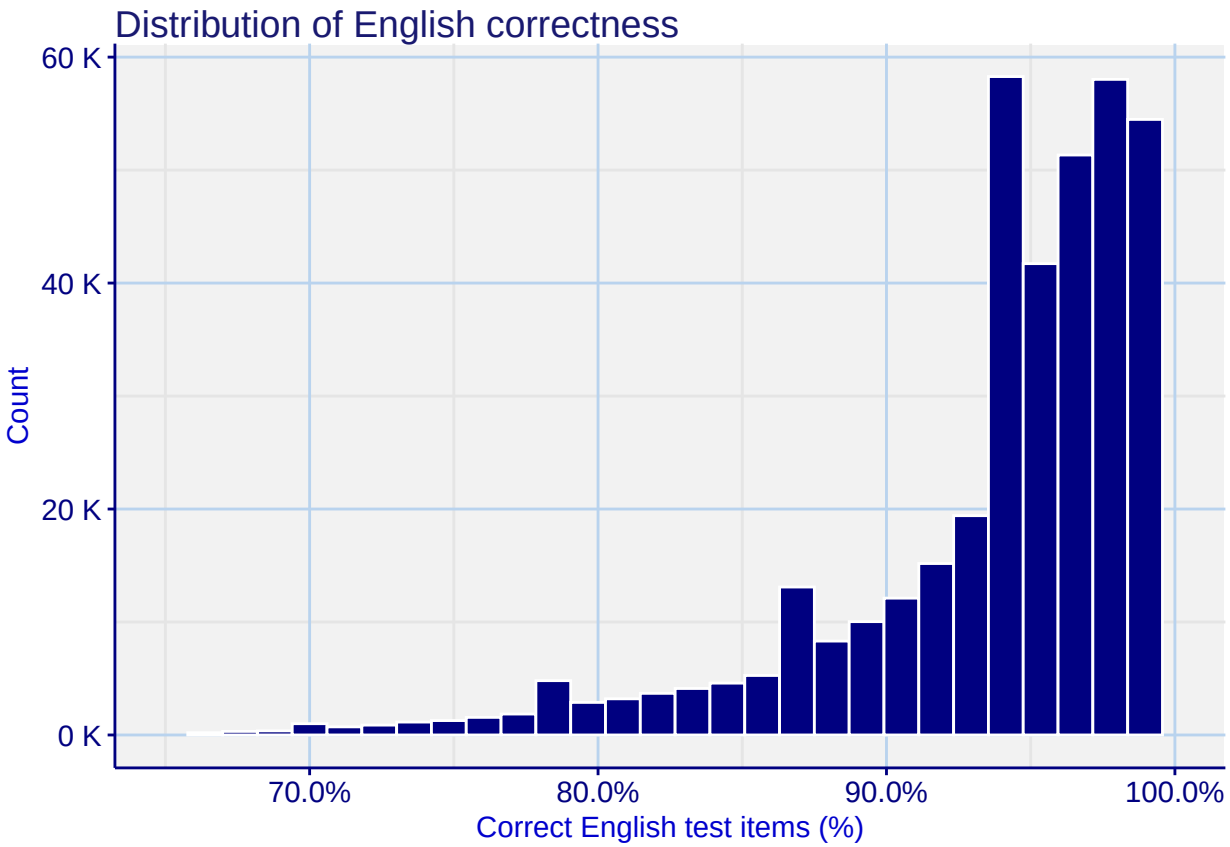
Table 1: Descriptive statistics

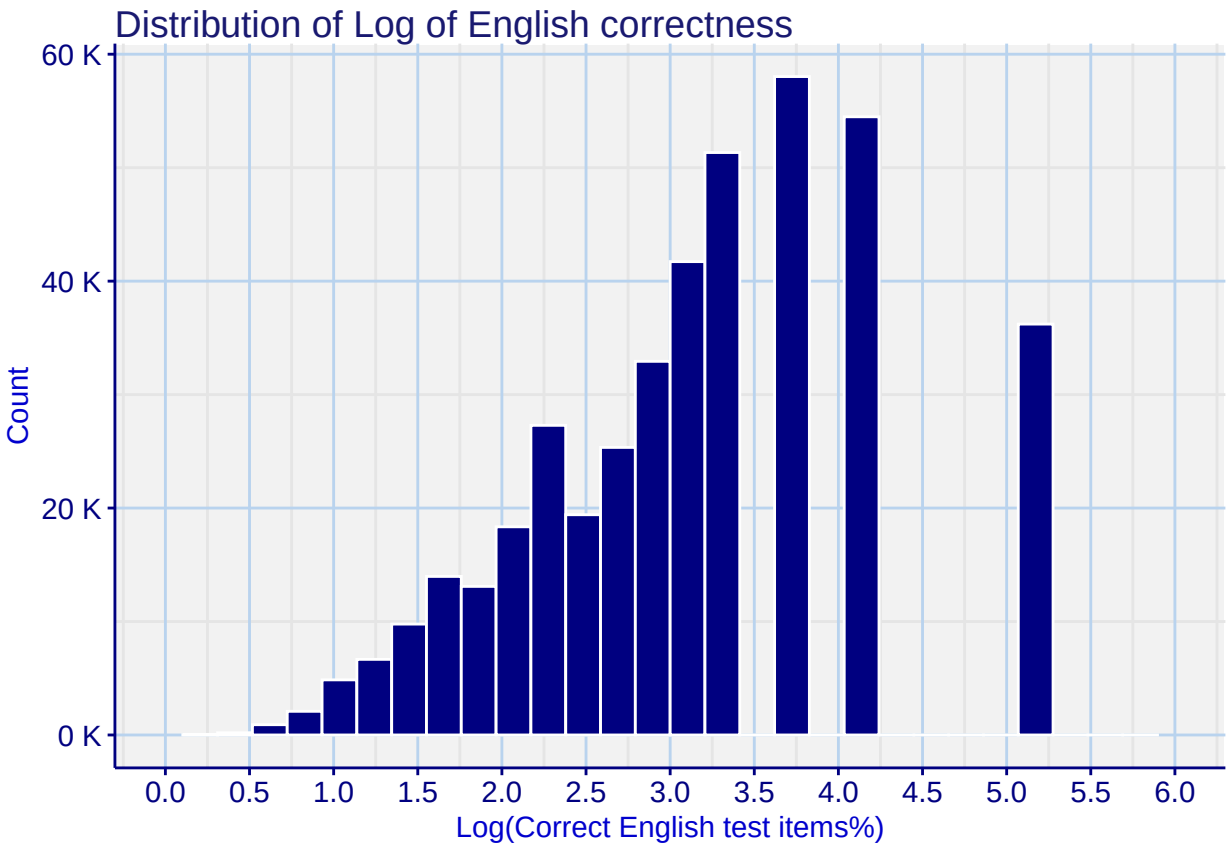
	Mean	Median	SD	Min	Max	P05	P95
Age	30.43	27.00	11.65	7.00	89.00	17.00	55.00
Age at start of English learning	4.01	0.00	5.42	0.00	69.00	0.00	13.00
Psychiatric disorder	0.03	0.00	0.18	0	1	0.00	0.00
Critical items correct (%)	0.94	0.96	0.06	0.06	1.00	0.81	1.00
Log(correct)	3.13	3.01	1.02	−2.62	5.25	1.43	5.25

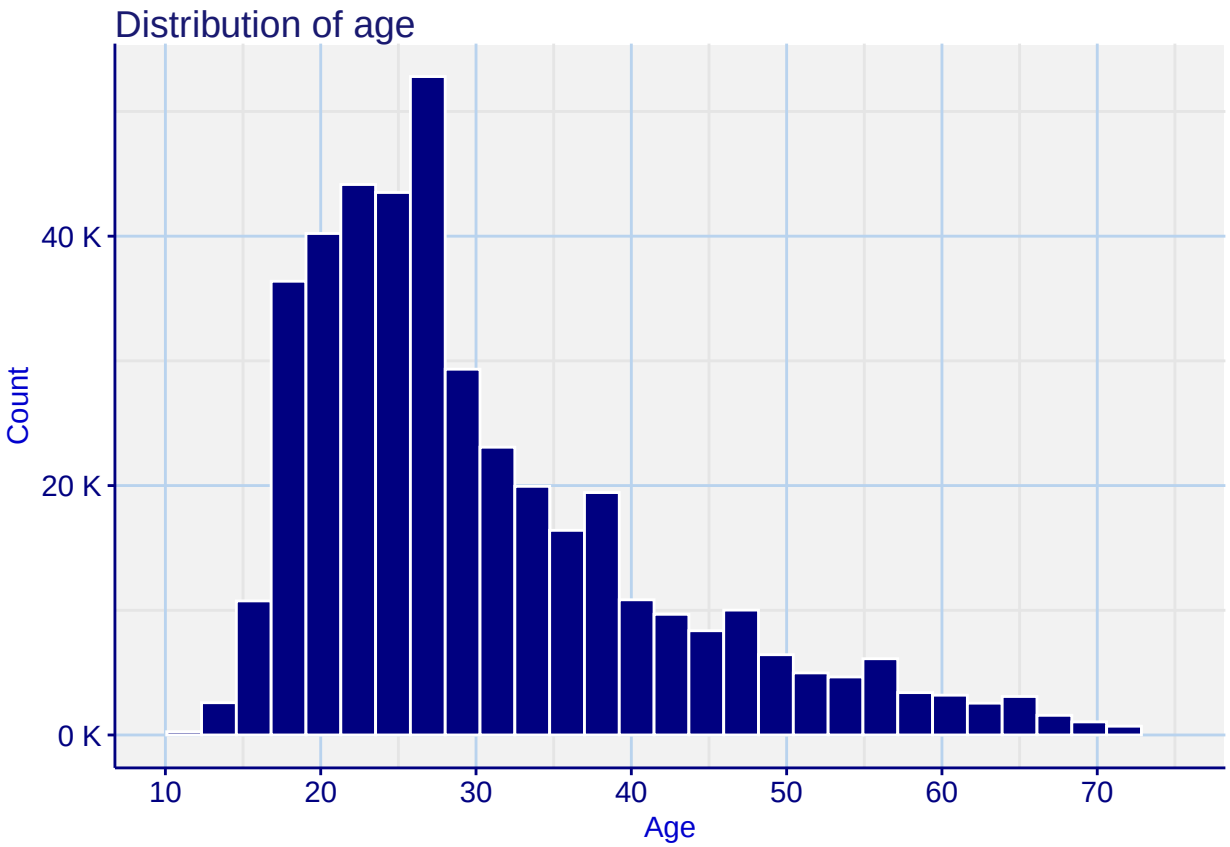
The number of observations is 0 for all of our key variables.

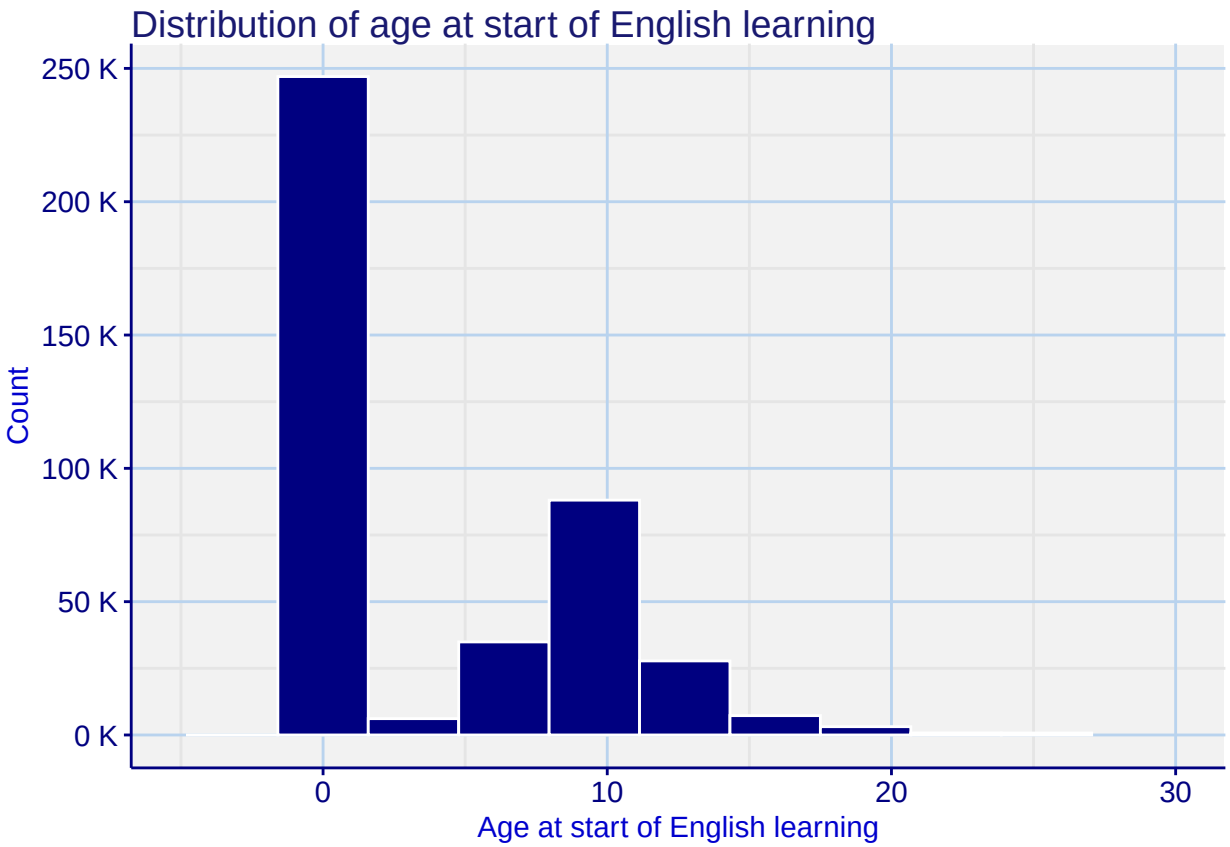
DESCRIPTION OF THE SUMMARY STATS: WHAT CAN WE LEARN FROM THEM?

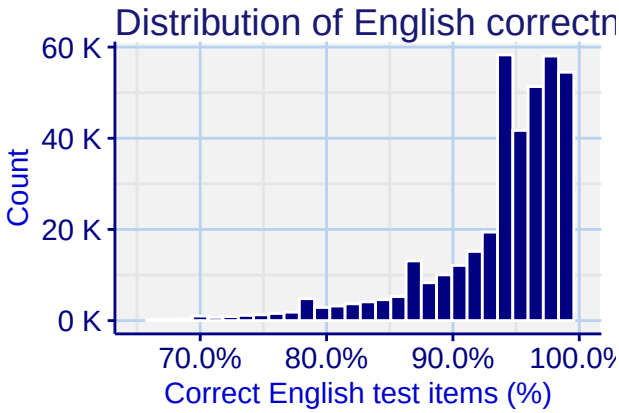
As the focus is the price difference, the next Figure shows the histogram for this variable.

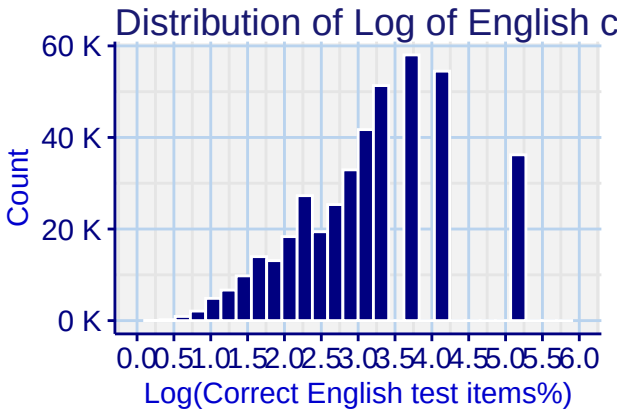

Distribution of English correctness
60 K
40 K
20 K
0 K
Count
70.0%
80.0%
90.0%
100.0%
Correct English test items (%)

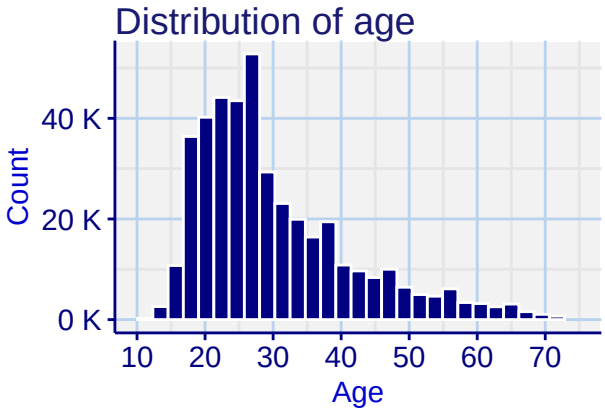

Distribution of Log of English correctness
60 K
40 K
20 K
0 K
Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Log(Correct English test items%)

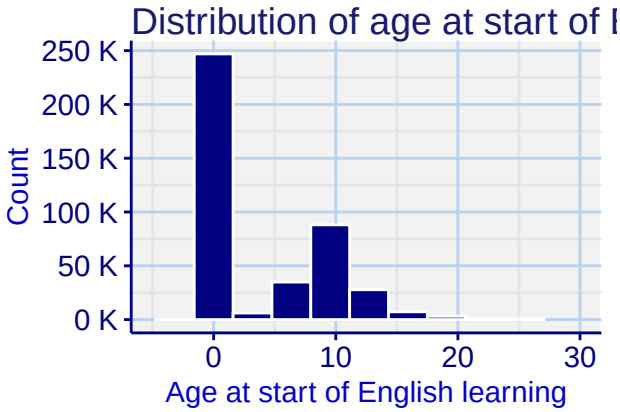

Distribution of age
40 K
20 K
0 K
Count
10
20
30
40
50
60
70
Age

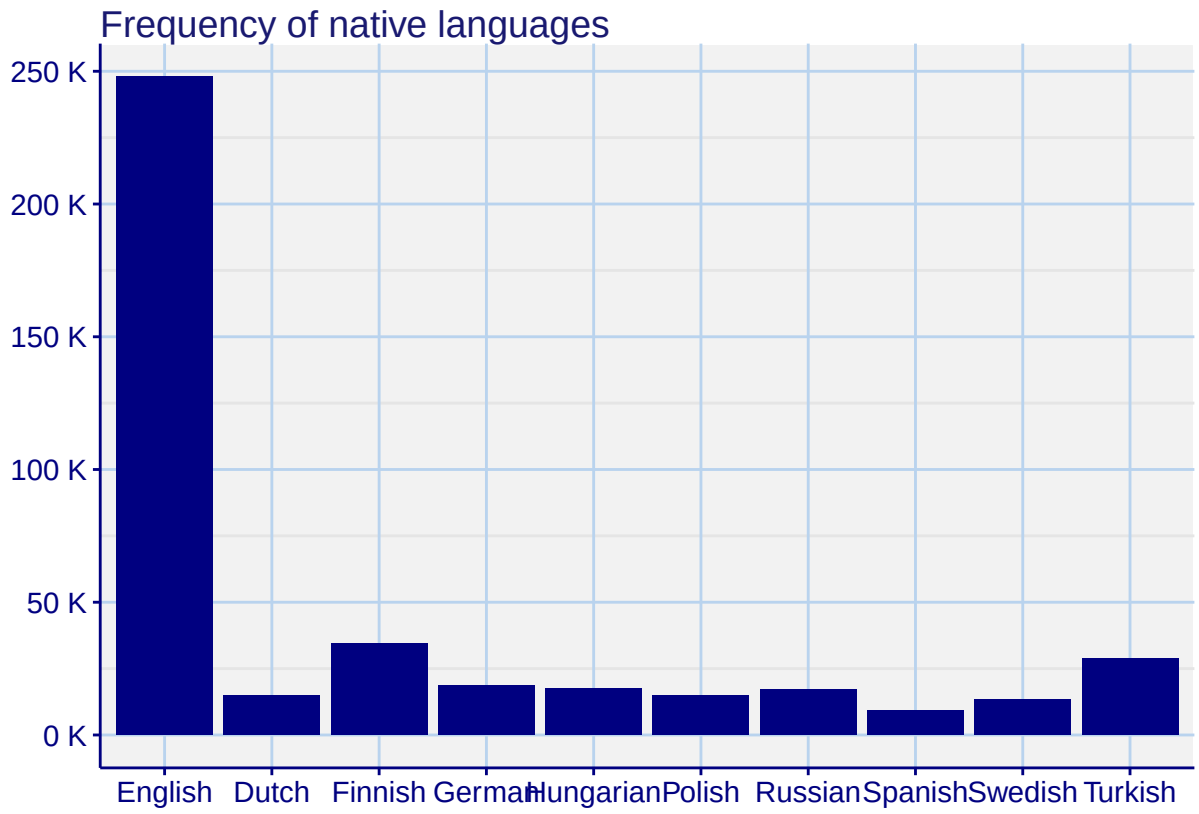

Distribution of age at start of English learning
250 K
200 K
150 K
100 K
50 K
0 K
Count
0
10
20
30
Age at start of English learning


Distribution of English correctn
60 K
40 K
20 K
0 K
Count
70.0%
80.0%
90.0%
100.0%
Correct English test items (%)


Distribution of Log of English c
60 K
40 K
20 K
0 K
Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Log(Correct English test items%)


Distribution of age
40 K
20 K
0 K
Count
10
20
30
40
50
60
70
Age

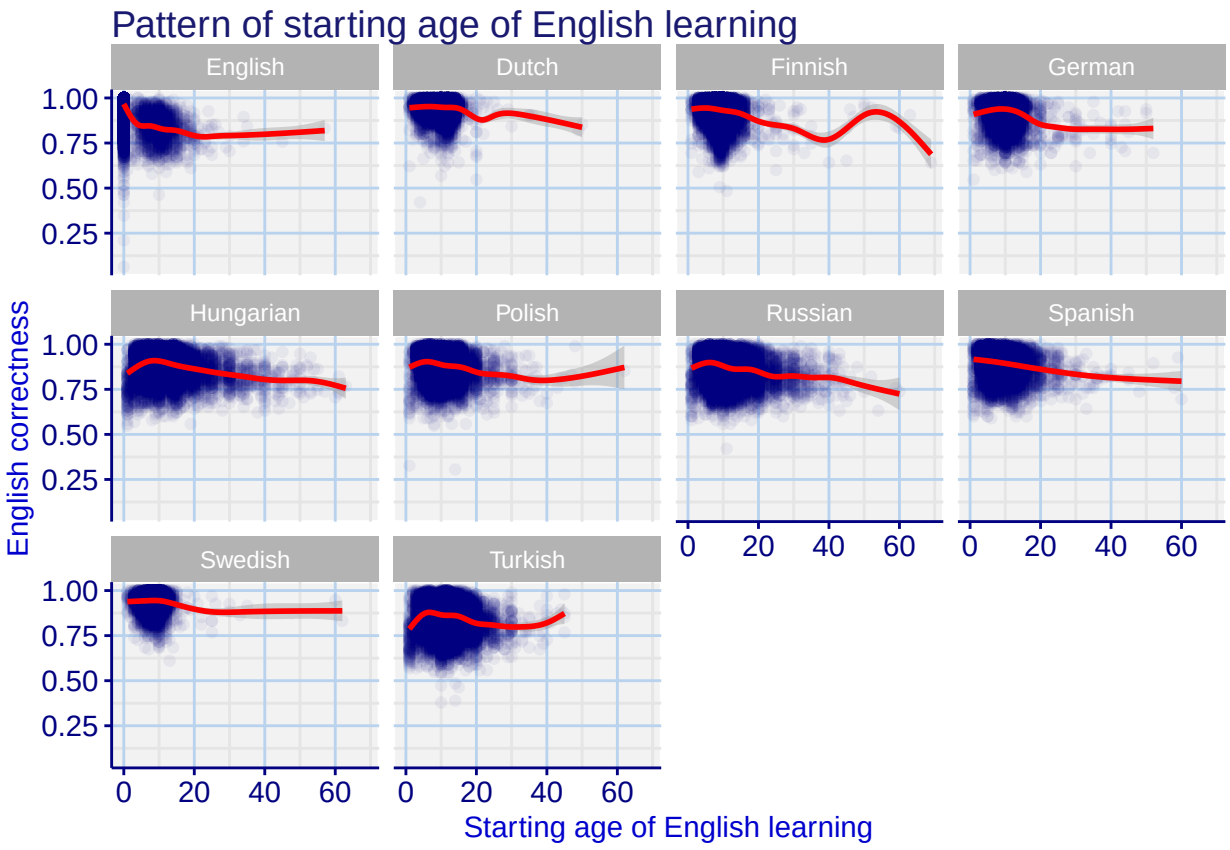

Distribution of age at start of l
250 K
200 K
150 K
100 K
50 K
0 K
Count
0
10
20
30
Age at start of English learning

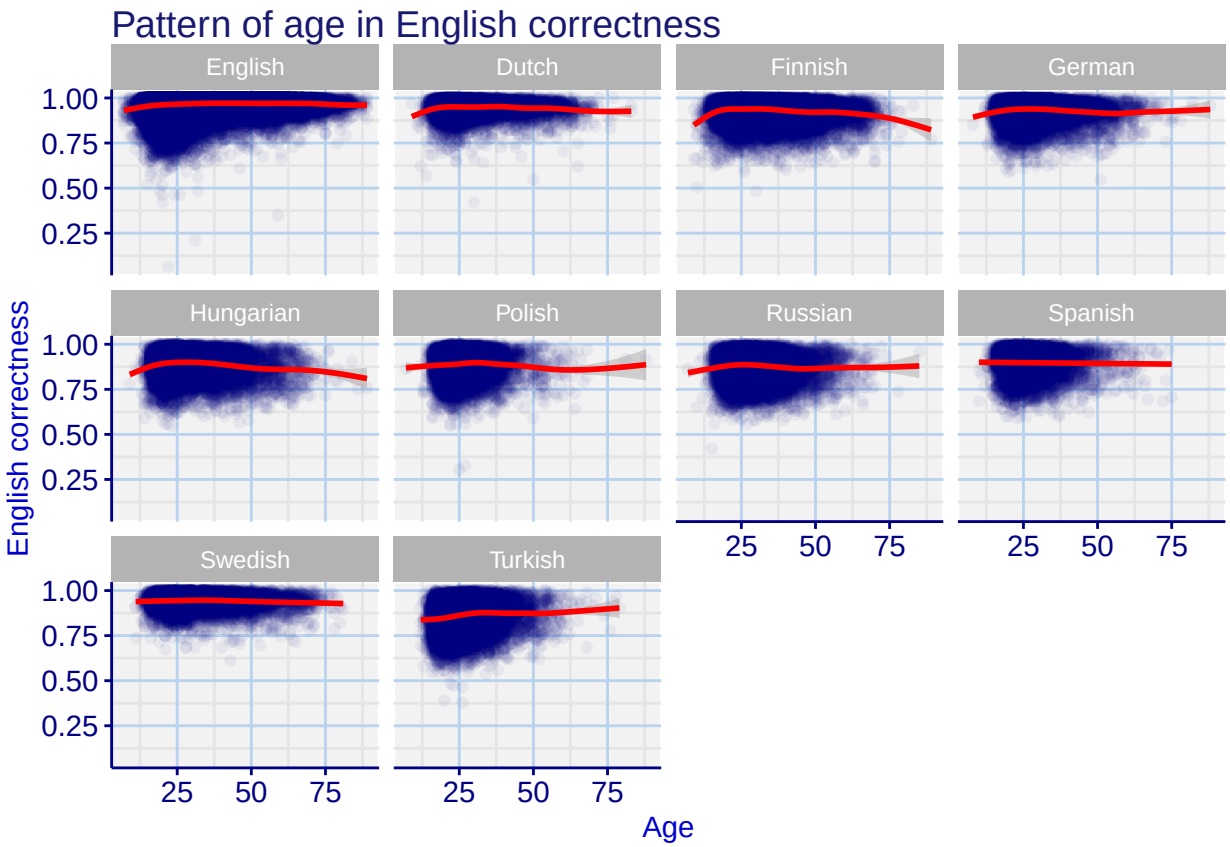


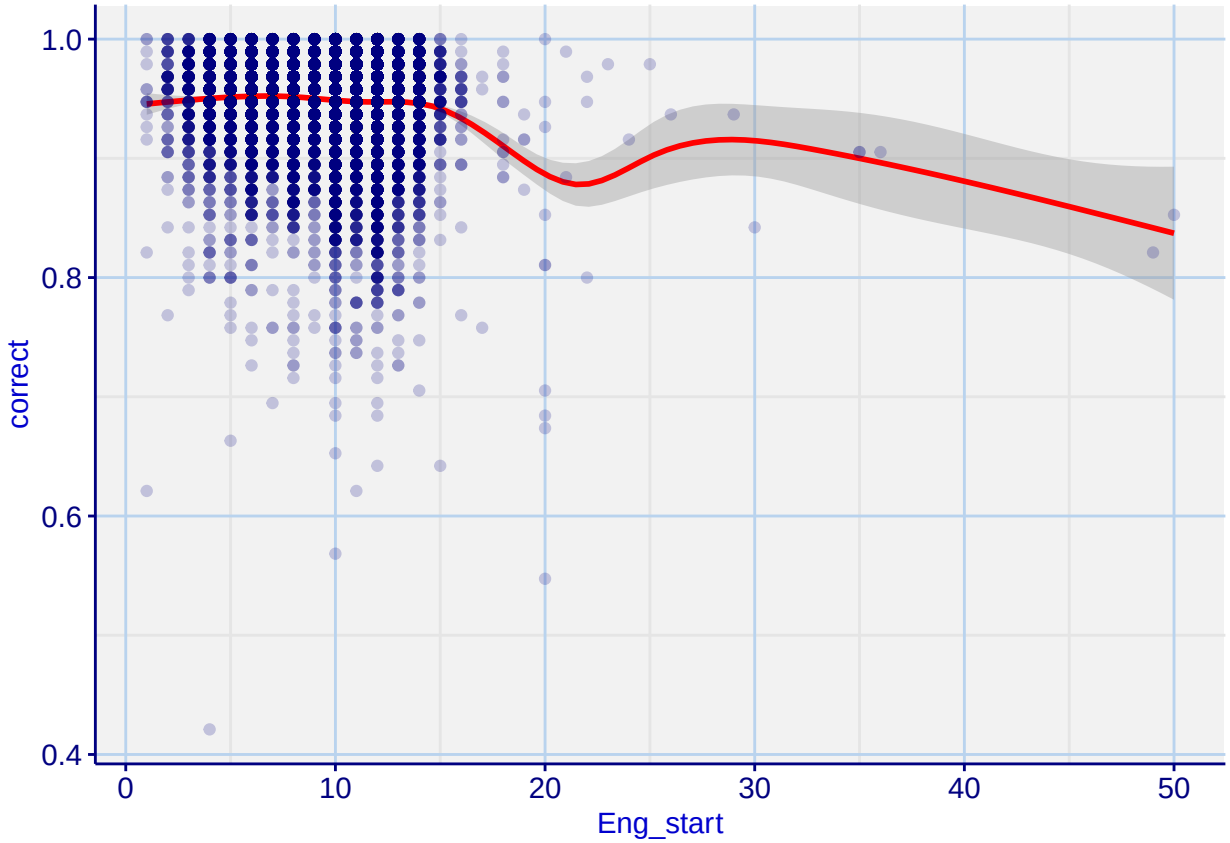
DESCRIPTION OF THE FIGURE. WHAT DOES IT TELS US?

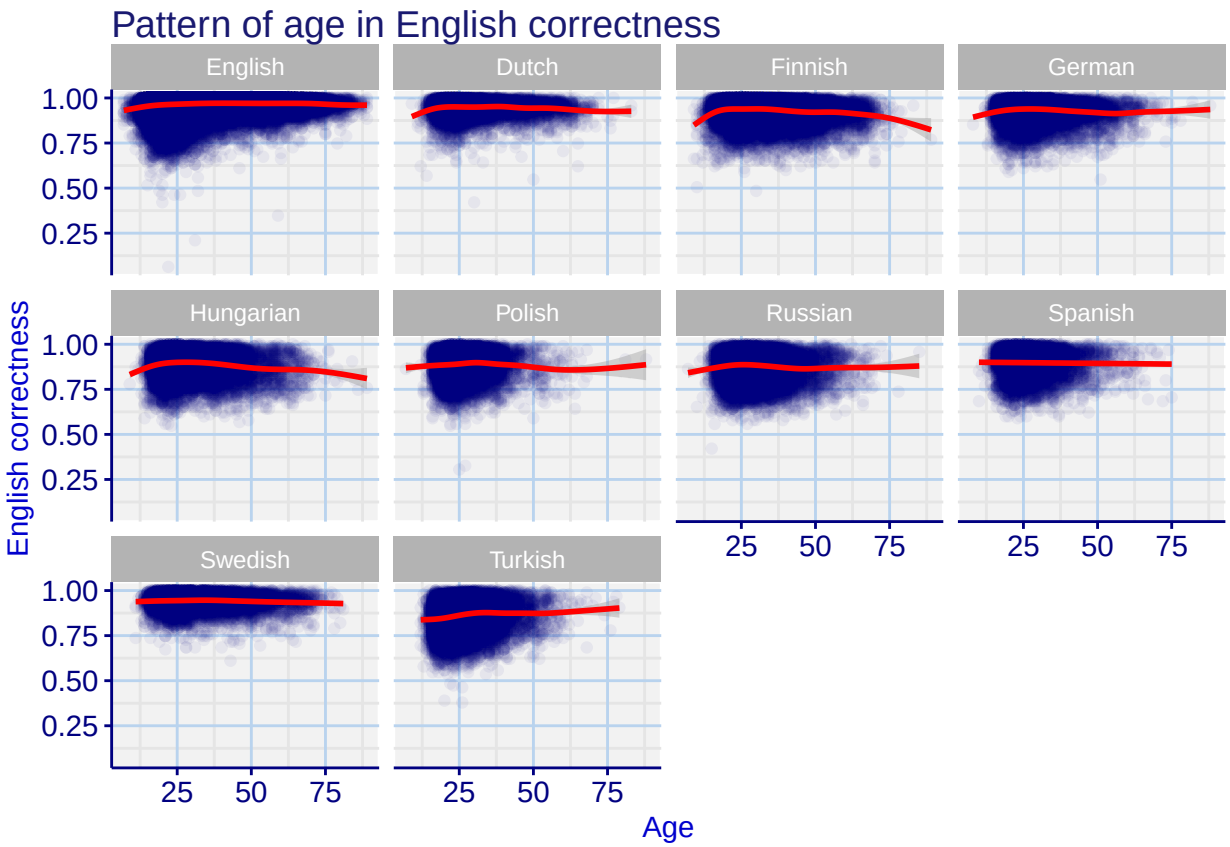
(May change the order of descriptive stats and graph.)

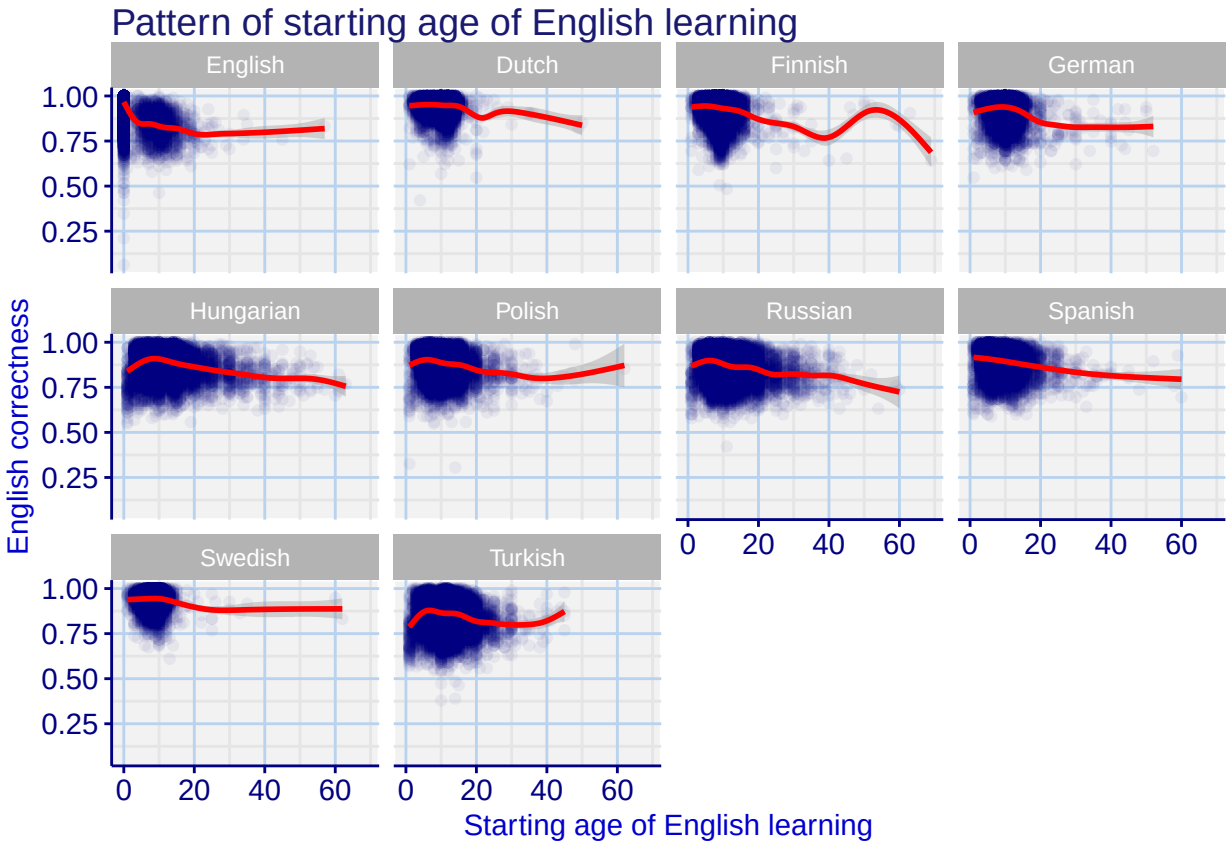
The key pattern of association is:

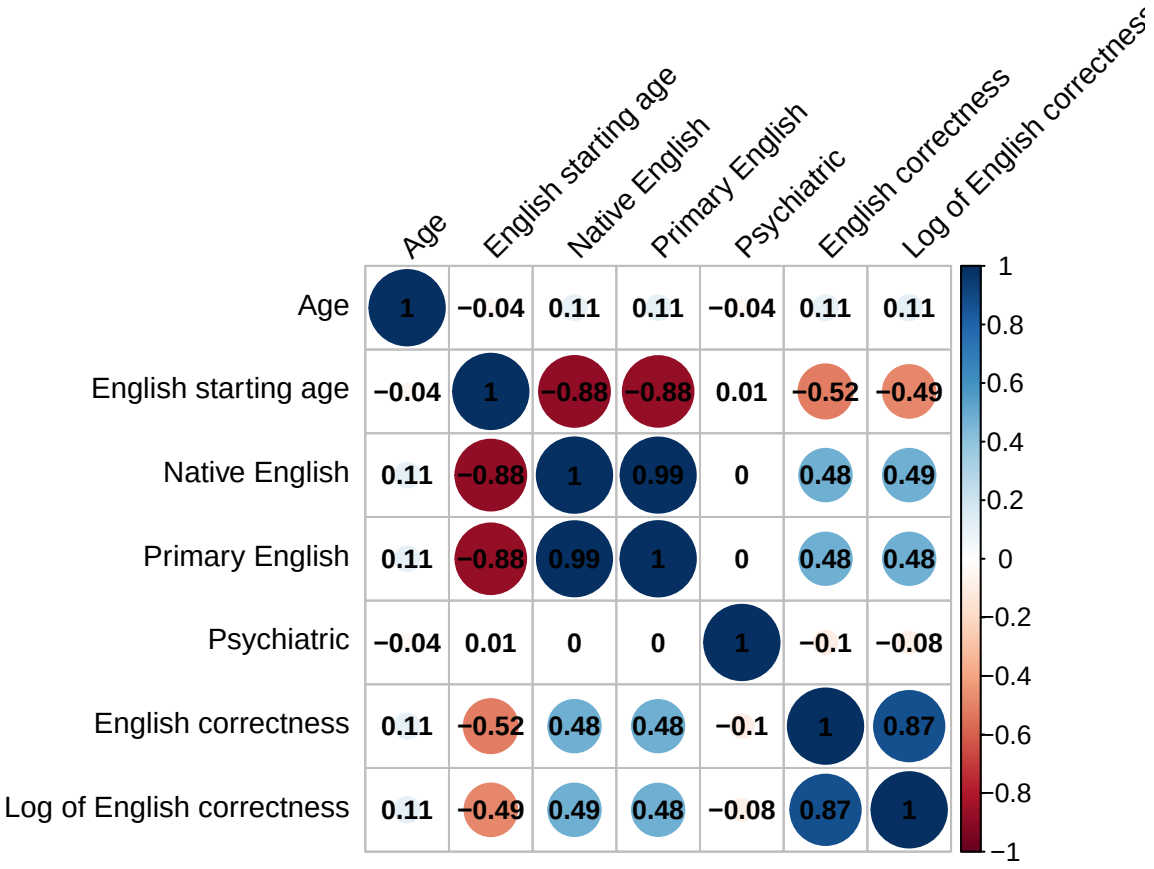

Pattern of starting age of English learning
English
Dutch
Finnish
German
Hungarian
Polish
Russian
Spanish
Swedish
Turkish
English correctness
1.00
0.75
0.50
0.25
0
20
40
60
Starting age of English learning

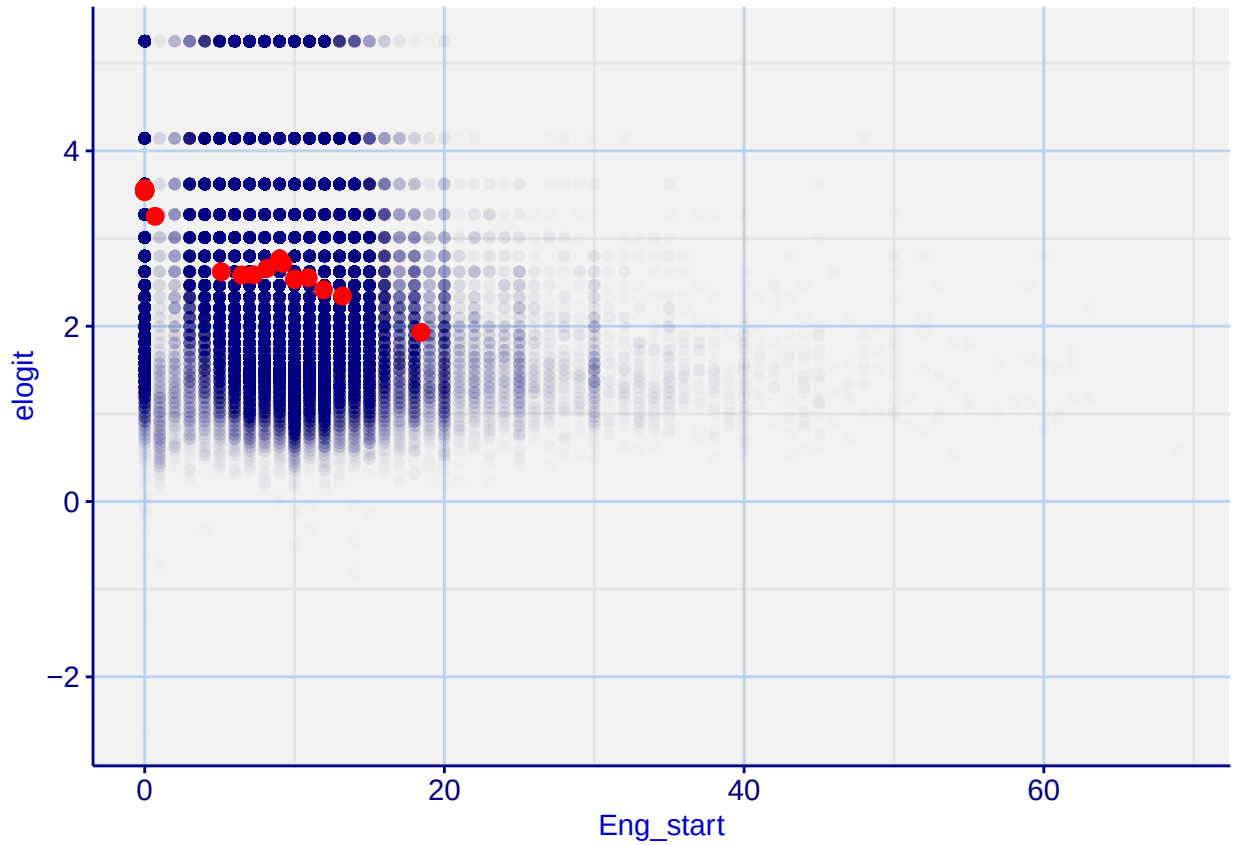

Pattern of age in English correctness
English
Dutch
Finnish
German
Hungarian
Polish
Russian
Spanish
Swedish
Turkish
English correctness
1.00
0.75
0.50
0.25
25
50
75
Age


1.0
0.8
0.6
0.4
correct
0
10
20
30
40
50
Eng_start


Pattern of age in English correctness
English
Dutch
Finnish
German
Hungarian
Polish
Russian
Spanish
Swedish
Turkish
English correctness
1.00
0.75
0.50
0.25
25
50
75
Age


Pattern of starting age of English learning
English
Dutch
Finnish
German
Hungarian
Polish
Russian
Spanish
Swedish
Turkish
English correctness
1.00
0.75
0.50
0.25
0
20
40
60
Starting age of English learning


Age
English starting age
Native English
Primary English
Psychiatric
English correctness
Log of English correctness
Age
English starting age
Native English
Primary English
Psychiatric
English correctness
Log of English correctness
1 −0.04 0.11 0.11 −0.04 0.11 0.11
−0.04 1 −0.88 −0.88 0.01 −0.52 −0.49
0.11 −0.88 1 0.99 0 0.48 0.49
0.11 −0.88 0.99 1 0 0.48 0.48
−0.04 0.01 0 0 1 −0.1 −0.08
0.11 −0.52 0.48 0.48 −0.1 1 0.87
0.11 −0.49 0.49 0.48 −0.08 0.87 1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



How will you include this in your model?

Short description on the other variables: 2-10 sentence depends on the amount of variables you have. You should reference your decisions on the graphs/analysis which are located in the appendix.

Models

My preferred model is:

score $= 0.95\ 0\ (student/teacher < 18)\ -0.01\ (student/teacher \geq 18) + \delta Z$

where Z are standing for the controls, which includes controlling for english language, lunch, other special characteristics and wealth measures. From this model we can infer:

- when every covariates are zero, students expected to have grade score of 0.95
- when the student to teacher is one unit larger, but below the value of 18, we see students to have on average 0 smaller grades.
- when the student to teacher is one unit larger, with the value above or equal to 18, we see students to have on average 0.01 smaller grades.

However, based on the heteroskedastic robust standard errors, these results are statistically non different from zero. To show that, I have run a two-sided hypothesis test:

$$H_0 := \beta_1 = 0$$

$$H_A := \beta_1 \neq 0$$

I have the t-statistic as 5.58 and the p-value as 0, which confirms my conclusion.

We compare multiple models to learn about the stability of the parameters. Bla-bla:

Table 2: Estimated models

	Simple	Multiple	Multiple with starting age	Splines	Splines
Intercept	0.966***	0.955***	0.952***	0.954***	0.954***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
natlangsDutch	−0.017***	−0.017***	0.016***	0.088***	0.112***
	(0.000)	(0.000)	(0.001)	(0.002)	(0.002)
natlangsFinnish	−0.032***	−0.031***	−0.002***	0.071***	0.095***
	(0.000)	(0.000)	(0.001)	(0.002)	(0.002)
natlangsGerman	−0.031***	−0.031***	0.002***	0.074***	0.098***
	(0.000)	(0.000)	(0.001)	(0.002)	(0.002)
natlangsHungarian	−0.073***	−0.074***	−0.037***	0.035***	0.059***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
natlangsPolish	−0.076***	−0.074***	−0.042***	0.030***	0.054***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
natlangsRussian	−0.085***	−0.086***	−0.053***	0.019***	0.043***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
natlangsSpanish	−0.068***	−0.069***	−0.041***	0.032***	0.056***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
natlangsSwedish	−0.023***	−0.021***	0.006***	0.080***	0.104***
	(0.000)	(0.000)	(0.001)	(0.002)	(0.002)
natlangsTurkish	−0.104***	−0.103***	−0.066***	0.005**	0.029***
	(0.000)	(0.000)	(0.001)	(0.002)	(0.002)
Age		0.000***	0.000***	0.000***	0.000***
		(0.000)	(0.000)	(0.000)	(0.000)
Male		−0.007***	−0.006***	−0.006***	−0.006***
		(0.000)	(0.000)	(0.000)	(0.000)
Other gender		−0.008***	−0.007***	−0.008***	−0.008***
		(0.001)	(0.001)	(0.001)	(0.001)
educationGraduate Degree		0.020***	0.020***	0.021***	0.021***
		(0.000)	(0.000)	(0.000)	(0.000)
educationHigh School Degree (12-13 years)		0.008***	0.007***	0.008***	0.008***
		(0.000)	(0.000)	(0.000)	(0.000)
educationSome Graduate School		0.011***	0.011***	0.011***	0.012***
		(0.000)	(0.000)	(0.000)	(0.000)
educationSome Undergrad (higher ed)		0.014***	0.013***	0.013***	0.013***
		(0.000)	(0.000)	(0.000)	(0.000)
educationUndergraduate Degree (3-5 years higher ed)		0.019***	0.018***	0.018***	0.018***
		(0.000)	(0.000)	(0.000)	(0.000)
psychiatric		−0.029***	−0.029***	−0.029***	−0.028***
		(0.001)	(0.001)	(0.001)	(0.001)
Starting age of English learning			−0.003***		
			(0.000)		
lspline(Eng_start, 2)1				−0.042***	
				(0.001)	
lspline(Eng_start, 2)2				−0.003***	
				(0.000)	
lspline(Eng_start, 1)1					−0.106***
					(0.002)
lspline(Eng_start, 1)2					−0.003***
					(0.000)
Num.Obs.	416 738	416 738	416 738	416 738	416 738
R2	0.349	0.369	0.389	0.395	0.399
R2 Within					
R2 Pseudo					
BIC	−1 363 916.1	−1 377 114.4	−1 390 021.3	−1 394 382.8	−1 397 280.0
Log.Lik.	682 022.741	688 680.140	695 140.042	697 327.286	698 775.882
Std.Errors	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust	Heteroskedasticity-robust

** $p < 0.05$, *** $p < 0.01$

Robustness check / 'Heterogeneity analysis'

Task: calculate and report t-tests for each countries.

Conclusion

HERE COMES WHAT WE HAVE LEARNED AND WHAT WOULD STRENGHTEN AND WEAKEN OUR ANALYSIS.

Appendix

Here comes all the results which are referenced and not essential for understanding the MAIN results.