**Report of Multilingualism in Basque Country**

Yuehan (Hailey) Liu

Division of the Humanities, The University of Chicago

DIGS 20002 1: Data Analysis for the Humanities

December 10, 2021

The current analysis aims to provide insights of the multilingual situation including language use, proficiency, dominance and exposure in Basque Country. Four research questions are addressed in the following analysis:

1. Is there a difference in language proficiency between the three languages spoken in Basque Country for all four measurements?
2. Is there a correlation between the proficiency in each language spoken in Basque Country and its respective percentage exposed?
3. Is there a correlation between the proficiency in each language spoken in Basque Country and its respective age of acquisition?
4. Is there a difference in perceived English proficiency between people with different educational level?

The **B**asque, **E**nglish, and **S**panish **t**ests (BEST) dataset includes information from four different subtests divided into three language blocks. A group of 650 (435 female) participants completed various language proficiency measurements: picture naming test, LexTALE test, interview, and self-perceived level. For the purpose of this report, the procedures and rubrics of these measurements will not be discussed. Participants’ ages ranged from 18 to 50 years (mean = 25.02, *SD* = 5.58). The maximum level of education achieved at the time of testing ranged from high school to university, although the majority of participants (80%) reached a higher level of education (professional training, university, or a postgraduate degree). All participants were Spanish-Basque-English trilinguals and they had acquired Basque and Spanish before the age of six (mean AoASpanish = 0.67, *SD* = 1.55; mean AoABasque = 1.68, *SD* = 1.81). On average, English was acquired at a later age (mean AoAEnglish = 6.37, *SD* = 2.49), but all participants reported having acquired English at or before age 12.

To discern whether there are real differences between people’s proficiencies in Spanish, Basque and English in Basque Country, a one-way ANOVA test was performed four times for the four measurements. For the picture naming test, the boxplots (Figure 1) suggest that the results for Spanish are quite unanimous with a few outliers scoring between 50 and 60. The mean and median are both around 64, and the interquartile range is extremely small. This is because Spanish is the native language for most people in Basque Country, therefore the test score is intensively high. Scores for Basque and English show a little more variation with wider interquartile range and more extended whiskers. Among all, English score has the largest spread and the lowest mean and median. This remains consistent across four measurements. The major cause of this result is that English is the learned foreign language for Basque Country people, and most speakers have shorter years of acquisition and less exposure of English.

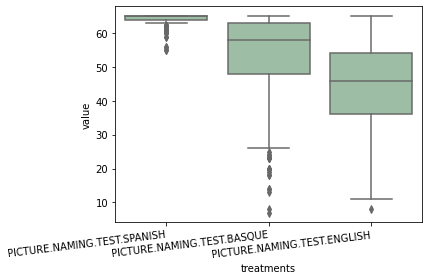


Figure 1: Language proficiencies in Spanish, Basque, and English (Picture naming test)

The ANOVA output for picture naming test (*df*G = 2, *df*E = 647) gives a p-value of 7.275071e-229. The p-value is smaller than 0.01, indicating the evidence is strong enough to reject the null hypothesis at a significance level of 0.01. So, in this test, the differences between Basque Country people’s language proficiencies in Spanish, Basque and English are statistically significant.

For the LexTALE test, the boxplots (Figure 2) follow similar pattern with wider spread and generally lower means and medians as compared to the maximum score for all three languages.

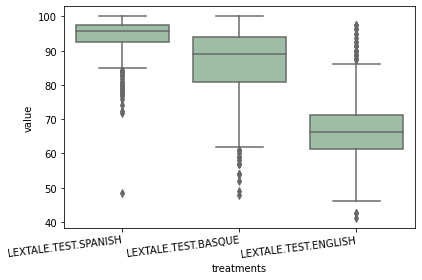


Figure 2: Language proficiencies in Spanish, Basque, and English (LexTALE test)

The ANOVA outcome for LexTALE test (*df*G = 2, *df*E = 647) gives a p-value of 0.0. In practice, the p-value could never be exactly 0.0[[1]](#footnote-1), so the output value might be due to rounding error. This p-value announces overwhelming evidence to reject the null hypothesis at a significance level of 0.01. So, in LexTALE test, the differences between Basque Country people’s language proficiencies in Spanish, Basque and English are statistically significant.

For the interview, since the full mark is only 5, the boxplots (Figure 3) for the three languages show little variation and all participants score 5 for Spanish. Probably due to the same reason, fewer outliers are detected.

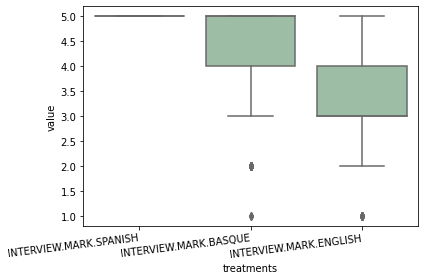


Figure 3: Language proficiencies in Spanish, Basque, and English (Interview)

The ANOVA output for interview (*df*G = 2, *df*E = 647) yields a p-value of 1.372281e-258. The p-value is also smaller than 0.01 thus the null hypothesis is rejected and statistically significant differences are found in proficiency between languages.

For the self-perceived level, the full mark is 10, which also has a relatively small range, so the boxplots’ (Figure 4) pattern follows the previous one but with lower mean and median, revealing people’s diverse perception of their language proficiencies due to possible confidence reasons. the ANOVA output (*df*G = 2, *df*E = 647) has a p-value of 1.891469e-226. Again, the p-value is smaller than 0.01 and signifies statistically significant differences between the means of test scores for the three languages.

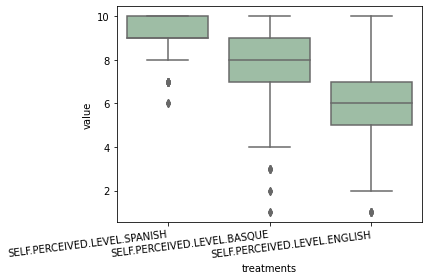


Figure 4: Language proficiencies in Spanish, Basque, and English (Self-perceived level)

To sum up the results of the first research question, ANOVA outcomes of all four measurements appear to be smaller than 0.01, declaring statistical differences in Basque Country people’s proficiencies across three languages. This conclusion is not surprising since a discrepancy is expected in speakers between their native, second, and foreign language.

For the second research question, a linear regression was conducted between the percentage exposed of a language and its proficiency. Before going into one particular measurement, a correlation matrix of all proficiency measurements was created (Figure 5). The matrix displays all possible Pearson’s correlation coefficient (r) between each two measurement variables. Spanish interview scores are not included as all participants scored the maximum score. The correlations between the different tests in each language revealed that despite their similar aim, the indications offered are complementary and that the additional information provided by each indicator is necessary. Although some tests correlated quite highly, it is worth noting that none of the correlations were close to ceiling. Furthermore, correlations between self-ratings and several of the objective tests were relatively low, suggesting that self-ratings alone are not an optimal reflection of proficiency. Hence, taking multiple objective and subjective indices together provides a more complete understanding of the participants’ language proficiencies.

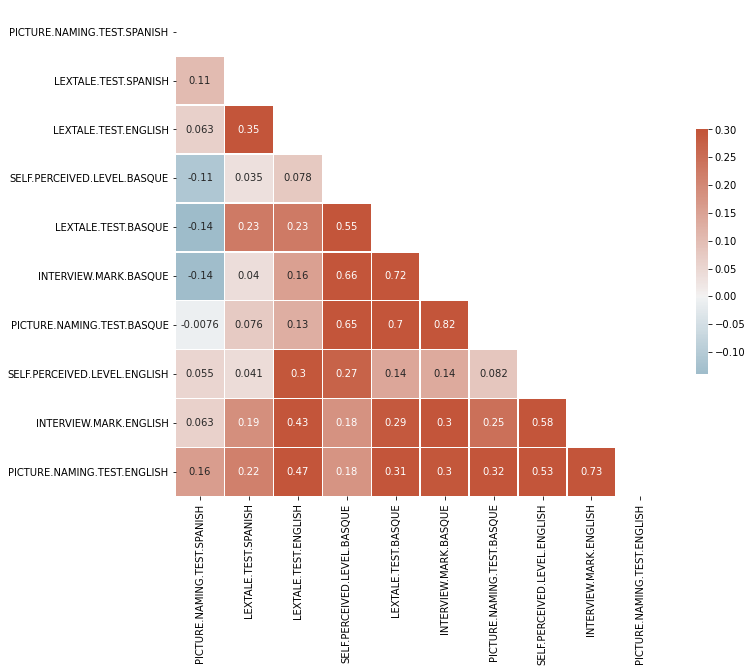


Figure 5: Correlation matrix showing the Pearson’s correlation coefficient (r) between Spanish, Basque, and English self-rated proficiency, interview scores, picture naming scores, and LexTALE scores

For the purpose of plotting a linear regression to see specific trends in the correlation between percentage exposed and language proficiency, test scores of the picturing naming test and LexTALE test are chosen considering they have the highest maximum score and thus the largest range. Also worth mentioning, the variable *Percentage Exposed* is a categorical variable (the values are 10, 20, 30...till 100), making the data points in regression models discrete with stop at each scale instead of being continuous.

For English and taking the picture naming test scores as the representation of language proficiency, the overall regression is statistically significant (R2 = 0.116, F (1, 648) = 84.62, p < .001), and the fitted regression model (Figure 6) is: English proficiency = 40.0199 + 0.4295 \* (percentage exposed).

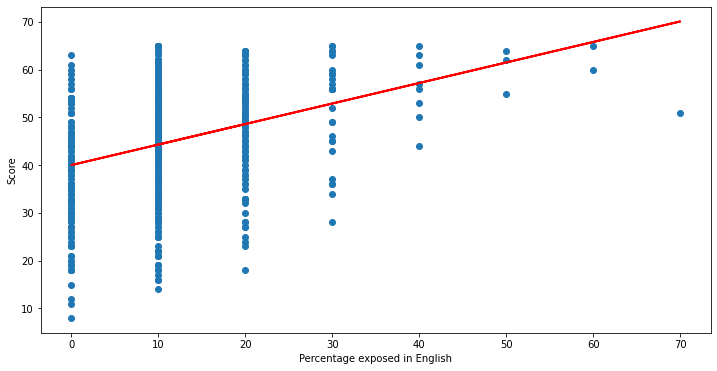


Figure 6: A linear model representing the relationship between percentage exposed in English and English test score (Picture naming test)

For Basque, the overall regression is statistically significant (R2 = 0.253, F (1, 648) = 219.0, p < .001), and the fitted regression model (Figure 7) is: Basque proficiency = 44.0964 + 0.3081 \* (percentage exposed).

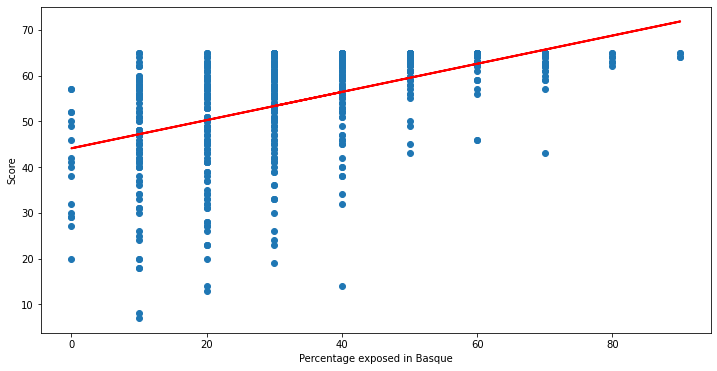


Figure 7: A linear model representing the relationship between percentage exposed in Basque and Basque test score (Picture naming test)

For Spanish, the overall regression is statistically significant (R2 = 0.040, F (1, 648) = 27.20, p < .001), and the fitted regression model is (Figure 8): Spanish proficiency = 63.8523 + 0.0119 \* (percentage exposed).

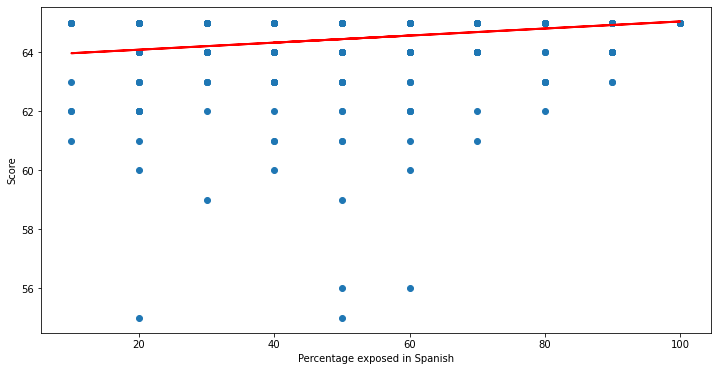


Figure 8: A linear model representing the relationship between percentage exposed in Spanish and Spanish test score (Picture naming test)

Turning into another measurement LexTALE Test, the results for English suggests a statistically significant regression model (R2 = 0.083, F (1, 648) = 58.85, p < .001), and the prediction equation is (Figure 9): English proficiency = 63.5756 + 0.2845 \* (percentage exposed).

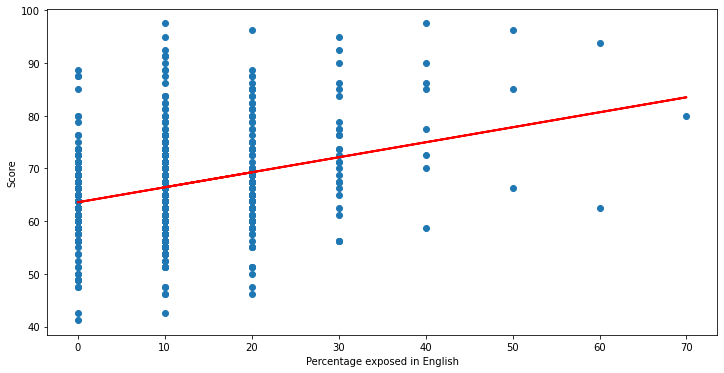


Figure 9: A linear model representing the relationship between percentage exposed in English and English test score (LexTALE test)

For Basque, a significant regression (Figure 10) is also found (R2 = 0.221, F (1, 648) = 184.2, p < .001), with an equation of Basque proficiency = 77.9315 + 0.2571 \* (percentage exposed).

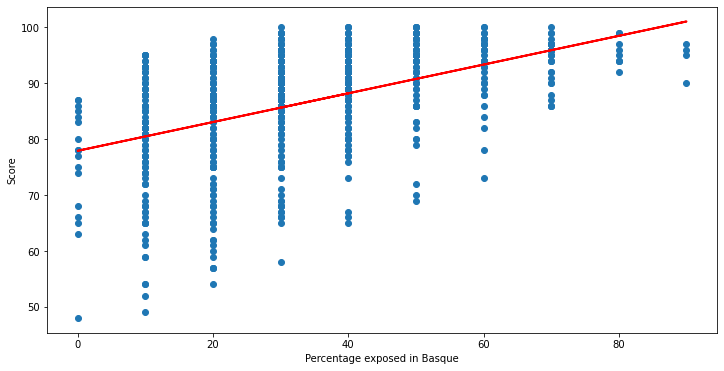


Figure 10: A linear model representing the relationship between percentage exposed in Basque and Basque test score (LexTALE test)

For Spanish, the situation is distinct, the outcome of the regression model (Figure 11) pronounces no significant result to interpret (R2 = 0.00, F (1, 648) = 0.04306, p > .005), and the figure implies that no shift in language proficiency is detected with change in percentage exposed.

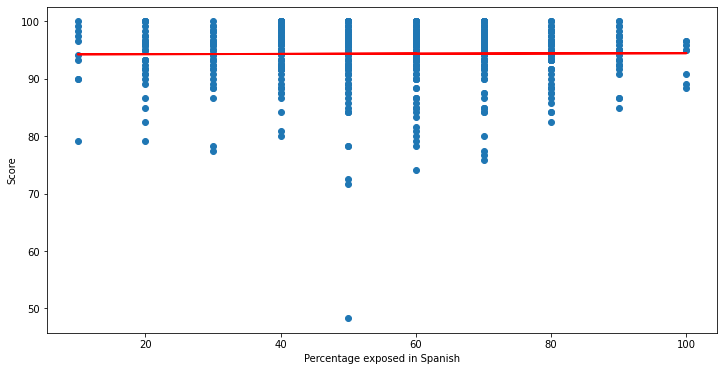


Figure 11: A linear model representing the relationship between percentage exposed in Spanish and Spanish test score (LexTALE test)

Conclusively speaking, the more a speaker is exposed to a language, the more proficient they get. Percentage expose also determines the lower limit of language proficiency: with increase in exposure, the minimum proficiency score increases steadily. The only exception under both measurements is Spanish, the regression model conveys no apparent pattern in proficiency with change in exposure. This result is anticipated with the descriptive information of Spanish test grades listed previously: people score concertedly high in their native language. From the F statistics, it can be inferred that unlike in Basque and English, occasional low scores in Spanish happened by chance owing to possible reasons like carelessness or distraction during test.

Similar to previous analysis, another linear regression was conducted to see whether year of acquisition significantly predicted language proficiency. A new variable *Years of Acquisition* was created by subtracting *Age of Acquisition* from *Age in Years* to denote how many years one has acquired the particular language. Both scores of picture naming test and LexTALE test are also selected for the same reason in modeling, but the results for picture naming test are omitted since all R-squared for the three languages are lower than that of LexTALE test given same p-value outcomes. For English, the regression appears (Figure 12) to be statistically significant (R2 = 0.033, F (1, 648) = 21.98, p < .001), and the fitted line is: English proficiency = 60.1843 + 0.3520 \* (years of acquisition).

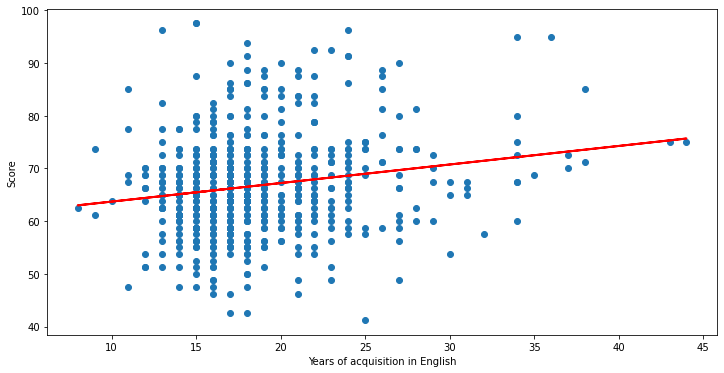


Figure 12: A linear model representing the relationship between years of acquisition in English and English test score (LexTALE test)

For Basque, a significant regression (Figure 13) is also found (R2 = 0.041, F (1, 648) = 27.67, p < .001), and the predicted regression model is: Basque proficiency = 78.1919 + 0.3494 \* (years of acquisition).

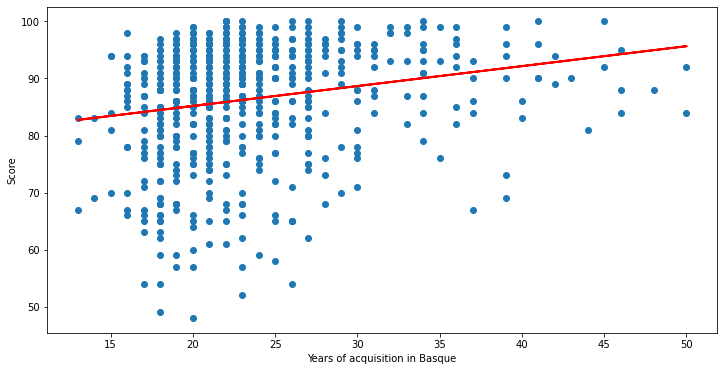


Figure 13: A linear model representing the relationship between years of acquisition in Basque and Basque test score (LexTALE test)

For Spanish, the overall regression (Figure 14) is statistically significant (R2 = 0.048, F (1, 648) = 32.91, p < .001), and the equation goes: Spanish proficiency = 89.4767 + 0.2011 \* (years of acquisition).

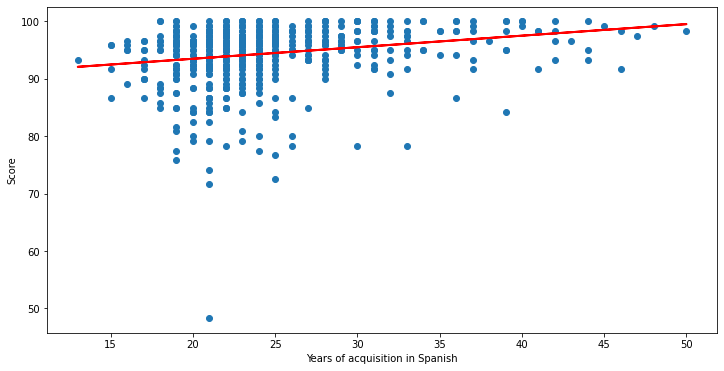


Figure 14: A linear model representing the relationship between years of acquisition in Spanish and Spanish test score (LexTALE test)

From the figures of the three regression models, we can speculate that *Language Proficiency* and *Years of Acquisition* are also positively correlated. It can also be observed that the points distribution is more tightly clustered than previous analysis. Furthermore, English scores seem to be more scattered given *Years of Acquisition* since there tend to be more variation in learned language due to disparity in learning capacity, educational background, and other potential parameters. And Spanish, as the participants’ native language, shows greatest performance regardless of their acquisition length.

However, the reported results for research question 3 should be approached with caution since the R-squared for the three models are relatively low and thus only few proportions of variance in the dependent variable are predicted by the independent variable in practice. Various factors might have contributed to the limited explanatory power. First, standard of R-squared is subject to specific field of study and objective of research. Any field that attempts to predict human behavior, like the present one for language aptitude, is inclined to have lower expected R-squared[[2]](#footnote-2). Secondly, the regression has only one independent variable, in other words, it is hard to account for much variability in the dependent variable when there are still many intricacies being left out from the analysis. And a low R-squared should also not be the sole determinant in deciding whether the predictor is significant. Thirdly, the heterogeneity of the large sample tested may render one single measurement of language proficiency unconvincing, therefore a precise prediction is not the inherent requirement for language aptitude investigations.

Last but not least, to decide whether the English proficiencies of people with different educational level in Basque Country is statistically distinct, p-value for hypothesis test is calculated based on merged variables. The categorical variable *Maximum Educational Level* has four types of values: High School, Professional Training, University, and Postgraduate. To make the difference in educational level more salient and number of each category more balanced, data for High School and Professional Training are combined into a new category of *Before University* (n = 201), and data for University and Postgraduate into *After University* (n = 449). Also, the LexTALE test is selected in consistent with previous explanation. The T-test outcome indicates that there is a significant difference in English proficiency between *Before University* (M = 65.3855, *SD* = 8.7369) and *After University* (M = 67.3608, *SD* = 9.3873); t (648) = -2.5321, p = 0.0115.

There are still so much more to be explored in the dataset that is not included in the current analysis, and here are some possible contributions that would make the study more valuable. To begin with, more variables can be included in regression models to generate more well-rounded and interpretable results. For example, gender can be added in the examination of research question 2 and 3 to see how different in trends are male and female in their language capacity given changes in percentage exposed and years in acquisition. Secondly, no classification of sample is conducted in this study, and all participants are taken homogenous in statistical calculations. A clustering analysis can be performed using diagnostic indices (i.e., interview, self-perceived proficiency, LexTALE tests and picture-naming tests) to establish the potential subgroups of people with respect to their English and Basque linguistic skills (as Spanish proficiency was close to ceiling for all participants). Last but not least, more data triangulation can be carried out to increase the credibility and validity of previous results.

Despite a multi-measure approach to better estimate multilinguals’ language skills is adopted in this dataset, certain ambiguities and deficiencies are still identified. First, it remains questionable to what degree can language test score represent the true proficiency of speakers. In language acquisition studies, proficiency and performance are generally perceived as two separate concepts: proficiency refers to the measurement of how well an individual has mastered a language, whereas performance is the ability to use a language in real-world, limited, or controlled situations[[3]](#footnote-3). Language tests used in data collection of this dataset definitely belong to the latter scenario, where participants are placed in a controlled situation-based test environment. Therefore, test scores of these kinds of measurements do not necessarily reflect speakers’ knowledge and mastery of a language, rather, only their performances in the given situation. With this in mind, the validity of the data becomes somehow suspicious. In addition, the variable *Percentage Exposed* should also be applied critically. The proportion of each language used in real-world setting is hard to be quantified into clear-cut numerical terms. Also, Basque is a contextually present language in the Basque Country while English is a foreign language whose presence is mainly restricted to academic contexts. Considering all these restrictions, the values for *Percentage Exposed* might not be sufficiently valid and accurate.

In conclusion, this study offers a partial delineation of the linguistic reality in Basque Country. The present analysis discovered statistically significant differences in people’s language proficiencies of Spanish, Basque and English in Basque Country under multiple language proficiency measurements. Furthermore, positive correlations were detected between language proficiency and percentage exposed, as well as between language proficiency and years of acquisition. However, due to practical reasons and extraneous factors, these results should be taken with a grain of salt. In addition, significant difference in subsequently learned language (English) proficiency between groups of people with distinct educational level was observed, highlighting the impact of education on language aptitude and learning capacity. Future efforts can be made to incorporate more parameters in model construction, and closer calibration in model execution.

1. Statology. (November 27, 2018). Here is How to Interpret a P-Value of 0.000. https://www.statology.org/here-is-how-to-interpret-a-p-value-of-0-000/ [↑](#footnote-ref-1)
2. Minitab Blog Editor. (2013, May 30). Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? Retrieved from https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit [↑](#footnote-ref-2)
3. Elder, C (1994) Performance testing as benchmark for foreign language teacher education. *Babel: Journal of the Australian Federation of Modern Language Teachers Associations* *29*(2): 9–19. [↑](#footnote-ref-3)