Cognate-effect Final Report

Ziyi Bai; Yu Du; Xiaozhou Lu; Jinzhe Zhang

12/18/2020

Abstract

In this project, we explored the "cognate effect" in bilingual aphasia patients. After including the information of patient's accuracy from the information of our client Manuel Marte, we layered the difficulty of the words from 1 to 6 and took each patient's age of acquisition of non native language into consideration. Based on the characteristic of our dataset, we built multilevel logistic regression models with random intercept and multinomial logistic regression models with random intercept to interpret the cognate effect. The multilevel logistic regression shows that cognate words tend to more likely to be answered correctly and more difficult words tend to have higher level of accuracy. The multinomial logistic regression models are convergence for all variables. Then, in the following part, we will discuss the results of each model in detail.

Introduction

Our client, Manuel Marte,is studying Spanish - English bilingual individuals with aphasia and seeks to understand the "cognate effect" for these subject and how it compares with healthy subjects which have been previously studied. Our dataset based on Boston Naming Test: patients are asked to name an object both in English and Spanish.

We have 27 patients in the dataset, 23 of them are Spanish dominant patient and 4 of them are English dominant patient. In accuracy column, we coded people correctly name the object into 0 and people incorrectly name the object into 1. In cognate column, we coded cognate word into 1 and non-cognate word into 0. In diff column, we leveled our word from difficult 1 to difficult 6 based on the article *Gollan 2007*. Because each patient is unique in our analysis, so we took subject as random intercept. L2AoA indicates the age of a patient acquire the non native language. In language column, 0 means Spanish is this patient's native language and English is non-native language, 1 is the opposite.

In the following part, we mainly interpret the results of multilevel logistic regression models and multinomial logistic regression models.

Multilevel Logistic Regression Model

We first fit the model using logistic regression with each subject as mixed effect for both English words and Spanish words, because the performance of different subjects can be somehow different. In the first model. We set cognateness and difficulty levels of word ranging from 1 to 6 as predictors to fit a model predicting the accuracy.

The coefficients of logistic mixed model for English words are as follows.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: accuracy ~ cognate + (1 | subject) + diff
##    Data: eng3
```

```
##
##      AIC      BIC   logLik deviance df.resid
##    947.9    968.2   -469.9    939.9     1184
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.9507 -0.3735 -0.0971  0.3559  5.2318
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  subject (Intercept) 6.617    2.572
## Number of obs: 1188, groups:  subject, 27
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.08652    0.56691   3.681 0.000233 ***
## cognate      0.74409    0.17385   4.280 1.87e-05 ***
## diff        -0.88052    0.07006 -12.568  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) cognat
## cognate -0.096
## diff    -0.410 -0.153
```

```
## $subject
##           (Intercept)   cognate       diff
## subject1   -2.0224583 0.7440878 -0.8805246
## subject10   2.2441637 0.7440878 -0.8805246
## subject11  -1.0960686 0.7440878 -0.8805246
## subject12   4.2160135 0.7440878 -0.8805246
## subject13   5.6408731 0.7440878 -0.8805246
## subject14   1.9736556 0.7440878 -0.8805246
## subject15   5.8950960 0.7440878 -0.8805246
## subject16  -2.0224583 0.7440878 -0.8805246
## subject17   2.2441637 0.7440878 -0.8805246
## subject18   5.0212992 0.7440878 -0.8805246
## subject19   0.1502718 0.7440878 -0.8805246
## subject2    2.2441637 0.7440878 -0.8805246
## subject20   4.8442144 0.7440878 -0.8805246
## subject21   5.6408731 0.7440878 -0.8805246
## subject22   1.5384150 0.7440878 -0.8805246
## subject23   1.6886274 0.7440878 -0.8805246
## subject24   2.5054783 0.7440878 -0.8805246
## subject25   2.1103523 0.7440878 -0.8805246
## subject26   5.2105076 0.7440878 -0.8805246
## subject27   2.1103523 0.7440878 -0.8805246
## subject3   -2.0224583 0.7440878 -0.8805246
## subject4    4.3639150 0.7440878 -0.8805246
## subject5    1.3814412 0.7440878 -0.8805246
## subject6    1.9736556 0.7440878 -0.8805246
## subject7    0.1502718 0.7440878 -0.8805246
## subject8   -0.5536111 0.7440878 -0.8805246
```

```
## subject9    2.7616180 0.7440878 -0.8805246
##
## attr(,"class")
## [1] "coef.mer"
```

The intercept of regression functions are different for each subject. Let's take coefficients for subject 10 as an example to interpret these coefficients. The probability of a correct naming for subject 10 is

$$p_{10,English}(accuracy = 1) = \frac{1}{1 + e^{-(2.24+0.74*cognate-0.88*difficulty)}}$$

That is to say, if subject 10 is trying to name an English cognate with difficulty level 1, the probability of naming it correctly is 89.1%. While if subject 10 is trying to name an English non-cognate with difficulty level 1, the probability of naming it correctly is non-cognate is 79.6%. Although the performance between different subjects can vary greatly, the coefficient of cognate is positive, so the patients are more likely to name a cognate correctly than to name a non-cognate.

The same is true for Spanish words as follows.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: accuracy ~ cognate + (1 | subject) + diff
##    Data: spa3
##
##      AIC      BIC   logLik deviance df.resid
##    965.1    985.4   -478.5    957.1     1183
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0504 -0.4155 -0.1654  0.3529  6.0991
##
## Random effects:
##  Groups  Name        Variance Std.Dev.
##  subject (Intercept) 4.1      2.025
## Number of obs: 1187, groups:  subject, 27
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.84823    0.46092   1.840   0.0657 .
## cognate      1.16385    0.17533   6.638 3.18e-11 ***
## diff        -0.83898    0.06712 -12.500  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) cognat
## cognate -0.112
## diff    -0.402 -0.256
```

```
## $subject
##           (Intercept)  cognate       diff
## subject1    0.8234535 1.163846 -0.8389805
## subject10   2.0966214 1.163846 -0.8389805
## subject11   1.8397349 1.163846 -0.8389805
## subject12   1.2917109 1.163846 -0.8389805
```

```
## subject13   1.7079875 1.163846 -0.8389805
## subject14  -0.2327676 1.163846 -0.8389805
## subject15  -0.2327676 1.163846 -0.8389805
## subject16  -2.2137818 1.163846 -0.8389805
## subject17   2.9742448 1.163846 -0.8389805
## subject18   1.4347627 1.163846 -0.8389805
## subject19   0.6488115 1.163846 -0.8389805
## subject2    1.8397349 1.163846 -0.8389805
## subject20   0.2554344 1.163846 -0.8389805
## subject21   1.4347627 1.163846 -0.8389805
## subject22   2.9742448 1.163846 -0.8389805
## subject23  -2.2137818 1.163846 -0.8389805
## subject24  -1.4242125 1.163846 -0.8389805
## subject25   3.7765635 1.163846 -0.8389805
## subject26   3.1020819 1.163846 -0.8389805
## subject27  -2.2137818 1.163846 -0.8389805
## subject3   -2.2097871 1.163846 -0.8389805
## subject4    2.2227877 1.163846 -0.8389805
## subject5   -1.4242125 1.163846 -0.8389805
## subject6    2.5974038 1.163846 -0.8389805
## subject7    3.1020819 1.163846 -0.8389805
## subject8    0.6488115 1.163846 -0.8389805
## subject9    2.4727815 1.163846 -0.8389805
##
## attr(,"class")
## [1] "coef.mer"
```

We also take subject 10 as an example. The possibility for a correct naming is

$$p_{10,Spanish}(accuracy = 1) = \frac{1}{1 + e^{-(2.09+1.16*cognate-0.84*difficulty)}}$$

So if subject 10 is trying to name a Spanish cognate with difficulty level 1, the probability of naming it correctly is 91.8%. And if subject 10 is trying to name a Spanish non-cognate with difficulty level 1, the probability of naming it correctly is 77.8%.

Then let's take demographic information into consideration. For English words, the result is shown as follows.

```
## $subject
##           (Intercept)   cognate      diff      L2AoA   Language
## subject1     2.889011 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject10    3.951891 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject11    1.465202 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject12    6.079535 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject13    7.086070 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject14    3.440221 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject15    7.896240 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject16    1.421710 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject17    5.863749 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject18    5.892370 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject19    3.656779 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject2     3.288777 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject20    5.715279 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject21    6.999351 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject22    2.920794 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject23    3.157637 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject24    3.968626 0.7459406 -0.881119 -0.1700228 -0.9408839
```

```
## subject25    3.487458 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject26    6.081626 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject27    3.653156 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject3    -1.467054 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject4     5.400373 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject5     2.271005 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject6     6.497537 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject7     3.174719 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject8     1.527195 0.7459406 -0.881119 -0.1700228 -0.9408839
## subject9     5.794239 0.7459406 -0.881119 -0.1700228 -0.9408839
##
## attr(,"class")
## [1] "coef.mer"
```

We also have similar interpretation, that for subject 10, the possibility of naming it correctly is

$$p_{10,English}(accuracy = 1) = \frac{1}{1 + e^{-(3.95 + 0.75 * cognate - 0.89 * difficulty - 0.17 * L2AoA - 0.94 * Language)}}$$

And for Spanish words with demographic information:

```
## $subject
##            (Intercept) cognate       diff      L2AoA  Language
## subject1  -0.922628765  1.1634 -0.8384205 0.05205551 -1.585225
## subject10  1.578335702  1.1634 -0.8384205 0.05205551 -1.585225
## subject11  1.072655997  1.1634 -0.8384205 0.05205551 -1.585225
## subject12  0.729691977  1.1634 -0.8384205 0.05205551 -1.585225
## subject13  3.066816495  1.1634 -0.8384205 0.05205551 -1.585225
## subject14  1.073130082  1.1634 -0.8384205 0.05205551 -1.585225
## subject15 -0.812571145  1.1634 -0.8384205 0.05205551 -1.585225
## subject16 -1.608239799  1.1634 -0.8384205 0.05205551 -1.585225
## subject17  3.676984283  1.1634 -0.8384205 0.05205551 -1.585225
## subject18  1.171078054  1.1634 -0.8384205 0.05205551 -1.585225
## subject19 -0.398793199  1.1634 -0.8384205 0.05205551 -1.585225
## subject2   1.523564956  1.1634 -0.8384205 0.05205551 -1.585225
## subject20  0.002580854  1.1634 -0.8384205 0.05205551 -1.585225
## subject21  1.021232266  1.1634 -0.8384205 0.05205551 -1.585225
## subject22  2.550907541  1.1634 -0.8384205 0.05205551 -1.585225
## subject23 -1.162422559  1.1634 -0.8384205 0.05205551 -1.585225
## subject24 -0.233864160  1.1634 -0.8384205 0.05205551 -1.585225
## subject25  3.347115807  1.1634 -0.8384205 0.05205551 -1.585225
## subject26  2.828278116  1.1634 -0.8384205 0.05205551 -1.585225
## subject27 -2.482030892  1.1634 -0.8384205 0.05205551 -1.585225
## subject3  -2.133772581  1.1634 -0.8384205 0.05205551 -1.585225
## subject4   1.904390477  1.1634 -0.8384205 0.05205551 -1.585225
## subject5  -1.617391445  1.1634 -0.8384205 0.05205551 -1.585225
## subject6   1.222910515  1.1634 -0.8384205 0.05205551 -1.585225
## subject7   2.176638795  1.1634 -0.8384205 0.05205551 -1.585225
## subject8   0.045698218  1.1634 -0.8384205 0.05205551 -1.585225
## subject9   1.550684708  1.1634 -0.8384205 0.05205551 -1.585225
##
## attr(,"class")
## [1] "coef.mer"
```

For subject 10, the possibility of naming it correctly is

$$p_{10,Spanish}(accuracy = 1) = \frac{1}{1 + e^{-(1.58+1.16*cognate-0.84*difficulty+0.05*L2AoA-1.59*Language)}}$$

We also notice that, although the performance between different subjects can vary greatly, the coefficients of cognateness are both positive for English and Spanish, no matter we include demograpic information or not. So the patients are more likely to name a cognate correctly than to name a non-cognate.

Multinomial Logistic Regression Model

As shown in the previous part, the result from multilevel model tells us that patients are more likely to name a cognate correctly than to name a non-cognate. We then want to consider the patients' responses in the English test and Spanish test jointly, so we choose to fit mixed-effects multinomial logistic regression models using the brms package. We create the responses in 4 levels: 1. the patient named a word correctly in both English and Spanish. 2. the patient only named a word correctly in English. 3. the patient only named a word correctly in Spanish. 4. the patient named a word incorrectly in both English and Spanish. We will assess the relationship between these 4 responses and cognateness of words based on the result from the multinomial model.

The first multinomial logistic regression model takes cognateness and difficulty levels of words as fixed effects and subjects as random effects to assess the effects of cognateness and difficulty levels at varied levels of random effects.

The summary output of multinomial logistic regression model has a block of coefficients. Each of these blocks has one row of values corresponding to a model equation. The baseline outcome here is **response="both correct"**. For example, the model equation for **response="both incorrect"** comparing to the baseline outcome would be

$$ln(\frac{Pr(response = bothincorrect)}{Pr(response = bothcorrect)} = \boldsymbol{\alpha}_i + \boldsymbol{\beta}_1(cognate = 1) + \boldsymbol{\beta}_2 difficulty_j$$

where i indexes patients (1 to 27) and j indexes words (1 to 44).

The second multinomial logistic regression model takes the demographic information into consideration, so we add L2AoA and Language into fixed effects.

After adding the demographic information, the model equation for **response="both incorrect"** comparing to the baseline outcome would be

$$ln(\frac{Pr(response = bothincorrect)}{Pr(response = bothcorrect)} = \boldsymbol{\alpha}_i + \boldsymbol{\beta}_1(cognate = 1) + \boldsymbol{\beta}_2 difficulty_j + \boldsymbol{\beta}_3 L2AoA_i + \boldsymbol{\beta}_4 Language_i$$

where i indexes patients (1 to 27) and j indexes words (1 to 44).

The coefficient estimates for fixed effects "cognate" and "difficulty levels" changes slightly after we add L2AoA and Language in the model.

Model Comparison

To investigate which multinomial model gives a better fit for the data, we compare two models via the approximate LOO (leave-one-out cross-validation).

```
##    elpd_diff se_diff
## b2  0.0       0.0
## b1 -0.1       1.3
```

As shown above, the second multinomial model has a better fit. The elpd_diff=0 in the first row tells us that the difference between the preferred model and the second multinomial model itself is 0.

Figure 1. The plot to visualize the relationship between the cognateness and responses.

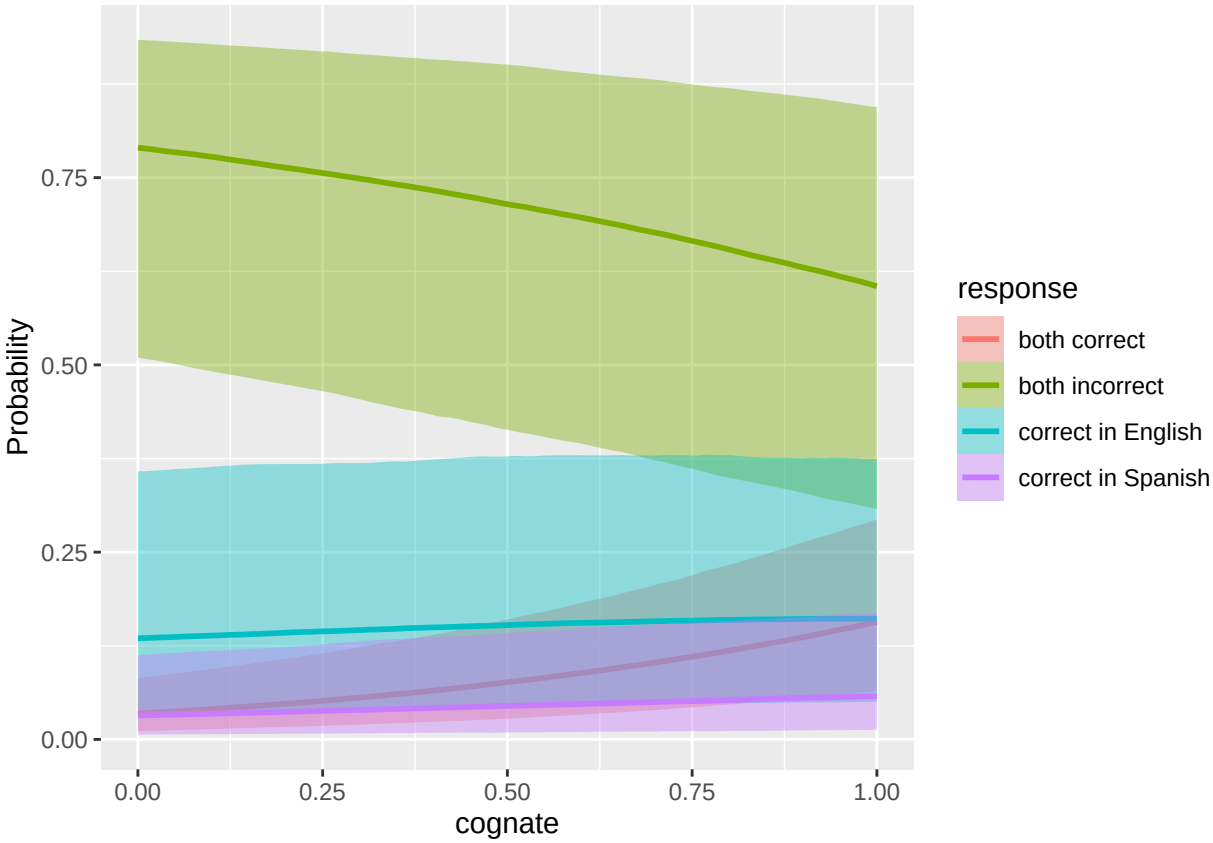

Figure 1: When patients are testing on cognates, the probability of answering words incorrectly in both English and Spanish decreases and the probability of answering correctly in both English and Spanish increases.

Figure 2. The plot to visualize the relationship between difficulty levels and responses.

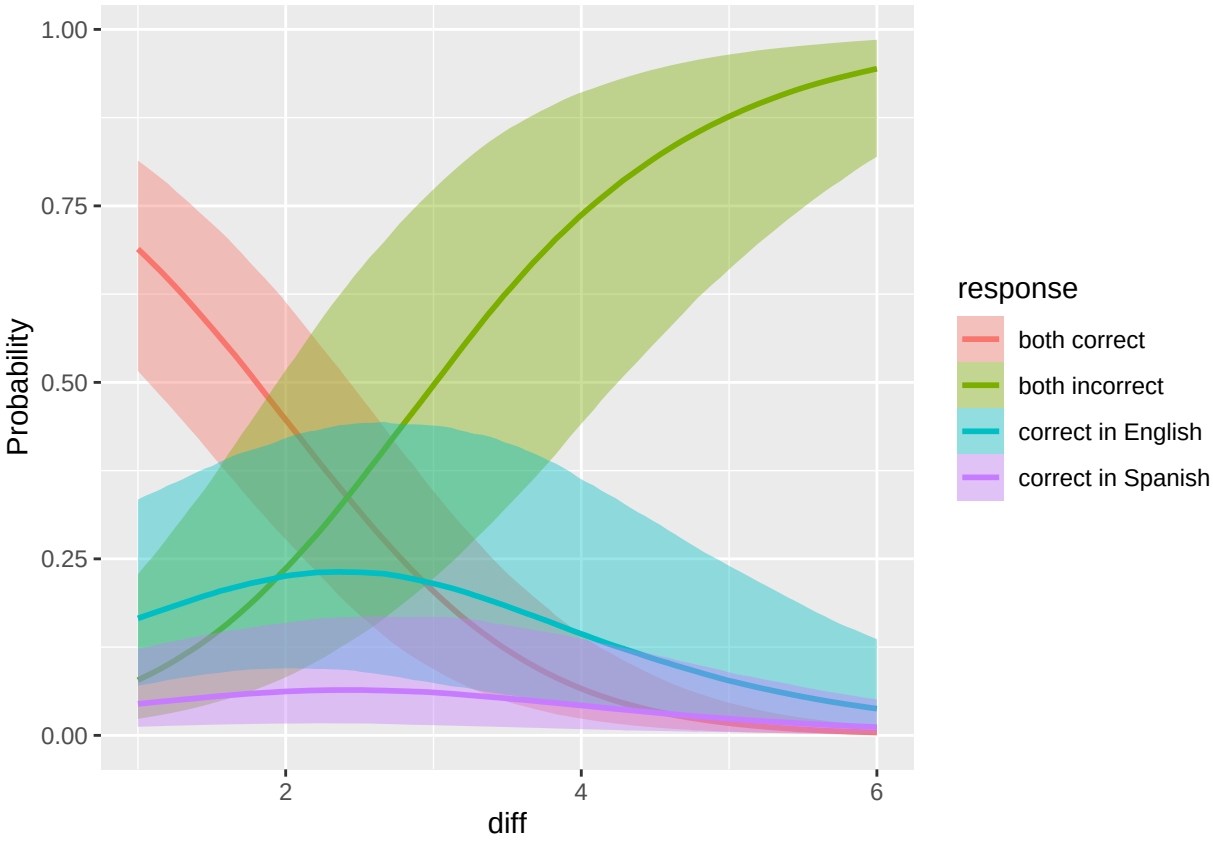

Figure 2: When patients are testing on words with higher difficulty levels, the probability of answering the words correctly in both English and Spanish decreases and the probability of answering incorrectly in both English and Spanish increases.

Conclusion

We tried multilevel logistic regression. We still divide the data into two groups: Spanish and English. We first included only cognate as the only factor, as our baseline to find the optimal model by comparing it. By using the lme4 package, we divide the words into different subject levels. When the words are in English and have the same difficulty level, cognate words will have a higher probability (94.43% higher) to answer correctly. When the words are in Spanish and the difficulty level is the same, cognate words will have a higher probability (76.195.% higher) to answer correctly.

Before the fit model, we processed the missing data in the data, and explored the influence of Spanish and English cognate words on the accuracy rate. We divided the data of all subjects into four categories, 1. Spanish and English is correct, 2. Spanish is correct, English is wrong, 3. Spanish is wrong, and English is correct. 4. Both languages are wrong. At the same time, considering the differences of each participant to words, we will use random intercept to express this phenomenon. Our group fits two mutinomial equations, and compares the influence of cognate words through the LOO equation. In the first equation, our group did not include any demographic information. We will use it as a baseline to compare with other models to arrive at a relatively optimal model. In the second multinomial logistic regression model, we included demographic information AOA and language as variables. We use cognate as the independent variable and the probability of being correct as the dependent variable. It can be seen that with the inclusion of demographic data, as the degree of cognate becomes stronger, the possibility of simultaneous errors in two languages gradually

decreases. The probability that the pair and all are correct gradually rises. Therefore, according to the results of the graph, we can conclude that the influence of cognate has a positive influence on the correctness rate. When we replace cognate with difficulty, the possibility of correctness decreases as the difficulty increases, and the possibility of error increases, which is in line with our expectations. It also proved the reliability of the model.

Therefore, from the above regression model, we can see that the influence of cognate on the correct green is still very significant. Of course, our research still has certain limitations, such as insufficient data and insufficient diversity and comprehensiveness in the optimization and comparison of models.

Appendix

Figure 2. The summary output of the first multinomial logistic regression model.

	Estimate	Est.Error	Q2.5	Q97.5
mubothincorrect_Intercept	-2.8780905	0.6992858	-4.2407486	-1.5208553
mucorrectinEnglish_Intercept	-1.4300957	0.5283754	-2.5040667	-0.4148390
mucorrectinSpanish_Intercept	-2.9674472	0.7155234	-4.4634832	-1.6728305
mubothincorrect_cognate	-1.7778614	0.2505733	-2.2628824	-1.2815957
mubothincorrect_diff	1.5247753	0.1065956	1.3189172	1.7367686
mucorrectinEnglish_cognate	-1.3235572	0.2393682	-1.7945225	-0.8642788
mucorrectinEnglish_diff	0.7318452	0.0945260	0.5497572	0.9200266
mucorrectinSpanish_cognate	-0.9158925	0.2778460	-1.4615867	-0.3638269
mucorrectinSpanish_diff	0.7579998	0.1130596	0.5377400	0.9804502

Figure 3. The summary output of the second multinomial logistic regression model.

	Estimate	Est.Error	Q2.5	Q97.5
mubothincorrect_Intercept	-4.3469839	1.2576175	-6.9253406	-1.9082222
mucorrectinEnglish_Intercept	-0.5603854	0.9143785	-2.3979533	1.1885129
mucorrectinSpanish_Intercept	-5.4828347	1.2030360	-8.1244396	-3.3103240
mubothincorrect_cognate	-1.7912996	0.2464905	-2.2702457	-1.2965688
mubothincorrect_diff	1.5298057	0.1065451	1.3248454	1.7416835
mubothincorrect_L2AoA	0.0998388	0.0848044	-0.0676188	0.2706126
mubothincorrect_Language	2.0847398	1.5917323	-0.9951459	5.2887104
mucorrectinEnglish_cognate	-1.3295231	0.2359588	-1.7886207	-0.8792227
mucorrectinEnglish_diff	0.7318716	0.0939587	0.5490810	0.9137388
mucorrectinEnglish_L2AoA	-0.1124073	0.0695143	-0.2547937	0.0215807
mucorrectinEnglish_Language	1.1786755	1.1441863	-0.9629676	3.5318323
mucorrectinSpanish_cognate	-0.9262772	0.2732924	-1.4614459	-0.3978927
mucorrectinSpanish_diff	0.7627460	0.1113066	0.5545200	0.9916779
mucorrectinSpanish_L2AoA	0.2328166	0.0753159	0.0923555	0.3970610
mucorrectinSpanish_Language	-0.2684152	1.5737596	-3.4933127	2.7035040

Figure 4. The trace and density plot of the second multinomial logistic regression model for fixed-effect "cognate".

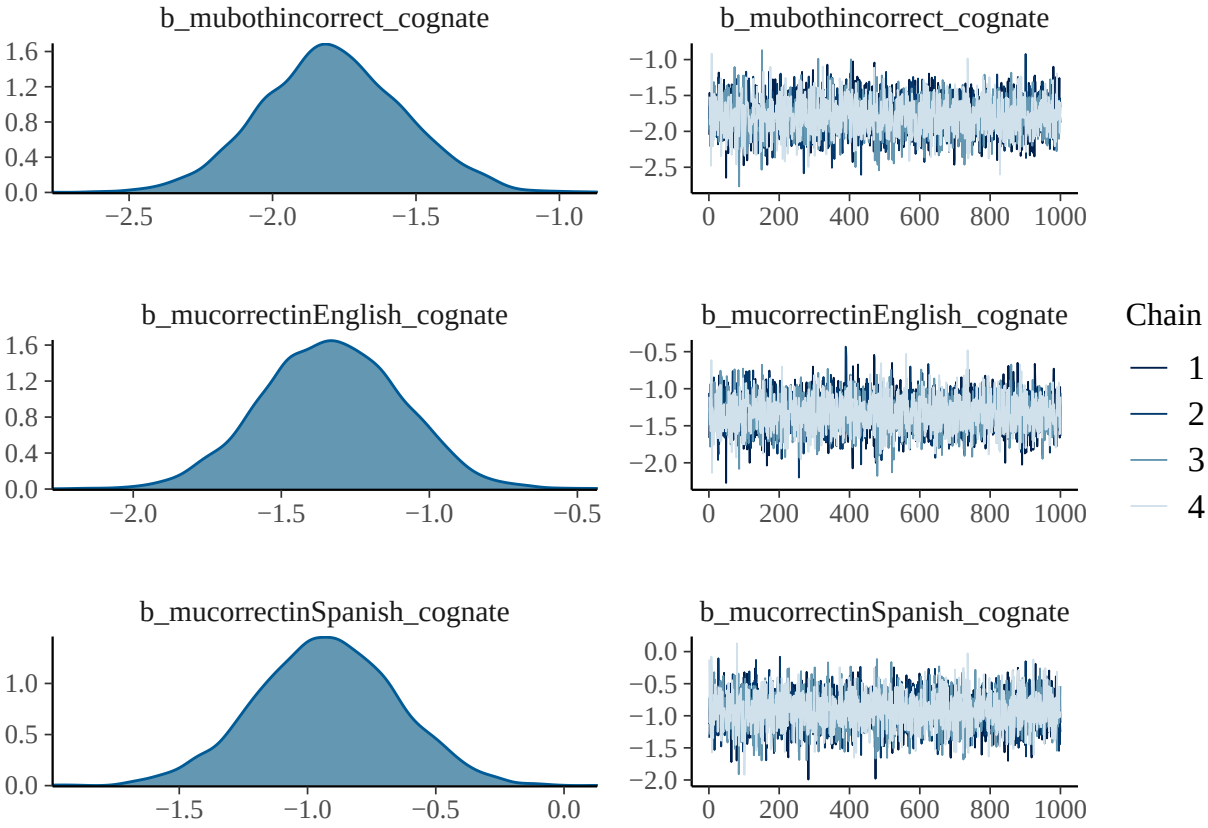

Figure 5. Posterior Predictive Check for the second multinomial logistic regression model.