Analysis of Sociocultural Adaptation

Leung, Yvonne
Shih, Hsiu-Yu (Sherry)
Northeastern University
Apr 23, 2025

Northeastern University

Data Overview-Total 60 Entries

Variable Name	Description
participant.ID	A unique identifier assigned to each student
Native.language	The student's native language (‘Chinese’, ‘Spanish’, ’Vietnamese’, ‘Ukrainian’ , ’ran’ , ‘Portuguese’ , ‘oriya’ , ‘Korean’ , ‘Russian’ , ‘Greek’ , ‘French’)
High.School.Language	The primary language of instruction at the student’s high school('Chinese', 'English', 'Spanish', 'Vietnamese', 'Greek’)
Time.in.the.US	How long the student has lived in the U.S. ('Less than one month’, 'One to three months’, ‘Three months to one year’, ‘One to two years’, ‘More than two years’)
X.Pre.1 to X.Pre.21	Self-rated adaptation items before the experience (Likert scale: 1 = Strongly Disagree to 7 = Strongly Agree))
X.Post.1 to X.Post.21	Same items as Pre-Survey, rated after the experience (Likert scale: 1 = Strongly Disagree to 7 = Strongly Agree)

Data Cleaning and Preprocessing

- Native.language was recoded into two categories: "Chinese" vs. "Other" for Native Language. Factors were set with "Other" as the reference.
- High.School.Language was recoded into two categories: "English" vs. "Other”. Factors were set with "Other" as the reference.
- Time.in.the.US was grouped into: "Less than 1 year", "1-2 years", and "More than 2 years". ”. Factors were set with "Less than 1 year” as the reference.

Missing Value Imputation

- Missing values were thoughtfully filled in using each student's own average—unless they skipped all pre-survey items, in which case we couldn’t include them.

Summative Score Creation and Reshaping

- We gave each student a "Pre-Sum" and a "Post-Sum" score by adding up how they rated their adaptation on a series of questions before and after a period of time.
- From this, we calculated a simple but powerful number: the "Change Score" = Post-Sum − Pre-Sum. This told us how much each student improved.
- We reshaped the data to allow comparison at both the individual and group levels across time. (long format)

	participant.ID	Native.language	High.School.Language	Time.in.the.US	Time	Score
1	45111	Chinese	Other	Less than 1 year	0	59.00000
2	45111	Chinese	Other	Less than 1 year	1	95.55000

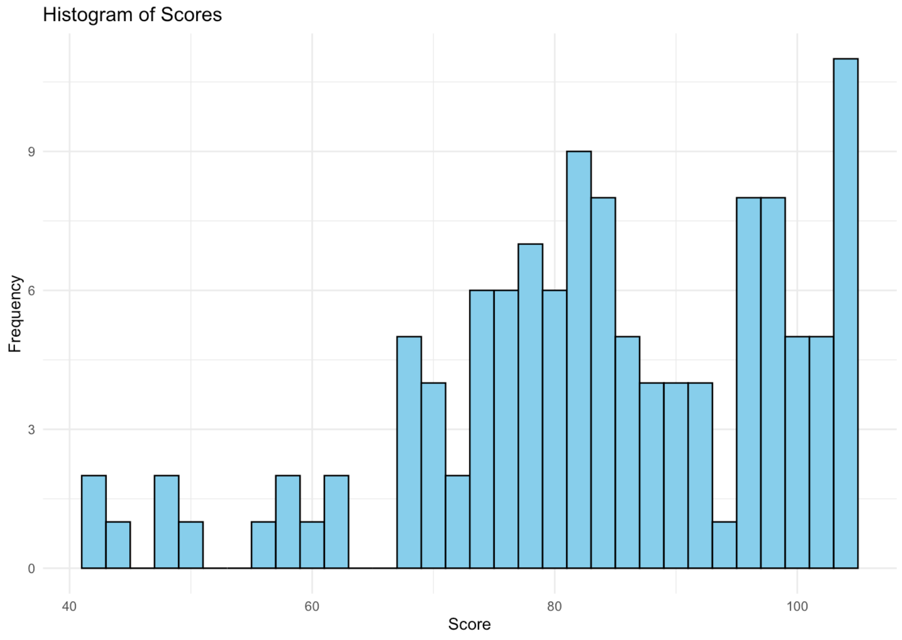
Comparison between original dataset and cleaned dataset

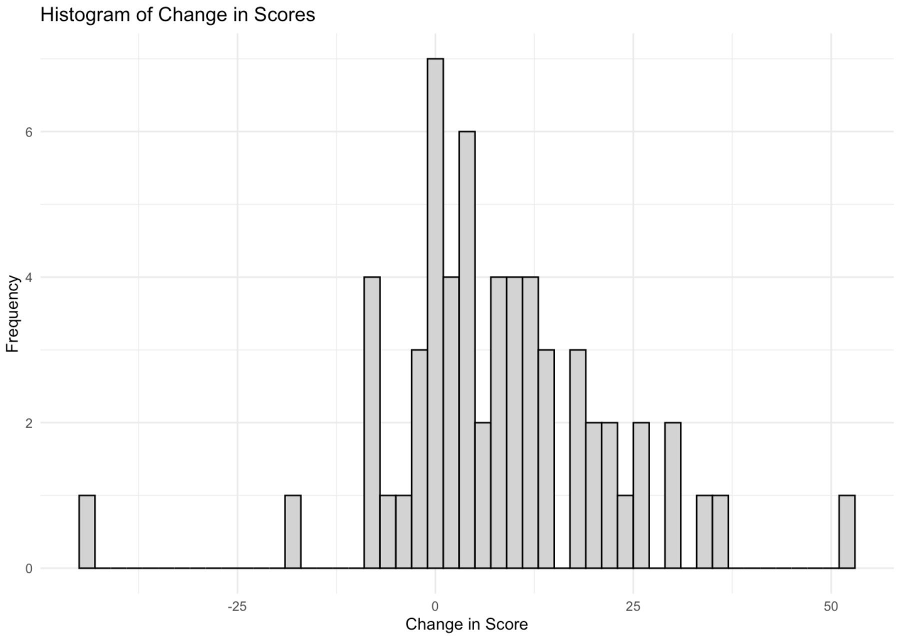
Aspect	Original Dataset	Cleaned Dataset
Participants	60 entries	59 entries (1 removed with all missing Pre responses)
Variables	54 columns	54 columns
Key Variables	Pre & Post scores across 21 adaptation items	Same
Missing Data	Present in multiple Pre/Post items	Imputed using each participant’s row-wise mean
Categorical Variables	Text-based (with inconsistent formatting)	Recoded (e.g., “chinese” → “Chinese”, grouped into factors)
Derived Variables	Not included	Pre_Sum, Post_Sum, and change_score added
Data Structure	Wide format	Wide and reshaped into long format for time-based modeling
Skewness	Not assessed	Examined for raw, log, sqrt, and change score distributions

LVX VERITAS VIRTVS

Score Transformation and Skewness Comparison

- Before modeling, we examined the distribution of the scores:

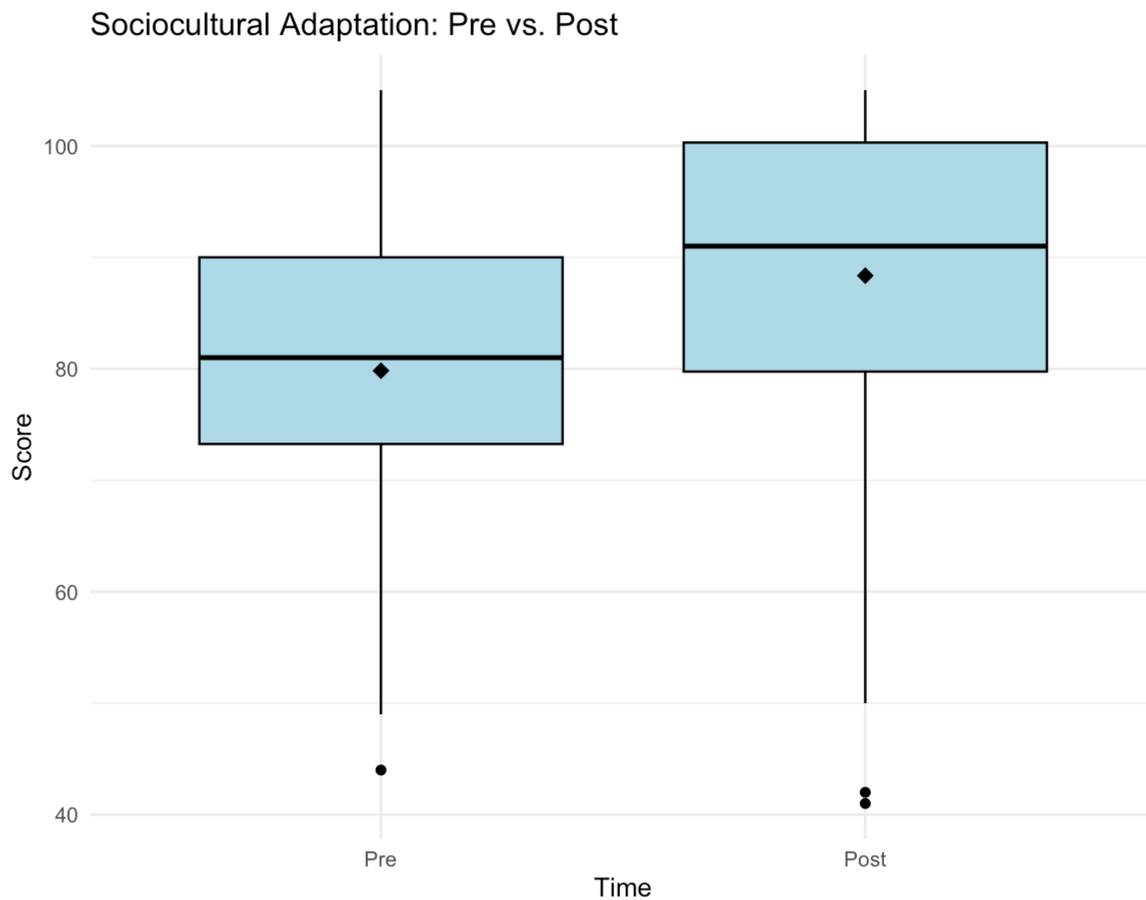




	Score	Change Score
Skewness	-0.688	-0.147
Interpretation	moderately left-skewed, meaning more students rated themselves higher with fewer low scores. not severely skewed and still acceptable for analysis	**very close to symmetric**

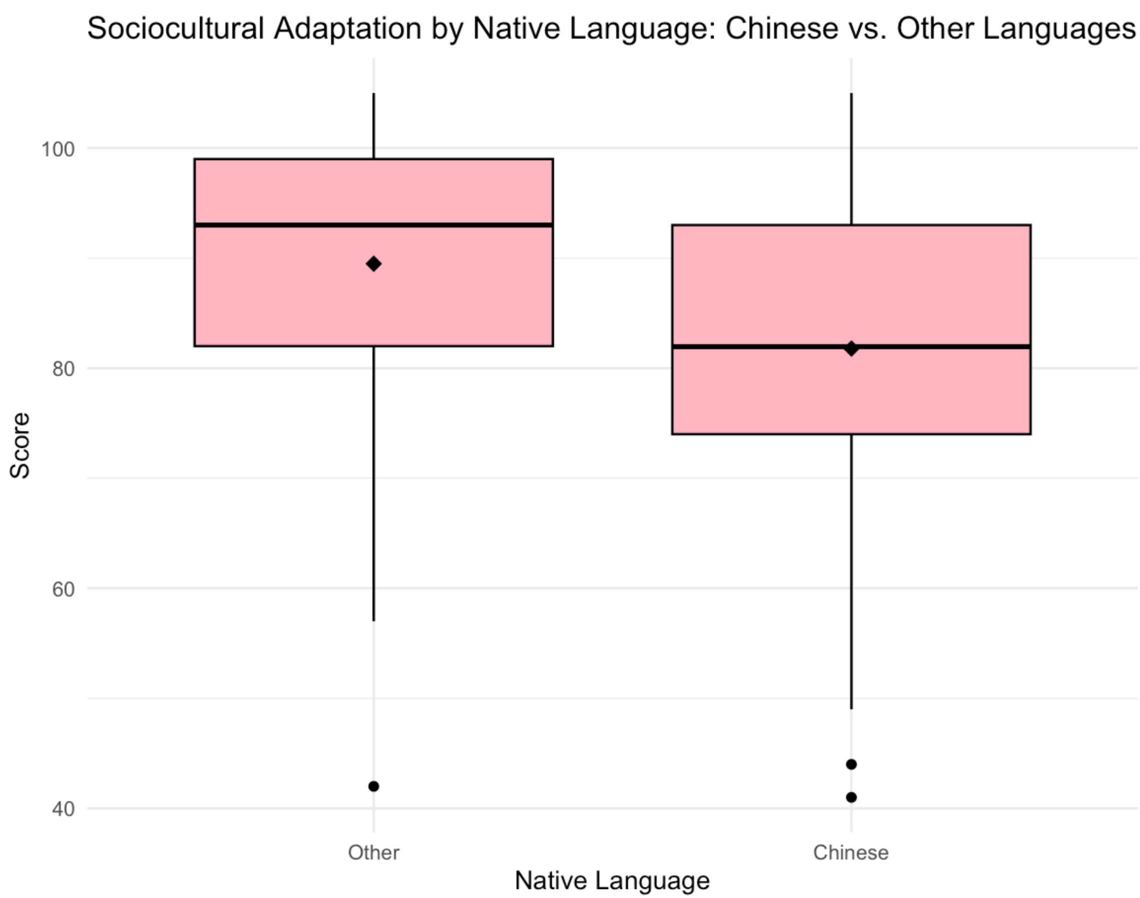
Social Adaptation Score: Pre vs. Post

- Most students feel more socioculturally adapted over time.



Social Adaptation Score by Native Language: Chinese vs. Other Languages

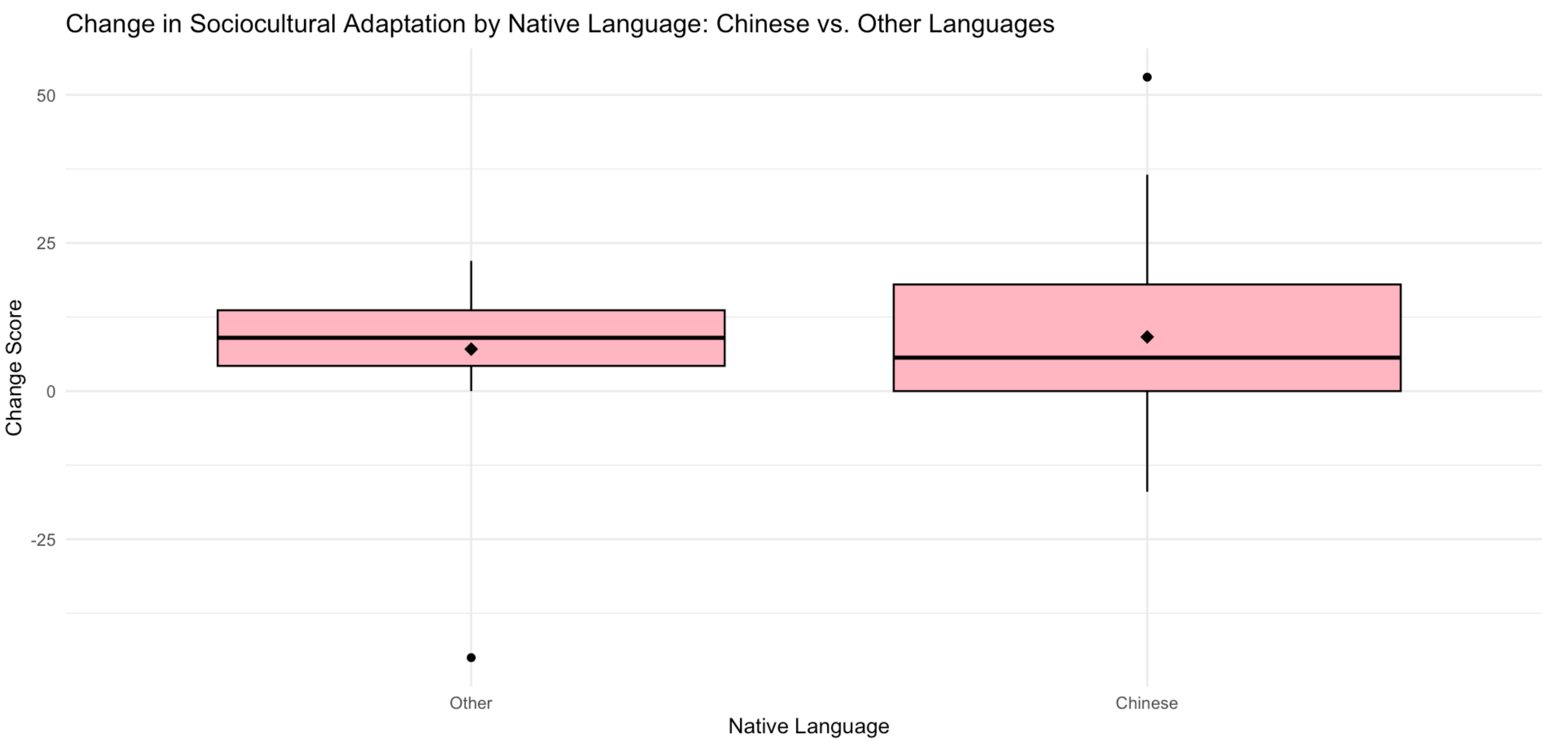
- On average, students from other language backgrounds rated themselves more adapted than their Chinese-speaking peers.
- Chinese-speaking students may face more challenges in adapting socially or culturally.



Social Adaptation Change by Native Language: Chinese vs. Other Languages

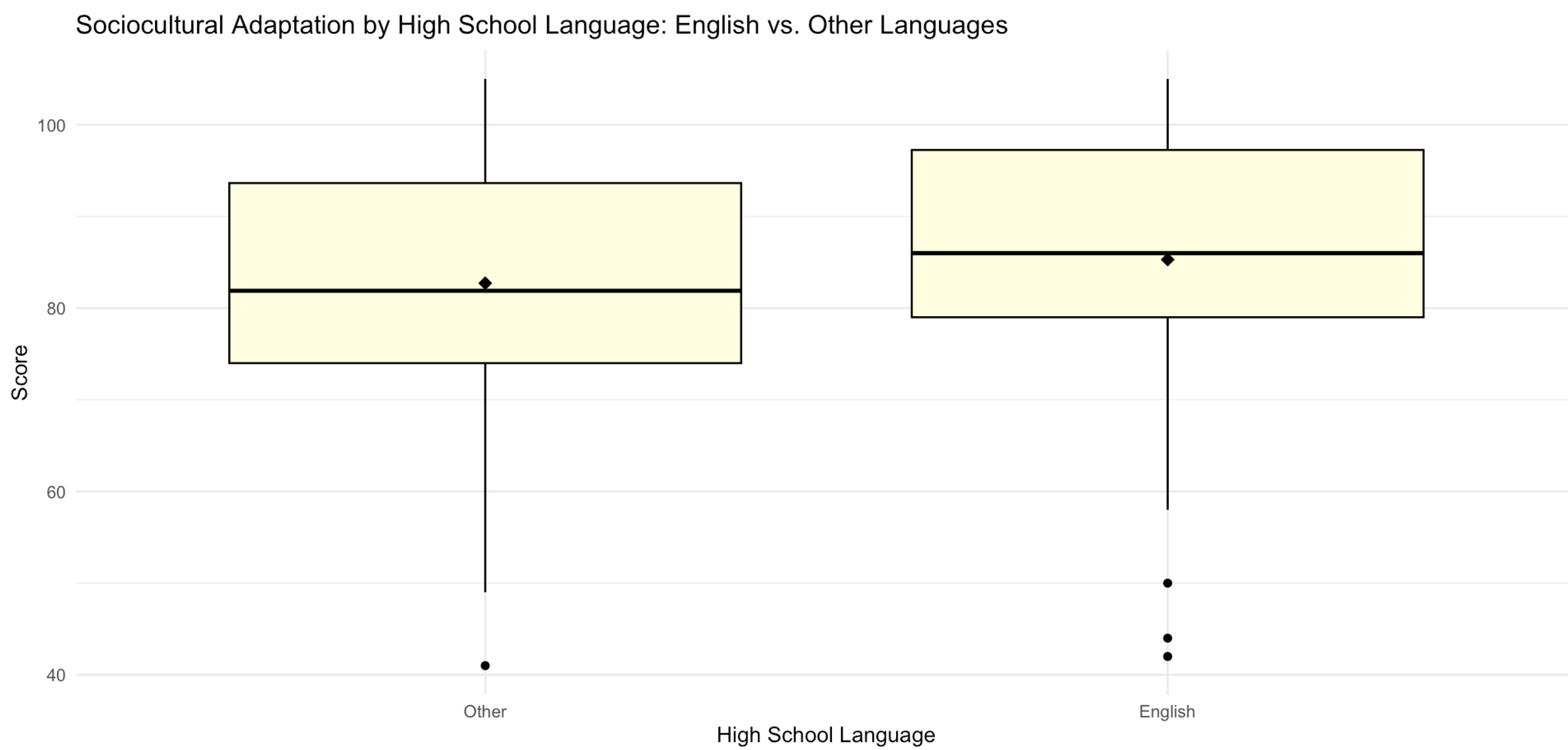
- Students from non-Chinese language backgrounds tend to show consistent, positive growth in adaptation. Their change scores cluster closely together, suggesting a relatively uniform experience of improvement.

- Students who speak Chinese as their native language demonstrate a much wider range of outcomes. While some report strong gains, others show little improvement—or even a decline—in adaptation scores over time.



Social Adaptation Score by High School Language: English vs. Other Languages

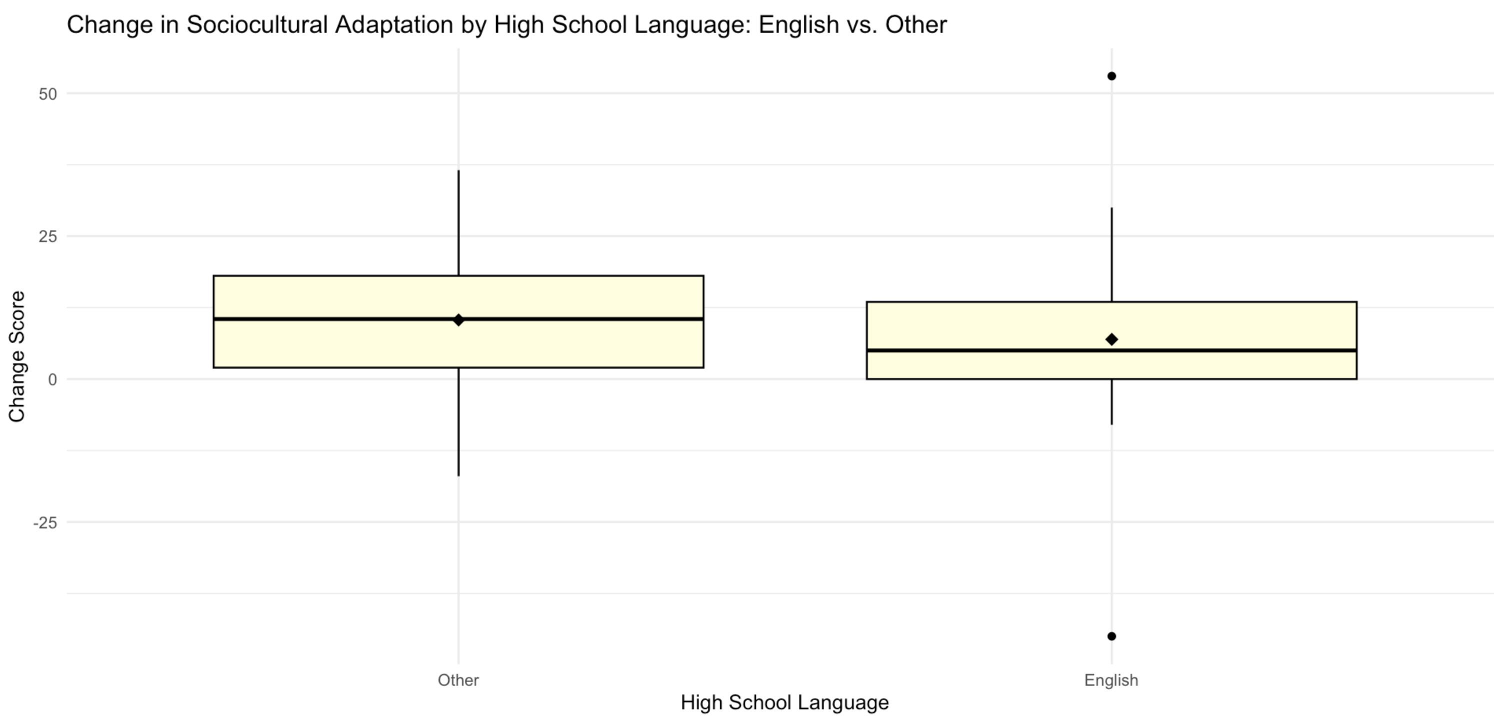
- The difference between students who went to high schools where English or not is mild. In short, this factor alone isn't driving big differences.



Social Adaptation Change by High School Language: English vs. Other Languages

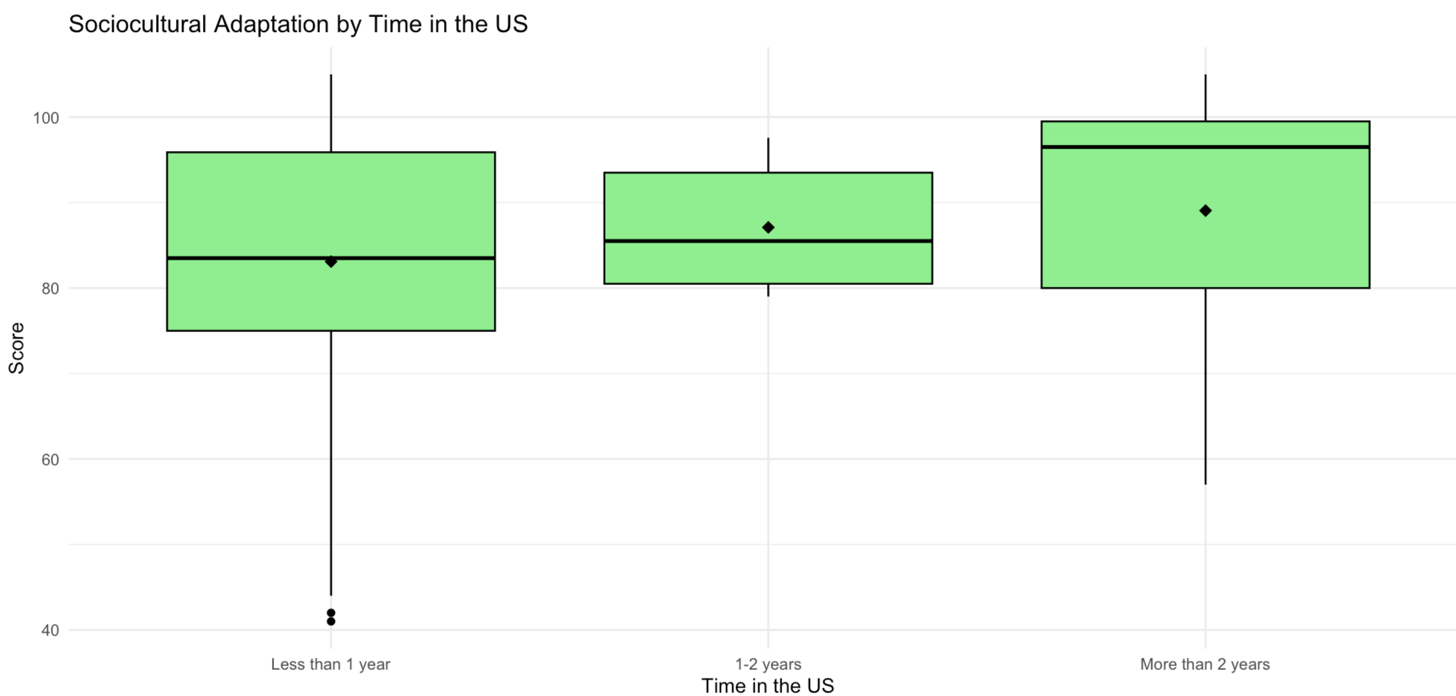
- The median change score is slightly higher for students who studied in non-English high schools. This suggests that these students, on average, reported more growth in sociocultural adaptation over time.

- However, the distribution of scores is relatively similar between the two groups. Both show a wide range of experiences, from students with strong gains to those with minimal or even negative changes.



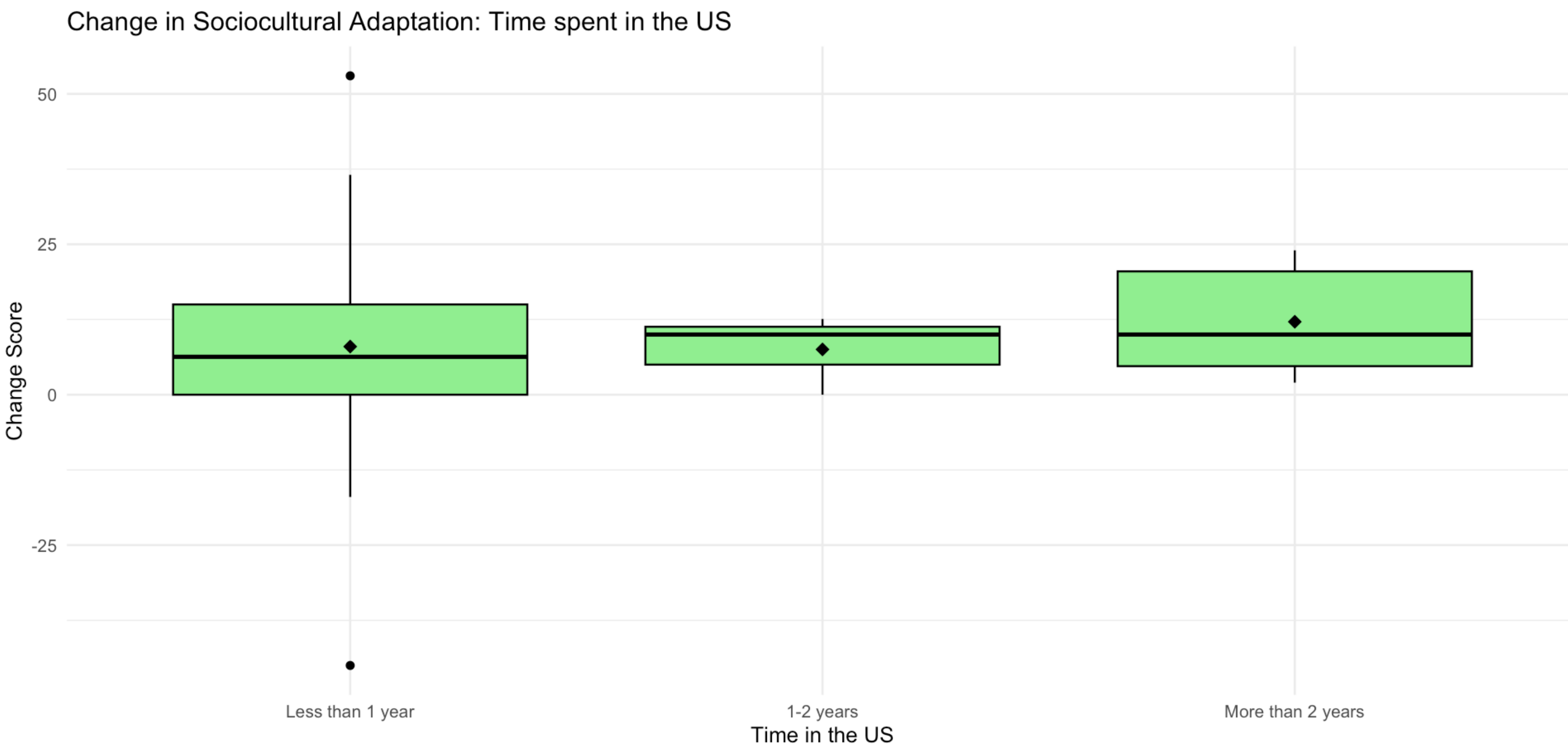
Social Adaptation Score by Time in the US

- The difference of social adaptation between the time in the US were not dramatic.
- Time in the US ‘more than 2 years’ group scoring a bit higher — the difference isn’t dramatic.



Social Adaptation Change by Time in the US

- Students who have been in the U.S. for more than 2 years show the highest median improvement in adaptation scores.
- Those in the 1–2 years category show a modest gain, with a relatively narrow range of improvement.



Modeling Sociocultural Change

We ran five models to identify what predicted change in adaptation.

1) Model 1:Linear Regression on Change Score
2) Model 2: Linear Regression on Change Score with Interaction
3) Model 3: Linear Regression on Scores (long format)
4) Model 4: Linear Regression on Scores with Interaction (long format)
5) Mixed-Effects Model

Model 1:Linear Regression on Change Score

What predicts someone improving more over time?

- Students who started off lower (lower Pre_Sum) showed more growth.
- Language background or time in the U.S. didn't strongly explain who improved.
- R^2 was 0.138, meaning about 14% of variation in improvement is explained by these factors.

```
Call:
lm(formula = change_score ~ Native.language + High.School.Language +
    Time.in.the.US + Pre_Sum, data = socialcultural)

Residuals:
    Min      1Q  Median      3Q     Max
-47.788  -7.755   0.626   8.207  30.377

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      43.8668    12.4930   3.511 0.000909 ***
Native.languageChinese            0.6429     5.0787   0.127 0.899734
High.School.LanguageEnglish      -2.2491     3.5610  -0.632 0.530319
Time.in.the.US1-2 years           2.3552     8.9687   0.263 0.793858
Time.in.the.USMore than 2 years   6.9090     6.2268   1.110 0.272106
Pre_Sum                          -0.4463     0.1328  -3.362 0.001428 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.38 on 54 degrees of freedom
Multiple R-squared:  0.2115,    Adjusted R-squared:  0.1385
F-statistic: 2.897 on 5 and 54 DF,  p-value: 0.02178
```

LVX VERITAS VIRTVS

Model 2: Linear Regression on Change Score with Interaction

Does Language Moderate Starting Score?

We added an interaction: does the effect of Pre_Sum depend on whether you speak Chinese?

- The interaction wasn't significant, so Chinese vs. Other didn't change the relationship between starting point and improvement.
- Model fit stayed about the same.

```
Call:
lm(formula = change_score ~ Native.language + High.School.Language +
    Time.in.the.US + Pre_Sum + Pre_Sum * Native.language, data = socialcultural)

Residuals:
    Min      1Q  Median      3Q     Max
-47.087  -7.112   1.425   8.661  28.966

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      20.0459    28.2173   0.710    0.481
Native.languageChinese           29.2025    30.7496   0.950    0.347
High.School.LanguageEnglish      -3.0864     3.6740  -0.840    0.405
Time.in.the.US1-2 years           3.7864     9.1058   0.416    0.679
Time.in.the.USMore than 2 years   8.1271     6.3661   1.277    0.207
Pre_Sum                          -0.1709     0.3212  -0.532    0.597
Native.languageChinese:Pre_Sum   -0.3320     0.3525  -0.942    0.351

Residual standard error: 13.39 on 53 degrees of freedom
Multiple R-squared:  0.2245,    Adjusted R-squared:  0.1367
F-statistic: 2.557 on 6 and 53 DF,  p-value: 0.03004
```

Model 3: Linear Regression on Scores (long format)

We then looked at all scores across time (Pre and Post together), asking: are scores higher after time passes?

- Time" variable was significant — students rated themselves higher after some time in the program.
- Chinese speakers still reported lower scores overall.
- R^2 was 0.102 — a bit less explanatory power than the earlier model but still meaningful.

```
Call:
lm(formula = Score ~ Time + Native.language + High.School.Language +
    Time.in.the.US, data = long_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-52.713  -6.776   1.285  10.181  28.142 

Coefficients:
                               Estimate Std. Error t value             Pr(>|t|)    
(Intercept)                    84.80539    3.94969  21.471 < 0.0000000000000002 ***
Time                            8.52792    2.60542   3.273              0.00141 ** 
Native.languageChinese         -7.94708    3.67950  -2.160              0.03288 *  
High.School.LanguageEnglish     1.37979    2.66655   0.517              0.60585    
Time.in.the.US1-2 years        -2.89117    6.74456  -0.429              0.66897    
Time.in.the.USMore than 2 years -0.04831    4.67507  -0.010              0.99177    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.27 on 114 degrees of freedom
Multiple R-squared:  0.1399,	Adjusted R-squared:  0.1022 
F-statistic: 3.708 on 5 and 114 DF,  p-value: 0.003798
```

LVX VERITAS VIRTVS

Model 4: Linear Regression on Scores with Interaction (long format)

Do Chinese Students Improve Differently?

We added an interaction between Time and Native Language. Asking: do Chinese speakers grow differently than others?

- The interaction wasn't significant, and the model fit didn't improve much.

```
Call:
lm(formula = Score ~ Time + Native.language + High.School.Language +
    Time.in.the.US + Native.language * Time, data = long_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.993  -7.187   1.288  10.019  28.450 

Coefficients:
                                Estimate Std. Error t value            Pr(>|t|)    
(Intercept)                     85.52523    4.43963  19.264 <0.0000000000000002 ***
Time                             7.08824    4.77509   1.484               0.140    
Native.languageChinese          -8.97542    4.66757  -1.923               0.057 .  
High.School.LanguageEnglish      1.37979    2.67678   0.515               0.607    
Time.in.the.US1-2 years         -2.89117    6.77045  -0.427               0.670    
Time.in.the.USMore than 2 years -0.04831    4.69301  -0.010               0.992    
Time:Native.languageChinese      2.05669    5.70732   0.360               0.719    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.33 on 113 degrees of freedom
Multiple R-squared:  0.1409,	Adjusted R-squared:  0.09526 
F-statistic: 3.088 on 6 and 113 DF,  p-value: 0.007743
```

LVX VERITAS VIRTVS

Mixed-Effects Model

We accounted for the fact that each participant gave two scores (Pre and Post).

We used a model that lets each person have their own baseline.

- "Time" was again significant — students improved in social adaption
- This model confirmed that the improvement wasn't just due to chance or specific individuals. It was consistent across participants.

```
Linear mixed model fit by REML ['lmerMod']
Formula: Score ~ Time + Native.language + High.School.Language + Time.in.the.US +
    (1 | participant.ID)
   Data: long_data

REML criterion at convergence: 934.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4322 -0.4597  0.0747  0.5306  1.7414

Random effects:
 Groups         Name        Variance Std.Dev.
 participant.ID (Intercept)  97.02    9.85
 Residual                   111.33   10.55
Number of obs: 120, groups:  participant.ID, 59

Fixed effects:
                                Estimate Std. Error t value
(Intercept)                      84.8361     4.6712  18.161
Time                              8.5279     1.9264   4.427
Native.languageChinese           -7.9882     4.5147  -1.769
High.School.LanguageEnglish       1.3205     3.2908   0.401
Time.in.the.US1-2 years          -2.8824     8.2593  -0.349
Time.in.the.USMore than 2 years  -0.0294     5.7264  -0.005

Correlation of Fixed Effects:
            (Intr) Time   Ntv.lC H.S.LE T...Uy
Time        -0.206
Ntv.lnggChn -0.836  0.000
Hgh.Schl.LE -0.409  0.000  0.087
Tm...US1-2y -0.433  0.000  0.450 -0.034
Tm...USMt2y -0.522  0.000  0.546 -0.106  0.324
```

Model Comparison

Model	Purpose	Significant Findings	Adjusted R^2
Model 1 Linear Regression on Change Score	Examine what predicts improvement over time	**Lower Pre_Sum → more improvement ($p < .01$)**	0.138
Model 2 Linear Regression + Interaction	Check if Pre_Sum effect differs by language	Interaction not significant	0.137
Model 3 Linear Regression on Raw Scores	Check effect of time (Pre vs. Post) on scores	**Post scores significantly higher ($p < .01$)**	0.102
Model 4 Interaction Between Time & Language	Test if improvement over time differs by language	Interaction not significant	0.095
Model 5 Mixed-Effects Model	Account for repeated measures within participants	Time significant (**$p < .001$**)	-

LVX VERITAS VIRTVS

Integrating 5 Sociocultural Adaptation Categories

1. Grouped 21 pre/post survey items into five meaningful domains based on their content:

- Interpersonal
- Academic
- Interest
- Ecological
- Language

2. Created Composite Scores

For each participant, calculated sum scores for:

- Pre (e.g., Interpersonal_Pre)
- Post (e.g., Interpersonal_Post)
- This was done by summing selected items for each category.

Integrating 5 Sociocultural Adaptation Categories

3. Reshaped Dataset for MANOVA

- Created a new dataset with two rows per participant (Pre and Post), with each of the five category scores as columns:

	participant.ID	Native.language	High.School.Language	Time.in.the.US	Time	Interpersonal	Academic	Interest	Ecological	Language
1	45111	Chinese	Other	Less than 1 year	0	18.00000	10.00000	14.00000	11.00000	6.000000
2	45111	Chinese	Other	Less than 1 year	1	33.00000	18.55000	19.00000	16.00000	9.000000
3	45711	Other	Other	More than 2 years	0	29.00000	17.00000	14.00000	15.00000	6.000000
4	45711	Other	Other	More than 2 years	1	34.00000	20.00000	19.00000	18.00000	10.000000

```r
#Create wide format with one row per participant per time
manova_ready <- socialcultural %>%
  select(participant.ID, Native.language, High.School.Language, Time.in.the.US,
         Interpersonal_Pre, Academic_Pre, Interest_Pre, Ecological_Pre, Language_Pre,
         Interpersonal_Post, Academic_Post, Interest_Post, Ecological_Post, Language_Post) %>%
  pivot_longer(
    cols = ends_with("_Pre") | ends_with("_Post"),
    names_to = c("Category", "Time"),
    names_pattern = "(.*)_(Pre|Post)",
    values_to = "Score"
  ) %>%
  pivot_wider(
    names_from = Category,
    values_from = Score,
    values_fn = mean  # ← This ensures numeric output
  ) %>%
  mutate(Time = ifelse(Time == "Pre", 0, 1))

View(manova_ready)
```

Multivariate Effects of Time and Language Background on Sociocultural Adaptation

- **Time** significantly influenced adaptation scores across five domains, indicating strong program impact.
- **Native language** was a significant predictor, with Chinese-speaking students differing in overall adaptation profiles.
- **High school language** showed no significant effect, while **time spent in the U.S.** showed a near-significant trend, suggesting a possible influence worth further investigation.

```
> # MANOVA with Time and background variables
> manova_model <- manova(cbind(Interpersonal, Academic, Interest, Ecological, Language) ~
+                          Time + Native.language + High.School.Language + Time.in.the.US,
+                        data = manova_ready)
> summary(manova_model, test = "Pillai")
```

```
> summary(manova_model, test = "Pillai")
                      Df  Pillai approx F num Df den Df  Pr(>F)  
Time                   1 0.12306   3.0312      5    108 0.01339 *
Native.language        1 0.11552   2.8211      5    108 0.01960 *
High.School.Language   1 0.07369   1.7183      5    108 0.13648  
Time.in.the.US         2 0.14350   1.6851     10    218 0.08553 .
Residuals            112                                          
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Category-Specific Impacts of Time and Language on Sociocultural Adaptation

```
> summary.aov(manova_model)
 Response Interpersonal :
                     Df  Sum Sq Mean Sq F value   Pr(>F)
Time                  1  264.79 264.786 10.1097 0.001908 **
Native.language       1  220.71 220.708  8.4268 0.004454 **
High.School.Language  1   25.46  25.463  0.9722 0.326259
Time.in.the.US        2   13.57   6.785  0.2591 0.772239
Residuals           112 2933.42  26.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Academic :
                     Df Sum Sq Mean Sq F value  Pr(>F)
Time                  1  54.33  54.334  6.2146 0.01413 *
Native.language       1  45.23  45.233  5.1737 0.02484 *
High.School.Language  1   0.10   0.097  0.0111 0.91640
Time.in.the.US        2   2.25   1.127  0.1289 0.87921
Residuals           112 979.21   8.743
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Interest :
                     Df Sum Sq Mean Sq F value    Pr(>F)
Time                  1 100.14 100.137 11.4823 0.0009706 ***
Native.language       1  43.93  43.927  5.0369 0.0267784 *
High.School.Language  1   0.47   0.474  0.0544 0.8159908
Time.in.the.US        2  16.80   8.398  0.9629 0.3849145
Residuals           112 976.75   8.721
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
 Response Ecological :
                     Df Sum Sq Mean Sq F value  Pr(>F)
Time                  1  71.43  71.427  8.5049 0.00428 **
Native.language       1  31.27  31.272  3.7236 0.05618 .
High.School.Language  1   0.36   0.358  0.0426 0.83690
Time.in.the.US        2   1.47   0.737  0.0877 0.91607
Residuals           112 940.62   8.398
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Language :
                     Df  Sum Sq Mean Sq F value    Pr(>F)
Time                  1  27.857 27.8569 11.9836 0.0007608 ***
Native.language       1  25.964 25.9636 11.1692 0.0011311 **
High.School.Language  1   2.207  2.2072  0.9495 0.3319453
Time.in.the.US        2   2.659  1.3296  0.5720 0.5660535
Residuals           112 260.353  2.3246
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

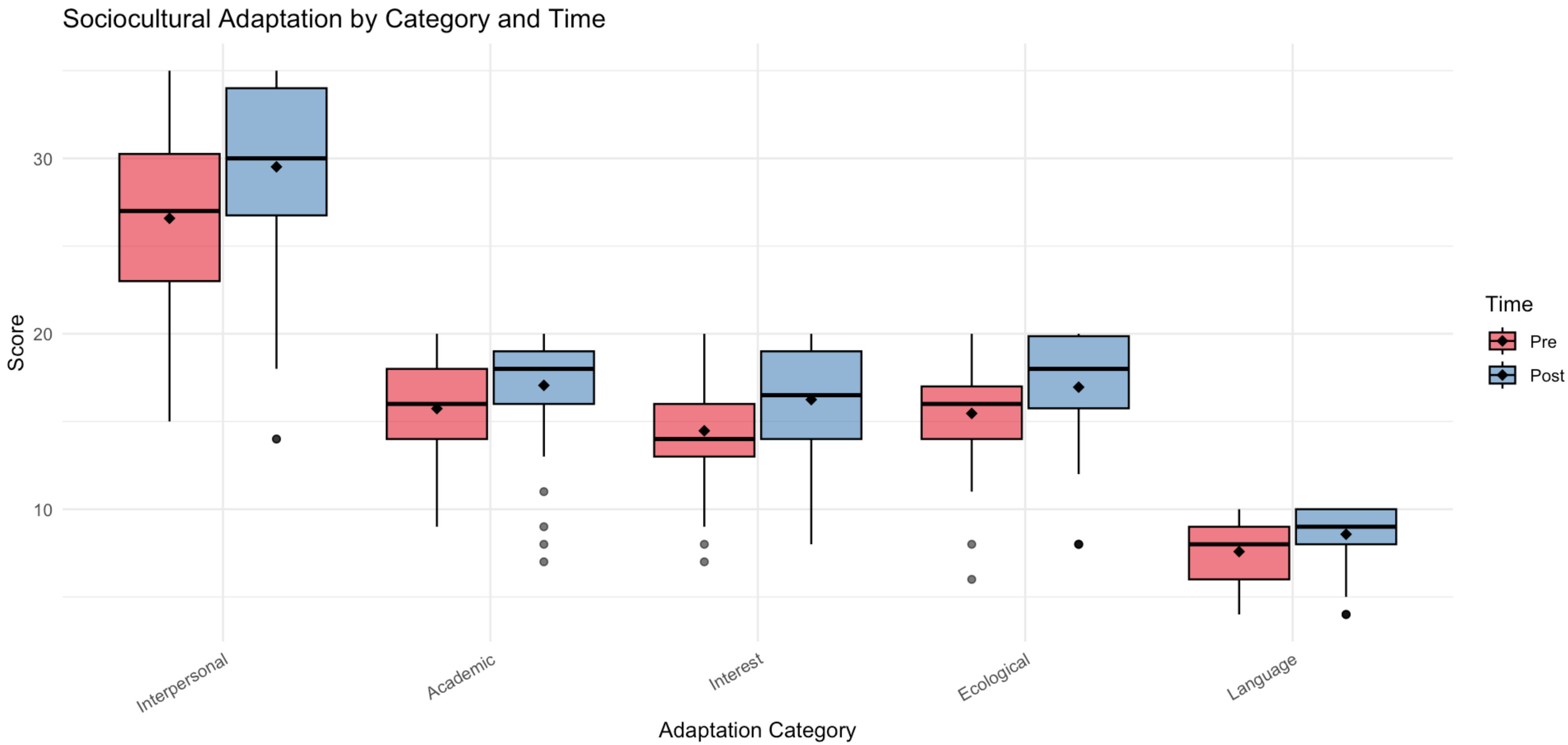
LVX VERITAS VIRTVS

Category-Specific Impacts of Time and Language on Sociocultural Adaptation

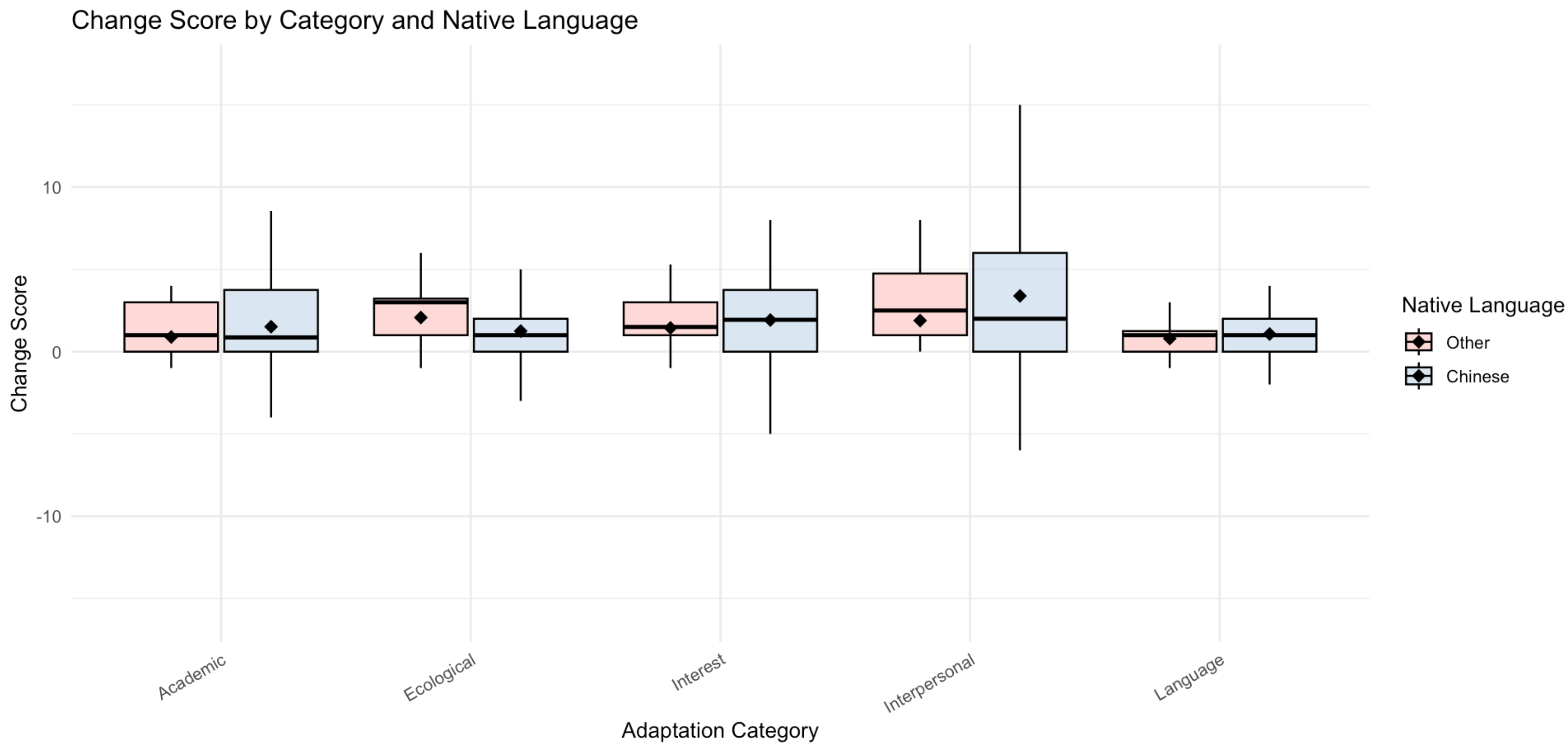
- Using univariate ANOVA results from a MANOVA model, we found Time (Pre vs. Post) had a statistically significant positive effect across all five domains.
- Native language background (Chinese vs. Other) also significantly influenced adaptation outcomes in four domains (except Ecological).
- High school language and time in the U.S. did not significantly influence most outcomes in this model.
- These findings highlight the importance of tailored cultural support, especially for students from linguistically and culturally distinct backgrounds.

LVX VERITAS VIRTVS

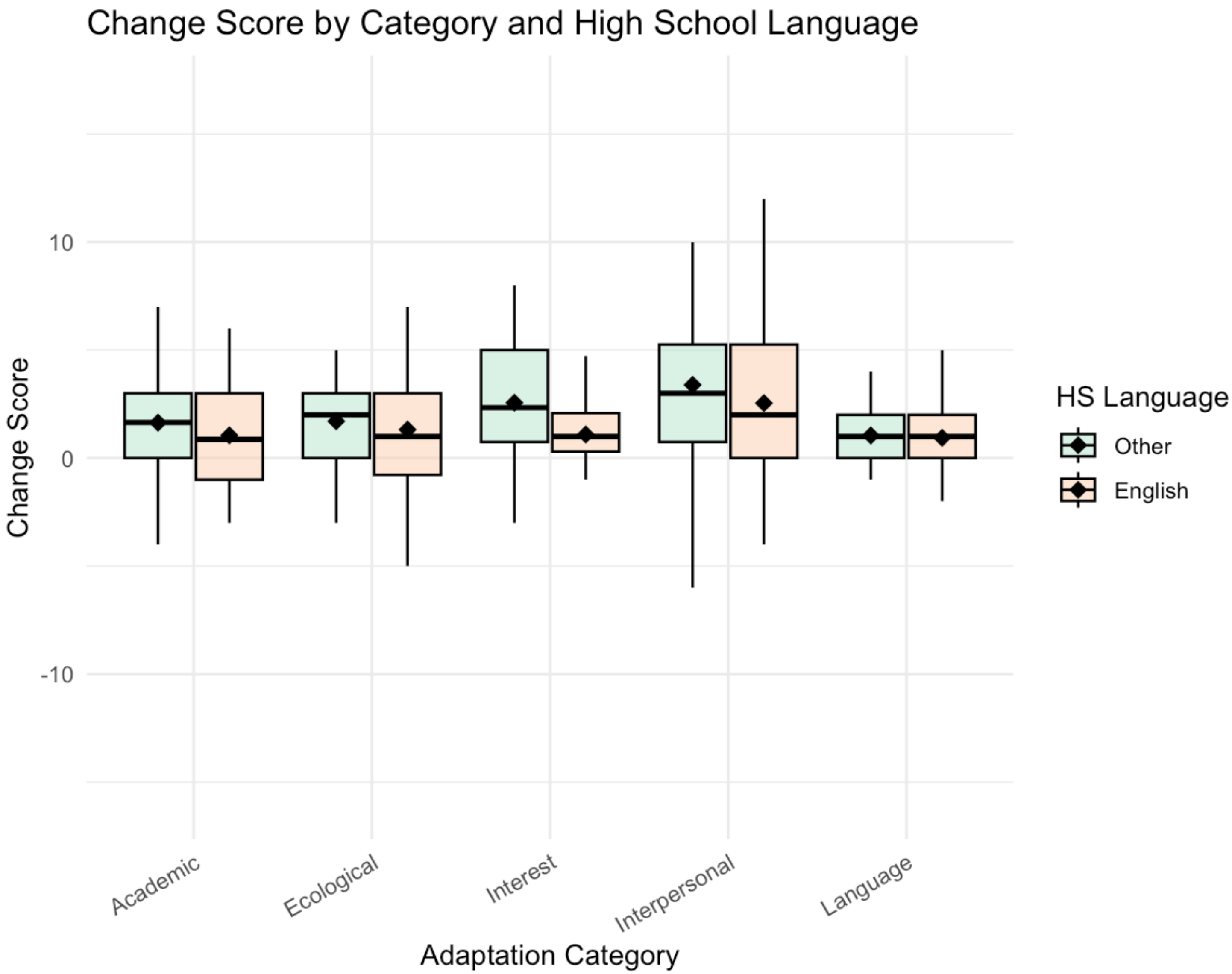
Sociocultural Adaptation Across Five Domains by Time (Pre vs. Post)



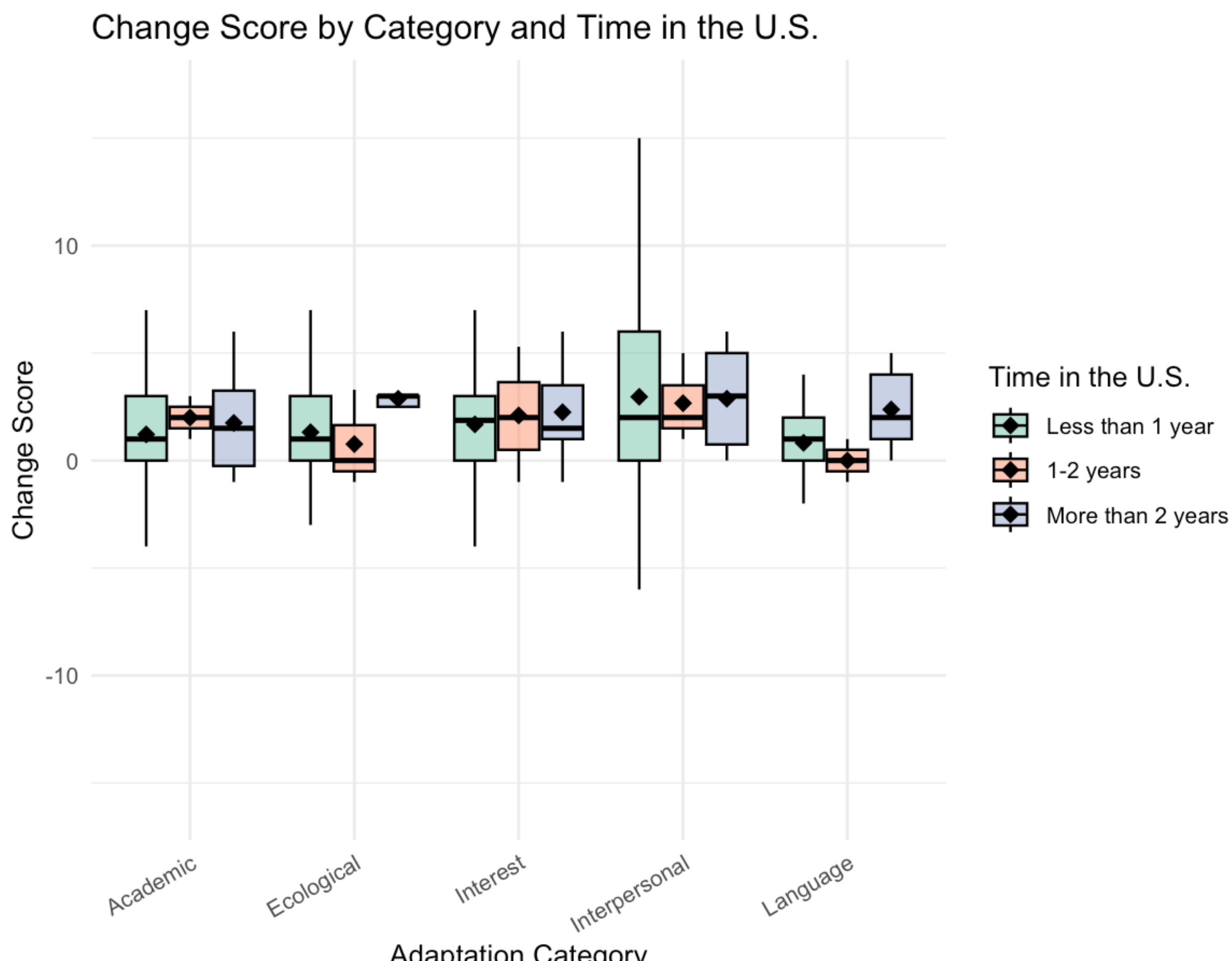
Change Scores by Adaptation Categories and Native Language



Change Scores by Adaptation Categories and High School Language



Change Scores by Adaptation Categories and Time in the US



Conclusion

- This study explored sociocultural adaptation among students by analyzing both overall scores and domain-specific outcomes. Multiple modeling approaches were used to assess the influence of language background, time spent in the U.S., and high school language of instruction.

1. Overall Change Patterns
   Linear models using total scores showed that:

- Students significantly improved from pre to post assessment (*Time* effect).
- Chinese-speaking students had lower average adaptation scores across time points.

2. Change Score Model
   When explicitly modeling the difference between post and pre (i.e., change scores), no statistically significant difference in overall improvement was found between Chinese-speaking students and others.

3. Domain-Specific Insights (MANOVA)
   Despite the lack of group-level differences in total change scores, the MANOVA model, which assessed each adaptation domain separately, revealed a more nuanced pattern:

- Significant gains over time were detected in several domains (e.g., Language, Interpersonal, Interest).
- Chinese-speaking students showed greater improvement in key domains, especially Language and Interpersonal adaptation.

Northeastern University

Thank you!