Descubriendo la Lectura (DLL) Improves Spanish Literacy Skills of Spanish-speaking First-grade Students who Have the Lowest Literacy Test Scores in Texas, Illinois and Arizona*

Jia Jia Ji

27 April 2021

Abstract

Since many Spanish-speaking elementary schoolers in U.S. have weakness in developing Spanish literacy skills, Descubriendo la Lectura (DLL) is one of the approaches aiming to solve this problem. To explore the effect of DLL program on Spanish literacy skills of these students, this paper analyzed the dataset collected from a randomized controlled trial where DLL program was the intervention and the first-grade Spanish-speaking students who had the lowest literacy test scores were randomly assigned to a control group receiving the regular education and a treatment group received the DLL training. I used the Logramos literacy test to measure the Spanish literacy skills of students, then I visualized the differences in posttest scores between 2 groups, and fitted a linear mixed model for each subtest (i.e. Reading, Language, Vocabulary, and overall score) that treated the random assignment and school as the main fixed effect and random effect respectively; and I found that DLL program significantly affected all posttest scores and the students receiving DLL training achieved higher scores than those receiving the regular teaching, so this program could improve the Spanish literacy skills of target students. The government can use these findings to implement more effective literacy programs to support the academic development of these Spanish-speaking students who had poor literacy skills.

1 Introduction:

Countries in the North America, such as Canada and U.S., often accept immigrants from foreign countries and these people bring various languages, culture and technologies together to create an open and diversified community. Meanwhile, they have to adapt to the new environment, especially with respect to learn English, receive the high-quality education and seek for employment opportunities. Among these English language learners, children have been thrown into focus because building a solid foundation in literacy skills for elementary schoolers is of importance. Since Spanish has been the second frequently used language around the world and there are many people from Spanish-speaking countries moving to the U.S. each year, many Spanish-speaking children encounter this problem. Usually, the differences between English-taught knowledge and Spanish culture may make them have difficulty in developing Spanish literacy skills, accordingly, the U.S. schools have tried different approaches to help to improve the Spanish literacy skills of these students, and Descubriendo la Lectura (DLL), which is a one-on-one tutoring program implemented in Spanish language (Trisha H. Borman; Geoffrey D. Borman 2019), is among them. Therefore, exploring whether DLL program affects the Spanish literacy skills of these struggling students is crucial, and the findings can guide the educators to utilize these effective programs to advance the Spanish literacy skills and potentially improve the long-term academic performance of these students.

This paper aimed to investigate the effect of DLL program on Spanish literacy skills of the Spanish-speaking early elementary schoolers who have poor literacy skills in the U.S.. The sample was 152 Spanish-speaking first-grade students whose idO (i.e. a standard assessment of literacy skills) test score is within the lowest 25% of their schools in Texas, Illinois and Arizona (Trisha H. Borman; Geoffrey D. Borman 2019). The

*Code and data are available at: https://github.com/jiaj6/effect_of_DLL_on_spanish_literacy_skills.

scores of Logramos, a Spanish literacy test, were used as the response variables to measure the students' Spanish literacy skills. The dataset was from OPENICSPER (Borman 2020) and it contained 152 records of the participated students' pretest and posttest scores in Reading, Language, Vocabulary and Total. Besides, a randomized controlled trial that used DLL program as the intervention was conducted to conclude the causal relationship between DLL program and Spanish literacy skills, and 78 sample students were randomly assigned to the treatment group enrolling in the DLL program and the others were in the control group receiving the regular teaching. The main fixed effect was DLL program and the random effect was the qualified schools. Therefore, the main research question was to estimate the differences in the average posttest scores between control group and treatment group. By addressing this question, we can know whether these Spanish-speaking students can take advantage of DLL to improve their literacy skills and whether the government can implement this program to a larger population to eliminated the problem of poor literacy skills faced by many Spanish-speaking children.

The remaining paper was organized into 5 sections. Firstly, I introduced the DLL program, and talked about the original paper and the extensions that I made. Then, in the Data section, I discussed the data collection process, along with the selection bias, ethical issues and non-responses of the experiment. I also gave an overview of the data variables and made graphs to compare the outcomes between control group and treatment group. Afterwards, in the Model section, I fitted a linear mixed model for each subtest (i.e. Reading, Language, Vocabulary and Total) of Logramos, and I explained the fixed effects, random effect, model assumptions and validation. In the Results section, I interpreted the estimates and p-values of each fitted model to see whether the effect was statistically significant and how DLL affected the outcomes. Then, in the Discussion section, I summarized this paper and gave the key findings, reviewed some previous works, and discussed the limitations and suggestions of this paper regarding the internal and external validity, ethics and choice of literacy skills measures.

2 DLL & Original Paper:

2.1 Descubriendo la Lectura (DLL):

DLL is a one-on-one literacy training program that aims to improve the literacy skills of first-grade Spanish-speaking English language learners who have poor literacy skills, and the students enrolling in this program received 30-minute daily literacy tutoring in Spanish from well-trained teachers (Trisha H. Borman; Geoffrey D. Borman 2019). There are 3 common assessments to test the literacy skills of students: Instrumento de Observación (IdO) and Logramos that both measure the Spanish literacy skills (and IdO is the most systematic assessment for DLL), and Iowa Test of Basic Skills (ITBS) that measures the English literacy skills; also, each of them have some subtests and an overall score that combines the scores of subtests (Trisha H. Borman; Geoffrey D. Borman 2019). DLL has been implemented in some U.S. states and it has the potential to improve the literacy skills of enrolled students.

2.2 Original Paper:

My paper was a replication and an extension of the original paper (Trisha H. Borman; Geoffrey D. Borman 2019) that collected the experimental data and analyzed it. The original paper introduced the background and the instruction mode of DLL program, also it explained the design of this randomized controlled trial with respect to the sample, random assignment and the assessments to measure students' literacy skills. To analyze the impact of DLL on literacy skills, for each assessment, the original paper fitted a hierarchical linear model that included the pretest measures and DLL assignment as predictors and also included the school-specific error. It concluded that the positive impact of DLL on literacy skills was statistically significant for all IdO and Logramos measures, and the positive impact of DLL on literacy skills was educationally significant for the ITBS measures.

I mostly referred to the original paper for the design and implementation process of this experiment, and I also discussed the strengths, weaknesses (regarding the biases) and ethical issues of this data collection process and the intervention. Besides, I only focused on the Logramos test (i.e. a Spanish assessment), because the data on this measure was publicly available and I also consider it as a standard and equally important indicator to measure the Spanish literacy skills of participated students. In terms of the data

cleaning, besides removing the null values of each variable as the original paper did, I also examined the extreme values in the data and replaced each outlier with the average of the corresponding variable and the assigned group (of this instance). Moreover, I displayed the summary statistics and graphs to visualize the differences between 2 groups (i.e. control group and treatment group) in both pretest measures and posttest measures to see whether DLL was possible to be effective based on the descriptive statistics. The final extension was that for each subtest, I explained all aspects of its linear mixed model regarding the fixed and random effects, relationships among variables, assumptions, strengths, weaknesses and validation of model; and I interpreted the estimates and p-values (together with the hypothesis test) of each fitted model to conclude the causal relationship between DLL program and Spanish literacy skills.

3 Data:

I used `R` (R Core Team 2019), the `tidyverse` package (Wickham et al. 2019), the `ggplot2` package (Wickham 2016),the `gridExtra` package (Auguie 2017), and the `knitr` package (Xie 2019) (Xie 2015) (Xie 2014) to analyze the data and make plots and tables.

Also, I made a data sheet[1] explaining the data information in detail based on this paper (Gebru, n.d.).

3.1 Intervention & Data Collection Methodology:

The data was downloaded from the website OPENICPSR (Borman 2020) that is a public repository where people can share the research data, and it can be freely accessed[2] as long as the future users have created a free account on this website. The data was collected using a Randomized Controlled Trial (RCT) that was designed by the authors of the original paper (Trisha H. Borman; Geoffrey D. Borman 2019) and aimed to explore the effect of DLL program on the literacy skills of first-grade Spanish-speaking students who performed poor in literacy. This experiment was conducted in the 2016-2017 school year (Trisha H. Borman; Geoffrey D. Borman 2019), but this data only recorded the information of participants during Fall semester. I treated the DLL program as the intervention and the sample students were randomly assigned to a control group and a treatment group. The Students in control group received the regular education in school and the students in treatment group received the daily DLL training in school. The administrators monitored the experiment to ensure the separation of 2 groups and implementation of program (Trisha H. Borman; Geoffrey D. Borman 2019). The population was all Spanish-speaking students that were at the early stage of elementary education and had poor literacy skills in the U.S.. Since the RCT was conducted by recruiting the schools that had more than 1 year experience in implementing DLL program and the first-grade Spanish-speaking students whose IdO test scores were within the lowest 25% of their schools in the U.S. (Trisha H. Borman; Geoffrey D. Borman 2019), the frame was these qualified students educated in the qualified schools in the U.S., and the sample was 152 students among 22 qualified schools in Texas, Illinois and Arizona sampled within the frame (Trisha H. Borman; Geoffrey D. Borman 2019).

The random assignment procedure of this experiment used the random number generator, the students that got the lowest 78 numbers were in the treatment group, and the others were in the control group (Trisha H. Borman; Geoffrey D. Borman 2019). This random allocation of treatment ensured that each sample student had an equal chance of being assigned to the treatment group and receiving the treatment (ie. DLL program). Also, it avoided the counterfeit counterfactual estimate and selection bias caused by the self-selection of students (Gertler 2016), because students who had poorer literacy skills may tend to join the treatment group and enroll in the potential effective program. However, since no random sampling method, like Simple Random Sampling or Stratified Random Sampling, was used to generate a sample within the frame, this sample was not random. Accordingly, the selection biases, such as selection for convenience and self-selection bias, may exist in the sample, because the experimenters may choose the students that could participate in the experiment more conveniently and the students with lower test scores tended to participate in the experiment to have the chance to receive the literacy training. Besides, this non-random sampling method, along with the relatively small sample size (i.e. 152), lowered the representativeness of the sample

[1]Data sheet is available at: https://github.com/jiaj6/effect_of_DLL_on_spanish_literacy_skills.

[2]The data is available at: https://www.openicpsr.org/openicpsr/project/118041/version/V1/view?path=/openicpsr/118041/fcr:versions/V1/DLL_lgrm.csv&type=file

and restricted the external validity, so that the analysis results on the sample may not be appropriately generalized to a larger population. Moreover, since the demographic information of participated students was not available in the dataset, I could not ensure the observable characteristics between control group and treatment group were similar. But I could initially assume that the pre-test scores of students between groups were similar, because the sample students all had the lowest 25% pre-test scores of their schools. So, although the allocation of treatment was random, the non-random sampling method, small sample size and missing demographics made the similarity of observed and unobserved characteristics between groups uncertain (Gertler 2016). In other words, other factors, such as sex, race and the highest education levels of their parents, may also vary between groups other than the intervention (i.e. DLL program) and affect the comparisons of posttest scores between groups. It may also restrict our ability to conclude that the differences in posttest scores between the control group and treatment group were totally caused by DLL, which may further restrict the internal validity.

Furthermore, this sample included the students among 22 schools across 3 U.S. states, which reduced the biases due to school differences or geographical differences (i.e. the schools in some specific areas may have higher teaching quality than others) and then improved the generalization of analysis results. Also, all participated students took the pretest and posttest of Logramos before and after the implementation of treatment respectively at the same time (Trisha H. Borman; Geoffrey D. Borman 2019), which removed the bias due to timeframe difference. Otherwise, some students may have received the DLL training or regular teaching for a longer period and have their literacy skills advanced. More importantly, this RCT was performed for an entire school year, and the students in treatment group received DLL training in the first semester and the students in control group would receive the same training in the second semester (Trisha H. Borman; Geoffrey D. Borman 2019). So, all participants had the opportunity to enroll in this potentially effective program and improve their literacy skills, and this design eliminated the ethical concerns that in many clinical trials, the treatment was only open to the participants in treatment group even when the treatment was highly potentially to be beneficial.

3.2 Data Overview & Data Cleaning:

The dataset contained the information about school, allocation of treatment, pretest and posttest scores for 152 sample students participated in the RCT. There were 152 instances and 12 variables. Specifically, the variables were:

1. StudentId: It represented the unique id number for each student and it was a nominal, categorical variable.
2. SchoolId: It represented the unique id number for each of the 22 qualified schools and it was a nominal, categorical variable. It was used to identify the corresponding school for each student and was included in the models as the random effect to adjust for the school differences.
3. Group and T_assignment: Group was the nominal, categorical variable and T_assignment was the binary variable (that only had 2 values: 0 and 1). They both represented the results of random assignment that 0 indicated the control group (i.e. delayed group that received the DLL training in the second semester) and 1 represented the treatment group (i.e. immediate group that received the DLL training in the first semester).
4. Pretest_Reading, Pretest_Language, Pretest_Vocabulary and Pretest_ELA.Total: The first 3 variables represented the test scores of each subtest (i.e. Reading, Language and Vocabulary) in Logramos took by students before implementing the treatment, and the variable Pretest_ELA.Total represented the overall score that combined the scores of 3 subtests. They were all continuous variables and were treated as the explanatory variables in the models to adjust for the slight differences in pretest scores between groups.
5. Posttest_Reading, Posttest_Language, Posttest_Vocabulary and Posttest_ELA.Total: Similarly, these variables represented the test scores of 3 subtests in Logramos took by students after the treatment finished, as well as the overall score combining the subtests' scores. They were all continuous variables and were considered as the response variables.

The sample units were people in the data, but there was no confidential information included because each instance was represented by a unique id (instead of name) and no demographics in the data. However, the data recorded the academic performance, which may be considered as sensitive information for some people.

To clean the data, firstly I removed the variable Group, because it represented how each student was assigned to the group, which was the same as the variable T_assignment and was redundant in the data. I kept the variable T_assignment, because it was a binary variable and could be included in the models to distinguish the groups easily and compare the outcomes between groups. Secondly, Row 113 and row 145 were 2 instances that had extremely low posttest scores in all subtests (i.e. both rows had 94 for Posttest_Reading, 105 for Posttest_Language, 92 for Posttest_Vocabulary, and 99.16667 for Posttest.ELA_Total). Since their posttest scores were exactly the same for all subtests but their pretest scores seemed normal compared to other students' scores (i.e. not too low), I could appropriately assume that these extreme values were manually recorded incorrectly (i.e. measurement error). Thus, I considered them as the outliers and replaced the pretest scores with the average of the corresponding subtest for this instance's assigned group (when calculating the mean for a variable, I did not consider the null values for now) to make the data less biased. Moreover, non-responses existed in the experiment, because some sampled students may refuse to participate in or exit the experiment early. And these non-responses may lead to the over-estimate of the effect of DLL on Spanish literacy skills, because the students in treatment group that considered this program as useless during the training may exit early and not take the posttest. Also, these non-responses would lead to a smaller sample size and lower the generalization of analysis result. In the data, there were 30 instances containing missing values for the pretest or pos-test scores (28 null values for Pretest_Reading and Posttest_Reading, 14 for Pretest_Language and Posttest_Language, 11 for Pretest_Vocabulary and Posttest_Vocabulary, 29 for Pretest_ELA.Total and Posttest_ELA.Total). I ignored the null values for each pretest and posttest score, because the number of null values for each subtest was still within a reasonable range compared to the sample size (i.e. 152).

3.3 Summary Statistics & Graphs:

The variable T_assignment was a binary variable with 2 levels: 0 and 1 (0 indicated the control group and 1 indicated the treatment group). I used this variable to distinguish the control group and treatment group, and I visualized the differences in pretest scores and posttest scores between groups respectively by creating some graphs and summary statistics.

3.3.1 Comparing the pretest scores between groups: both of them shared similar Spanish literacy skills overall before the treatment

The variables Pretest_Reading, Pretest_Language, Pretest_Vocabulary and Pretest_ELA.Total were continuous. For each subtest and the overall score, I made a histogram showing the distribution of the score for 2 groups, and a boxplot showing the spread of the score for each group. Also, I calculated the average of each pretest score for each group to see whether students between groups shared similar pretest scores generally.

Firstly, Figure 1 illustrated that more students in the treatment group achieved median-to-high pretest score in Reading (i.e. scores higher than 150) than that for the control group, and the spread of this score for treatment group was also slightly higher regarding the quantiles (i.e. first quantile Q1, median, third quantile Q3). Besides, Table 1 showed that the average of pretest Reading score for the students in treatment group was 153.91, which was 2.22 higher than that for the students in control group (151.69 for control group). This indicated that the students in treatment group had slightly stronger Spanish literacy skills in Reading than those in control group before the treatment. However, Figure 1 also showed that more students in the treatment group had low-to-median pretest Language score (i.e. scores lower than 162.5) than those in the control group, and the spread of this score for treatment group was lower regarding the quantiles. Also, the students in treatment group achieved 166.21 for this score on average, which was 4.85 lower than those in control group (171.06 for control group). This showed that the students in treatment group had poorer Language skills, which was the opposite situation of pretest Reading score.

Besides, as the histograms in Figure 2 shown, for both pretest Vocabulary score and pretest overall score, the distributions between control group and treatment group were similar, in other words, the number of

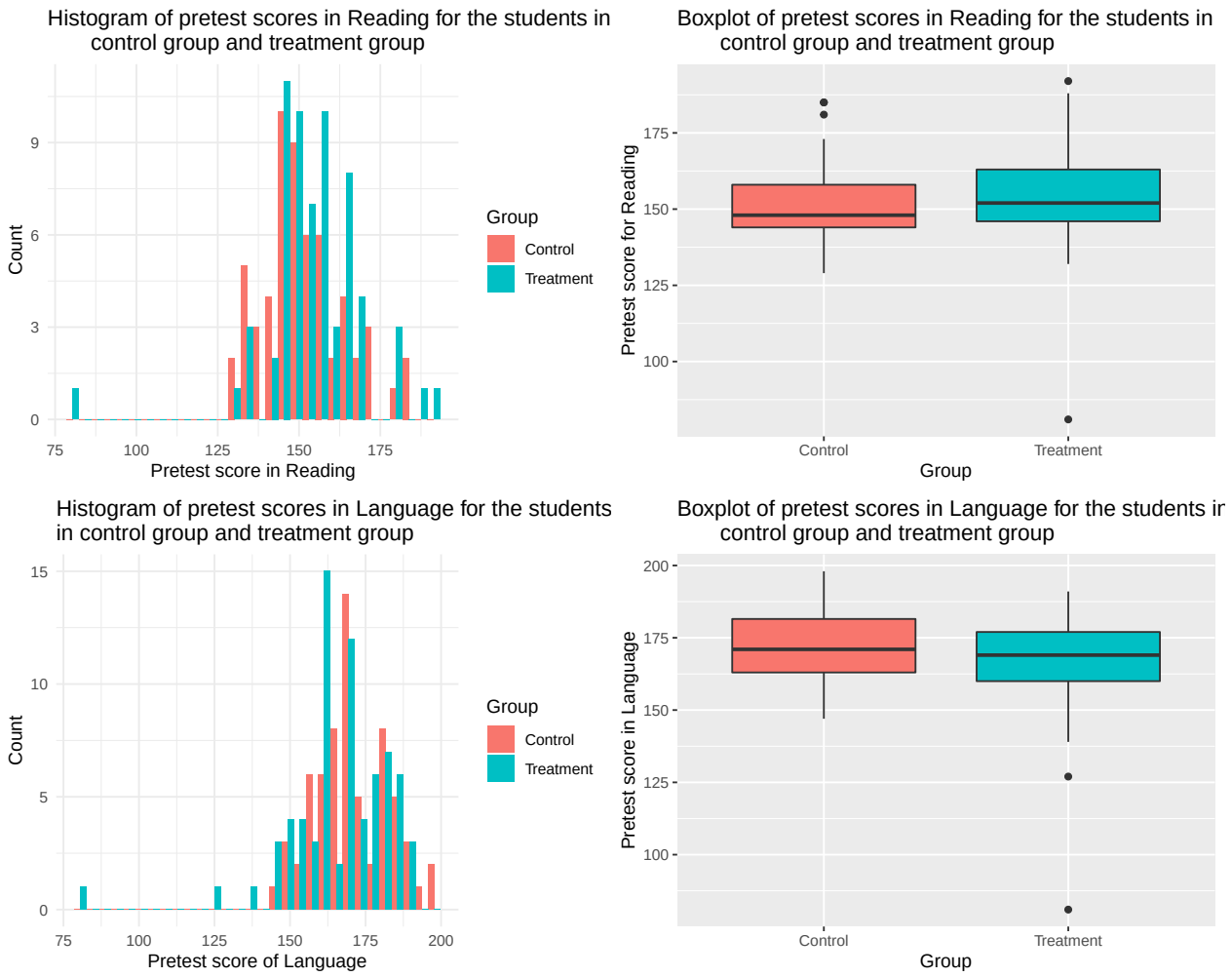

Figure 1: The histograms and boxplots of pretest scores in Reading and Language for the students in control group and treatment group

students distributed within a specific score range was similar. Also, these 2 groups had similar spreads in both the boxplot for pretest Vocabulary score and the boxplot for overall score, except that the control group had a slightly higher Q3 than the treatment group for the overall score. Table 1 showed that the students in control group and treatment group had very similar average score for both pretest Vocabulary score and overall score (pretest Vocabulary score: 160.28 for control group, 159.34 for treatment group, overall score: 162.98 for control group, 162.50 for treatment group). So, these graphs and numbers indicated that the students in control group and treatment group shared a similar level of Vocabulary skills and overall Spanish literacy skills before the treatment.

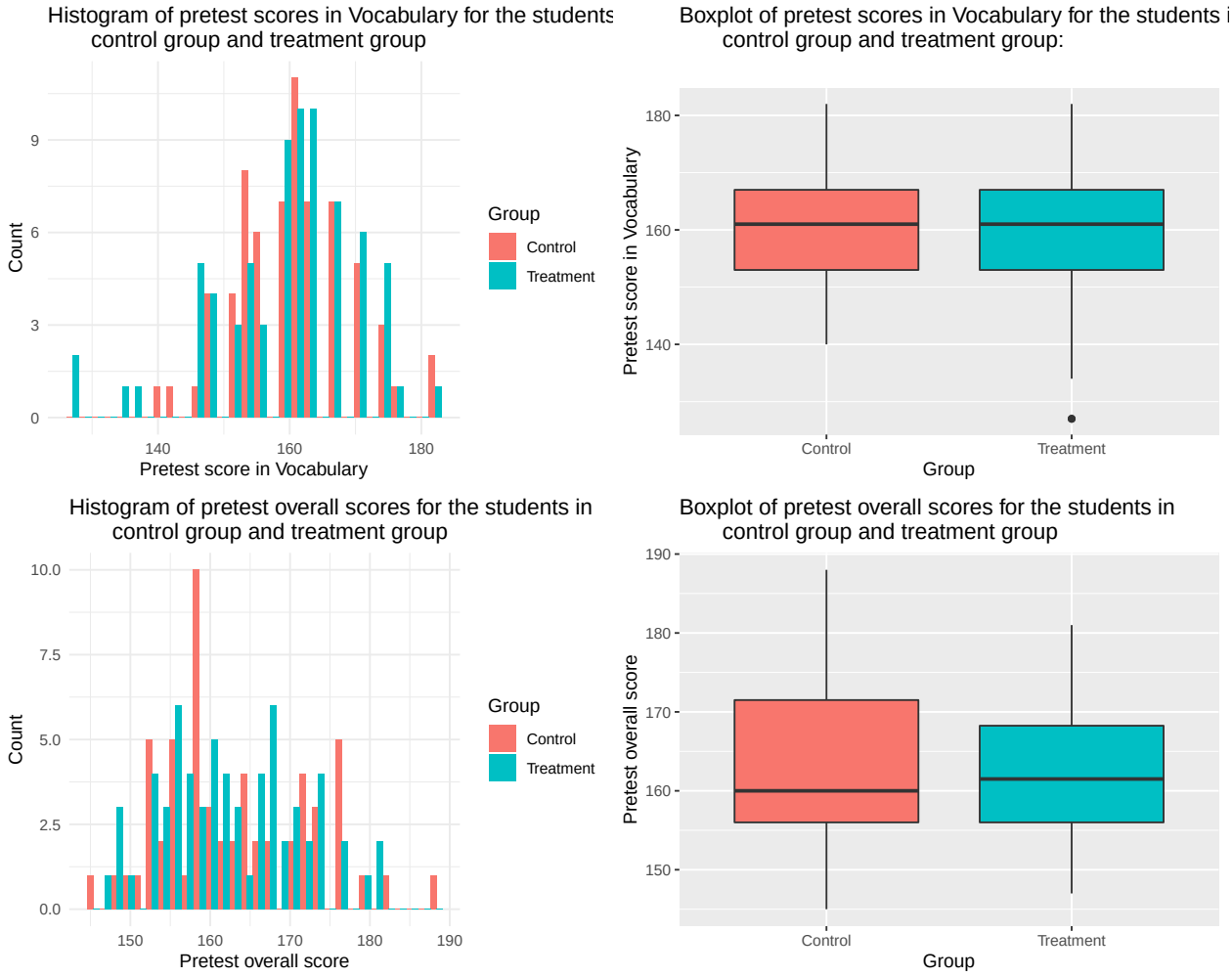

Figure 2: The histograms and boxplots of the pretest scores in Vocabulary and the overall score for the students in control group and treatment group

In the data, although the treatment group had slightly higher pretest Reading score and lower pretest Language score than the control group on average, the overall score balanced this opposite situation, so that the 2 groups had very similar overall score on average. Since the overall score combined all subtests' scores, using it to compare the students' overall Spanish literacy skills between groups before the treatment was more appropriate. So, I could assume that the observable characteristic of students' overall Spanish literacy skills before the treatment was controlled to be invariant between groups, and the internal validity was improved.

3.3.2 Comparing the posttest scores between groups: the students in treatment group had higher Spanish literacy skills on average after the treatment

Table 1: Average value of the 3 subtests' pretest scores and the overall score for the students in control group and treatment group

	Reading	Language	Vocabulary	Overall score
Control	151.69	171.06	160.28	162.98
Treatment	153.91	166.21	159.34	162.50

The variables Posttest_Reading, Posttest_Language, Posttest_Vocabulary and Posttest_ELA.Total were continuous. Similar to the pretest scores, for each subtest and the overall score, I made a histogram and a boxplot, and calculated the average for each group to compare the posttest scores between groups and see whether the treatment was effective (i.e. the students in treatment group tended to achieve higher posttest scores generally).

Figure 3 and Figure 4 illustrated that for each subtest's posttest score (i.e. Reading, Language and Vocabulary) and the overall score, more students in the treatment group achieved median-to-high score (i.e. scores higher than 175) than those in the control group. Meanwhile, for each posttest score, the treatment group had higher spread than that for the control group regarding the quantiles (i.e. Q1, median and Q3). Also, Table 2 showed that the average posttest overall score for the students in treatment group was 176.39, which was 4.44 higher than that for the students in control group (171.95 for control group). Similarly, the students in treatment group had higher posttest scores on average for all 3 subtests (i.e. Language: 177.62 and 180.42 for control group and treatment group respectively, Vocabulary: 164.38 and 169.10 for control group and treatment group respectively). Particularly, the average posttest Reading score for treatment group was 173.08, which was 6.1 higher than that for control group (166.98 for control group), probably because DLL is a Spanish-version Reading Recovery Program (Trisha H. Borman; Geoffrey D. Borman 2019). Therefore, the students receiving the DLL training tended to achieve stronger Spanish literacy skills than those receiving the regular teaching, and DLL program was possible to be effective.

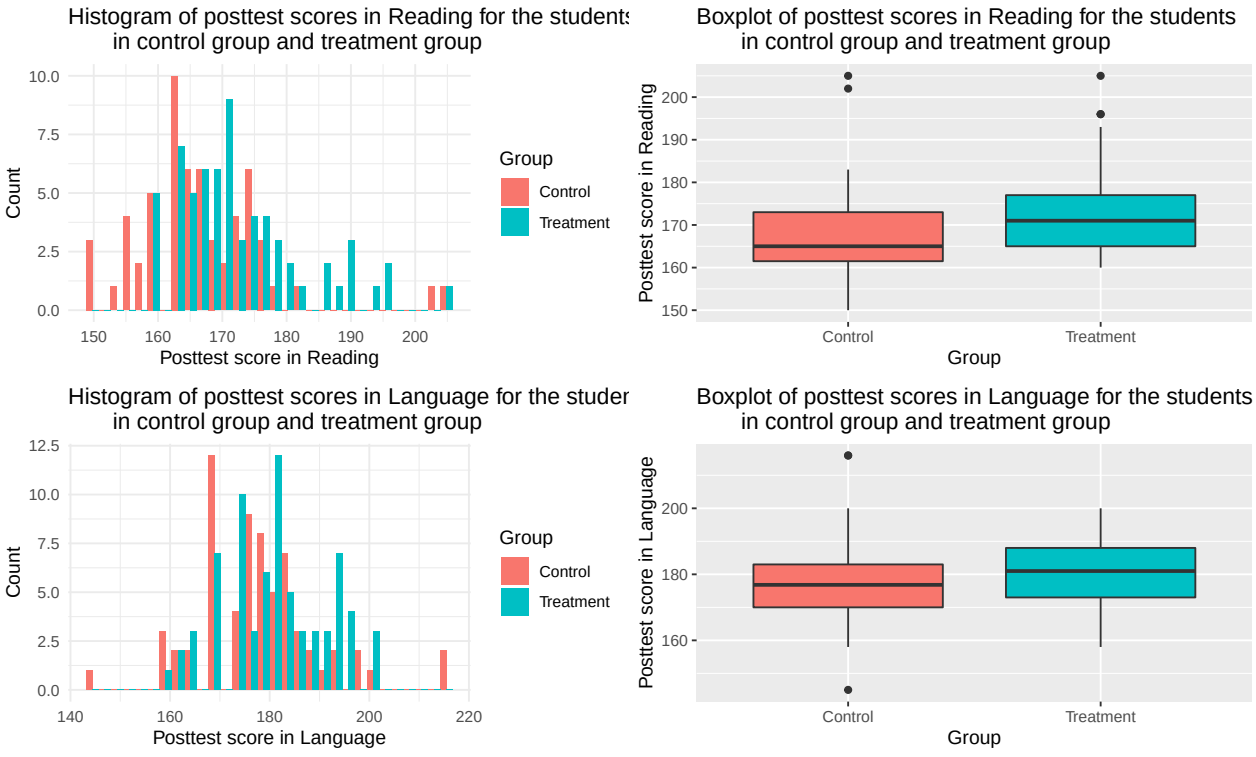

Figure 3: The histograms and boxplots of the posttest scores in Reading and Language for the students in control group and treatment group

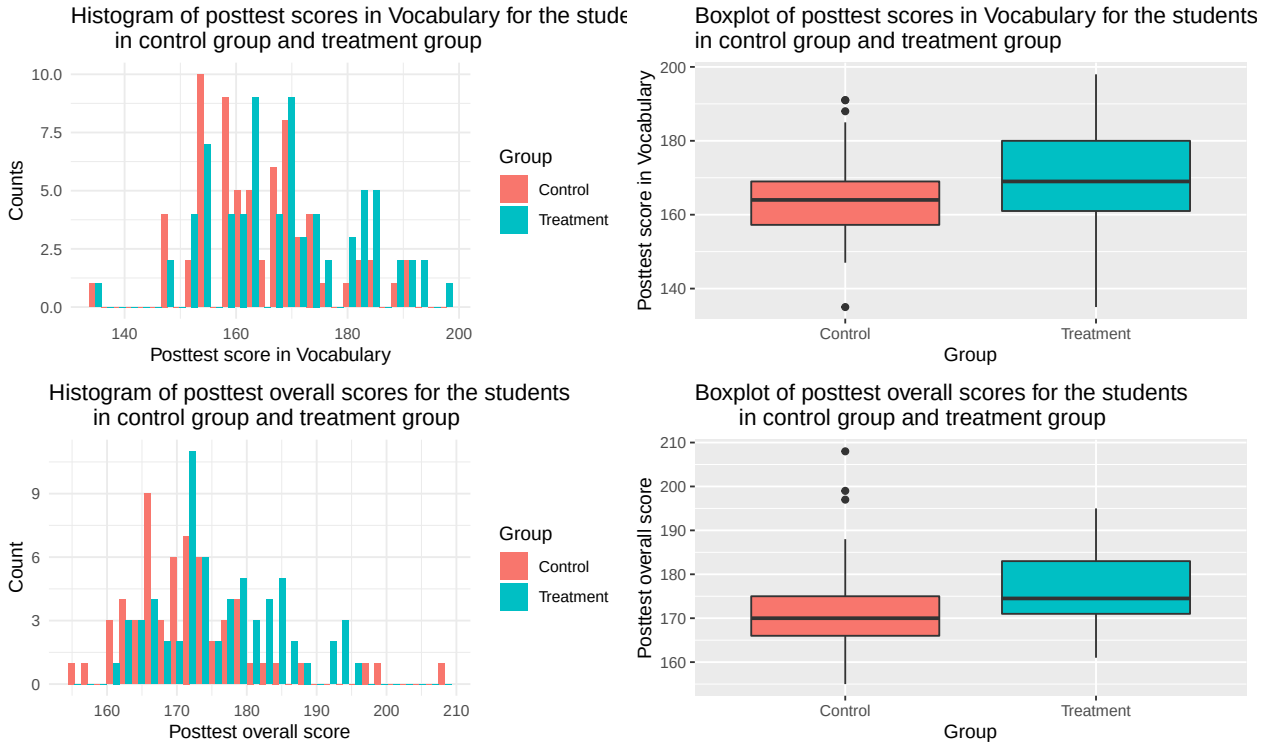

Figure 4: The histograms and boxplots of the posttest scores in Vocabulary and the overall score for the students in control group and treatment group

Table 2: Average value of the posttest scores for 3 subtests and the overall score for the students in control group and treatment group

	Reading	Language	Vocabulary	Overall score
Control	166.98	177.62	164.38	171.95
Treatment	173.08	180.42	169.10	176.39

Moreover, in Figure 5, most points for the treatment group were distributed in the upper part of the plot, also the linear best fit line for treatment group was above that for control group, which showed that most students that received DLL training tended to have higher posttest overall score than those that had similar pretest overall score but received the regular teaching. Thus, DLL program was possibly effective for the target students.

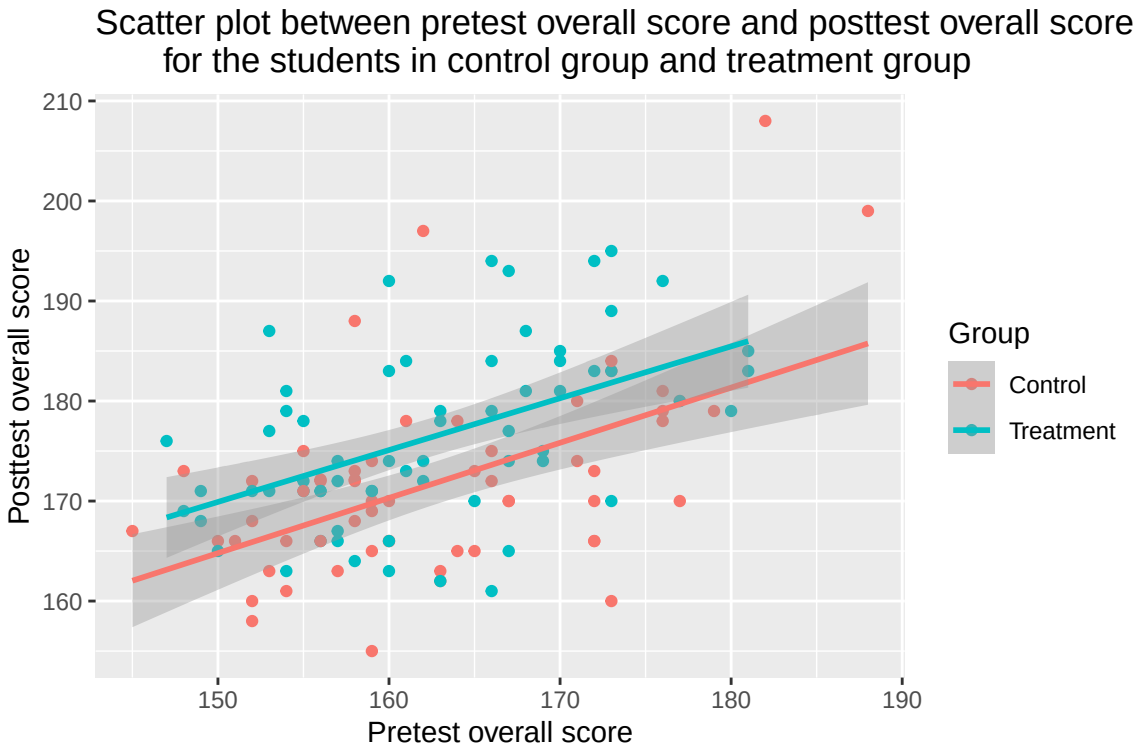

Figure 5: The scatter plot between pretest overall scores and posttest overall score for the students in control group and treatment group

4 Model:

I built a model for each subtest (including the overall score) to see the effect of DLL on the skills of each subtest more clearly. Also, I made a model card[3] explaining the model information in detail based on this paper (Mitchell, n.d.).

This paper aimed to investigate whether the intervention (i.e. DLL program) affected the Spanish literacy skills of students, accordingly, each posttest score of Logramos was treated as the outcome to measure the students' Spanish literacy skills after receiving the DLL training or regular teaching. So, the continuous variables Posttest_Reading, Posttest_Language, Posttest_Vocabulary and Posttest_ELA.Total were included in the models as the response variables. Since the posttest score was continuous, and motivated by the appropriate linear best fit line in the scatter plot between pretest overall score and posttest overall score in the Data section, I chose the linear relationship in this model. Besides, since students were sampled from schools and the repeated measurements existed, there may be correlations within each school and some schools may have higher teaching quality than others. Thus, I used the linear mixed model and treated the variable School_id as the random effect to represent each school and adjust for the school differences. Meanwhile, the random allocation of treatment was treated as the fixed effect (i.e. explanatory variable) to explore its relationship to the posttest scores, so I used the binary variable T_assignment in the model to easily distinguish the 2 groups and make predictions on the posttest scores based on the group. Besides, in the

[3]Model card is available at: https://github.com/jiaj6/effect_of_DLL_on_spanish_literacy_skills.

Data section, although I found that students between groups had similar pretest overall scores, the students in treatment group still had slightly higher and lower pretest score in Reading and Language respectively than those in control group. So, I included the pretest score, which belonged to the same subtest as the post-test score, as another fixed effect to adjust for the slight difference in this subtest's pretest score between groups and improve the accuracy of analysis results. I added the continuous variable Pretest_Reading, or Pretest_Language, or Pretest_Vocabulary or Pretest_ELA.Total as another explanatory variable in the model. By fitting this linear mixed model with the data, I could know how being assigned to different groups and receiving different forms of education would affect the posttest scores, and conclude that whether the intervention (i.e. DLL program) affected the Spanish literacy skills of target students and whether this program was effective.

I fitted 4 linear mixed models separately, each was for each subtest's pretest score and posttest score. Each model had the same form:

$(posttestscore)_{ij} = \beta_0 + \beta_1(T\_assignment)_{ij} + \beta_2(pretestscore)_{ij} + (School\_id)_j + (Residual)_{ij}$

where i was the student index, j was the school index; β_0 was the intercept or the post-test score for a student that was in the control group and had a pre-test score of 0, β_1 was the estimate of the effect of random assignment (that was, the intervention) on a subtest's posttest score of student i in school j, T_assignment was the random allocation of treatment for student i in school j and its reference level was 0 (i.e. control group), β_2 was the estimate of the effect of a subtest's pretest score on its posttest score of student i in school j, School_id was the random effect and it was used to identify the school, Residual was the student-level error term for student i in school j. Also, posttest score and pretest score were the score for each subtest before and after the treatment respectively for student i in school j.

For example, the model for the subtest Reading was:

$(Posttest\_Reading)_{ij} = \beta_0 + \beta_1(T\_assignment)_{ij} + \beta_2(Pretest\_Reading)_{ij} + (School\_id)_j + (Residual)_{ij}$, i = 1, 2, ..., 124, j = 1, 2, ..., 22

So, the model made predictions on the posttest score for each subtest based on the group assignment and this subtest's pretest score of each student, meanwhile it used the random effect School_id to consider the correlations within each school and the school differences. It also had a coefficient before each explanatory variable to estimate its effect on the outcome.

I did not include the interaction term between T_assignment and pretest score in the model, because this interaction term represented that the effect of being assigned to a group on the posttest score also depended on the pretest score, and then the RCT should have 4 groups: 1 control group included the students with high pretest scores, 1 treatment group included the students with high pretest score, 1 control group included the students with low pretest scores and 1 treatment group included the students with low pretest score. While the design of RCT in this paper ensured that the control group and treatment group only differed in the DLL program, and the pretest scores between these 2 groups were similar according to the qualifications of sample students and the graphs of pretest scores between groups in the Data section, thus the additive model was more appropriate here than the interaction model.

This linear mixed model involved the linear relationship, so it was straightforward and easy to interpret. And, this model was more appropriate than the Linear Regression Model, because it also considered the random effect based on the nature of this RCT (i.e. had repeated measurements) and simply using the linear model would be misleading. If the participated students were all sampled from the same school, then no random effect needed in the model and the linear regression model was enough to model the effect of DLL on posttest score. As I discussed in the Data section, the data did not have the demographics, so that I could not include the demographic variables as another fixed effects in the model and adjust for any differences in these variables between groups to improve the model performance.

Based on the aim of this paper (i.e. to explore the effect of DLL program on the posttest score), this model was equivalent to the hypothesis test:

Null hypothesis H_0: β_1 was 0, i.e. no effect of DLL on the posttest score

Alternative hypothesis H_1: β_1 was not 0, i.e. there is effect of DLL on the posttest score.

So, I could check the p-values for each variable after fitting the model to see whether I should reject the null hypothesis or not and whether an effect was statistically significant. If an effect was statistically significant, I could also check its estimated coefficient to explore how this variable affected the outcome (i.e. posttest score).

There were 3 main assumptions for the linear mixed model:

1. Random effect School_id followed the Normal distribution with a mean of 0 and constant variance.
2. Error term Residual followed the Normal distribution with a mean of 0 and constant variance.
3. Random effect and error term were independent.

Firstly, I made the histogram and QQ plot of School_id in the fitted model to check whether the random effect was normally distributed. Also, I made a scatter plot between School_id and fitted values to check whether the random effect had the constant variance. Similarly, I made the same 4 plots for errors to check whether the errors was normally distributed and had constant variance. Also, I made the QQ plot for marginal residuals and the scatter plot between marginal residuals and fitted values to see whether a linear relationship was appropriate. Besides, I computed the variance-covariance of the fitted model to check whether it was appropriate.

5 Results:

I used `R` (R Core Team 2019), the `tidyverse` package (Wickham et al. 2019), the `ggplot2` package (Wickham 2016), the `gridExtra` package (Auguie 2017), the `knitr` package (Xie 2019) (Xie 2015) (Xie 2014), the`emmeans` package (Lenth 2019), the `nlme` package (Pinheiro et al. 2019), the `Pmisc` package (Brown 2019), and the 'lme4' (Bates et al. 2015) to format and visualize the results from the linear mixed models.

I fitted the linear mixed model for each subtest and the overall score. For each fitted model, I displayed the summary statistics using tables and visualized the predicted outcomes using scatter plots. For the summary statistics, MLE (Maximum Likelihood Estimate) represents the estimate of the effect of a variable on the posttest score using Maximum Likelihood Estimation method, p-value is the probability of getting a value as extreme or more extreme than observed statistic if assuming that the null hypothesis is true. I used 0.05 significance level to determine whether to reject the null hypothesis or not, specifically, I would reject the null hypothesis if p-value was smaller than 0.05, while I would fail to reject the null hypothesis if p-value was larger than 0.05. But this significance level would not be used as a strict threshold to determine whether a variable significantly affected the outcome, the decision would also depend on the context. The emmeans (estimated marginal mean) is the predicted average outcome for a model, and 95% confidence interval represents that there is 95% probability that a range of value contains the true mean.

5.1 Result of the model for posttest score in Reading:

Table 3 showed that the MLE (i.e. estimated coefficient) for the intercept was 116.50, which represented that a student in treatment group with pretest Reading score of 0 was predicted to have a posttest score in Reading of 116.50. Besides, the MLE for the allocation of treatment (i.e. T_assignment) was 5.36, which showed that the posttest score of a student in treatment group was 5.36 higher than that of a student in control group with the same pretest Reading score. Also, the p-value of T_assignment was 0.0015, which was much smaller than 0.05, so I rejected the null hypothesis and considered the variable T_assignment as significantly affecting the posttest Reading score.

Moreover, Table 4 showed that the estimated marginal mean for the control group and treatment group was 167.37 and 172.73 respectively, which showed that the students in treatment group tended to achieve higher average posttest score in Reading. Also, the 95% confidence interval for the control group and treatment group was [165, 170] and [170, 175] respectively, so the treatment group had a higher upper bound and a higher lower bound than the control group, which indicated that the students in treatment group were more likely to achieve a higher average Reading score than those in control group.

5.2 Result of the model for posttest score in Language:

Table 3: The summary statistics for the fitted Reading model

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	116.4971	8.8710	102	13.1323	0.0000
T_assignment1	5.3603	1.6389	102	3.2707	0.0015
Pretest_READING	0.3328	0.0580	102	5.7421	0.0000
σ	0.0007	NA	NA	NA	NA
τ	9.0862	NA	NA	NA	NA

Table 4: The estimated marginal means for T.assignment in the fitted Reading model

T_assignment	emmean	SE	df	lower.CL	upper.CL
0	167.37	1.18	19	164.89	169.85
1	172.73	1.13	19	170.36	175.09

Table 5 showed that the MLE for the intercept was 156.57, which was the predicted posttest score in Language of the student in treatment group with pretest Language score of 0. Then, the MLE for T_assignment was 3.42, which showed that the posttest Language score of the student in treatment group was 3.42 higher than that of the student in control group with the same pretest score. Also, the p-value of T_assignment was 0.065 and it was slightly larger than 0.05, so I failed to reject the null hypothesis, which indicated that T_assignment did not statistically significantly affected the posttest score in Language.

Besides, Table 6 showed that the estimated marginal mean for control group and treatment group was 177.48 and 180.90 respectively, which showed that the students in treatment group tended to have higher average posttest Language score. Also, the 95% confidence interval for control group and treatment group was [174, 181] and [178, 184] respectively, so the students in treatment group were more likely to achieve a higher average Language score than those in control group.

5.3 Result of the model for post-test score in Vocabulary:

Table 7 showed that the MLE for the intercept was 107.26, which represented that the student in treatment group with pretest Vocabulary score of 0 was predicted to have a posttest score in Vocabulary of 107.26. Besides, the MLE for T_assignment was 4.98, which showed that the posttest score of a student in treatment group was 4.98 higher than that of a student in control group with the same pre-test score. Also, the p-value of T_assignment was much smaller than 0.05 (i.e. p-value was 0.008), so I rejected the null hypothesis and concluded that T_assignment significantly affected the post-test Vocabulary score.

Then, Table 8 showed that the estimated marginal mean for control group and treatment group was 164.30 and 169.28 respectively, which showed that the students in treatment group tended to achieve higher average posttest score in Vocabulary. Also, the 95% confidence interval for the control group and treatment group was [161, 168] and [166, 173] respectively, so the average Vocabulary score for students in treatment group was more likely to be higher than that for students in control group.

5.4 Result of the model for post-test overall total score:

Table 5: The summary statistics for the fitted Language model

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	156.5685	11.1753	115	14.0102	0.0000
T_assignment1	3.4193	1.8387	115	1.8597	0.0655
Pretest_LANGUAGE	0.1241	0.0648	115	1.9142	0.0581
σ	2.7661	NA	NA	NA	NA
τ	10.6240	NA	NA	NA	NA

Table 6: The estimated marginal means for T.assignment in the fitted Language model

T_assignment	emmean	SE	df	lower.CL	upper.CL
0	177.48	1.46	20	174.43	180.52
1	180.90	1.41	20	177.96	183.83

Table 7: The summary statistics for the fitted Vocabulary model

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	107.2563	16.6405	117	6.4455	0.0000
T_assignment1	4.9780	1.8497	117	2.6913	0.0082
Pretest_VOCABULARY	0.3570	0.1033	117	3.4558	0.0008
σ	3.9991	NA	NA	NA	NA
τ	10.9275	NA	NA	NA	NA

Table 9 showed that the MLE for the intercept was 85.92, which represented that a student in control group with pretest overall score of 0 was predicted to have a posttest overall score of 85.92. Also, the MLE for T_assignment was 4.64, which showed that the posttest overall score of a student in treatment group was 4.64 higher than that of a student in control group with the same pre-test overall score. And, the p-value of T_assignment was 0.0012, which was much smaller than 0.05, so I rejected the null hypothesis. Thus, the random allocation of treatment (i.e. T_assignment) significantly affected the posttest overall score.

Besides, Table 10 showed that the estimated marginal mean for the control group and treatment group was 171.96 and 176.61 respectively, which showed that the students in treatment group tended to achieve higher average posttest overall score. Also, the 95% confidence interval for the control group and treatment group was [170, 174] and [174, 179] respectively, so the treatment group had a higher upper bound and a higher lower bound than the control group, which also indicated that the students in treatment group were more likely to have higher overall scores than those in control group on average.

Moreover, the plots in Figure 6 compared the predicted posttest score between groups for each subtest and the overall score, and all of them shared the similar pattern that for the same x value, the points for treatment group were above those for control group. This feature illustrated that for each subtest's score and the overall score, the predicted posttest score for students in treatment group were higher than those for the students in control group with the same pretest score.

Then, I checked the graphs for random effect and errors to see whether the model assumptions were satisfied, I found that all 4 fitted models had similar patterns. (I included the figures for model assumptions in the Appendix.) Figure 7, Figure 10, Figure 13 and Figure 16 depicted that for each model, the random effect was roughly normally distributed and most points in the QQ plot were on/near the QQ line, so the random effect followed the normal distribution. Also, there was no apparent pattern in the scatter plot between random effect and fitted value, which indicated that the random effect had constant variance. So, the normality assumption and the constant variance assumption for random effect were satisfied. Similarly, Figure 8, Figure 11, Figure 14 and Figure 17 showed that errors was roughly normally distributed and most points in the QQ plot were on/near the QQ line, so that errors followed the normal distribution. Also, there was no apparent pattern in the plot of errors and the scatter plot between errors and fitted value, which showed that errors had constant variance. So, the normality assumption and the constant variance

Table 8: The estimated marginal means for T.assignment in the fitted Vocabulary model

T_assignment	emmean	SE	df	lower.CL	upper.CL
0	164.30	1.59	21	160.99	167.61
1	169.28	1.55	21	166.05	172.51

Table 9: The summary statistics for the fitted Overall Score model

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	85.9212	13.2759	101	6.4720	0.0000
T_assignment1	4.6420	1.3959	101	3.3253	0.0012
Pretest_ELA.TOTAL	0.5287	0.0812	101	6.5124	0.0000
σ	1.8006	NA	NA	NA	NA
τ	7.7154	NA	NA	NA	NA

Table 10: The estimated marginal means for T.assignment in the fitted Overall Score model

T_assignment	emmean	SE	df	lower.CL	upper.CL
0	171.96	1.09	19	169.67	174.26
1	176.61	1.05	19	174.40	178.81

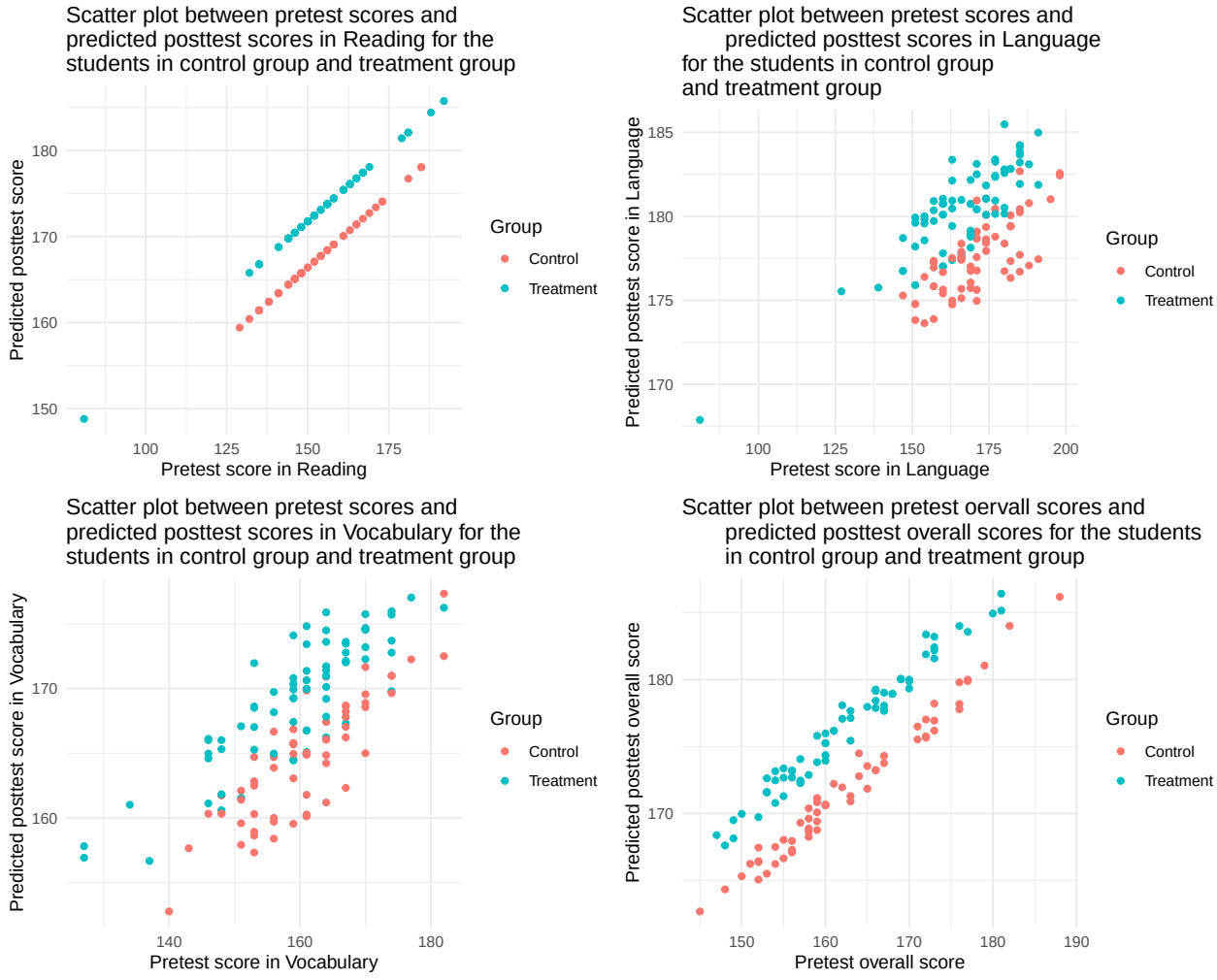

Figure 6: The scatter plot between pretest score and predicted posttest score for the students in control group and treatment group for each subtest and the overall score

assumption for errors were also satisfied. And, since the error term was independent of the other terms in the model, the assumption that random effect and errors were independent was satisfied. Moreover, Figure 8, Figure 12, Figure 15 and Figure 18 illustrated that there was no apparent pattern in the scatter plot between marginal residuals and fitted values, which showed that the linearity of the model was appropriate. Also, Table 11, Table 12, Table 13 and Table 14 showed that the variance-covariance matrix for T_assignment was appropriate. Thus, all assumptions of the model were satisfied and the linear mixed model was appropriate here.

6 Discussion:

6.1 Interpretation of Results & Key Findings:

For each fitted model except the model for Language, the estimated coefficient for T_assignment was positive and the p-value was very small (i.e. much smaller than 0.05), which indicated that the intervention significantly affected the post-test score in a positive direction. Also, the estimated marginal mean for treatment group was higher than that for control group, which showed that the students receiving DLL training tended to achieve higher posttest score on average than those receiving the regular education. For the model for Language, although the p-value was slightly larger than 0.05, its estimate for T_assignment was positive and the estimated marginal means for treatment group was much larger than that for control group, so the result could also reveal that DLL postively affected the posttest scores. Thus, I could conclude that DLL program improved the Spanish literacy skills of the first-grade Spanish-speaking students who had weak literacy skills, and this program was effective.

This paper could conclude a causal relationship between posttest score and DLL training, because a Randomized Controlled Trial (RCT) was conducted and it ensured the random allocation of treatment and separation of control group and treatment group. All the other factors were controlled to be constant (or at least similar) between groups (based on the given information), so that groups only differed in the intervention. I could then compare each subtest's posttest score between groups, and any differences existed were caused by the intervention. I could also fitted the model with a binary variable representing the groups and compare the predicted values for groups, and if the effect of intervention on the outcome was statistically significant and positive, then DLL program was proven to improve the Spanish literacy skills of the enrolled students.

Furthermore, because I used a different data cleaning method and I did not standardize the pretest and posttest scores before fitting the models, my model results were different from those in the original paper (Trisha H. Borman; Geoffrey D. Borman 2019) regarding the Logramos measure. But I draw the same conclusion that DLL was effective in improving these students' Spanish literacy skills.

6.2 Paper Overview & Literature Review:

This paper was to estimate the effect of DLL, which was a one-one-one literacy program designed for the Spanish-speaking students that had poor literacy skills, on the Spanish literacy skills of these students that was measured by the scores of logramos. I got the data from OPENICPSR (Borman 2020), and the data was collected using a Randomized Controlled Trial (RCT) conducted in 2016-2017 school year in the U.S.. I referred to the original paper (Trisha H. Borman; Geoffrey D. Borman 2019) for DLL introduction and data collection process. The intervention of this RCT was the DLL program, which generated a control group where students receiving the regular education during Fall semester, and a treatment group where the students receiving the DLL training. The population was all early elementary Spanish-speaking students that had weak literacy skills in the U.S., and the sample was 152 qualified first-grade Spanish-speaking students whose IdO was within the lowest 25% of their schools among the qualified 22 schools in Texas, Illinois and Arizona.

This sample was not randomly sampled from the frame, and the non-random sampling process led to the selection bias and reduced the representativeness of sample. However, the random allocation of treatment was ensured in the RCT and avoided the self-selection bias. Then, for each subtest's score and the overall score, I visualized the differences in pretest score between groups and found that these groups shared similar

pretest overall score, so that I could assume that this observable characteristic was similar between groups. But I still included each pre-test score as the fixed effect in its corresponding model to adjust for these slight differences and improve the accuracy of predictions. Besides, the non-random sampling method and relatively small sample size could not ensure the similarities of observable and unobservable characteristics between groups, and they lowered the internal validity of this RCT.

There were 12 variables and 152 rows in the data. I used the binary variable T_assignment (where 0 represented the control group and 1 represented the treatment group) to distinguish 2 groups and make comparisons. I cleaned the data by removing all missing values mostly due to non-responses and replacing the outliers with the average values. Then, using the clean data, I made the histogram and boxplot for each posttest score, and found that the students in treatment group had higher posttest score on average.

Then, I fitted a linear mixed model for each subtest (including the overall score), and the response variable was the posttest score. Since many students were sampled from the same school and there may be correlations within a school, I included the variable School_Id in the model as the random effect to adjust for these school differences. I also included the pretest score and random assignment (i.e. T_assignment) as the fixed effects in the model to explore whether the intervention (i.e. DLL program) affected the Spanish literacy skills. For each fitted model, I interpreted the estimates and p-values, and I found that for each subtest, DLL program positively affected the Spanish literacy skills and the students receiving DLL training achieved higher posttest scores on average. Therefore, DLL program could improve the Spanish literacy skills of enrolled students.

Since DLL program was like the Spanish-version of Reading Recovery program, I looked at a paper about the impacts of Reading Recovery program that helped to advance the (English) literacy skills of first-grade schoolers who had low literacy scores (Gray 2017). It performed a randomized controlled trial to show that the students receiving Reading Recovery services had higher scores of ITBS than others, so the Reading Recovery program advanced the English literacy skills of struggling students, and particularly, the students with poorer literacy skills benefited more. This study also used the RCT, which was the same as my paper, to be able to conclude that the literacy program caused the literacy test scores of participated students to increase.

A study explored the effect of Reading Recovery program and DLL program on literacy skills of the first-grade English Language Learners (ELL) and the first-grade Spanish-speaking students that performed bad in literacy respectively, from the collected data in 1993-1996 in California (Neal 1999) It found that both the ELL and DLL students had higher average posttest scores of the selected literacy measures, so the Reading Recovery program and the DLL program positively affected the literacy skills of the ELL and the DLL students respectively. This study and my paper together revealed that the early literacy interventions could benefit the target students regarding their English or Spanish literacy skills.

While my paper showed that DLL program improved the Spanish literacy skills of first-grade Spanish-speaking students in the short-term (i.e. immediately after they completed the program), there was another study exploring the long-term effect of DLL (Escamilla 1998). It compared the academic performance of second and third graders that had participated in DLL in the first grade and those that were randomly sampled from the classrooms, and it found that these DLL students performed equally well or even better than the sample students in reading and writing. So, DLL exerted long-term positive effect on the literacy skills of participated students. However, since this study did not use the RCT to control all other factors invariant except that some students had enrolled in DLL before, it could only conclude that DLL positively affected the literacy skills in the long-term, instead of concluding the causal relationship between them.

6.3 Limitations:

There were 3 main limitations with respect to the sampling method (and the sample size), experiment design, choice of response variable and ethical issues. Firstly, as discussed in the Data section, this RCT did not use a random sampling method, so that the sample units were not randomly sampled from the frame. Also, the sample size was small, so the selection bias may exist and the generalization of analysis results could be reduced. Accordingly, although this RCT used the random allocation of treatment, I could not ensure that all the other factors were the same and the internal validity were appropriate, which may further restrict the appropriateness of causal relationship. Secondly, since the students receiving DLL training obviously know

that they were treated, this RCT could not be designed as a single-blind study to remove the bias due to the different attitudes of participants (Bridgman 2003). So, the students in treatment group may work hard and try to perform well in the posttest, which may cause the over-estimate of effect. Thirdly, I used the test score of Logramos to measure the Spanish literacy skills of students, because this was also a standard and important measure for literacy skills. However, since the IdO test was the most systematic assessment for DLL (Trisha H. Borman; Geoffrey D. Borman 2019), using it as the response variable may have advantages over Logramos. And using it along with the Logramos to measure the literacy skills may be more appropriate to conclude that DLL could improve the Spanish literacy skills. Finally, although this RCT allowed all participated students to have the opportunity to receive the treatment, the students in control group received DLL training in the second semester, which was 1 semester later than the students in treatment group. Since the Reading Recovery program is very important for students at their early elementary stage, the earlier they receive the effective training, the more benefits for their long-term academic development. So, this RCT may make them fall behind the peers and may have negative effect on their long-term academic performance.

6.4 Future Directions:

Based on the limitations, I proposed some suggestions. Firstly, besides the random assignment, the RCT should also use a random sampling method, such as the Simple Random Sampling or Stratified Random Sampling, to ensure the randomness of sample and improve the generalization of results. Meanwhile, more students should be recruited and sampled to increase the sample size and make the control group and treatment group have similar observable and unobservable characteristics. Besides, some basic demographic variables should also be collected and recorded in the data, and then I could include them in the model to adjust for any slight differences in them between groups to improve the accuracy of predictions. Finally, people could perform an observational study where some qualified students received the DLL training, and then use the difference-in-difference method by ensuring that DLL students and control students would have consistent features, or the discontinuity regression method by finding a running variable and a threshold, to estimate the effect of DLL on students' literacy skills. This study design would eliminate the ethical issues and all participated students could receive the effective literacy program at the early elementary to build a solid foundation on literacy skills.

Table 11: The table of variance-covariance martix for 2 groups

98.76886	-12.05478
-12.05478	102.19712

Table 12: The table of variance-covariance martix for 2 groups

136.94908	21.66579
21.66579	101.17606

7 Appendix

7.1 Assumptions for Reading Model:

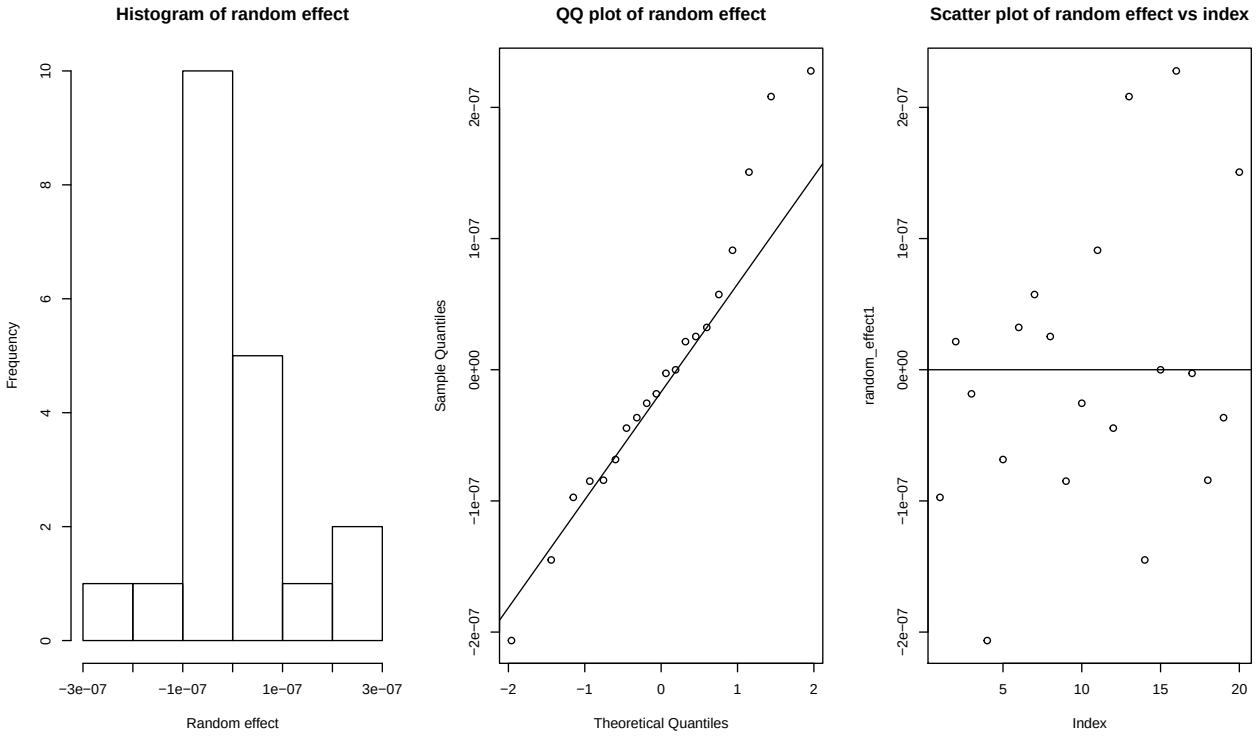

Figure 7: The graphs of checking assumptions for random effect in the Reading model

7.2 Assumptions for Language Model:

7.3 Assumptions for Vocabulary Model:

7.4 Assumptions for Overall Score Model:

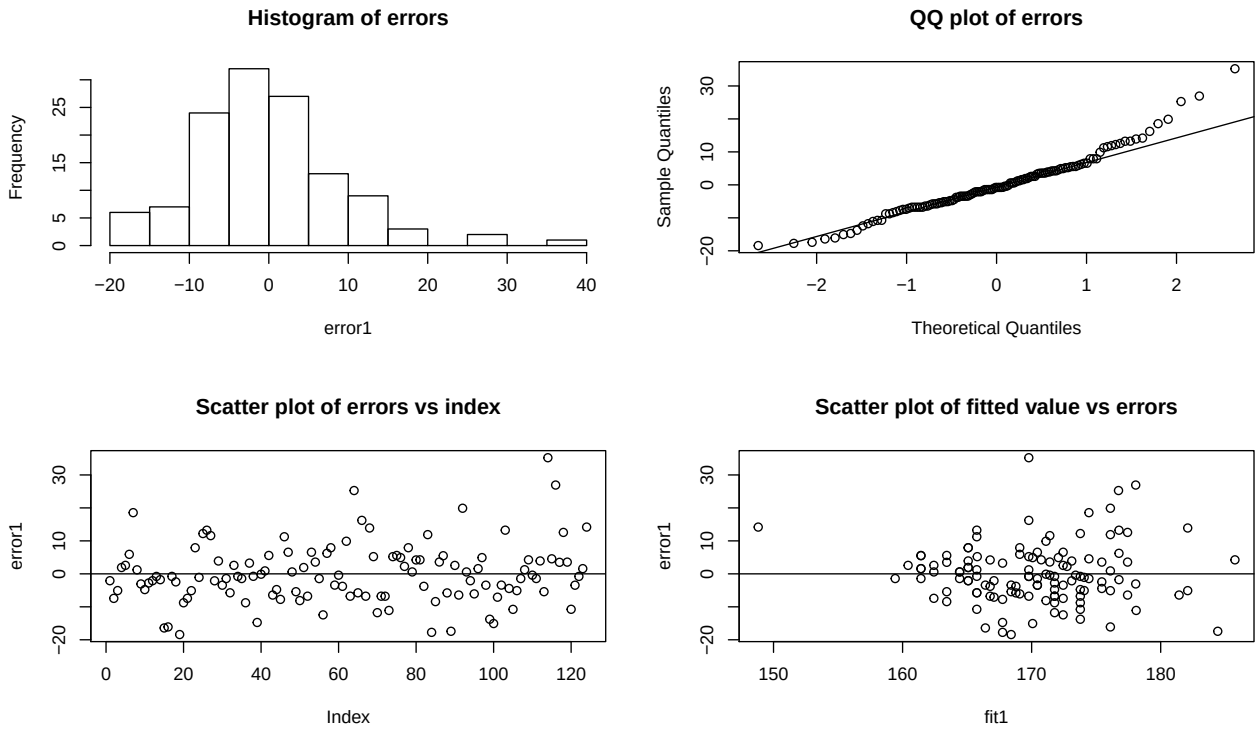

Figure 8: The graphs of checking assumptions for errors in the Reading model

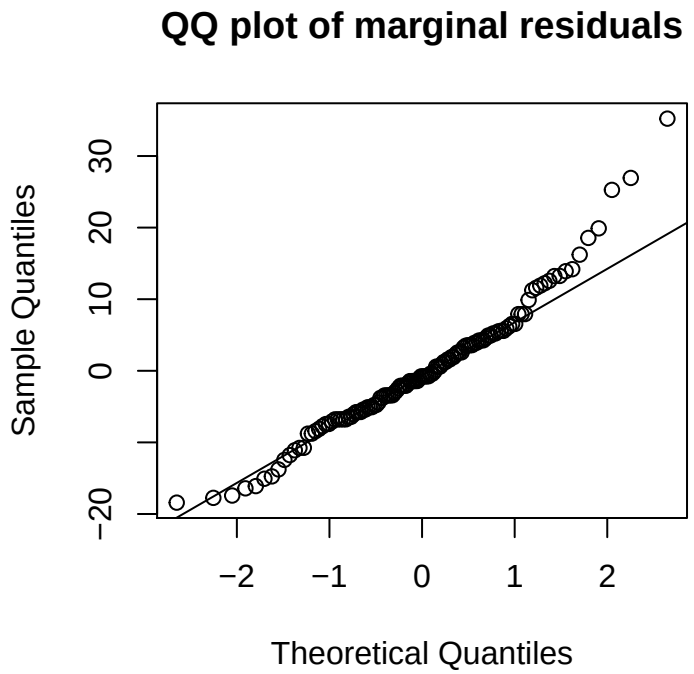

Figure 9: The graphs of checking assumptions for marginal residuals in the Reading model

Table 13: The table of variance-covariance martix for 2 groups

132.9890708	-0.7141305
-0.7141305	170.7823440

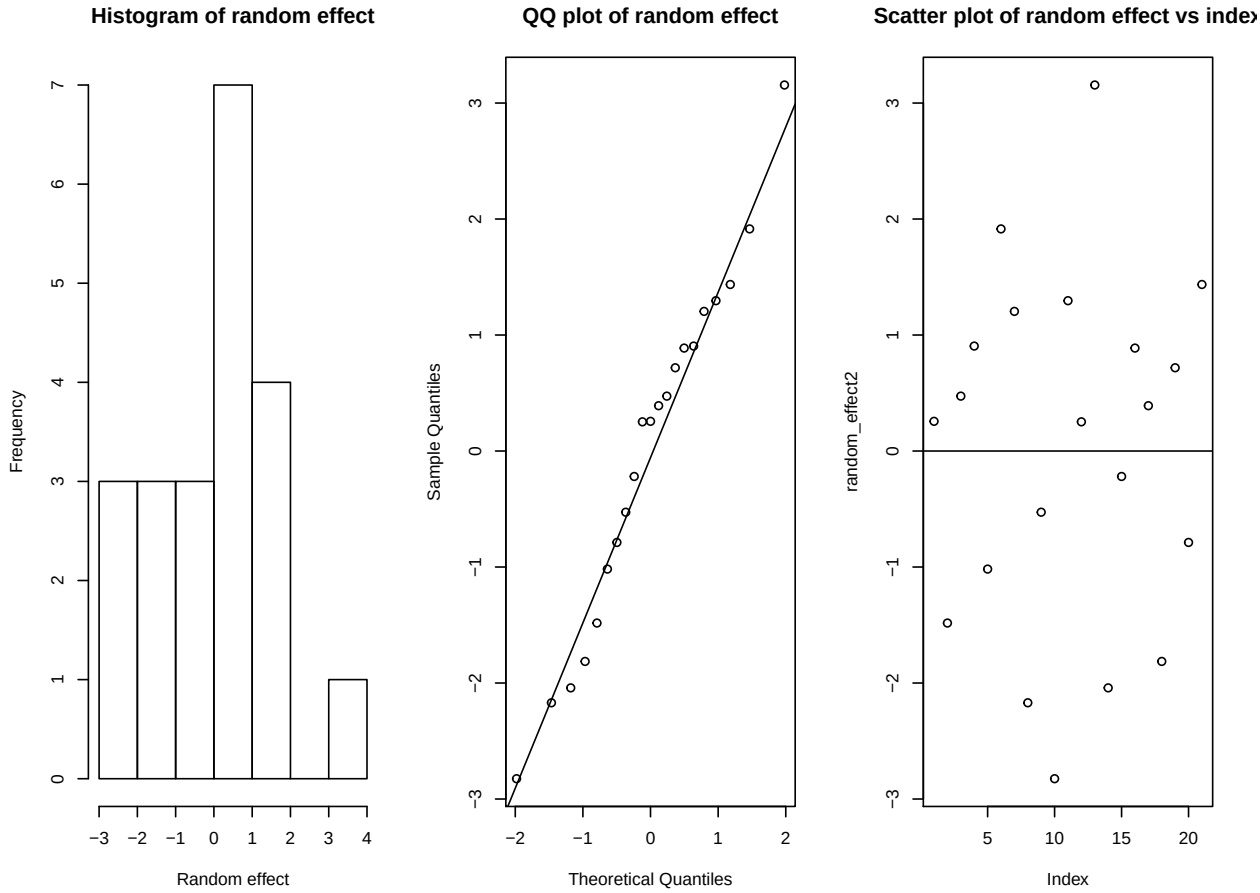

Figure 10: The graphs of checking assumptions for random effect in the Language model

Table 14: The table of variance-covariance martix for 2 groups

87.157668	6.396586
6.396586	77.575149

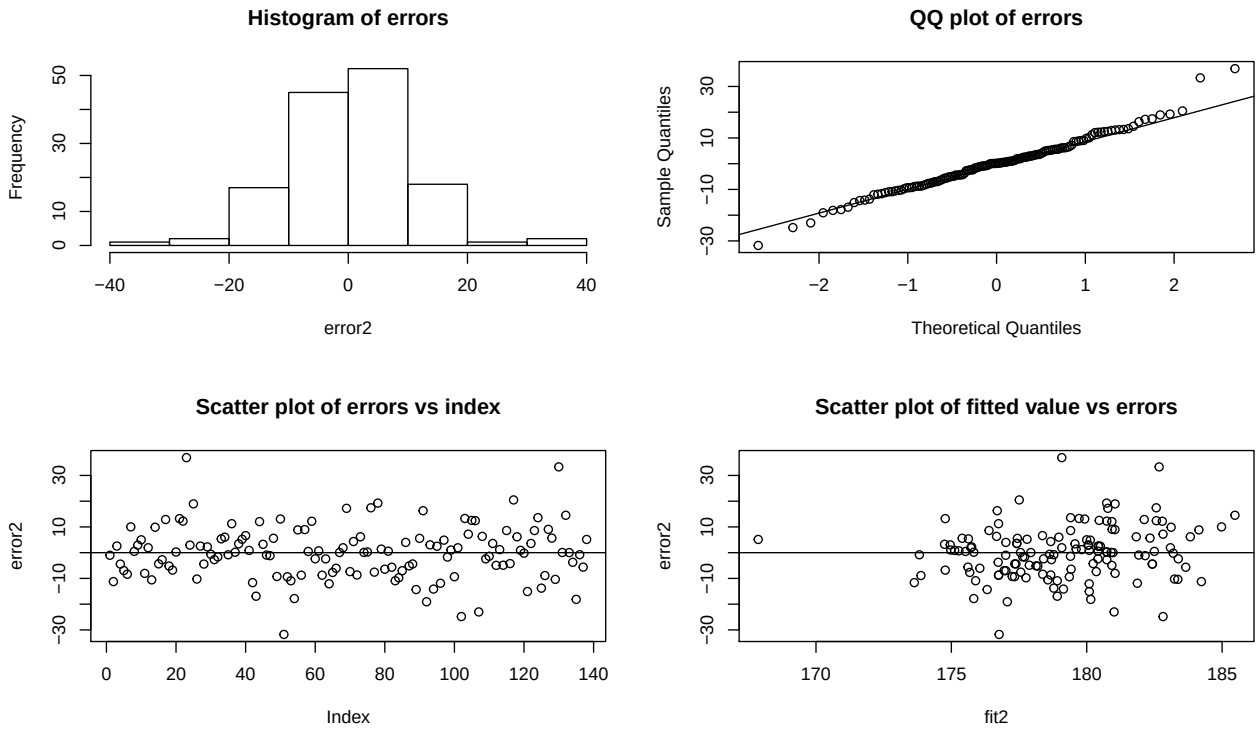

Figure 11: The graphs of checking assumptions for errors in the Language model

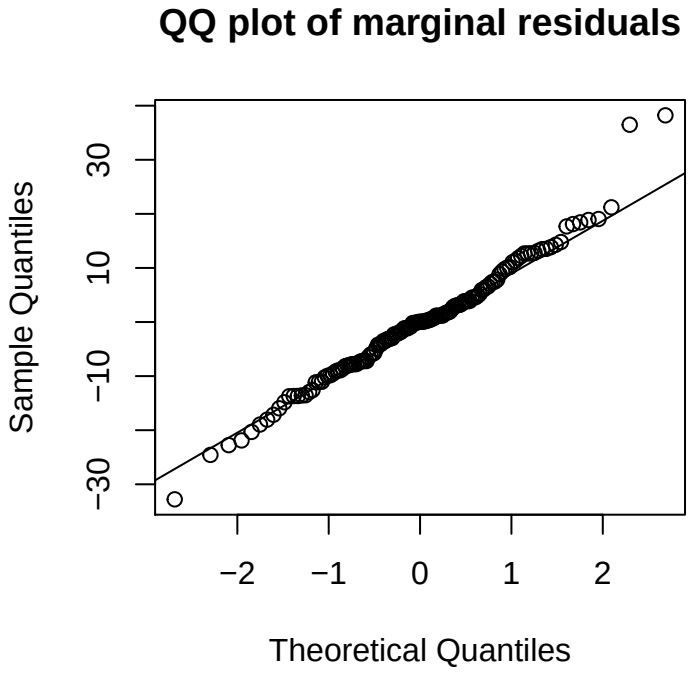

Figure 12: The graphs of checking assumptions for marginal residuals in the Language model

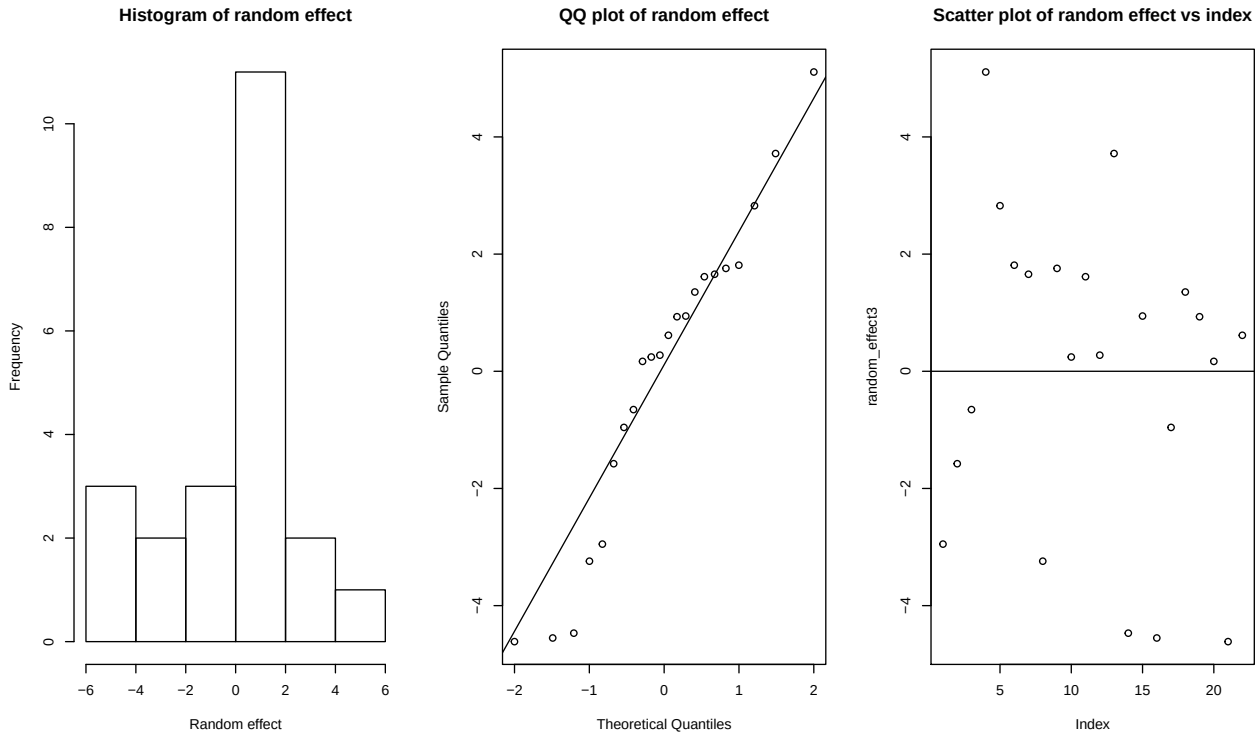

Figure 13: The graphs of checking assumptions for random effect in the Vocabulary model

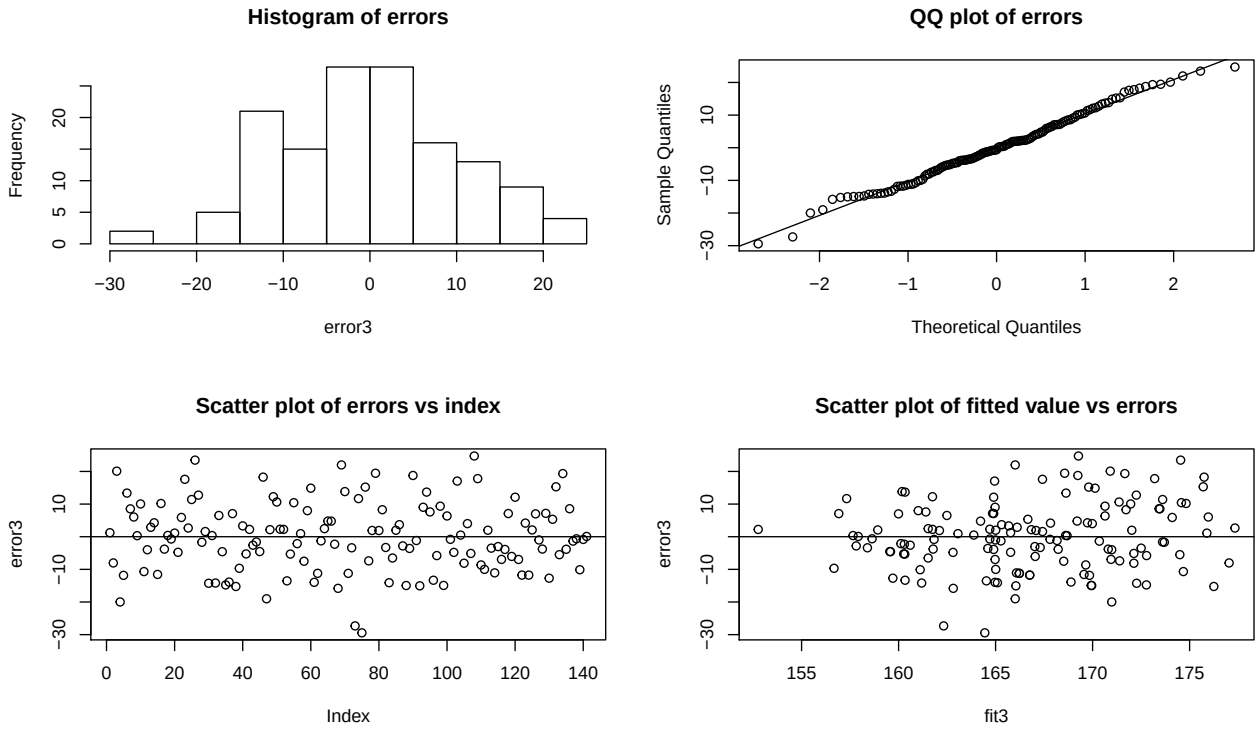

Figure 14: The graphs of checking assumptions for errors in the Vocabulary model

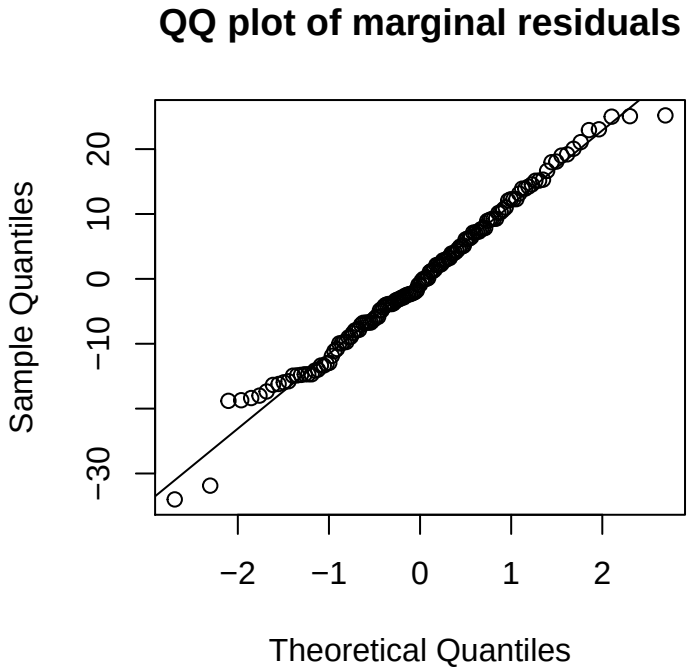

Figure 15: The graphs of checking assumptions for marginal residuals in the Vocabulary model

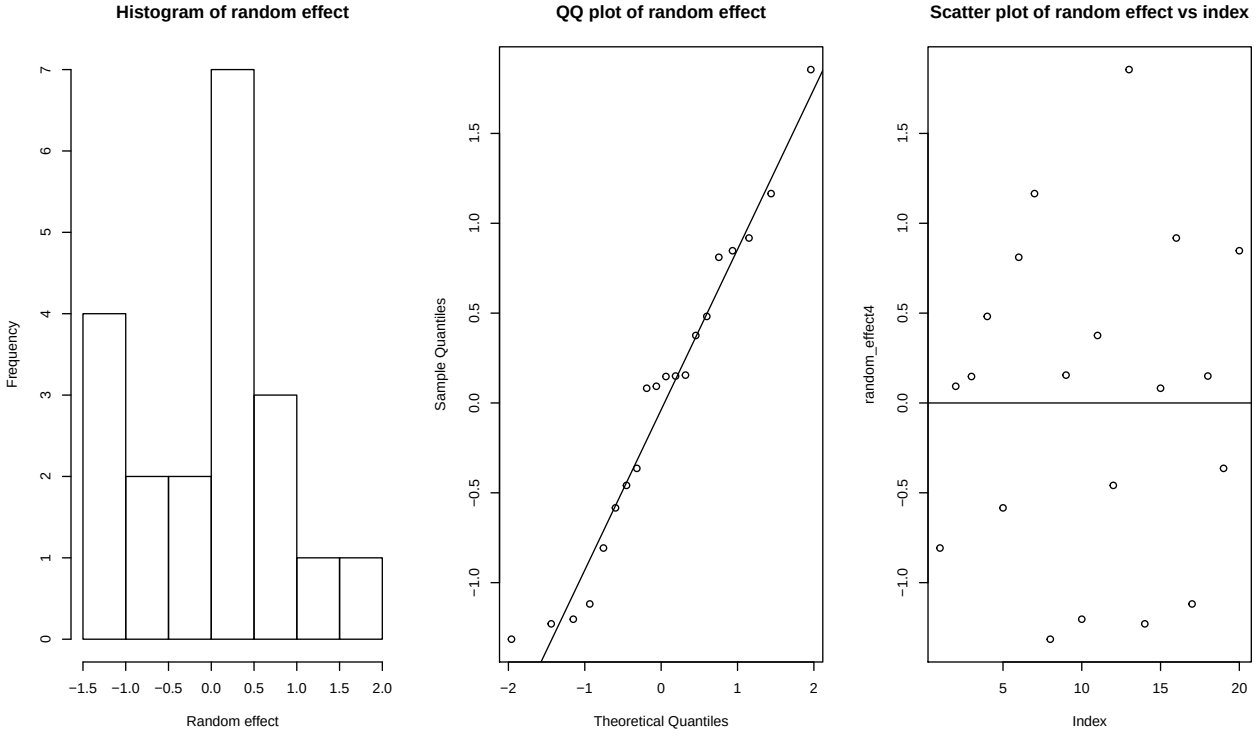

Figure 16: The graphs of checking assumptions for random effect in the Overall Score model

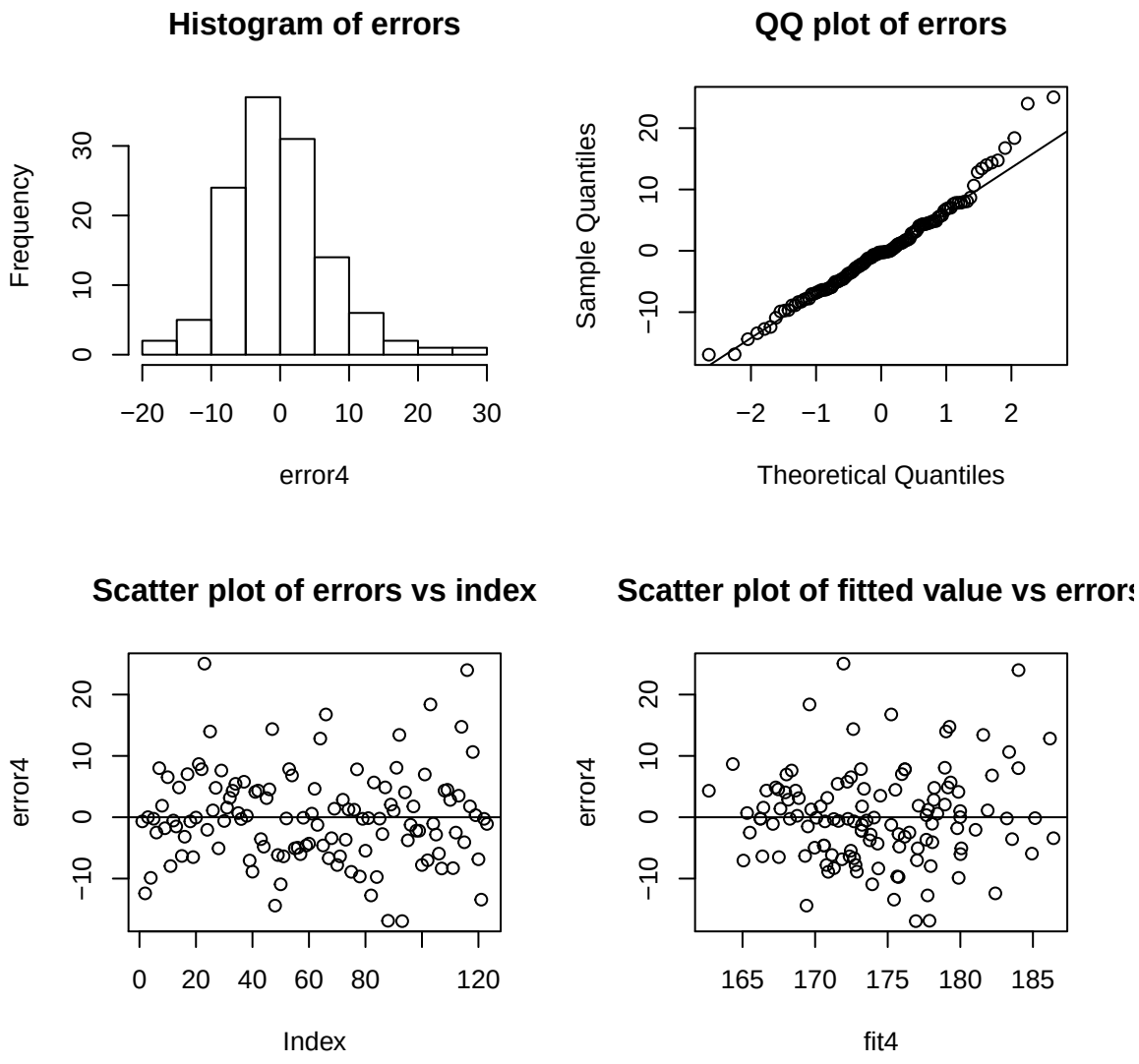

Figure 17: The graphs of checking assumptions for errors in the Overall Score model

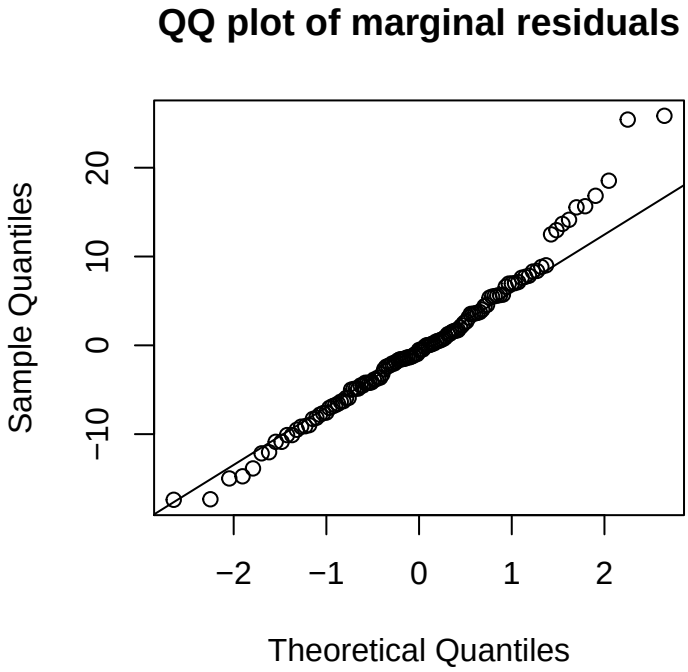

Figure 18: The graphs of checking assumptions for marginal residuals in the Overall Score model

References

Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. doi:10.18637/jss.v067.i01.

Borman, Trisha. 2020. "Addressing Literacy Needs of Struggling Spanish-Speaking First Graders: First-Year Results from a National Randomized Controlled Trial of Descubriendo La Lectura." https://doi.org/10.3886/E118041V1.

Bridgman, Katie; Kirkley, Stephen; Dainty. 2003. "Practical Aspects of Randomization and Blinding in Randomized Clinical Trials." *Arthroscopy: The Journal of Arthroscopic and Related Surgeryt*. https://www.researchgate.net/publication/9017497_Practical_Aspects_of_Randomization_and_Blinding_in_Randomized_Clinical_Trials.

Brown, Patrick. 2019. *Pmisc: Various Utilities for Knitr and Inla*. https://R-Forge.R-project.org/projects/diseasemapping/.

Escamilla, Martha; Ruiz, Kathy; Loera. 1998. "An Examination of Sustaining Effects in Descubriendo La Lectura Programs." *Literacy Teaching and Learning*. https://www.semanticscholar.org/paper/An-Examination-of-Sustaining-Effects-in-La-Lectura-Ruiz-Loera/4af1e0684cf63d98c54a0dbd829ddc64fff18d26.

Gebru, Jamine; Vecchione, Timnit; Monrgenstern. n.d. "Datasheets for Datasets." https://arxiv.org/pdf/1803.09010.pdf.

Gertler, Sebastian; Premand, Paul J.; Martinez. 2016. *Impact Evaluation in Practice, Second Edition*. World Bank. https://openknowledge.worldbank.org/handle/10986/25030 License: CC BY 3.0 IGO.

Gray, Heather; May, Abigail; Goldsworthy. 2017. "Evidence for Early Literacy Intervention: The Impacts of Reading Recovery." *CPRE Policy Briefs*. http://repository.upenn.edu/cpre_policybriefs/82.

Lenth, Russell. 2019. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. https://CRAN.R-

project.org/package=emmeans.

Mitchell, Simone; Zaldivar, Margaret; Wu. n.d. "Model Cards for Model Reporting." https://arxiv.org/pdf/1810.03993.pdf.

Neal, C.; Kelly, Judith. 1999. "The Success of Reading Recovery for English Language Learners and Descubriendo La Lectura for Bilingual Students in California." *Literacy Teaching and Learning*. http://www.literacylearning.net/uploads/3/7/8/8/37880553/ltl_4.2-neal-kelly.pdf.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. 2019. *nlme: Linear and Nonlinear Mixed Effects Models*. https://CRAN.R-project.org/package=nlme.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Trisha H. Borman; Geoffrey D. Borman, So Jung Park, Scott Houghton. 2019. "Addressing Literacy Needs of Struggling Spanish-Speaking First Graders: First-Year Results from a National Randomized Controlled Trial of Descubriendo La Lectura." *AERA Open*. https://doi.org/10.1177/2332858419870488.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain Fran<U+00E7>ois, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. doi:10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.name/knitr/.

———. 2019. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.name/knitr/.