## Section A

### Section A (a) – Description of Dataset

The Junior School Project dataset employed in this report is taken from the "faraway" package in R. The dataset is denoted as "jsp" in the R environment and it consists of data that were collected from elementary schools in inner London. There are 3236 records which are explained by a total of 9 variables in the dataset, with each record representing a student (micro level) and each of them is further categorized into groups according to the different schools that they have attended (macro level). The 9 variables and their respective descriptions could be found in the table below. The purpose of this report is to investigate the influence of students' English scores (proxy for comprehension ability of students) on their Mathematics scores, and to assess if the impact is consistent across different school contexts. The explanatory variable selected is the *english* variable while the response variable chosen is the *math* variable.

Besides that, it is important to mention that while the dataset do allow for longitudinal analysis by examining the performance of students across three different schooling years (Year 1, Year 2, and Year 3), this report will specifically focus on the results of students from Year 2 only. The rationale for choosing such focus is because Year 2 represents a significant transitional phase in the United Kingdom's (UK) curriculum given that it is the year when the Standardised Assessment Tests (SATs) will be taken by those in Key Stage 1 (5 to 7 years old), before the students transition to Key Stage 2 (7 to 11 years old) under the UK's education system, essentially making it pivotal in the evaluation of the impact of students' English comprehension on their performances in Mathematics. In essence, concentrating on Year 2 solely would allow for an in-depth investigation into how foundational English skills, solidified by the end of the first year would affect Mathematics achievements of students.

The rest of the report follows the structure of firstly the output for each subsection, whenever required, followed by the formulation of equations and their respective interpretations, again, whenever applicable.

| **Explanatory Variables** | **Description** | **Sample Data** |
|---|---|---|
| school | A total of 49 schools with different identification numbers<br>(“1” – “42”; “44” – “50”)<br><br>*Remark:* The number “43” is not included in the list of identification numbers. | 1;<br>42;<br>50;<br>44 |
| class | The different identification number of the class in which each student belongs to<br>(“1”, “2”, “3” or “4”) | 1;<br>2;<br>3;<br>4 |
| gender | The gender of the student<br>(“girl” or “boy”) | girl;<br>boy |
| social | The social class of the student’s father<br>(“1” – I,<br>“2” – II,<br>“3” – non-manual,<br>“4” – manual;<br>“5” – IV;<br>“6” – V;<br>“7” – long-term unemployed;<br>“8” – not currently employed ; or<br>“9” – father is absent)<br><br>*Remark:* Even though there were no information on what the social classes of “1”, “2”, “5” and “6” represent, this lack of information will not affect the analysis of this report because the social variable will not be included in the analysis. | 1;<br>7;<br>9;<br>5 |
| raven | The Raven test score of the student<br>(Ranges between “0” to “40”) | 4;<br>36;<br>23;<br>12 |

| | | |
|---|---|---|
| id | The identification (ID) number of each student where these numbers are unique to each student<br>(ID numbers are coded from “1” to “1402” )<br><br>*Remark:* Certain students may not have complete records for all three years of the elementary school education (Year 0, Year 1, and/or Year 2). | 1;<br>1023;<br>98;<br>289 |
| english | The English score of the student according to the different year of schooling (Year 0, Year 1, and/ or Year 2)<br>(Ranges between 0 to 100) | 39;<br>18;<br>56;<br>98 |
| math | The Mathematics score of the student according to the different year of schooling (Year 0, Year 1, and/ or Year 2)<br>(Ranges between 0 to 40) | 1;<br>25;<br>31;<br>40 |
| year | Year of schooling<br>(Year 0 [Reception] - (age 4 – 5);<br>Year 1 - (age 5 - 6);<br>Year 2 - (age 6 - 7)) | 0;<br>1;<br>2 |

*Table 1: Variables Included in the Dataset*

**Section A (b) – Centralization of Explanatory Variable**

To centralize the *english* variable, two new variables namely the *english_mean* variable and the *english_dev* variable are defined. The former is defined as the group mean for each school while the latter is defined as the deviation of each individual English score of the students from their respective group means (*english_dev* = *english* – *english_mean*). The *english_dev* variable is the variable representing the centred values and centralization has been successfully implemented as the mean for the *english_dev* variable is now equals to zero, as pointed out in the figure below.

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | school | class | gender | social | raven | id | english | math | year | english_mean | english_dev | math_mean | math_dev |
| 2 | 1 | 1 | girl | 9 | 23 | 1 | 39 | 23 | 2 | 29 | 10 | 25.5384615 | -2.53846 |
| 3 | 1 | 1 | boy | 2 | 22 | 3 | 83 | 39 | 2 | 29 | 54 | 25.5384615 | 13.4615 |
| 4 | 1 | 1 | boy | 9 | 14 | 4 | 12 | 32 | 2 | 29 | -17 | 25.5384615 | 6.46154 |
| 5 | 1 | 1 | girl | 4 | 16 | 6 | 19 | 11 | 2 | 29 | -10 | 25.5384615 | -14.5385 |
| 6 | 1 | 1 | boy | 9 | 17 | 7 | 31 | 26 | 2 | 29 | 2 | 25.5384615 | 0.46154 |
| 7 | 1 | 1 | girl | 9 | 21 | 8 | 40 | 28 | 2 | 29 | 11 | 25.5384615 | 2.46154 |
| 8 | 1 | 1 | girl | 4 | 32 | 11 | 50 | 32 | 2 | 29 | 21 | 25.5384615 | 6.46154 |
| 9 | 1 | 1 | girl | 3 | 25 | 13 | 39 | 27 | 2 | 29 | 10 | 25.5384615 | 1.46154 |
| 10 | 1 | 1 | girl | 2 | 29 | 14 | 80 | 36 | 2 | 29 | 51 | 25.5384615 | 10.4615 |
| 11 | 1 | 1 | girl | 7 | 34 | 15 | 37 | 33 | 2 | 29 | 8 | 25.5384615 | 7.46154 |
| 12 | 1 | 1 | boy | 5 | 27 | 16 | 4 | 33 | 2 | 29 | -25 | 25.5384615 | 7.46154 |
| 13 | 1 | 1 | boy | 4 | 21 | 17 | 29 | 30 | 2 | 29 | 0 | 25.5384615 | 4.46154 |

JSP Data

Ready Accessibility: Good to go Average: 0 Count: 953 Sum: 0

*Figure 1: Centralization of the Independent Variable (english)*

## Section A (c) – Random Intercept Model

The Random Intercept Model (RIM) Analysis will be discussed in this subsection.

## Section A (c) (i) – Output

The detailed output from the RIM analysis is presented below.

**The Mixed Procedure**

| Model Information | |
|---|---|
| Data Set | MDA.JSP |
| Dependent Variable | math |
| Covariance Structure | Variance Components |
| Subject Effect | school |
| Estimation Method | REML |
| Residual Variance Method | Profile |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Between-Within |

**Class Level Information**

| Class | Levels | Values |
|---|---|---|
| school | 48 | 1 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 44 45 46 47 48 49 5 50 6 7 8 9 |

| Dimensions | |
|---|---|
| Covariance Parameters | 2 |
| Columns in X | 3 |
| Columns in Z per Subject | 1 |
| Subjects | 48 |
| Max Obs per Subject | 63 |

| Number of Observations | |
|---|---|
| Number of Observations Read | 953 |
| Number of Observations Used | 953 |
| Number of Observations Not Used | 0 |

| Iteration History | | | |
|---|---|---|---|
| Iteration | Evaluations | -2 Res Log Like | Criterion |
| 0 | 1 | 5816.40822872 | |
| 1 | 3 | 5776.41698566 | 0.00000365 |
| 2 | 1 | 5776.40944704 | 0.00000000 |

Convergence criteria met.

| Covariance Parameter Estimates | | |
|---|---|---|
| Cov Parm | Subject | Estimate |
| Intercept | school | 2.2178 |
| Residual | | 23.6051 |

| Fit Statistics | |
|---|---|
| -2 Res Log Likelihood | 5776.4 |
| AIC (Smaller is Better) | 5780.4 |
| AICC (Smaller is Better) | 5780.4 |
| BIC (Smaller is Better) | 5784.2 |

| Solution for Fixed Effects | | | | | |
|---|---|---|---|---|---|
| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
| Intercept | 22.0488 | 1.2037 | 46 | 18.32 | <.0001 |
| english_dev | 0.1979 | 0.008047 | 904 | 24.60 | <.0001 |
| english_mean | 0.1999 | 0.02772 | 46 | 7.21 | <.0001 |

| Type 3 Tests of Fixed Effects | | | | |
|---|---|---|---|---|
| Effect | Num DF | Den DF | F Value | Pr > F |
| english_dev | 1 | 904 | 604.94 | <.0001 |
| english_mean | 1 | 46 | 51.98 | <.0001 |

*Figure 2: Output from Random Intercept Model Analysis*

Note that reasonings as to why there were only 48 schools in the analysis as opposed to 49 schools which was previously mentioned in *Section A (a)* will be detailed in *Appendix A*. Similar reasonings would apply for the rest of the sections to come.

**Section A (c) (ii) – General Regression Equation**

$$Y = 22.0488 + 0.1999\ english\_mean + 0.1979\ english\_dev + U_{0j} + R_{ij}$$

Across the entire Section A (c), "$Y$" denotes the dependent variable, which is the *math* variable.

**Section A (c) (iii) – Top 2.5%**

$$Standard\ Deviation = \sqrt{2.2178} = 1.4892$$

$$Y = 22.0488 + 0.1999\ english\_mean + 0.1979\ english\_dev + 2(1.4892)$$

$$Y = 25.0272 + 0.1999\ english\_mean + 0.1979\ english\_dev$$

**Section A (c) (iv) – Bottom 2.5%**

$$Standard\ Deviation = \sqrt{2.2178} = 1.4892$$

$$Y = 22.0488 + 0.1999\ english\_mean + 0.1979\ english\_dev - 2(1.4892)$$

$$Y = 19.0704 + 0.1999\ english\_mean + 0.1979\ english\_dev$$

**Section A (c) (v) – Standardized Coefficient**

The MEANS Procedure

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| math | math | 953 | 30.4858342 | 6.6558607 | 5.0000000 | 40.0000000 |
| english_mean | english_mean | 953 | 42.6327387 | 9.6844779 | 23.8000000 | 66.7142857 |
| english_dev | english_dev | 953 | 3.653368E-16 | 19.5687278 | -47.3125000 | 65.8666667 |

*Figure 3: Output to Obtain Variables' Standard Deviation Values*

$$Standardized\ Coefficient = \frac{SD\ (X)}{SD\ (Y)}\gamma = \frac{19.5687278}{6.6558607}(0.1979) = 0.5818$$

Interpretation of the calculated standardized coefficient is that each additional standard deviation on the English scores of students leads to an increase in their Mathematics scores by 0.5818 standard deviation on average. 58.18% of the total variation in Mathematics scores is explained by the variation in the English scores of students. The remaining 41.82% is explained by other external factors.

**Section A (c) (vi) – Residual Intraclass Correlation**

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2} = \frac{2.2178}{2.2178 + 23.6051} = 0.0859$$

The residual intraclass correlation value of 0.0859 indicates that the value is on the extreme low end, suggesting that there is a very weak between-group effect that exists between the schools.

**Section A (d) – Hierarchical Linear Model**

For this section, data analysis is performed on the *jsp* dataset using the Hierarchical Linear Modelling (HLM) technique, encompassing both Random Slope Model and Random Intercept Model. The results are shown in the following figure.

The Mixed Procedure

| Model Information | |
|---|---|
| Data Set | MDA.JSP |
| Dependent Variable | math |
| Covariance Structures | Unstructured, Compound Symmetry |
| Subject Effects | school, school |
| Estimation Method | REML |
| Residual Variance Method | Profile |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Containment |

| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| school | 48 | 1 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 44 45 46 47 48 49 5 50 6 7 8 9 |

| Dimensions | |
|---|---|
| Covariance Parameters | 5 |
| Columns in X | 3 |
| Columns in Z per Subject | 2 |
| Subjects | 48 |
| Max Obs per Subject | 63 |

| Number of Observations | |
|---|---|
| Number of Observations Read | 953 |
| Number of Observations Used | 953 |
| Number of Observations Not Used | 0 |

| Covariance Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Cov Parm | Subject | Estimate | Standard Error | Z Value | Pr Z |
| UN(1,1) | school | 2.5547 | 0.8252 | 3.10 | 0.0010 |
| UN(2,1) | school | -0.07153 | 0.02474 | -2.89 | 0.0038 |
| UN(2,2) | school | 0.001929 | 0.001184 | 1.63 | 0.0517 |
| CS | school | 0.04194 | 0.002023 | 20.73 | <.0001 |
| Residual | | 22.8395 | 1.1016 | 20.73 | <.0001 |

| Fit Statistics | |
|---|---|
| -2 Res Log Likelihood | 5758.6 |
| AIC (Smaller is Better) | 5768.6 |
| AICC (Smaller is Better) | 5768.7 |
| BIC (Smaller is Better) | 5778.0 |

| Null Model Likelihood Ratio Test | | |
|---|---|---|
| DF | Chi-Square | Pr > ChiSq |
| 4 | 57.80 | <.0001 |

| Solution for Fixed Effects | | | | | |
|---|---|---|---|---|---|
| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
| Intercept | 23.8199 | 1.0823 | 46 | 22.01 | <.0001 |
| english_dev | 0.1931 | 0.01031 | 47 | 18.72 | <.0001 |
| english_mean | 0.1575 | 0.02465 | 857 | 6.39 | <.0001 |

| Iteration History | | | |
|---|---|---|---|
| **Iteration** | **Evaluations** | **-2 Res Log Like** | **Criterion** |
| 0 | 1 | 5816.40822872 | |
| 1 | 2 | 5759.82380961 | 2276.9725801 |
| 2 | 1 | 5758.96198217 | 4322.2011217 |
| 3 | 1 | 5758.62395276 | 3674.9574244 |
| 4 | 1 | 5758.61256763 | 10.26927372 |
| 5 | 1 | 5758.60679428 | 0.00546661 |
| 6 | 1 | 5758.60337228 | 0.00030903 |
| 7 | 1 | 5758.60328654 | 0.00000026 |
| 8 | 1 | 5758.60328651 | 0.00000000 |

Convergence criteria met but final Hessian is not positive definite.

*Figure 4: Output from Hierarchical Linear Model Analysis*

### Section A (d) (i) – Slope Standard Deviation

$$Slope\ Standard\ Deviation = \sqrt{0.001929} = 0.0439$$

### Section A (d) (ii) – Average Slope

$$Average\ Slope = 0.1931$$

### Section A (d) (iii) - Correlation Between Random Slope and Random Intercept

$$Correlation\ Between\ Random\ Slope\ and\ Random\ Intercept = -\frac{0.07153}{\sqrt{2.5547(0.001929)}}$$

$$= -1.0189 \approx -1$$

### Section A (d) (iv) - Interpretation

There is a negative correlation between the random slope and random intercept, suggesting that schools with higher baseline Mathematics scores exhibit a smaller increase in Mathematics scores for students of average English scores, thus showing a lower within-school effect of English score.

### Section A (d) (v) – Estimated Equation

$$Y_{ij} = 23.8199 + 0.1931\ english_{dev_{ij}} + 0.1575\ english\_mean_{.J} + U_{0j}$$
$$+ U_{1j}\ english\_dev_{ij} + R_{ij}$$

**Section A (d) (vi) – Residual Intraclass Correlation**

$$ICC = \frac{\tau^2}{\tau^2 + \sigma^2} = \frac{2.5547}{2.5547 + 22.8395} = 0.1006$$

The residual intraclass correlation value of 0.1006 indicates that the value is on the low end, suggesting that there is a weak between-group effect that exists between the schools.

**Section A (d) (vii) – Test of Coefficient**

$$H_0: Coefficient\ of\ english\_mean\ is\ equal\ to\ 0$$

$$H_1: Coefficient\ of\ english\_mean\ does\ not\ equal\ to\ 0$$

$$t - statistics = \frac{0.1575}{0.02465} = 6.3895\ (> 2)$$

Given that the value of the t-statistics is greater than 2, we reject the null hypothesis. There is sufficient evidence to conclude that the coefficient of *english_mean* does not equal to 0.

**Section A (e) – Discussion on Intraclass Correlation of Both Models**

The residual intraclass correlation values for the RIM and HLM are 0.0859 and 0.1006 respectively, implying that both models attribute only a small proportion of the total variance in the outcome to the differences between the schools. The slightly higher intraclass correlation value of the HLM proposes that the model may have captured a bit more of the between-group variability when compared to the RIM, possibly owing to its more complex structure.

**Section B**

**Section B (a) – Description of the Section**

In this section, data analysis is performed on the *jsp* dataset using multiple variations of the Regression Model.  Both explanatory (*english*) and response variable (*math*) remain the same. The respective results are shown in the following figures.

## Section B (a) (i) – Total Regression (Output and Interpretation)

The REG Procedure
Model: MODEL1
Dependent Variable: math math

| | |
|---|---|
| Number of Observations Read | 953 |
| Number of Observations Used | 953 |

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 17624 | 17624 | 682.70 | <.0001 |
| Error | 951 | 24550 | 25.81512 | | |
| Corrected Total | 952 | 42174 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 5.08086 | R-Square | 0.4179 |
| Dependent Mean | 30.48583 | Adj R-Sq | 0.4173 |
| Coeff Var | 16.66629 | | |

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 22.08462 | 0.36121 | 61.14 | <.0001 |
| english | english | 1 | 0.19706 | 0.00754 | 26.13 | <.0001 |

*Figure 5: Output from Total Regression Analysis*

$$Y_{ij} = 22.08462 + 0.19706X_{ij} + R$$

The regression equation above represents a Simple Linear Regression model. Interpretation of the coefficient of the *english* variable is that for each one-unit increase in the English score, the Mathematics score is expected to increase by 0.19706 units, on average.

## Section B (a) (ii) – Regression Between Group Means (Output and Interpretation)

The Mixed Procedure

| Model Information | |
|---|---|
| Data Set | MDA.JSP |
| Dependent Variable | math_mean |
| Covariance Structure | Diagonal |
| Estimation Method | REML |
| Residual Variance Method | Profile |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Residual |

| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| school | 48 | 1 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 44 45 46 47 48 49 5 50 6 7 8 9 |

| Dimensions | |
|---|---|
| Covariance Parameters | 1 |
| Columns in X | 2 |
| Columns in Z | 0 |
| Subjects | 1 |
| Max Obs per Subject | 953 |

| Number of Observations | |
|---|---|
| Number of Observations Read | 953 |
| Number of Observations Used | 953 |
| Number of Observations Not Used | 0 |

| Covariance Parameter Estimates | |
|---|---|
| Cov Parm | Estimate |
| Residual | 3.3244 |

| Fit Statistics | |
|---|---|
| -2 Res Log Likelihood | 3859.5 |
| AIC (Smaller is Better) | 3861.5 |
| AICC (Smaller is Better) | 3861.5 |
| BIC (Smaller is Better) | 3866.4 |

| Solution for Fixed Effects | | | | | |
|---|---|---|---|---|---|
| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
| Intercept | 22.2334 | 0.2668 | 951 | 83.35 | <.0001 |
| english_mean | 0.1936 | 0.006102 | 951 | 31.72 | <.0001 |

| Type 3 Tests of Fixed Effects | | | | |
|---|---|---|---|---|
| Effect | Num DF | Den DF | F Value | Pr > F |
| english_mean | 1 | 951 | 1006.36 | <.0001 |

*Figure 6: Output from Regression Between Group Means Analysis*

$$\overline{Y_{.J}} = 22.2334 + 0.1936\,\overline{X_{.J}} + R$$

The regression equation above describes a model where the average Mathematics scores of students $(\overline{Y_{.J}})$ within different schools are predicted according to their average English scores $(\overline{X_{.J}})$. Interpretation of the coefficient of the *english_mean* variable is that for every unit increase in the average English score within a school, the average Mathematics score in that school is expected to increase by approximately 0.1936 units.

### Section B (a) (iii) – Regression Within Group (Output and Interpretation)

The Mixed Procedure

| Model Information | |
|---|---|
| Data Set | MDA.JSP |
| Dependent Variable | math_dev |
| Covariance Structure | Diagonal |
| Estimation Method | REML |
| Residual Variance Method | Profile |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Residual |

| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| school | 48 | 1 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 44 45 46 47 48 49 5 50 6 7 8 9 |

| Dimensions | |
|---|---|
| Covariance Parameters | 1 |
| Columns in X | 2 |
| Columns in Z | 0 |
| Subjects | 1 |
| Max Obs per Subject | 953 |

| Number of Observations | |
|---|---|
| Number of Observations Read | 953 |
| Number of Observations Used | 953 |
| Number of Observations Not Used | 0 |

| Covariance Parameter Estimates | |
|---|---|
| Cov Parm | Estimate |
| Residual | 22.4893 |

| Fit Statistics | |
|---|---|
| -2 Res Log Likelihood | 5679.0 |
| AIC (Smaller is Better) | 5681.0 |
| AICC (Smaller is Better) | 5681.0 |
| BIC (Smaller is Better) | 5685.8 |

| Solution for Fixed Effects | | | | | |
|---|---|---|---|---|---|
| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
| Intercept | -177E-18 | 0.1536 | 951 | -0.00 | 1.0000 |
| english_dev | 0.1979 | 0.007854 | 951 | 25.20 | <.0001 |

| Type 3 Tests of Fixed Effects | | | | |
|---|---|---|---|---|
| Effect | Num DF | Den DF | F Value | Pr > F |
| english_dev | 1 | 951 | 634.96 | <.0001 |

*Figure 7: Output from Regression Within Group Analysis*

$$Y_{ij} = \overline{Y_{.J}} + 0.1979\left(X_{ij} - \overline{X_{.J}}\right) + R$$

The regression equation above describes a model aimed at analysing the within-group effects by focusing on how deviations of individual students from their school means relate to deviations in the math scores of those students. Variations in individuals' performances are isolated under this method, essentially shedding light on the influence of relative English comprehension on achievement in Mathematics within the context of each school. Interpretation of the slope coefficient is that for each one unit increase in the deviation of a student's English score from their respective school's average, the student's Mathematics score is expected to increase by 0.1979 units from the school's average Mathematics score.

## Section B (a) (iv) – Multilevel Regression (Output and Interpretation)

The Mixed Procedure

| Model Information | |
|---|---|
| Data Set | MDA.JSP |
| Dependent Variable | math |
| Covariance Structure | Diagonal |
| Estimation Method | REML |
| Residual Variance Method | Profile |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Residual |

| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| school | 48 | 1 11 12 13 14 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 3 30 31 32 33 34 35 36 37 38 39 4 40 41 42 44 45 46 47 48 49 5 50 6 7 8 9 |

| Dimensions | |
|---|---|
| Covariance Parameters | 1 |
| Columns in X | 3 |
| Columns in Z | 0 |
| Subjects | 1 |
| Max Obs per Subject | 953 |

| Number of Observations | |
|---|---|
| Number of Observations Read | 953 |
| Number of Observations Used | 953 |
| Number of Observations Not Used | 0 |

| Covariance Parameter Estimates | |
|---|---|
| Cov Parm | Estimate |
| Residual | 25.8409 |

| Fit Statistics | |
|---|---|
| -2 Res Log Likelihood | 5816.4 |
| AIC (Smaller is Better) | 5818.4 |
| AICC (Smaller is Better) | 5818.4 |
| BIC (Smaller is Better) | 5823.3 |

| Solution for Fixed Effects | | | | | |
|---|---|---|---|---|---|
| Effect | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
| Intercept | 22.2334 | 0.7437 | 950 | 29.89 | <.0001 |
| english_mean | 0.1936 | 0.01701 | 950 | 11.38 | <.0001 |
| english_dev | 0.1979 | 0.008419 | 950 | 23.51 | <.0001 |

| Type 3 Tests of Fixed Effects | | | | |
|---|---|---|---|---|
| Effect | Num DF | Den DF | F Value | Pr > F |
| english_mean | 1 | 950 | 129.47 | <.0001 |
| english_dev | 1 | 950 | 552.60 | <.0001 |

*Figure 8: Output from Multilevel Regression Analysis*

$$Y_{ij} = 22.2334 + 0.1979X_{ij} + 0.1936\overline{X_{.j}} + R$$

The regression equation above describes a multilevel regression model that is aimed at analysing the impact of English comprehension on the Mathematics scores at both the student's and school's levels. Interpretation of the student-level coefficient $(X_{ij})$ is that holding the school's average constant, every one unit increase in a student's English score is associated with an increase of approximately 0.1979 units in their Mathematics score. As for the school-level coefficient $(\overline{X_{.J}})$, the interpretation is that an increase of one unit in the school's average English score is associated with an increase of approximately 0.1936 units in an individual student's Mathematics score, independent of the student's own English score.

**Section B (a) (v) – Comparison of All Four Models**

Total Regression: $Y_{ij} = 22.08462 + 0.19706X_{ij} + R$

Regression Between Group Means: $\overline{Y_{.J}} = 22.2334 + 0.1936\,\overline{X_{.J}} + R$

Regression Within Group: $Y_{ij} = \overline{Y_{.J}} + 0.1979\left(X_{ij} - \overline{X_{.J}}\right) + R$

Multilevel Regression: $Y_{ij} = 22.2334 + 0.1979X_{ij} + 0.1936\overline{X_{.J}} + R$

The regression equations from the four distinct regression models consistently indicate a positive relationship between English comprehension and Mathematics scores, although the impact varies by model owing to the different analytical approaches and inherent assumptions within each model.

Among the four regression models, the Multilevel Regression Model is especially insightful as it addresses possible drawbacks associated with simpler analytical methods which often fails to accommodate for the hierarchical structure inherent in multilevel dataset, which is precisely the case for the analysis of this report. That is, the Multilevel Regression Model serves to mitigate the "shift in meaning" error which arises when important micro-level variables are inaccurately aggregated to represent macro-units of the dataset, thereby altering the initial implications of those micro-level variables. Taking the example of the educational study of this report, averaging the Mathematics scores of individuals students to represent the entire school's performance may conceal the fact that different groups within the schools may perform considerably better or worse depending on factors such as the teachers' teaching qualities or the access to resources that may be heterogenous across different classes. Referring back to the case of the analysis of this report, the "shift of meaning" error is said to have taken place when the aggregated data (school-level average score) no longer directly represent the performance nuances of individual students.

Besides that, the Multilevel Regression Model is also able to prevent researchers from committing the ecological fallacy. This fallacy refers to the wrongful assumption that findings obtained from group-level data does indeed apply to the individuals within that group. For instance, and in the case of an educational study, given that data shows that schools with certain characteristics (such as high-teacher-to-student ratios) tend to yield better student performance, this might lead researchers to wrongfully assume that any student from such a school will inherently excel than students from schools with lower teacher-to-student ratios. This generalization is a grievous mistake as it fails to consider variability among individuals where some students may still perform poorly due to factors like their individual aptitudes or a lack of personal support, despite the overall high performance of their schools.

Given the above discussion, utilizing a multilevel approach helps mitigate errors that arise from aggregation by providing a layered analysis that considers both individual abilities and group characteristics on educational outcomes. This method accommodates for more accurate and contextually relevant insights which are instrumental in the formulation of effective educational policies and interventions conducive to the different levels of the educational system. Therefore, for data structures that are nested just like the one examined in this report, multilevel models prove to be the most appropriate model for the analysis as it does allow for the analysing of both within-group and between-group effects simultaneously, ultimately allowing for a complete understanding of the full spectrum of factors influencing educational achievements.

**Section B (b) – Discussion on Random Intercept Model and Literature Review**

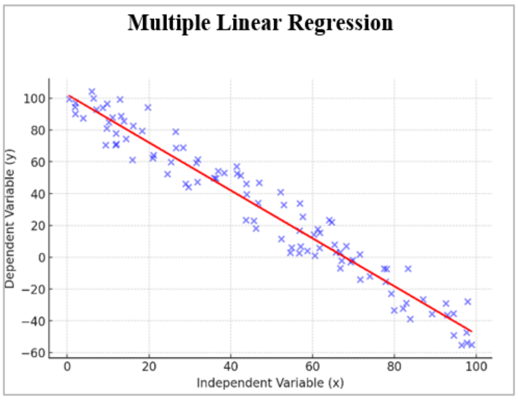


*Figure 9: Multiple Linear Regression*

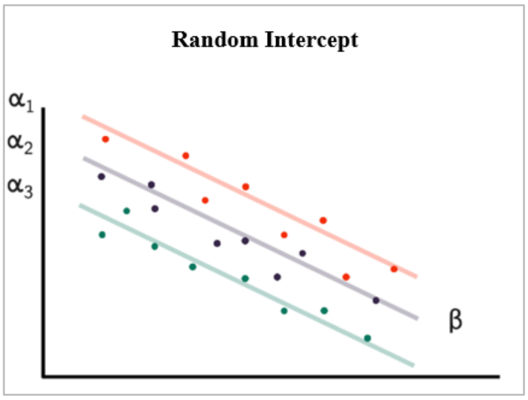


*Figure 10: Random Intercept Model (Midway, 2022)*

The Random Intercept Model is especially adept in analysing data with multilevel structures where observations are nested within hierarchical groups or clusters, allowing it to have the upper hand over Multiple Linear Regression (MLR). In the latter, the general model equation $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + R$ assumes that the residuals are independent. This assumption however, does not acknowledge the complex hierarchical structures that are prevalent in datasets like students clustered within schools, resulting in estimates that are biased and essentially inaccurate conclusions to be drawn. In essence, under the MLR method, crucial group-level structure and cross-levels interactions tends to be overlooked, thereby leading to the simplification of the variability and interdependencies across the different levels of the data.

In contrast, the Random Intercept Model (RIM), expressed as $Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + R_{ij}$ is distinctively designed to integrate these hierarchical structures through the addition of random effects. That is, RIM does not only provide estimates on the fixed effects which is common across all groups ($\gamma_{10}$), but does also account for the random variations between groups ($U_{0j}$). As seen in *Figure 10*, the Random Intercept Model features multiple lines with different intercepts for each group, demonstrating the individual starting positions of the different clusters, while keeping the slope consistent across the groups. This approach ensures that the model is able to adjust for the average effects within each group and subsequently to account for the variability that exists within those groups. Besides that, by considering the correlations among observations within the same group, the model is able to provide more precise estimates and robust inferences, essentially allowing the model to avoid the underestimation of standard errors which is common under the MLR method. This underestimation often leads to overly narrow confidence intervals where consequences include the invalidation of hypothesis tests and erroneous conclusions about the significance of the predictors, once again highlighting the superiority of the Random Intercept Model in effectively dealing with multilevel data.

To demonstrate the use of the Random Intercept Model in analysing complex hierarchical data, the study conducted by Sjögren et al. (2021) will be reviewed. Their study effectively establishes how the RIM's capabilities could be leveraged to obtain insights into behavioural dynamics within educational settings. The study examined how moral disengagement, defender self-efficacy, and the quality of student-teacher relationships influence bystander behaviour in situations of peer victimization. Data was obtained through self-reported surveys from 333 students, aged between 10 to 15, across four middle and junior high schools in Sweden. There are three explanatory variables namely moral disengagement, defender self-efficacy, and student-teacher relationship quality, while the response variable consisted of three types of bystander behaviours precisely, reinforcer, defender, and outsider. The objective of the study was to deepen the understanding of factors that influence students' 'reactions when encountered with peer victimization incidents.

To evaluate the interactions between the explanatory and response variables within the context of the data's nested structure (students within classrooms), the Random Intercept Model was employed. The motivation behind this choice of model was so that the regression intercepts are allowed to vary across the different classes. In the initial model, the gender and age variable were included as the controlled variables. In the second model, primary variables namely moral disengagement, defender self-efficacy, and student-teacher relationship qualities were subsequently integrated into the initial model. In the final and third model, interaction terms among the primary predictor variables were subsequently included. The findings of the study indicated that students with high levels of moral disengagement and low levels of defender self-efficacy tended to assume the roles of either reinforcers or outsiders. In contrast, students with high defender self-efficacy and established student-teacher relationships tended more to take on the roles of a defender. The researchers emphasized that their findings were observed both at the individual level and in comparisons among classmates, once again emphasizing the effectiveness and relevance of the Random Intercept Model in capturing the nuanced variations in behaviours across different classroom contexts.

**Section B (c) – Discussion on Random Slope Model and Literature Review**

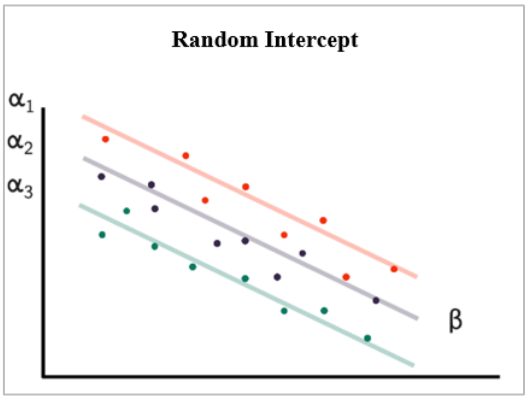


*Figure 11: Random Intercept Model (Midway, 2022)*

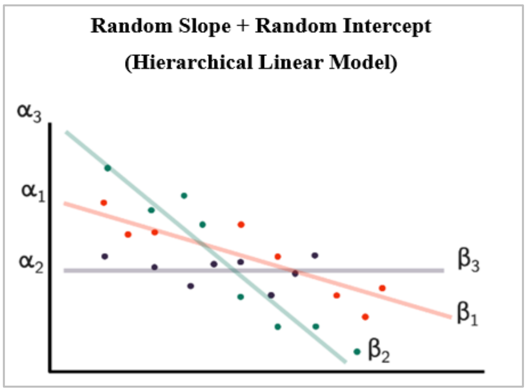


*Figure 12: Hierarchical Linear Model (Midway, 2022)*

The general equation for the Random Slope Model is expressed as $Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + U_{0j}x_{ij} + R_{ij}$, where $U_{0j}$ represents the random slope for group $j$, demonstrating the model's advanced capability in accounting for group-specific variability. As illustrated in *Figure 12* above, the Random Slope Model allows the slopes (the rate at which the response variable changes with respect to the explanatory variables) to differ across groups, essentially reflecting the complex dynamics of multilevel data structures. This flexibility is critical for analyzing data where interactions between explanatory variables and the response variable do vary significantly across groups or clusters. By allowing the slopes to differ, the model provides a more in-depth analysis of how predictor effects are influenced by different contextual factors within each group, essentially enhancing both model fitness and the accuracy of estimates for fixed effects, standard errors, and model predictions. *Figure 12* above illustrates this variability with divergent lines, each representing a different slope per group, allowing for a nuanced interpretation of fixed effects based on the interactions between predictors and group-specific characteristics.

Conversely, ignoring these group-specific slope variations, as illustrated in *Figure 11* above, may lead to biased estimates for fixed effects due to the oversimplified assumption of homogenous predictor effects across all groups. Such biases may significantly distort the results and lead to erroneous conclusions about the influence of the explanatory variables. Given the discussion thus far, inclusion of random slopes is therefore crucial for the purpose of accurately capturing the complex dynamics of multilevel data, ensuring that the analysis robustly reflects the true nature of the data structure and relationship.

Similar to the structure in *Section B (b)*, but this time is to demonstrate the inclusion of the Random Slope Model coupled with the Random Intercept Model in analyzing complex hierarchical data, the study conducted by Dusen and Nissen (2019) will be reviewed. This combined approach of integrating random slopes and random intercepts makes up the Hierarchical Linear Model. The study involved comparing the use of Multiple Linear Regression (MLR) and Hierarchical Linear Modelling (HLM) to investigate student learning improvement in introductory Physics courses. The data consisted of variables at the student-level (Level 1) precisely their pretest and post-test scores, coupled with variables at the course-level (Level 2) precisely the implementation of learning assistants and collaborative learning practices. The response variable was the improvement students made, calculated as the post-test scores minus the pretest scores. The explanatory variables however were students' individual pretest scores, average course pretest scores, as well as the incorporation of learning assistants and collaborative learning techniques as predictor variables. The aim of the study was to demonstrate the relevancy of employing the HLM technique when handling hierarchical data structures, like the data examined in their study, where students are nested within courses.

In conducting their study, the researchers developed several models using both the HLM and MLR, where they have progressively added predictors to evaluate their influence on the model fitness. HLM was chosen owing to its ability to handle the nested data structure coupled with its lack of reliance on the independence assumption required by MLR. The results revealed that both MLR and HLM suggested conflicting conclusions regarding the effectiveness of teaching practices. That is, while the former suggested that collaborative learning without the support of learning assistants resulted in the largest student improvement, the latter suggested that no statistically significant differences were found between collaborative learning with and without learning assistants. Moreover, the MLR approach produced standard errors that are smaller, implying that there is an increased risk of committing Type I errors. These findings of the MLR method thereby highlight the significance of employing suitable modelling techniques like HLM when analyzing hierarchical data, to avoid biased findings and erroneous conclusions.

**References**

Dusen, B.V., & Nissen, J. (2019). Modernizing use of regression models in physics education research: A review of hierarchical linear modeling. *Physical Review Physical Education Research*, *15.* 020108. https://doi.org/10.1103/PhysRevPhysEducRes.15.020108

Midway, S. (2022). Random effects. Data analysis in R (pp. 141-158). Bookdown. https://bookdown.org/steve_midway/DAR/random-effects.html

Sjögren, B., Thornberg, R., Wänström, L., & Gini, G. (2021) Bystander behaviour in peer victimisation: Moral disengagement, defender self-efficacy and student-teacher relationship quality. *Research Papers in Education*, *36*(5), 588-610. https://doi.org/10.1080/02671522.2020.1723679

## Appendices

### Appendix A

While the *jsp* dataset consists of student's data taken from a total of 49 schools, there were only 48 schools in the analysis because recall that the focus of this report is only on the analysis of students' Year 2 educational data. As previously mentioned as well, not all students will have data for all three years of schooling (Year 0, Year 1, and/ or Year 2). Specifically, the school with the ID number "10" does not have any data for its Year 2 students, hence justifying for why there were only 48 schools included in all of the analyses carried out in this report.