**Research Q3 part 3 phase 3 Regression Analysis**

**RQ3 part two of phase three: Linear regression**

Step three involved conducting linear regression tests first with total language scores of 19 questions to see if any of the factors were identified via literature reviews first, then assessing via t-test results to know if they predicted children's difficulties. Two of these variables (namely, deprived environments and chaotic environments) showed significant differences in pre-regression t-tests, thus in order to determine whether they predict language difficulties in vulnerable children, linear regression model analysis was run if any of the variables showed either significant or with an effect size of >d=0.5 (medium effect size). As the literature confirms that these variables affect children's language development and skills. The children's total language scores were the dependent variable, and demographic and environmental factors were independent variables. The useful function of linear regression is to predict the dependent variable's values that attempt to describe the relationship between two or more variables with a linear or straight-line function (Gujarati, 2019; Marill, 2004). Thus, the linear regression analysis was considered as the suitable analysis method. In this study, the linear regression analysis evaluated the relationships between total language scores and language scores of models of language development areas (e.g., syntactic language scores, semantic and social language) which were a continuous dependent variable, and 5 independent variables which were potential predictor variables including, deprived environments and chaotic environments, poor living conditions, low SES conditions and time spent in school.

A standard linear regression model was used to examine factors that predict language difficulty in LAC. In this model, Y resented total language scores (DV), and X represented independent variables (IV) which are mentioned above. With this regression model, both DVs and all IVs are entered in the equation model (Y = β0 + β1x demographic factors). The linear regression model was chosen rather than the multiply regression. This is because it suited best with the research question, i.e., the relative importance of potential predictors of language difficulty in vulnerable children has not been previously researched. Thus, this method was used to examine expected relationships between IVs and the total language scores.

**Processes of selecting predictors to enter in the regression model**

In Step 1 of the regression analysis, all the IVs that were considered to be related to vulnerable children's language difficulties were entered in the equation model. As mentioned earlier, these IVs were identified based on the literature review, which they were confirmed to be strongly related to vulnerable children's life trajectories. The final decision for entering potential predictors in this model was based on the results of the pre-regression parametric t-tests results where if they were significant or the calculated effect sizes over >d=0.5 entered in the regressions. Following the t-test analysis and effect size calculation for indicated which IVs to enter as potential predictor variables into the model.

First, to determine which IVs to enter in the model, only those with justification based on the literature were considered, i.e., variables shown to have a relationship with vulnerable children's language difficulties. Following, before the linear regression analysis was conducted, correlation analyses were also conducted to determine which IVs would be entered as predictors in the model. Correlation analyses examined the relationship between children's total language scores and all the IVs, for the whole sample and separately for each of the models of language development groups (syntax, semantic and social language). Only the IVs significantly associated with total language scores (DVs) whole cohorts and for the language development model, three language groups were entered in the equation, which were subject to linear regression assumptions. The only exception to these were poor living conditions, low living in SES conditions and time spent in schools, and the researcher had both theoretical and evidence-based reasons for those IVs. As the literature clearly indicates, poor living conditions and low living in SES conditions are known to affect children's development and life trajectories significantly. Thus, the final decision for entering it or not as a potential predictor in the model was based on the whole sample correlation results. Following the correlation analyses, which indicated which IVs to enter as potential predictor variables into the model, regression assumptions were tested before regression analyses were carried out, linear regression assumptions were tested. The regression assumptions were checked and followed this process, which involved several steps: checking, Durbin-Watson statistic, outliers, and homoscedasticity residuals (errors). Also, although children were grouped under three categories based on the traffic light system in RQ2 (red, amber and green groups), the linear regression model was first run for the whole cohort given that the three-language group exhibited different language mean scores in post hoc tests (see the results of RQ2), involving total language scores of 19 questions (see the result section).

In Step 2, in order to examine which of the factors predict the language difficulties in three language development areas, namely syntax, semantic and social language, standard linear regression analyses were also run to which of these identified factors predicts this cohorts' difficulties in these areas. The same IVs (e.g., deprived environments, chaotic environments, poor living conditions, low SES living conditions, time spend in schools) were used as predictors and syntax, semantic and social language scores were used as DVs. Before running the linear regressions, procedures in step 1 followed: correlations between variables run and assumptions of the regressions were check-in a similar manner. Once the all step was completed, the regression model was then run which the results were presented in results sections.

Relevant results are presented in this section, but the methods behind these assumptions are described in detail in the section called research method.

**The results of standard linear regression analysis**

Given that children's language scores differed significantly between three language groups in RQ2 (green, amber and red), to investigate consistent predictors of language difficulties across the whole cohort, in the step, total language scores of 19 questions and identified predictors (IVs) were entered in the regression. In step two involved running the regression analyses between the DVs involved in language syntax, semantic and social language scores and the same IVs used in step one to see if they were shown any significance between them. .........

**Percentages of difficulties observed in language areas in vulnerable children.**

In addition to the above statists, percentages of difficulties in areas of syntax, sematic and social language were explored by ching each children total language scores in these aspects of language. This process involved summing up the answers given for each question based on the coding used system used (0, 1 and 2) and dividing them.

It consists of adding the answers given for each question according to the coding system used (0, 1 and 2), dividing them by the total language scores and multiplying them by one hundred. For example, the total language scores for semantic was established in RO3 part as 2 as 16, and the formula used to gather percentages of difficulties children presented in 'listening information' (q1) was 5/16 x 100 and based on that, the sematic difficulties for child number 1 were found as 31.5%. The same process was followed for all three language areas. Results showed that percentages of difficulties presented in syntaxes were found to be between 25% to 85 %; sematic 22.5% to 78%, while social language difficulties were between 34.2% to 65.3%. The difficulties presented to vulnerable children related to these three areas were found to be considerably high in caparison to the general population.