Investigation Vocabulary Learning Pattern*

My subtitle if needed

Yongqi Liu

November 21, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section **??**….

The data used in this study comes from Braginsky (2024), and all analysis are conducted in R Core Team (2023)

*Code and data are available at: https://github.com/Cassieliu77/Vocabulary_Learning_Pattern.git

2 Data

2.1 Overview

This study uses R packages (R Core Team 2023) to clean and analyze the dataset, including libraries from tidyverse (Wickham et al. 2019), ggplot2 (**ggplot2?**), knitr (**knitr?**), arrow (**arrow?**). We use the statistical programming language R (R Core Team 2023)…. Our data (Braginsky 2024)

After cleaning the data, which included grouping and removing missing values, the following analysis focuses on category, age, comprehension, production, is_norming, broad_category columns in the analysis dataset. Table **??** shows the overview of the dataset.

Table 1: Summary Table for the Word Bank Dataset

Data_ID	Language	Age	Is_Norming	Broad_Category	Production	High_Vocabulary
396587	English (American)	25	FALSE	Sensory Words	658	1
396588	English (American)	26	FALSE	Sensory Words	552	1
396589	English (American)	24	FALSE	Sensory Words	504	1
396590	English (American)	26	FALSE	Sensory Words	272	0
396591	English (American)	24	FALSE	Sensory Words	350	0
396592	English (American)	25	FALSE	Sensory Words	580	1
396593	English (American)	22	FALSE	Sensory Words	351	1
396594	English (American)	24	FALSE	Sensory Words	310	0
396595	English (American)	25	FALSE	Sensory Words	257	0
396596	English (American)	26	FALSE	Sensory Words	188	0

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome Variable

2.3.1 High Vocabulary Score

The outcome variable in this study, High Vocabulary Score, is a binary indicator designed to identify individuals with advanced vocabulary proficiency. This variable is derived from two key measures:

1. Comprehension: This variable represents the ability to understand words and phrases, reflecting the receptive language skills of individuals. Comprehension scores are numerical and vary across the dataset.
2. Production: This variable captures the ability to produce words, reflecting expressive language skills. Like comprehension, production scores are numerical and provide the standard into verbal articulation capabilities.
3. The High Vocabulary Score is calculated using the average of comprehension and production scores for each individual. This average is represented as: prod_comp_mean = $\frac{\text{Comprehension}+\text{Production}}{2}$

To classify individuals, a threshold value of 350 is applied to `prod_comp_mean`: - Individuals with `prod_comp_mean > 350` are classified as having a high vocabulary score (outcome = 1). - Those with `prod_comp_mean <= 350` are classified as not having a high vocabulary score (outcome = 0).

This approach ensures that both receptive (comprehension) and expressive (production) skills are considered in defining advanced vocabulary. The threshold of 350 was chosen based on exploratory analysis of the dataset, reflecting a meaningful distinction between individuals with high and low vocabulary abilities. The High Vocabulary Score serves as the dependent variable in the following data analysis part. Its binary nature makes it suitable for modeling with a binomial family distribution, allowing for the estimation of factors that influence advanced vocabulary acquisition.

2.4 Predictor variables

2.4.1 Age

And also planes (Figure **??**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

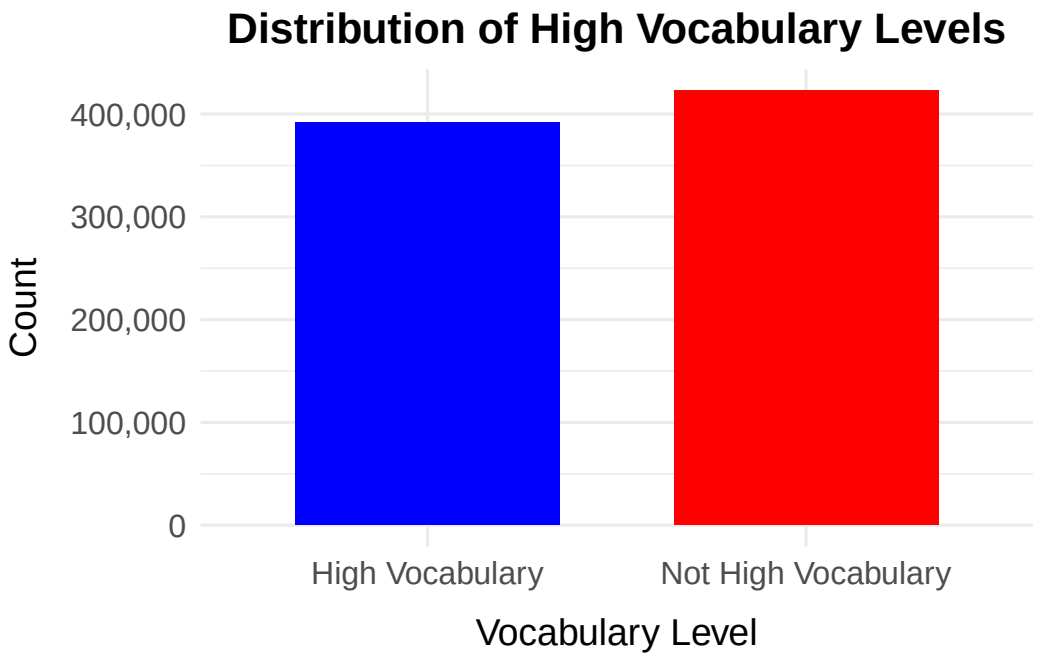

Figure 1: Distribution of the outcome variable, showing the counts of children classified as having "High Vocabulary" and "Not High Vocabulary" based on their comprehension and production scores. The bar plot illustrates the balance between the two categories in the dataset, which is important for modeling purposes.

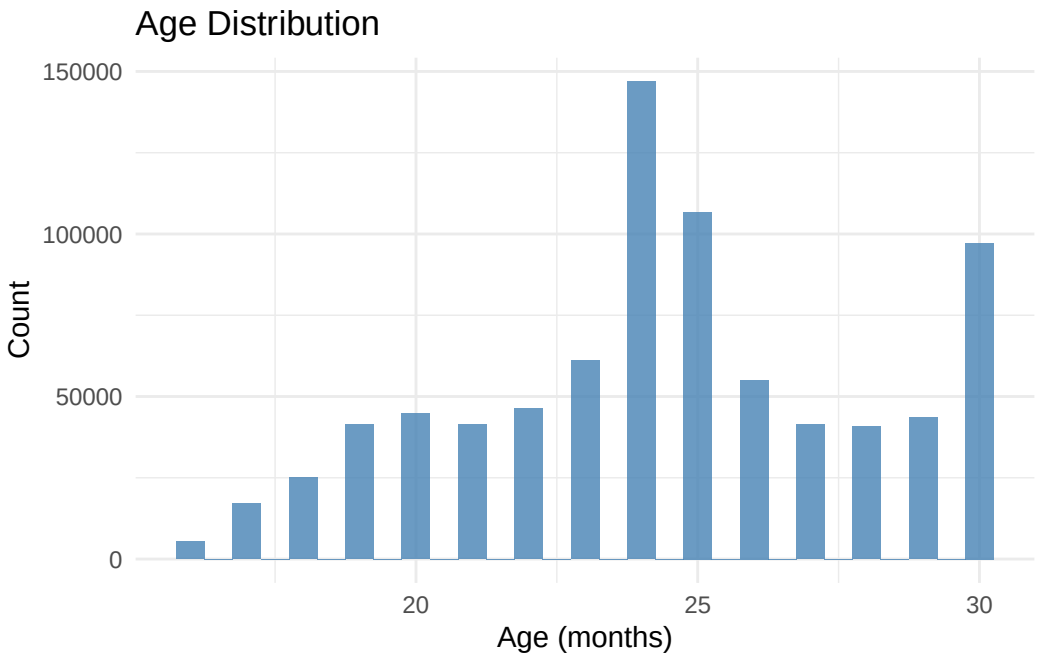

Figure 2: Age Distribution

2.4.2 Broad Category

2.4.3 Is Norming or Not

3 Model

3.1 Model Selection

To investigate the relationship between children's vocabulary acquisition and their demographic and linguistic characteristics, we constructed a logistic regression model. By examining key demographic and linguistic predictors, we aim to identify how characteristics like age, norming status, and word categories influence vocabulary development. The dependent variable, high_vocabulary, is a binary outcome indicating whether a child's average production and comprehension score (denoted as prod_comp_mean) exceeds 350. This threshold was chosen to distinguish children with relatively advanced vocabulary levels. More background details and diagnostics are included in Appendix- **??**.

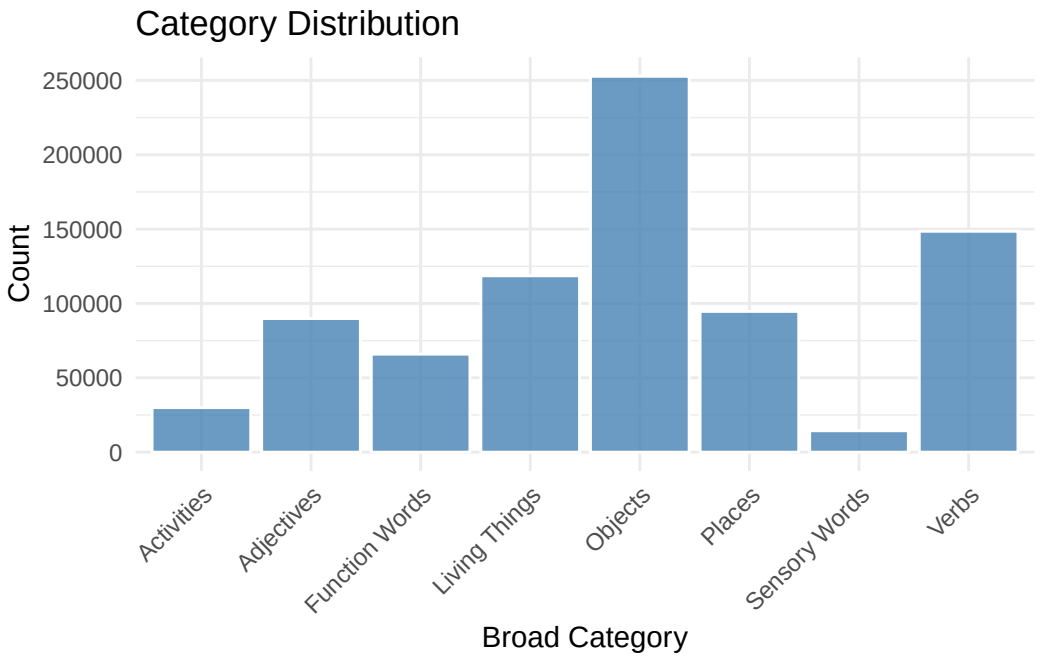

Figure 3: Distribution of Word Categories. This bar plot illustrates the distribution of items across broad linguistic categories in the dataset. "Objects" dominate the dataset, followed by "Verbs" and "Living Things," indicating that the dataset is rich in concrete nouns and action-related words, with smaller proportions of sensory words and adjectives.

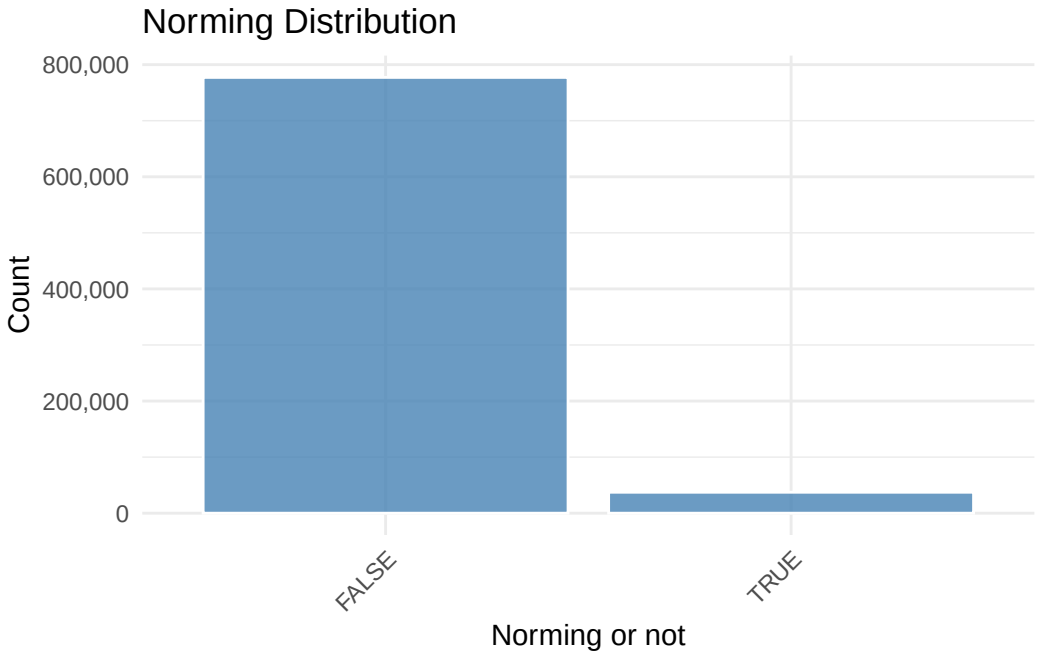

Figure 4: Norming Group Distribution. This plot shows the distribution of children in the dataset categorized by their norming status. The majority of the observations are from non-norming children, with only a small fraction representing norming children.

3.2 Logistic Regression Model Overview

- High Vocabulary: The outcome variable, high_vocabulary, is a binary indicator that takes the value of 1 if the average production and comprehension score (prod_comp_mean) exceeds 350 and 0 otherwise. This threshold was selected to represent children with relatively advanced vocabulary skills, determined through exploratory data analysis.
- Scaled Age (age_scaled): This continuous variable represents the child's age, standardized to ensure the model coefficients reflect changes per standard deviation in age. Standardization improves numerical stability and aids interpretability.
- Norming Status (is_norming): A binary indicator denoting whether a child is part of the norming dataset (TRUE) or not (FALSE). This variable accounts for potential differences in data collection or assessment protocols.
- Broad Category (broad_category): A categorical variable grouping words into four broad linguistic categories: nouns, verbs, adjectives, and function_words. The reference category is function_words.

The model takes the form:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \text{age\_scaled}_i + \beta_2 \cdot \text{is\_normingTRUE}_i \tag{1}$$

$$+ \beta_3 \cdot \text{broad\_categoryAdjectives}_i \tag{2}$$

$$+ \beta_4 \cdot \text{broad\_categoryFunction\_Words}_i \tag{3}$$

$$+ \beta_5 \cdot \text{broad\_categoryLiving\_Things}_i \tag{4}$$

$$+ \beta_6 \cdot \text{broad\_categoryObjects}_i \tag{5}$$

$$+ \beta_7 \cdot \text{broad\_categoryPlaces}_i \tag{6}$$

$$+ \beta_8 \cdot \text{broad\_categorySensory\_Words}_i \tag{7}$$

$$+ \beta_9 \cdot \text{broad\_categoryVerbs}_i \tag{8}$$

Where: - p_i represents the probability that person i has a high vocabulary - β_0 is the intercept, capturing the baseline log-odds when all predictors are at their reference or mean levels - β_1: Effect of age (standardized) - β_2: The effect of whether the individual belongs to the norming group - β_3, β_4, β_5, etc.: The effects of being in the respective broad word categories (nouns, function words, or verbs), compared to the reference category (likely "adjectives").

3.3 Model Assumptions

- Linearity of the Logit: The model assumes a linear relationship between the log-odds of the outcome (high vocabulary) and the independent variables. For example, the

standardized age variable (age_scaled) assumes that for every one standard deviation increase in age, the log-odds of achieving a high vocabulary score increase by a constant amount. Standardizing age ensures that the variable is centered and scaled, making it easier to meet this linearity assumption and interpret its effect across the dataset.

- Independent Observations: The model assumes that all data points are independent. This assumption holds because each observation represents data from a unique child, with no repeated measurements for the same individual. For instance, there are no longitudinal observations or nested data structures (e.g., children grouped by classrooms or schools) that could violate independence. If dependence were present, a more complex model like a mixed-effects logistic regression would be necessary.
- Categorical Variable Encoding: The broad_category variable, which includes categories such as "Adjectives," "Verbs," and "Living Things," was encoded using sum contrasts. This approach ensures that the coefficients for each category represent the deviation of that category's effect from the overall mean effect across all categories. For example, the coefficient for "Verbs" indicates how the log-odds of achieving a high vocabulary differ for "Verbs" compared to the average effect of all other categories. Sum contrasts are particularly useful for understanding relative effects and ensure that the intercept reflects the overall mean effect when all predictors are at their reference or average levels.

3.4 Interpretation of Coefficients

The logistic regression coefficient (β) represents the change in the log-odds of the dependent variable (high vocabulary) for a one-unit change in the predictor variable, holding all other variables constant.

- Intercept (β_0): Represents the log-odds of high vocabulary when all predictors are at their reference or mean levels. If $\beta_0 > 0$, the baseline odds of high vocabulary are greater than 50%.
- Scaled Age (β_1): For each one standard deviation increase in age, the log-odds of high vocabulary increase by β_1. If $\beta_1 = 0.5$, then $\exp(0.5) \approx 1.65$, meaning the odds increase by 65% for every one standard deviation increase in age.
- Norming Status (β_2): If a child belongs to the norming group, the log-odds of high vocabulary increase by β_2 compared to non-norming children. If $\beta_2 = 0.1$, then $\exp(0.1) \approx 1.11$, meaning being in the norming group increases the odds of high vocabulary by 11%.
- Broad Category (β_3, β_4, β_5, etc.): The coefficients for `broad_category` represent the difference in log-odds compared to the reference category ("Adjectives").

Interpretation:
- β_3 (Function Words): A positive β_3 indicates higher odds of having a high vocabulary for function words compared to adjectives. For instance, if $\beta_3 = 0.002$, then $\exp(0.002) \approx 1.002$, meaning the odds of having a high vocabulary for function words are 0.2% higher than for

adjectives.
- β_4 (Living Things): A negative β_4 indicates lower odds of having a high vocabulary for living things compared to adjectives. For example, if $\beta_4 = -0.007$, then $\exp(-0.007) \approx 0.993$, meaning the odds of having a high vocabulary for living things are 0.7% lower than for adjectives.
- β_5 (Objects): If $\beta_5 = -0.008$, then $\exp(-0.008) \approx 0.992$, indicating that objects are associated with 0.8% lower odds of having a high vocabulary compared to adjectives.
- β_6 (Verbs): If $\beta_6 = -0.005$, then $\exp(-0.005) \approx 0.995$, meaning verbs are associated with 0.5% lower odds of having a high vocabulary compared to adjectives.

General Example for Broad Categories:
If a coefficient $\beta_k = 0.01$, $\exp(0.01) \approx 1.01$, meaning the corresponding category increases the odds of having a high vocabulary by 1% compared to the reference category (Adjectives). Conversely, if $\beta_k = -0.01$, $\exp(-0.01) \approx 0.99$, indicating a 1% decrease in odds compared to the reference category.

3.5 Model Justification

The model was trained on 80% of the dataset, with the remaining 20% reserved for testing. Model performance was assessed using confusion matrices and overall accuracy. Further evaluation included the analysis of residual deviance, AIC, and interpretation of individual coefficients.

4 Results

4.1 Average Production by Broad Category and Age

Figure **??** shows..

4.2 Prediciton for the Probability of High Vocabulary Level

```
      Predicted
Actual     0     1
     0 59837 25062
     1 26778 51251
```

```
[1] "Accuracy: 0.68182264558578"
```

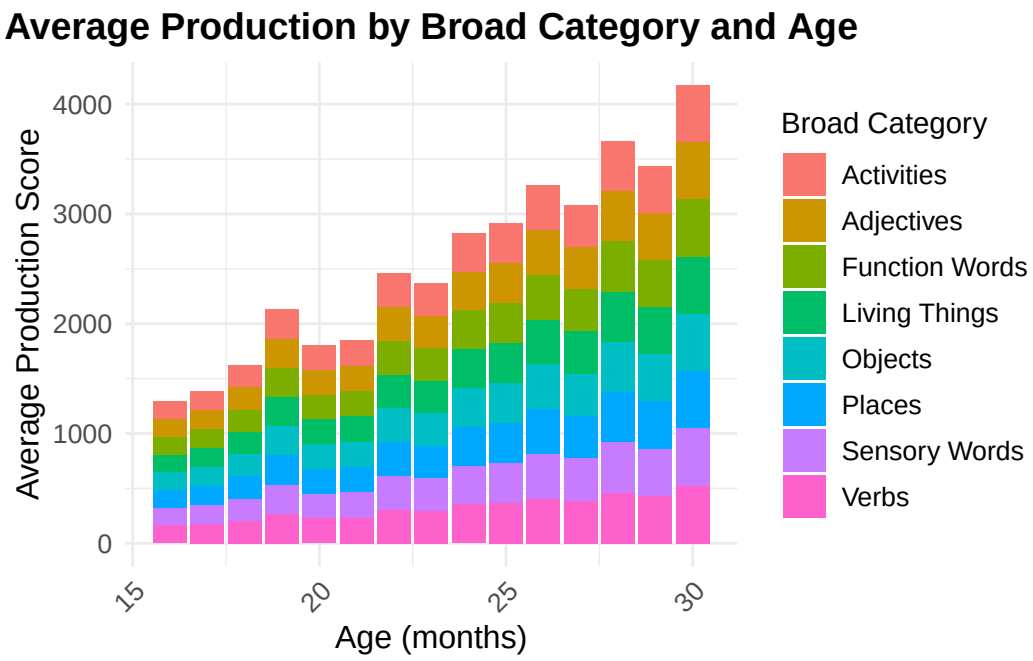

Figure 5: xx

4.2.1 Distribution of Predicted Probabilities by Broad Category

Figure **??** illustrates the distribution of predicted probabilities for high vocabulary levels across various broad categories. The density curves show distinct peaks and variability, reflecting the differences in how well each category predicts high vocabulary. Notably, "Sensory Words" and "Objects" exhibit higher density in the middle range of predicted probabilities, suggesting moderate association with high vocabulary. In contrast, categories like "Activities" and "Function Words" have wider, flatter distributions, indicating greater uncertainty or diversity in prediction. This variability highlights the nuanced role of word categories in predicting high vocabulary acquisition. Further analysis could explore why certain categories contribute more consistently to predictions than others. The density plot highlights distinct patterns in predicted probabilities, with nouns and verbs showing higher peaks, indicating a stronger likelihood of high vocabulary acquisition in these categories. This suggests that broad categories contribute differently to vocabulary development, with nouns and verbs potentially playing a more significant role.

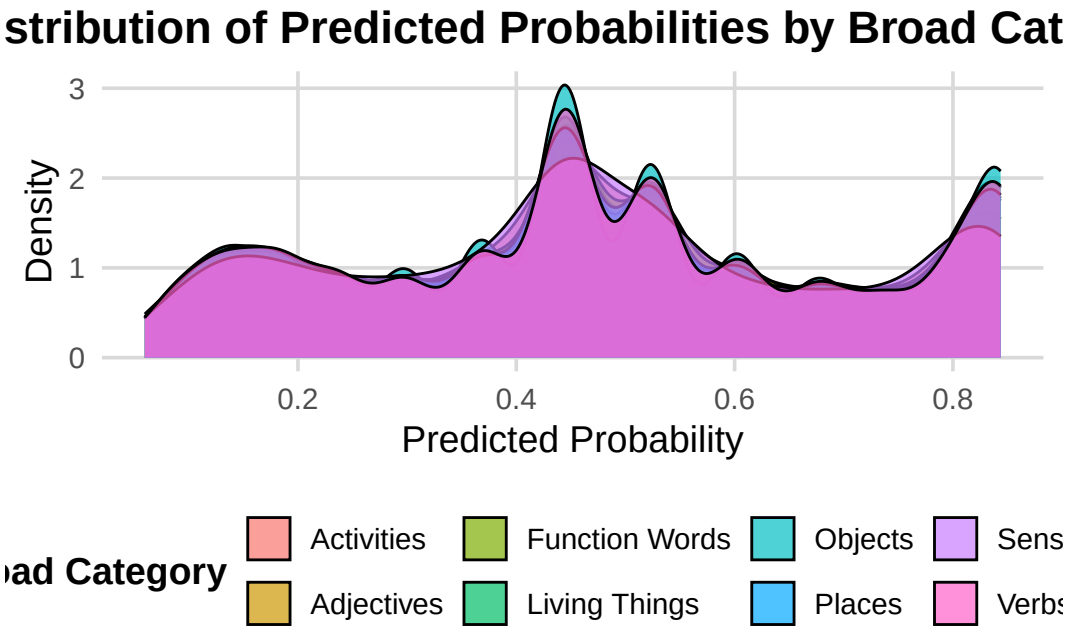

Figure 6: This density plot illustrates the predicted probabilities of having a high vocabulary across different broad lexical categories. Each curve represents the density of predicted probabilities within a category, showcasing the variation in predicted outcomes for categories such as Activities, Adjectives, Function Words, Living Things, Objects, Places, Sensory Words, and Verbs. The plot highlights overlapping patterns and areas of divergence in vocabulary acquisition likelihood across different types of words.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.