Pathways to Early Vocabulary Acquisition*

Predicting English Language Development Trajectories in Children Aged 16–30 Months

Yongqi Liu

November 27, 2024

This study examines vocabulary acquisition in children aged 16–30 months using logistic regression models and data from the MacArthur-Bates Communicative Development Inventories. Age is the strongest predictor, with older children showing higher vocabulary proficiency, while foundational word categories like Function Words are acquired more consistently than complex ones such as Sensory Words. Children in norming groups outperform non-norming peers, highlighting the impact of structured linguistic environments. The findings support targeted interventions to address variability in complex vocabulary acquisition and promote balanced language development.

Table of contents

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

1 Introduction

Understanding how children acquire vocabulary is a cornerstone of developmental linguistics and early childhood education. Vocabulary growth is not only a key indicator of cognitive development but also serves as a foundation for future linguistic and academic success. The

interplay of factors such as age, linguistic environments, and word categories influences children's vocabulary acquisition in complex and dynamic ways. By analyzing patterns and predictors of vocabulary growth, researchers can illuminate the developmental trajectories that underpin language acquisition, offering a clearer understanding of how children progress in their linguistic abilities.

This study utilizes a dataset from the Braginsky (2024) database, which aggregates data from the MacArthur-Bates Communicative Development Inventories (CDI), to investigate vocabulary development in children aged 16–30 months. By employing a logistic regression model, the research estimates the likelihood of achieving a high vocabulary score using predictors such as age, norming status, and lexical categories. This model serves as a tool to explore how these factors collectively influence children's vocabulary acquisition trajectories.

The estimand of interest in this study is the conditional probability that a child possesses advanced vocabulary proficiency, defined as a high combined comprehension and production score, given specific predictors. These predictors include the child's age in months, their norming status (whether they belong to a standardized sample used for benchmarking), and the lexical category of the words (e.g., Function Words, Verbs, Adjectives). By employing a logistic regression model, the research estimates how these variables individually and interactively influence the likelihood of achieving advanced vocabulary levels. The model demonstrates developmental trends, such as how median vocabulary proficiency increases with age, and reveals how specific word categories contribute uniquely to language growth. For example, foundational categories like Function Words and Objects may show consistently higher probabilities of acquisition, while more abstract categories, such as Verbs and Sensory Words, exhibit greater variability due to their complexity or context-specific nature.

Key findings reveal distinct trajectories of vocabulary acquisition across different word categories, with age emerging as the strongest predictor of vocabulary growth. The analysis demonstrates that as children age, their likelihood of achieving high vocabulary scores consistently increases, reflecting the natural progression of language development. Categories such as "Living Things" and "Function Words" display stable and high predicted probabilities, likely due to their frequent occurrence in everyday communication and their role in forming early linguistic structures. Conversely, categories such as "Sensory Words" and "Adjectives" show more variability, highlighting the challenges posed by their later acquisition and the contextual understanding they require.

Additionally, children in the norming group, who are assessed in structured linguistic environments, consistently achieve higher vocabulary scores compared to non-norming children. This difference underscores the influence of systematic linguistic exposure in shaping vocabulary development. The results not only provide a nuanced understanding of vocabulary acquisition but also emphasize the importance of targeted language interventions. They suggest that focusing on underdeveloped categories, such as Sensory Words and Adjectives, could help bridge gaps in vocabulary learning. Furthermore, these findings set the stage for future research into the broader contextual factors, such as socio-economic status and bilingualism, that may further explain variability in language acquisition trajectories.

The remainder of this paper is structured as follows: Section 1 provides an introduction and decribe the estimand for the paper. Section 2 describes the dataset, highlighting the variables used and their distributions. Section 3 provides an overview of the logistic regression model and its assumptions. Section 4 presents the visualized results, including variability in vocabulary acquisition, the role of predictors, and predicted probabilities. Finally, Section 5 discusses the implications of the findings and outlines limitations and directions for future research. The data used in this study all comes from Braginsky (2024), and the whole paper is conducted and analyzed in R Core Team (2023).

2 Data

2.1 Overview

The original dataset was obtained from Braginsky (2024). After undergoing a thorough cleaning process—including grouping related items and removing missing values—the analysis focuses on the key variables: category, age, comprehension, production, is_norming, and broad_category. These variables form the foundation of the analysis dataset. An overview of the cleaned dataset is presented in Table 1.

Table 1: Cleaned Word Bank Dataset

Language	Age	Is_Norming	Broad_Category	Production	High_Vocabulary
English (American)	25	FALSE	Sensory Words	658	1
English (American)	26	FALSE	Sensory Words	552	1
English (American)	24	FALSE	Sensory Words	504	1
English (American)	26	FALSE	Sensory Words	272	0
English (American)	24	FALSE	Sensory Words	350	1
English (American)	25	FALSE	Sensory Words	580	1
English (American)	22	FALSE	Sensory Words	351	1
English (American)	24	FALSE	Sensory Words	310	1
English (American)	25	FALSE	Sensory Words	257	0
English (American)	26	FALSE	Sensory Words	188	0

2.2 Measurement

The objective of measurement in this study is to translate raw parental reports into reliable indicators of vocabulary acquisition patterns in children. The data is derived from the MacArthur-Bates Communicative Development Inventories (CDI), a widely used tool that collects information on children's vocabulary comprehension and production through structured parental surveys. These surveys allow parents to report on their child's understanding and

use of specific words, grouped into lexical categories such as nouns, verbs, and adjectives. The raw data collected through the CDI forms the basis for creating the study's dependent and independent variables.

- Standardized Format and Structure: The CDI employs predefined response categories, which help to minimize ambiguity in reporting and ensure consistency across respondents. This structured approach mitigates some reporting variability but may not fully capture nuances in vocabulary acquisition.
- Words Categorization: Vocabulary items are grouped into meaningful lexical categories, allowing for a more nuanced understanding of children's vocabulary development across different word types.
- Norming Group Representation: To improve validity, a subset of children from norming groups is included as a benchmark for comparison. While useful, this raises concerns about whether the norming group adequately represents the population's diversity in language development.
- Variable Standardization: Continuous variables, such as age, are standardized (e.g., scaled) to reduce variability and improve the interpretability of statistical models. This ensures that coefficients reflect meaningful changes in relation to standardized measures.
- Bias Mitigation: The CDI's structured responses and norming benchmarks help reduce bias but cannot fully eliminate inaccuracies from parental reporting. Parents may overestimate or underestimate their child's abilities due to subjective perceptions or limited observations, which remains a limitation of the self-reported data.
- Missing Data Handling: Observations with incomplete or invalid responses were excluded from the analysis to maintain the integrity and reliability of the dataset.
- Temporal Limitations: The CDI data represents snapshots of vocabulary development at specific ages, which may not account for rapid changes or variations over time in a child's language acquisition process.
- Social Desirability Bias: Responses may be influenced by parents' desire to portray their child's language development favorably.

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 = $\frac{\text{Comprehension}+\text{Production}}{2}$

Based on Taintor and LaMarr (2023), I looked into the normal vocabulary size change among the children aged 16 to 30 months. To classify individuals, a threshold value of 300 is applied to `prod_comp_mean`: - Individuals with `prod_comp_mean > 300` are classified as having a high vocabulary score (outcome = 1). - Those with `prod_comp_mean <= 300` 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 300 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.

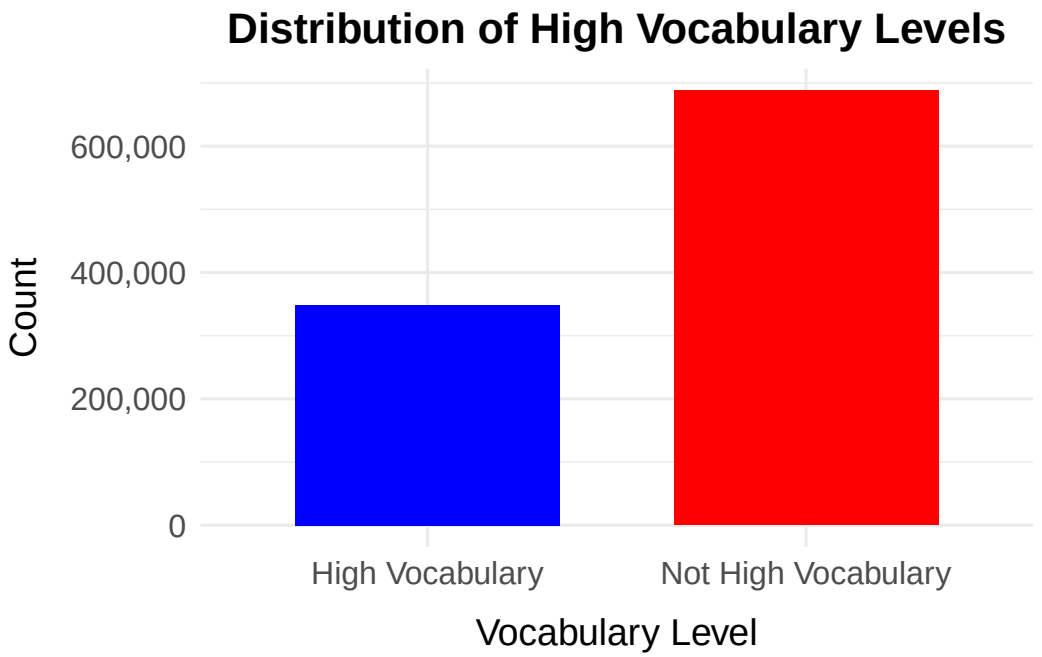


Figure 1: Distribution of the outcome variable, showing the counts of children classified as having "High Vocabulary" and "Not High Vocabulary" based on their comprehension and production scores. The bar plot illustrates the balance between the two categories in the dataset, which is important for modeling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

2.4 Predictor variables

2.4.1 Age

Figure 2 displays the distribution of children's ages (in months) within the dataset, highlighting key patterns in the sample's demographic structure. A notable concentration of data is observed among children aged between 24 and 30 months, reflecting an emphasis on capturing vocabulary development during critical periods of language acquisition. These age ranges are known to mark significant milestones in linguistic growth, which could explain their higher representation. Conversely, younger age groups (below 20 months) are underrepresented, likely due to the challenges of assessing vocabulary at earlier stages of development, where verbal communication is less pronounced and parental reporting is more variable.

The dataset also shows distinct peaks at specific ages, such as 25 and 30 months. These sharp spikes may reflect intentional focus points for testing or developmental benchmarks tied to standardized assessments like the MacArthur-Bates Communicative Development Inventories (CDI). This uneven age distribution underscores the importance of age as a critical factor in analyzing vocabulary acquisition. While the high concentration of data at older ages enhances the analysis into advanced vocabulary development, it also necessitates caution in generalizing findings to underrepresented age groups. This observation emphasizes the need to standardize age in statistical models to account for variability across different age groups.

2.4.2 Lexical Category

The words in the dataset were grouped into broad lexical categories to facilitate the analysis of vocabulary acquisition patterns. These categories include Activities, Adjectives, Function Words, Living Things, Objects, Places, Sensory Words, and Verbs. The classification was based on the semantic and functional roles of words, with nouns subdivided into more specific groups such as Living Things, Objects, and Places to capture distinct trends in vocabulary acquisition. For instance, Function Words include pronouns and question words, reflecting grammatical development, while Verbs and Adjectives capture action and descriptive words, essential for sentence construction and expression.

The bar graph illustrates the distribution of items across these categories, highlighting significant variation in word frequency. Objects constitute the largest category, suggesting a focus on tangible and concrete items, which are likely easier for children to recognize and recall. This is followed by Verbs and Living Things, categories that are fundamental to communication but slightly less prevalent. In contrast, Sensory Words and Activities are sparsely represented, possibly reflecting their specialized and context-dependent nature. The distribution underscores the importance of concrete and functional words in early vocabulary development while highlighting potential gaps in underrepresented categories. This variation provides a foundation for exploring how lexical diversity influences vocabulary acquisition patterns.

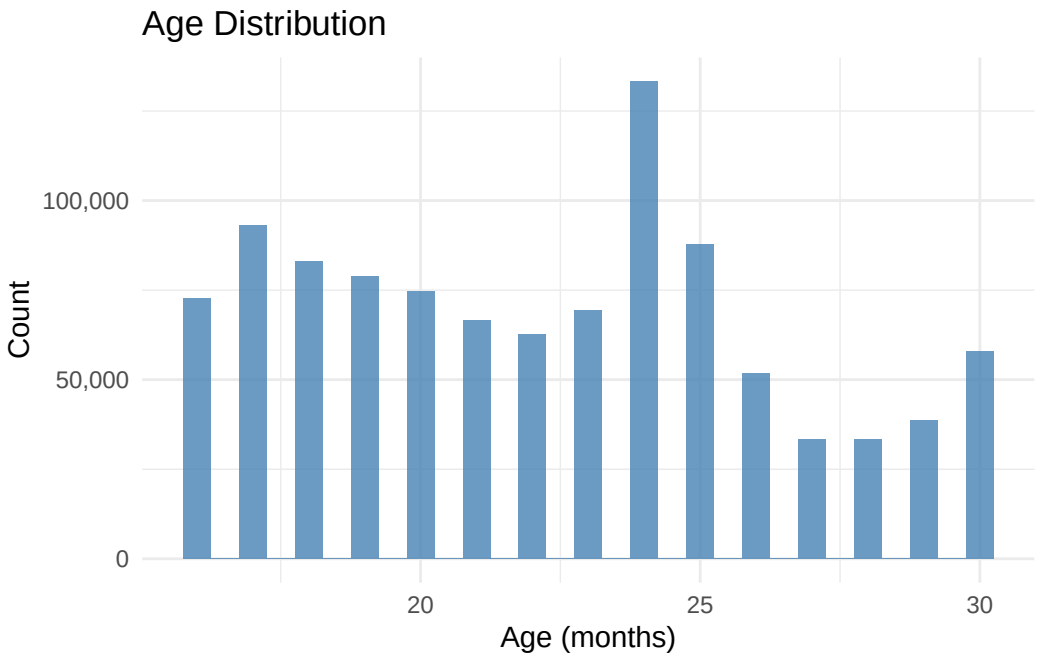


Figure 2: It shows the distribution of children's ages (in months) within the dataset. The majority of observations fall between 24 and 30 months, with noticeable peaks at 25 and 30 months, reflecting a focus on key developmental periods. Younger age groups are underrepresented, highlighting the need to account for variability in age when analyzing vocabulary acquisition patterns.

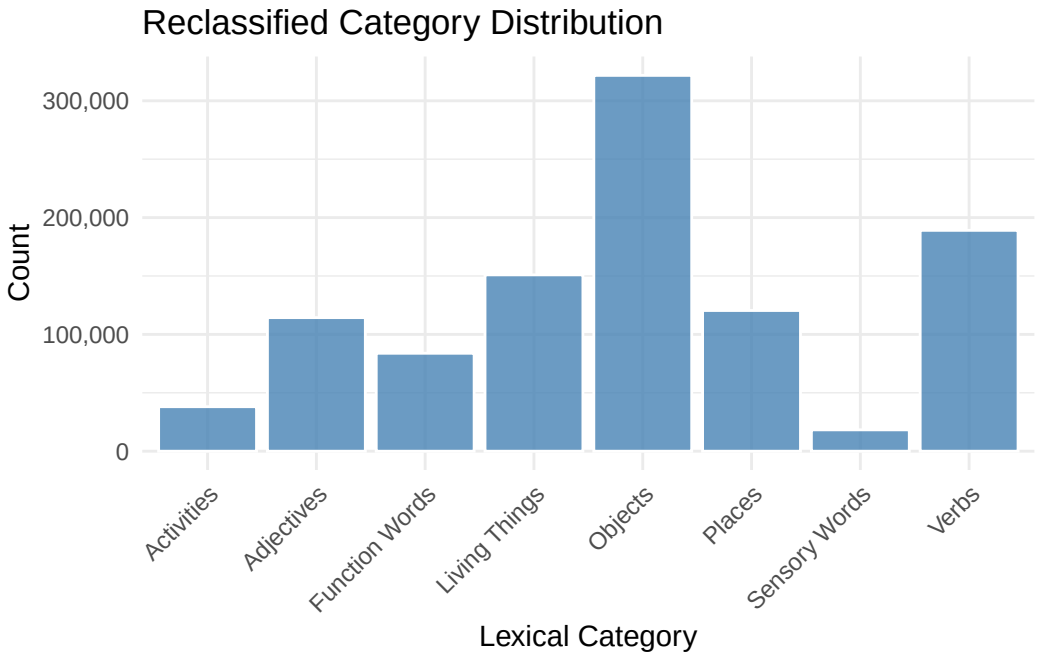


Figure 3: The figure shows objects dominate the vocabulary, reflecting an emphasis on concrete and tangible terms, while categories like Sensory Words and Activities are less frequently represented, indicating the relative complexity or specificity of these word types in early language acquisition

2.4.3 Norming Status

The is_norming variable categorizes children into two groups: the norming group and the non-norming group. The norming group serves as a standardized sample, providing a benchmark for assessing vocabulary development and enabling reliable comparisons across the dataset. This group is essential for ensuring the validity of the analysis by offering a consistent reference point for evaluating individual and group-level differences in vocabulary acquisition.

The dataset is predominantly composed of non-norming children, with only a small proportion belonging to the norming group. This imbalance reflects the inherent challenges of including a sufficiently broad and representative norming sample in large-scale assessments. While the non-norming group captures diverse linguistic and demographic contexts, the norming group remains vital for calibrating and interpreting vocabulary development patterns, particularly in early language acquisition studies. Table 2 illustrates this distribution, highlighting the predominance of non-norming children in the dataset.

The is_norming variable is included in the logistic regression model to account for potential systematic differences between the norming and non-norming groups. As the norming group represents a standardized sample used as a benchmark for vocabulary development, its inclusion ensures that variations in the likelihood of achieving a high vocabulary score are not confounded by differences in group composition or assessment protocols. It allows the model to estimate whether membership in the norming group significantly influences vocabulary acquisition outcomes. This is particularly important given the inherent imbalance in the dataset, where the non-norming group constitutes the majority. By controlling for is_norming, the model can differentiate between developmental patterns attributable to broader population diversity versus those arising from the structured selection of the norming sample.

Table 2: The dataset is primarily composed of non-norming children, with a smaller subset belonging to the norming group, serving as a standardized benchmark for assessing vocabulary development

Table 2: Summary of Norming Status in the Dataset

Norming Status	Count
FALSE	1031560
TRUE	5440

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 300. This threshold was chosen to distinguish children with relatively advanced vocabulary levels. More background details and diagnostics are included in Appendix- B.

3.2 Logistic Regression Model Overview

- High Vocabulary: A binary indicator where 1 represents a high vocabulary score (combined comprehension and production > 300), and 0 otherwise.
- Scaled Age (age_scaled): The child's age, standardized to reflect changes per standard deviation. Standardization aids in interpretability and ensures numerical stability.
- 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 lexical categories, such as adjectives, verbs, and nouns. The reference category for comparison is Function Words.

The model is specified as:

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

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

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

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

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

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

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

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

Where: - p_i represents the probability that child i has a high vocabulary score - 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, β_6, β_7, β_8, β_9: 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 proficiency) and the independent variables. For instance, the standardized age variable (age_scaled) ensures that for every one standard deviation increase in age, the log-odds of achieving high vocabulary change by a constant amount. Standardization centers the variable around zero and scales it to have a standard deviation of one, improving both interpretability and adherence to the linearity assumption.
- Independence of Observations: Each observation corresponds to a unique child, and the data assumes no repeated measures or nested structures (e.g., grouping by classrooms or schools). This independence ensures the validity of the logistic regression framework. If dependencies, such as repeated measures or clustering, were present, a mixed-effects logistic regression or other hierarchical modeling techniques would be required.
- Categorical Variable Encoding: The categorical variable broad_category (e.g., "Adjectives," "Verbs," "Living Things") is encoded using sum contrasts. This ensures that each coefficient reflects the deviation of a given category from the overall mean effect across all categories. For example, the coefficient for "Verbs" represents the difference in log-odds for this category compared to the average log-odds across all categories. Sum contrasts further allow the model's intercept to represent the overall mean effect when all predictors are at their reference or mean levels, facilitating clear interpretation.
- Binary Nature of the Outcome: The dependent variable, high_vocabulary, is binary (1 = high vocabulary, 0 = not high vocabulary). Logistic regression is appropriate for modeling binary outcomes, as it assumes the binomial distribution of the data, which aligns with the structure of the outcome variable.
- No Perfect Multicollinearity: The predictors are assumed to be non-perfectly correlated. High multicollinearity would lead to unreliable coefficient estimates and obscure the individual effects of predictors. Standardizing continuous predictors (e.g., age) and using appropriate encoding for categorical variables help minimize this risk and ensure stable model estimation.

3.4 Interpretation of Coefficients

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

- Intercept (β_0): Represents the log-odds of high vocabulary proficiency when all predictors are at their reference levels. If $\beta_0 > 0$, the baseline odds of high vocabulary are greater than 50%. Refering to the modelsummary in the Section B, we knwo that the baseline odds of high vocabulary proficiency are less 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): Indicates the effect of belonging to the norming group, a positive β_2 suggests higher odds of high vocabulary compared to non-norming children. If $\beta_2 = 0.1$, then $\exp(0.1) \approx 1.11$, meaning norming group children have 22% higher odds of high vocabulary.
- Broad Category (β_3, β_4, β_5, β_6, β_7, β_8, β_9): The coefficients for the broad word categories (e.g., Function Words, Living Things) represent their effect on the log-odds of high vocabulary compared to the reference category (Adjectives). A positive β_k indicates higher odds compared to Adjectives. For example, if $\beta_4 = 0.3$ (Function Words), then $\exp(0.3) \approx 1.35$, meaning Function Words increase the odds by 35% compared to Adjectives. A negative β_k implies lower odds. For instance, if $\beta_7 = -0.008$ (Places), then $\exp(-0.008) \approx 0.992$ shows a minor decrease in odds for words in the Places category.

3.5 Model Justification

Logistic regression is well-suited for this study, given its ability to model binary outcomes like high vocabulary proficiency (1 = high vocabulary, 0 = not high vocabulary) while maintaining interpretability. By constraining predicted probabilities between 0 and 1, logistic regression ensures meaningful predictions. Its coefficients offer clear investigation into the magnitude and direction of effects, enabling an understanding of how predictors like age and norming status influence the likelihood of high vocabulary acquisition. For instance, odds ratios derived from the model allow straightforward interpretation of the impact of each variable, such as the increased likelihood of high vocabulary with a one-standard-deviation increase in age.

Although more advanced machine learning models like random forests or neural networks could provide slight improvements in predictive accuracy, these approaches lack the transparency needed to understand the underlying relationships between predictors and outcomes. Given this study's focus on uncovering developmental patterns rather than solely optimizing prediction accuracy, logistic regression offers the necessary balance between interpretability and

performance. Moreover, the dataset size and structure favor logistic regression, which is less prone to overfitting than more complex models that often require larger datasets to generalize effectively.

To ensure model robustness, the dataset was split into training and testing subsets to validate the model and minimize overfitting. Additionally, predictors like age were standardized to ensure comparability and prevent dominance by variables with larger numerical ranges. Categorical variables, such as word categories, were encoded using sum contrasts, allowing meaningful comparisons and ensuring that coefficients reflect deviations from the overall mean effect. These steps, combined with logistic regression's simplicity and explanatory power, make it an optimal choice for investigating vocabulary acquisition patterns.

4 Results

4.1 Variability in Production Vocabulary

Figure 4 visualizes the relationship between age and production vocabulary scores, focusing on different percentiles of the distribution. The scatterplot shows individual production scores as gray dots, while overlaid lines represent percentiles (10th, 25th, 50th, 75th, and 90th), capturing central tendencies and variability across ages. The 50th percentile (median) line provides a benchmark for the typical vocabulary production score at each age, whereas the 10th and 90th percentiles outline the lower and upper ranges of vocabulary development. The gradual upward trend of the median line reflects consistent growth in production vocabulary as children age, with a widening gap between the percentiles at later ages. This widening suggests increasing variability in vocabulary acquisition, with some children advancing much faster than others in production abilities.

The data indicates that children in the 90th percentile acquire vocabulary at a significantly faster rate than their peers, as evidenced by the steeper slope of the topmost line. Conversely, the 10th and 25th percentiles show more gradual, stable growth, suggesting slower development for children in these groups. The broader range of scores at older ages emphasizes the heterogeneity of developmental trajectories, with some children reaching vocabulary sizes substantially larger than the median while others remain below average. These findings underscore the diversity in early language acquisition and highlight the importance of considering individual differences when evaluating children's vocabulary development.

4.2 Median Vocabulary Size Change by Age

Figure 5 illustrates the median comprehension vocabulary scores across different ages, focusing on the central tendency of children's comprehension development between 15 and 30 months. The gray dots represent individual data points, capturing the variability in comprehension

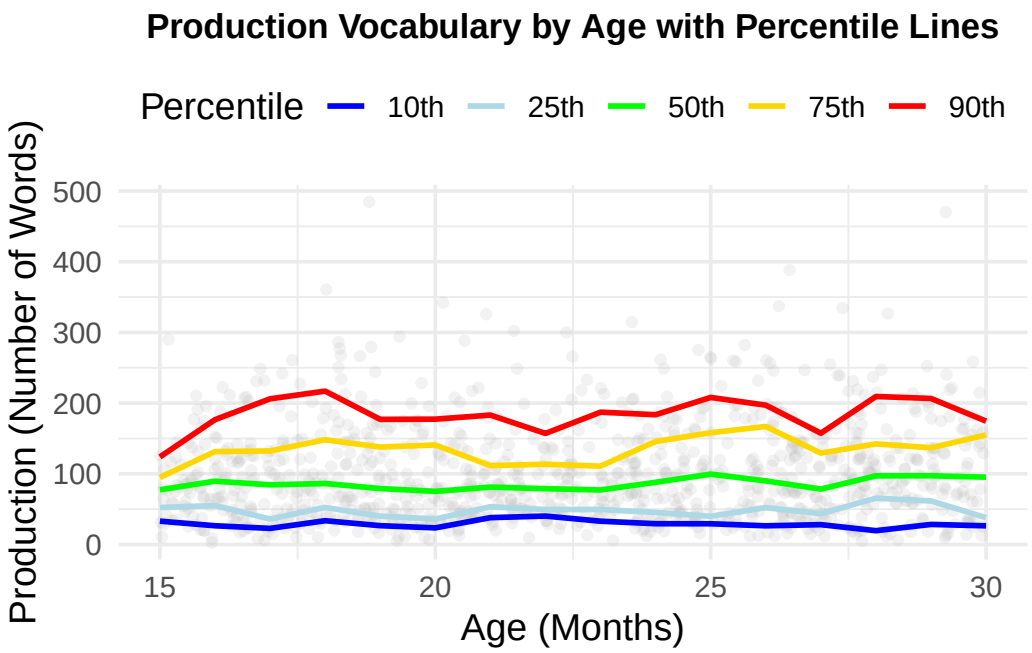


Figure 4: Production Vocabulary by Age with Percentile Lines. The graph illustrates the production scores of children across different ages (in months). Individual data points (gray dots) represent raw production scores for each participant. Colored lines correspond to standardized percentiles—10th (blue), 25th (light blue), 50th (green), 75th (yellow), and 90th (red)—showing trends in vocabulary production distribution over time.

scores, while the blue line highlights the median score for each age group. The graph shows a clear upward trajectory, with median comprehension steadily increasing with age, particularly after 18 months. This suggests a critical developmental period between 18 and 30 months during which children experience significant growth in comprehension vocabulary. The density and spread of gray points around the median line indicate individual variability, emphasizing that while the general trend is one of growth, some children exhibit slower or faster development compared to their peers. The visualization underscores the importance of age as a determinant of vocabulary comprehension while highlighting the diverse range of learning patterns among children.

Between 15 and 18 months, the median comprehension score remains relatively stable, indicating slower growth in vocabulary during early stages of language acquisition. A noticeable increase in vocabulary size is observed after 18 months, suggesting that children begin to acquire words more rapidly as their cognitive and linguistic abilities develop. The most significant growth occurs between 24 and 30 months, where the median comprehension score consistently rises. This period aligns with critical developmental milestones, such as the expansion of receptive language and comprehension skills.Around 30 months, the upward slope of the median line begins to level off slightly, suggesting that comprehension growth may slow down or stabilize as children approach the end of the observed range.

The use of the median instead of the mean ensures that the central trend is not skewed by outliers (e.g., extremely high or low comprehension scores). This choice provides a robust summary of comprehension at each age, especially in datasets with large variability or non-normal distributions.These findings highlight the critical window between 21 and 26 months for comprehension vocabulary growth. Interventions or language exposure strategies during this period may be particularly effective in enhancing language development. The observed variability suggests that individual-level factors (e.g., family environment, exposure to language) play a significant role in shaping comprehension scores, warranting further investigation into these influences.

4.3 Prediciton for the Probability of High Vocabulary Level

4.4 Predicted Probabilities by Age

Figure 7 illustrates the relationship between predicted probabilities of achieving high vocabulary and age, with percentile trends (10th, 25th, 50th, 75th, and 90th percentiles) overlaid to highlight the variability in predictions. Based on the resample of the test data, the scatterplot points represent individual predicted probabilities, while the percentile lines depict the progression of predictions across age groups.

The median predicted probability (50th percentile, solid blue line) steadily increases with age, reflecting the model's growing confidence in high vocabulary acquisition as children get

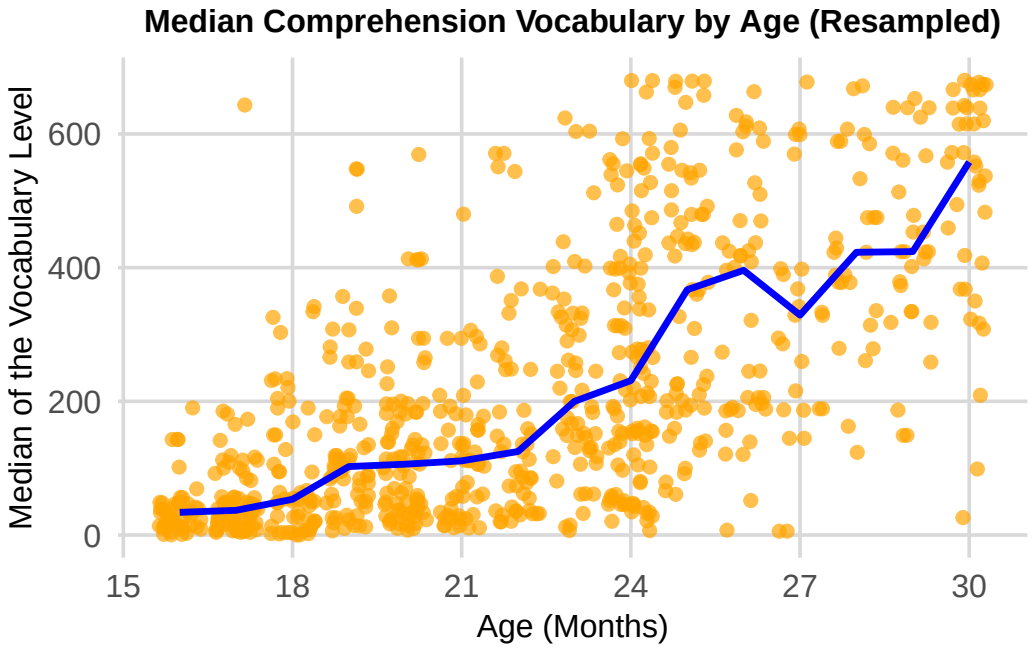


Figure 5: This figure illustrates the relationship between age (in months) and the median comprehension vocabulary level. Each orange point represents a resampled data point, while the blue line depicts the median vocabulary level at each age. The trend highlights a steady increase in vocabulary comprehension as children age, with noticeable variability across individual data points.

age	broad_category	is_norming	mean_predicted_prob
Min. :16.00	Activities :22	Mode :logical	Min. :0.02989
1st Qu.:20.00	Adjectives :22	FALSE:120	1st Qu.:0.18404
Median :23.00	Function Words:22	TRUE :56	Median :0.54167
Mean :23.27	Living Things :22	NA	Mean :0.49883
3rd Qu.:27.00	Objects :22	NA	3rd Qu.:0.78088
Max. :30.00	Places :22	NA	Max. :0.97843
NA	(Other) :44	NA	NA

Figure 6: How the probability of high vocabulary varies with age by aggregating the predictions for each age group

older. The 10th and 25th percentiles (dashed and dotted red/purple lines) remain relatively low at younger ages but show a notable rise after 20 months, suggesting greater variability in predictions among younger children. In contrast, the 75th and 90th percentiles (dashed and dotted purple/red lines) start higher and climb more sharply, indicating that some children exhibit advanced vocabulary skills even at younger ages.

Overall, the increasing spread between the percentile lines with age highlights a broader range of vocabulary acquisition patterns as children develop. This visualization underscores the model's ability to account for individual differences in learning trajectories while confirming the strong association between age and predicted probability of high vocabulary.

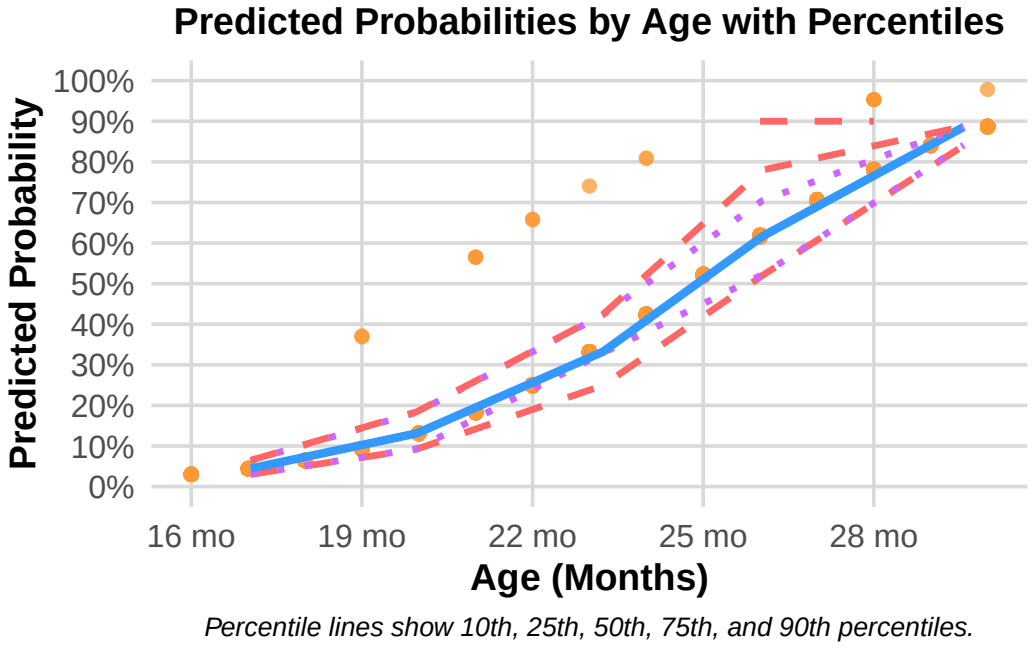


Figure 7

4.5 Distribution of Predicted Probabilities by Word Category

The distribution of predicted probabilities for achieving high vocabulary varies across word categories, as shown in Figure 8. Categories such as Function Words, Living Things, and Objects display a concentration of higher predicted probabilities, indicating consistent acquisition patterns and the foundational role of these words in early communication. In contrast, categories like Sensory Words and Verbs show a higher density of lower probabilities, reflecting their later emergence in language development and the contextual complexity they require.

Categories such as Adjectives and Places exhibit broader distributions, spanning a wider range of predicted probabilities. This variability highlights differences in children's exposure to and

use of these words, suggesting that external factors like linguistic environments may play a significant role. Overall, these results emphasize how word categories shape the model's predictions and highlight the importance of addressing variability in complex word types for targeted interventions.

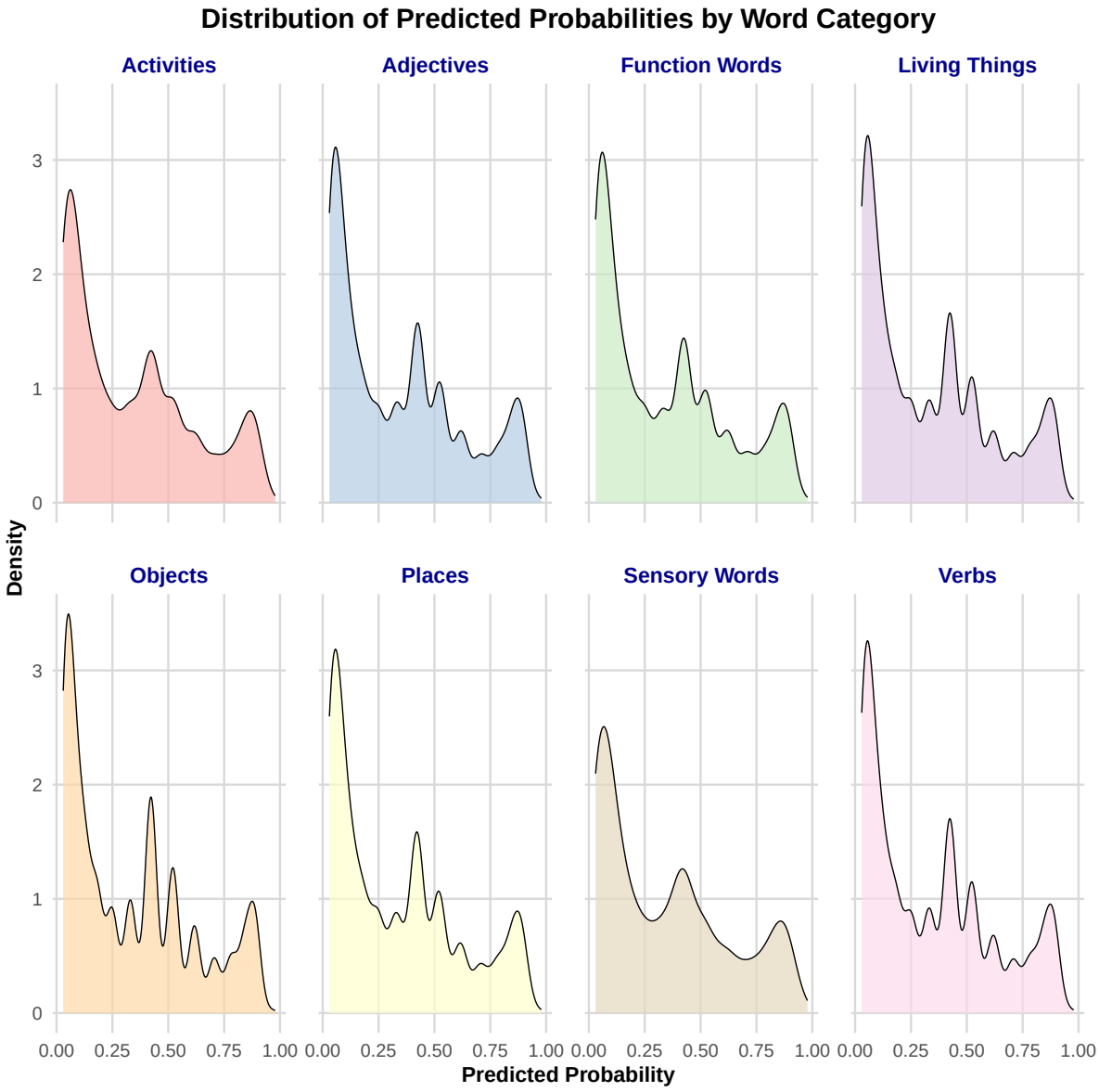


Figure 8: This figure displays the density distributions of predicted probabilities for achieving high vocabulary across different word categories. Categories such as Function Words and Living Things show concentrated higher probabilities, reflecting consistent acquisition patterns, while categories like Verbs and Sensory Words exhibit broader distributions and lower probabilities.

5 Discussion

5.1 Age and Developmental Trajectories

Age is unequivocally the strongest predictor of vocabulary acquisition, underscoring the rapid cognitive and linguistic development that occurs between 18 and 30 months. This developmental window represents a critical phase for linguistic foundation building, as evidenced by steep increases in both vocabulary production and comprehension scores. However, the variability among children's learning trajectories highlights a compelling dimension: not all children progress uniformly. Socio-economic factors, linguistic environments, and cognitive diversity significantly influence these individual paths, warranting further exploration into how specific conditions amplify or hinder vocabulary acquisition.

Interestingly, the plateau in comprehension growth observed beyond 30 months suggests a transition in linguistic acquisition dynamics. It reflects a possible shift from acquiring foundational vocabulary to mastering more nuanced, context-dependent language elements. This shift implies that while early interventions should focus on fundamental vocabulary, sustained linguistic engagement becomes crucial for later stages of development, particularly in supporting complex lexical categories.

5.2 Lexical Categories: Patterns and Variability

The distinct acquisition patterns across lexical categories reveal the interplay between cognitive accessibility and environmental exposure. Categories such as "Function Words" and "Living Things" exhibit high predictability and consistency, indicative of their essential role in early communication. In contrast, categories like "Verbs" and "Sensory Words" remain less consistently acquired, highlighting their reliance on contextual and experiential learning. These categories are often tied to abstract reasoning or specific social interactions, which tend to develop later.

This variability is a double-edged sword. While it underscores the adaptability of language acquisition processes to diverse environments, it also points to systemic gaps in early linguistic exposure. For instance, the underrepresentation of "Sensory Words" in routine interactions may delay their acquisition, underscoring the need for targeted strategies to integrate such categories into everyday linguistic exchanges.

5.3 Broader Implications for Language Development

The findings from this study have far-reaching implications for language development frameworks. First, they highlight the importance of designing tailored interventions that align with the linguistic needs of children at varying stages of development. For instance, younger children with slower vocabulary growth could benefit from high-frequency exposure to basic categories

like “Objects” and “Function Words”. Conversely, older children might require more nuanced approaches that incorporate storytelling, play-based learning, and experiential activities to enhance their grasp of complex categories such as “Verbs” and “Sensory Words”. It also suggests a way on how to develop some educational activities during the early childhood.

Furthermore, the results challenge the conventional focus of early childhood programs. They suggest a paradigm shift: instead of prioritizing sheer vocabulary size, programs should aim for a balanced acquisition across diverse lexical categories. Such an approach could bridge developmental gaps and promote linguistic versatility.

5.4 Limitations and Future Directions

Song et al. (2015) also point out in their study on 264 typically developing Chinese children (145 boys and 119 girls) were included from a longitudinal study of language. While the findings offer valuable directions, several limitations should be acknowledged. First, the reliance on parental reports introduces potential biases, including over- or underestimation of children's abilities. Future studies could complement CDI data with observational or experimental measures to enhance reliability. Second, the cross-sectional nature of the data limits the ability to track individual developmental trajectories. Longitudinal studies are needed to capture within-child variability and the dynamics of vocabulary growth over time.Familial factors and reading or language related cognitive skills were found to be associated with these developmental subgroups.

Additionally, the dataset we have now is limited on the word categories. Future data collection could focus more on the word types and provides more feasible analysis on the lexcial category learning pattern. Besides, Future research should also explore the influence of environmental and contextual factors, such as language exposure, educational interventions, and socio-economic status, on vocabulary acquisition. These factors could provide a more comprehensive understanding of the mechanisms underlying linguistic development.

The strongest developmental inferences can be made by the examination of longitudinal data, in which children's individual development is measured multiple times using the same instrument. Unfortunately, relatively little of our CDI data comes from this type of repeated administration. There is a substantial amount of two-administration longitudinal data for several languages, but only a few have more than two observations for individual children.In general, this aspect of our data is a consequence of the fact that, for normative datasets, pure cross-sectional data collection is used to ensure statistical independence between datapoints. Thus, we must typically settle for using the large amount of available cross-sectional data to average out individual variability.

A Appendix

A.1 Survey Design: The CDI Framework

The MacArthur-Bates Communicative Development Inventories (CDI) form the foundation of the Wordbank dataset used in this study. The CDI is a parent-report instrument that collects data on children's vocabulary comprehension and production through structured surveys. This methodology allows for large-scale data collection across diverse populations, balancing cost-efficiency and practicality. However, as a parent-reported tool, the CDI is subject to several biases. For instance, parental interpretations of word comprehension or usage can vary, and factors like educational background, socio-economic status, and linguistic environment may influence the accuracy of responses.

The design of the CDI ensures comprehensive coverage of early vocabulary acquisition by categorizing words into broad lexical groups (e.g., Function Words, Verbs, Sensory Words). This categorization supports nuanced analysis but introduces challenges. For instance, frequently observed word types like Function Words may be more accurately reported than abstract categories like Sensory Words, leading to potential underestimation in less salient categories. These characteristics highlight the trade-off between the CDI's extensive reach and the subjective variability in parental reporting.

A.2 Sampling Framework

The Braginsky (2024) dataset includes norming and non-norming samples, which together capture a broad spectrum of linguistic development. Norming samples are designed to represent a balanced, standardized population, serving as benchmarks for vocabulary acquisition. These data allow for cross-study comparisons and robust generalizations but may not fully capture real-world variability. In contrast, non-norming samples encompass a wider range of linguistic environments, including underrepresented groups. This dual approach helps investigate effects of contextual diversity on language acquisition but also introduces challenges in integrating findings from norming and non-norming subsets.

One critical consideration in observational datasets like Wordbank is the potential for sampling biases. For example, children from bilingual households or lower socio-economic backgrounds may be underrepresented in the norming sample, leading to an incomplete picture of vocabulary acquisition patterns. Future surveys could benefit from targeted recruitment strategies to ensure more comprehensive sampling of diverse populations.

A.3 Observational Nature of the Data

The Wordbank dataset is observational, capturing vocabulary acquisition as it naturally occurs in diverse contexts. While this design enhances ecological validity, it limits causal inference.

Factors such as parental education, linguistic exposure, and cultural norms may confound the observed relationships between predictors (e.g., age, norming status) and vocabulary outcomes. For instance, children in norming samples might have greater exposure to structured learning environments, inflating their vocabulary scores compared to non-norming counterparts.

To address these limitations, simulation methods were applied to assess how sampling variability influences model estimates. Random subsamples were drawn from the data, and the logistic regression model was refitted to evaluate the consistency of predictor effects. Key predictors like age and norming status remained stable, confirming the robustness of the model. However, coefficients for categories like Sensory Words and Places exhibited higher variability, reflecting their smaller sample representation and the challenges of capturing abstract vocabulary through parent-reported data.

A.4 Connecting to the Literature

The CDI framework and its use in Wordbank align with best practices in developmental research, as outlined in studies from Mayor and Mani (2018). The test-retest reliability of the CDI supports its validity, although the variability in parental reporting—especially for comprehension items—has been noted. Advanced statistical techniques like Item Response Theory (IRT) have been employed to evaluate the discriminative power of individual CDI items, revealing strengths in measuring commonly acquired words and limitations for abstract concepts. These findings inform the design of future surveys, emphasizing the need for complementary methods, such as direct assessments or longitudinal designs, to address the limitations of observational data.

A.5 Simulation and Recommendations

To further validate the findings, simulations could be expanded to incorporate potential sampling biases. For example, generating synthetic datasets that account for underrepresented groups (e.g., bilingual or low socio-economic households) could show how these populations might influence observed patterns. Additionally, calibration exercises—where parents undergo brief training on how to interpret survey items—could enhance the consistency of responses and reduce reporting variability.

This appendix underscores the importance of survey design, sampling methodology, and observational data analysis in understanding early vocabulary acquisition. By addressing inherent limitations and leveraging advanced methodologies, future research can refine the robustness and applicability of findings, ensuring a more inclusive representation of children's linguistic development.

B Model details

B.1 Model Summary

B.2 Diagnostics

B.2.1 Confusion Matrix

Metric	Value
Accuracy	0.77
Sensitivity (Recall)	0.59
Specificity	0.86
Precision	0.67

B.2.2 ROC Curve and AUC

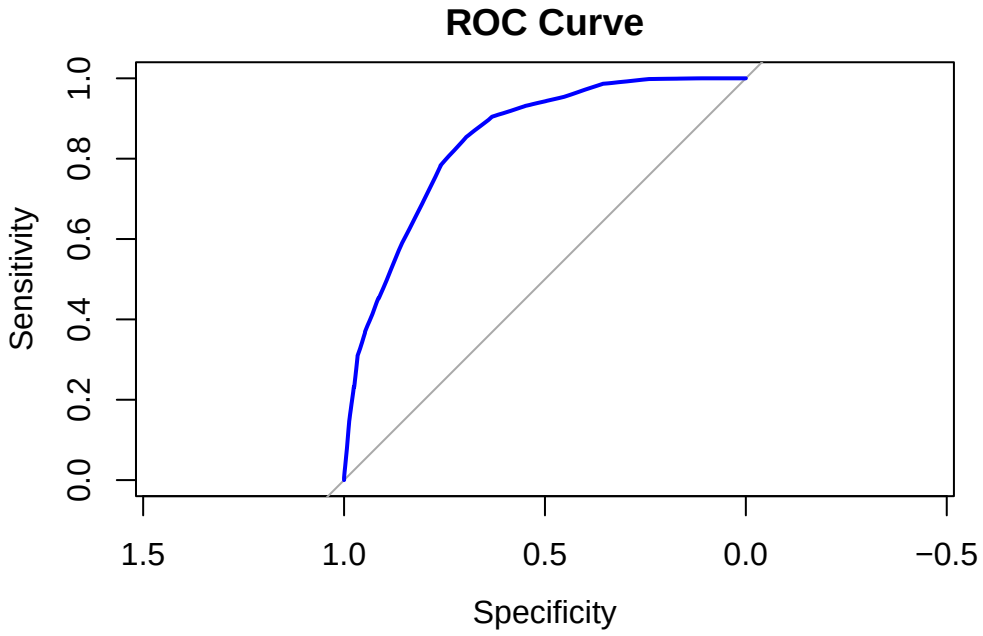
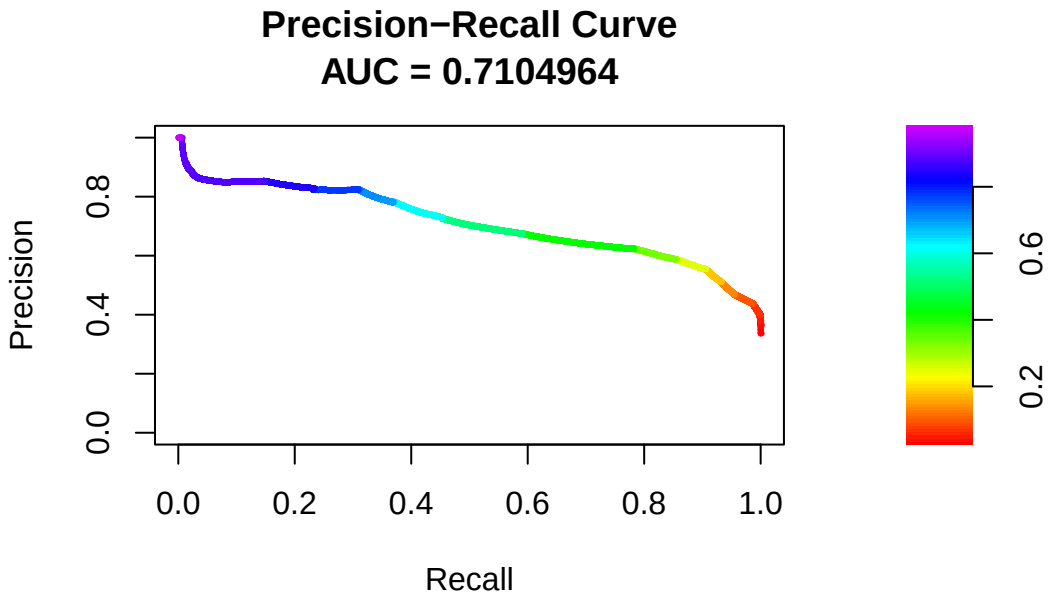


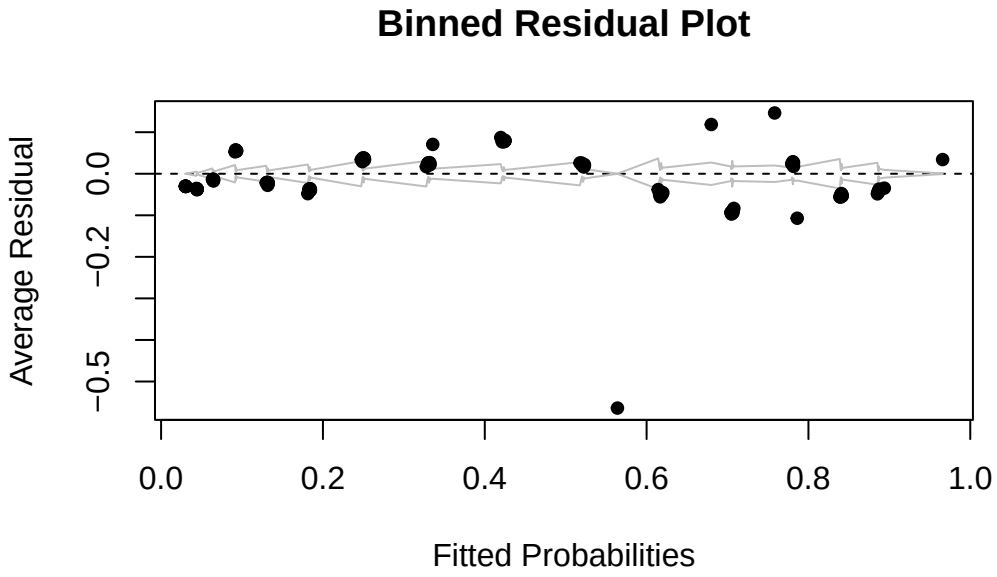
Table 4: This table provides coefficients summary for the logistic model used in this paper.

	(1)
(Intercept)	−1.014
	(0.015)
age_scaled	1.605
	(0.004)
is_normingTRUE	1.749
	(0.040)
broad_categoryAdjectives	0.006
	(0.017)
broad_categoryFunction Words	0.005
	(0.018)
broad_categoryLiving Things	0.005
	(0.017)
broad_categoryObjects	−0.005
	(0.016)
broad_categoryPlaces	−0.008
	(0.017)
broad_categorySensory Words	−0.015
	(0.026)
broad_categoryVerbs	0.004
	(0.016)
Num.Obs.	829 600
AIC	754 890.6
BIC	755 006.9
Log.Lik.	−377 435.316
RMSE	0.39

B.2.3 Precision-Recall Curve



B.2.4 Binned Residual Plot



C Acknowledgements

This project was conducted under the help of OpenAI's ChatGPT 4.0, which provided invaluable assistance in drafting and refining the paper. The analysis was conducted using a suite of packages from R Core Team (2023), which offered robust functionality for data manipulation, visualization, and storage. We extend our gratitude to the teams behind the Wickham et al. (2019), Wickham (2016), Wickham, Pedersen, and Seidel (2023), Wickham et al. (2023), Wickham, Vaughan, and Girlich (2024), Gelman and Su (2024), Arel-Bundock (2022) and Xie (2024) packages, whose tools were instrumental in streamlining the data cleaning, analysis, and graphing processes. Additionally, Richardson et al. (2024) played a critical role in efficient data handling and storage through Parquet files. Thanks to Robin et al. (2011) and Keilwagen, Grosse, and Grau (2014) for supporting model diagnostics in this study.

A special acknowledgment goes to the Braginsky (2024) team for providing the extensive dataset that forms the foundation of this research. Their contribution enabled a comprehensive exploration of vocabulary learning patterns in children. We are deeply grateful to the developers and maintainers of these open-source tools and datasets for their efforts in advancing research and accessibility in the data science community.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Braginsky, Mika. 2024. *wordbankr: Accessing the Wordbank Database*. https://github.com/langcog/wordbankr.

Gelman, Andrew, and Yu-Sung Su. 2024. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. https://CRAN.R-project.org/package=arm.

Keilwagen, Jens, Ivo Grosse, and Jan Grau. 2014. "Area Under Precision-Recall Curves for Weighted and Unweighted Data." *PLOS ONE* 9 (3). https://doi.org/10.1371/journal.pone.0092209.

Mayor, Julien, and Nivedita Mani. 2018. "A Short Version of the MacArthur–Bates Communicative Development Inventories with High Validity." *Behavior Research Methods* 51 (October). https://doi.org/10.3758/s13428-018-1146-0.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: An Open-Source Package for r and s+ to Analyze and Compare ROC Curves." *BMC Bioinformatics* 12: 77.

Song, Suiping, Mengmeng Su, Chunyan Kang, Hongyun Liu, Yan Zhang, Catherine McBride-Chang, Twila Tardif, et al. 2015. "Tracing Children's Vocabulary Development from Preschool Through the School-Age Years: An 8-Year Longitudinal Study." *Developmental Science* 18 (1): 119–31. https://doi.org/10.1111/desc.12190.

Taintor, Sue, and Wendy LaMarr. 2023. "Charting Language Growth in Infants and Toddlers." LibreTexts. https://socialsci.libretexts.org/Bookshelves/Early_Childhood_Education/Infant_and_Toddler_Care_and_Development_%28Taintor_and_LaMarr%29/11%3A_Overview_of_Language_Development/11.09%3A_Charting_language_growth_in_infants_and_toddlers.

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 François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *tidyr: Tidy Messy Data*.

https://CRAN.R-project.org/package=tidyr.
Xie, Yihui. 2024. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.