Children English Vocabulary Learning Pattern*

My subtitle if needed

Yongqi Liu

November 23, 2024

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

Table of contents

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

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

The remainder of this paper is structured as follows. Section 2 descirbe the dataset in detail, and shows the distribution of variables. The data used in this study 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	0
English (American)	25	FALSE	Sensory Words	580	1
English (American)	22	FALSE	Sensory Words	351	1
English (American)	24	FALSE	Sensory Words	310	0
English (American)	25	FALSE	Sensory Words	257	0
English (American)	26	FALSE	Sensory Words	188	0

2.2 Measurement

2.2.1 Data Collection and Quality Control

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. - Structured Response Formats: The CDI employs predefined response options for comprehension and production, reducing ambiguity and enhancing consistency in parental reporting. - 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 Comparison: The inclusion of norming groups as benchmarks helps to ensure the validity of reported vocabulary scores and allows for

cross-child comparison. These groups provide a standardized reference for analyzing individual differences in language acquisition. - 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: By structuring responses and including norming benchmarks, the CDI minimizes some of the biases inherent in self-reported data, such as over- or underestimation by parents. - Missing Data Handling: Observations with incomplete or invalid responses were excluded from the analysis to maintain the integrity and reliability of the dataset.

2.2.2 Reporting Bias

However, there are several considerations regarding the data collection process:

- Parental Reporting Bias: The reliance on parental reports introduces the potential for bias, including overestimation or underestimation of a child's abilities. This is inherent to self-reported data and can affect the accuracy of the results.
- 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.
- 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.
- 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.

Despite its standardized structure, the CDI is subject to biases inherent in parental reporting, including over or underestimation bias. Parents may unintentionally overestimate or underestimate their child's skills due to subjective perceptions or limited observations. 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 = $\frac{\text{Comprehension}+\text{Production}}{2}$

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

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 insights into advanced vocabulary development, it also necessitates caution in generalizing

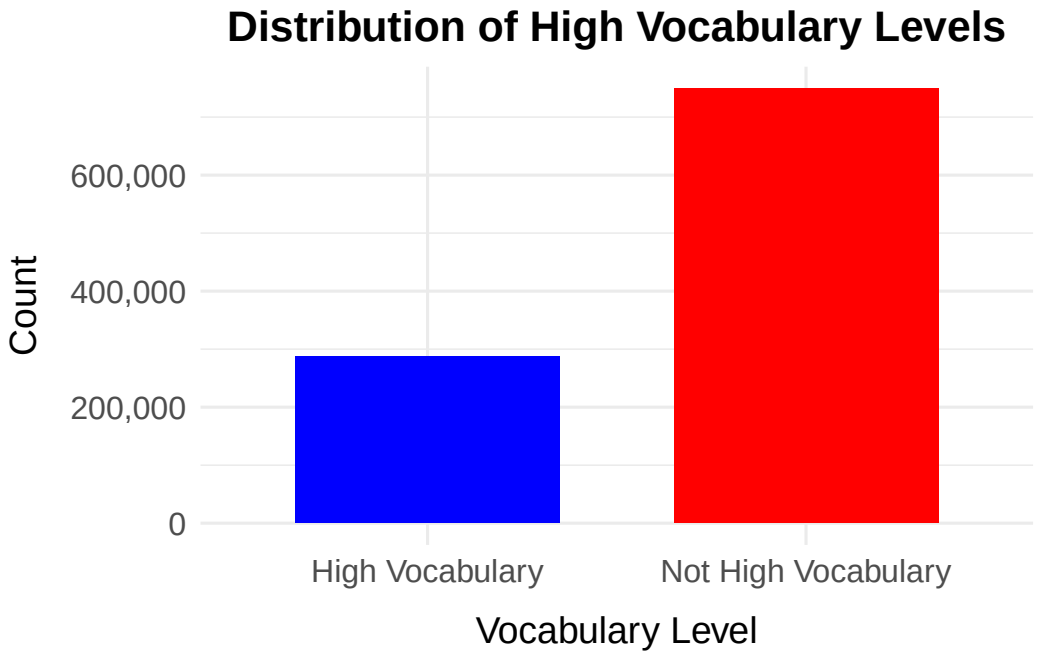


Figure 1: Distribution of the outcome variable, showing the counts of children classified as having "High Vocabulary" and "Not High Vocabulary" based on their comprehension and production scores. The bar plot illustrates the balance between the two categories in the dataset, which is important for modeling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

findings to underrepresented age groups. This observation emphasizes the need to standardize age in statistical models to account for variability across different age groups.

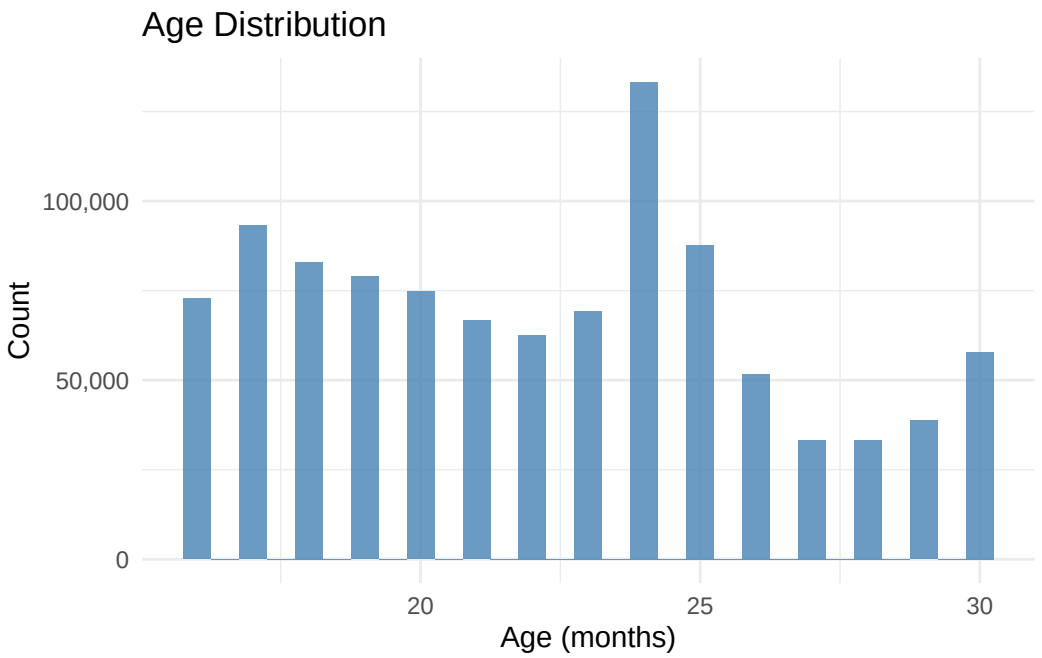


Figure 2: It shows the distribution of children's ages (in months) within the dataset. The majority of observations fall between 24 and 30 months, with noticeable peaks at 25 and 30 months, reflecting a focus on key developmental periods. Younger age groups are underrepresented, highlighting the need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

2.4.2 Broad 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.

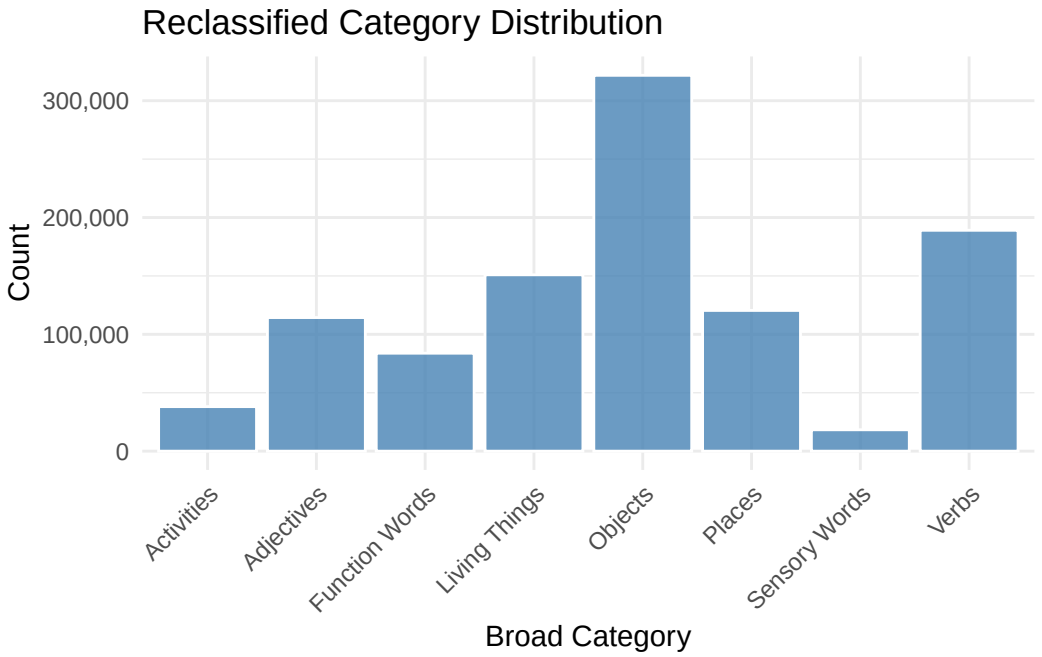


Figure 3: The figure shows objects dominate the vocabulary, reflecting an emphasis on concrete and tangible terms, while categories like Sensory Words and Activities are less frequently represented, indicating the relative complexity or specificity of these word types in early language acquisition

2.4.3 Norming Status

Norming status categorizes children into two groups: those included in the norming group and those who are not. The norming group represents a standardized sample used as a benchmark for assessing vocabulary development, providing a reference point for evaluating other children in the dataset. This distinction is important for ensuring the validity and reliability of comparisons in the analysis.

The bar plot illustrates the distribution of children based on their norming status. The dataset is predominantly composed of non-norming children, with only a small fraction representing the norming group. This imbalance reflects the practical challenges of including a broad and representative norming sample in large-scale vocabulary assessments. While the non-norming group provides diverse data for analysis, the norming group offers a crucial baseline for calibrating and interpreting the results. The distribution highlights the need to account for this imbalance in the analysis to avoid potential biases and ensure robust conclusions.

Table 2: Summary of Norming Status in the Dataset

Norming Status	Count
FALSE	1031560
TRUE	5440

Figure 4: 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

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- B.

3.2 Logistic Regression Model Overview

- High Vocabulary: A binary indicator where 1 represents a high vocabulary score (combined comprehension and production > 350), 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, 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").
  - 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. 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

Logistic regression is widely used in predictive modeling for categorical outcomes due to its simplicity and robustness. It provides probabilities that are constrained between 0 and 1, ensuring meaningful interpretations for binary outcomes. Unlike other complex models, logistic regression allows for clear coefficient interpretation, offering insights into the magnitude and direction of predictor effects. For example, the odds ratios derived from logistic regression help explain how changes in variables like age or norming status influence the probability of high vocabulary acquisition.

Although advanced machine learning models like decision trees, random forests, or neural networks could be used, these methods often lack the interpretability of logistic regression. While these models might yield slightly better predictive performance, they are often considered "black boxes," making it difficult to identify specific relationships between predictors and outcomes. Given the study's focus on understanding developmental patterns rather than maximizing predictive accuracy, logistic regression is more appropriate.

Moreover, complex models require larger datasets to avoid overfitting and ensure generalizability, which might not be feasible given the sample size and the structure of the data in this study. Logistic regression strikes a balance between simplicity, interpretability, and predictive performance.

The dataset was split into training and testing subsets to ensure model validation and reduce overfitting. This allows the model to generalize better to unseen data, providing a more reliable assessment of its predictive accuracy. Standardizing continuous predictors, such as age, enhances interpretability and ensures that variables are on a comparable scale, preventing dominance by predictors with larger numerical ranges.

To ensure reproducibility, the preprocessed datasets and the trained model were saved. This practice facilitates verification of results and supports future analyses or extensions of the study. Overall, the logistic regression model offers a clear and interpretable framework for investigating vocabulary acquisition patterns, providing both explanatory power and practical insights.

4 Results

4.1 Variability in Production Vocabulary

Figure 5 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.

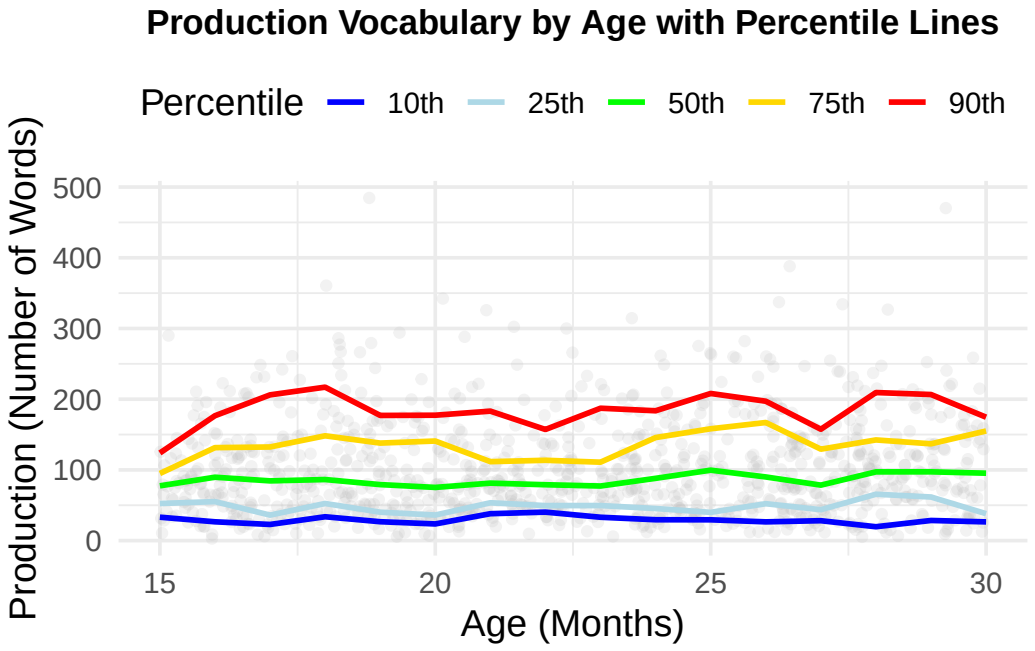


Figure 5: Production Vocabulary by Age with Percentile Lines. The graph illustrates the production scores of children across different ages (in months). Individual data points (gray dots) represent raw production scores for each participant. Colored lines correspond to standardized percentiles—10th (blue), 25th (light blue), 50th (green), 75th (yellow), and 90th (red)—showing trends in vocabulary production distribution over time.

4.2 Vocabulary Size Change by Age

Figure 6 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 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.

Steady Growth in Comprehension Vocabulary: 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.The distribution of points (grey scatter) highlights considerable variability in comprehension scores at each age. While the median line captures the central trend, some children show significantly higher or lower comprehension compared to their peers, reflecting individual differences in language learning rates.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

```
      Predicted
Actual      0      1
```

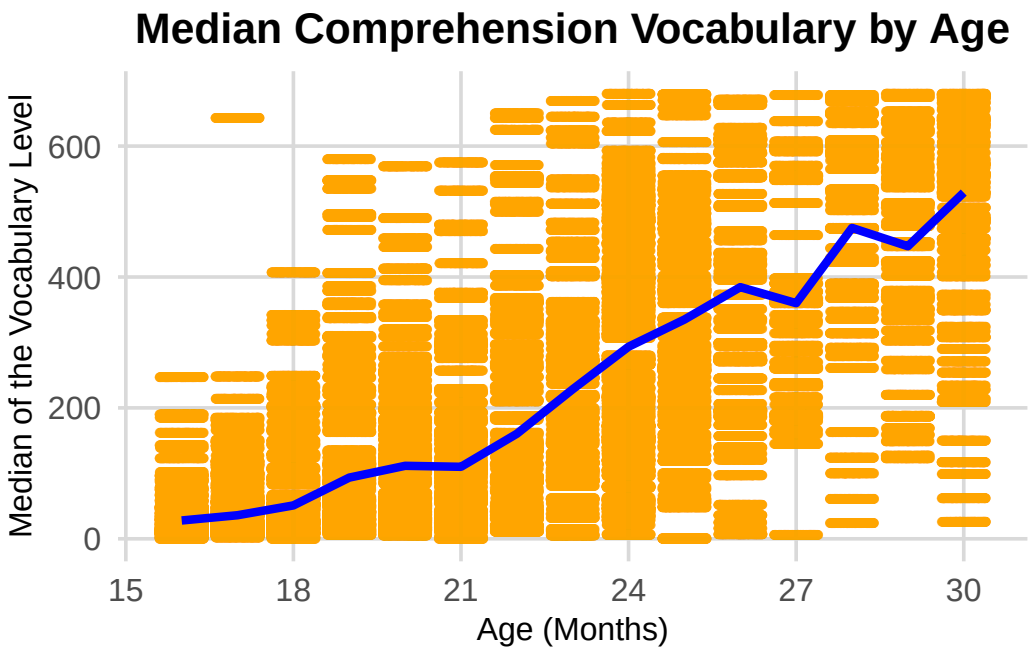


Figure 6: Illustrating central tendencies in comprehension vocabulary scores

```
  0 135148  14670
  1  29351  28231
```

```
[1] "Accuracy: 0.78774831243973"
```

4.4 Distribution of Predicted Probabilities by Broad Category

The Figure 7 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.

Each facet represents a unique category, with the x-axis denoting the predicted probability of high vocabulary and the y-axis representing density.

The plots highlight distinct patterns in the distributions of predicted probabilities for each category. Categories such as Adjectives and Function Words exhibit broader distributions, suggesting greater variability in the likelihood of achieving a high vocabulary score within these categories. In contrast, categories like Sensory Words and Activities show relatively concentrated distributions near lower probabilities, indicating more consistent outcomes within those lexical groups.

The graph underscores the heterogeneity of vocabulary acquisition probabilities across different lexical domains. Categories associated with more concrete or commonly encountered words, such as Objects and Living Things, tend to exhibit higher probabilities for vocabulary mastery, reflecting their early and frequent usage in language development. Conversely, categories like Sensory Words show lower probabilities, suggesting that they are acquired less consistently or later in development. This visualization offers insight into the differential patterns of lexical acquisition, emphasizing the varied developmental trajectories across word types.

5 Discussion

5.1 Vocabulary Development Patterns

The acceleration in early vocabulary is even clearer when looking at production reports from older children using Words & Sentences. Figure 5.3 shows this pattern. In every language, the median child is reported to produce 50 words between 16–20 months (dotted line), though – as we will see below – this analysis masks tremendous between-child variability during this period. In addition, languages vary considerably in the absolute number of words reported. (As it is a major outlier, we have discussed the Beijing Mandarin WS data in Chapter 3, section on difficult data). Nevertheless, there are still substantial consistencies in the shape and general numerical range across languages.

During the period of 24–30 months, we see curves leveling out. Presumably, this leveling does not reflect a slowing in the rate of acquisition, which most researchers assume continues unabated for many years (e.g., Bloom, Tinker, and Scholnick 2001). Instead, it reflects the limitations of the CDI instrument, in that there are many possible "more advanced" words that children are likely learning, of which only a small subset are represented on any form.

The results highlight distinct trajectories in vocabulary development among children. Production scores showed a steady increase with age, with notable accelerations between 18 and 30 months. This period is critical for vocabulary acquisition, as children rapidly expand their linguistic capabilities. The variability observed in percentile distributions emphasizes individual differences in learning rates, influenced by factors such as exposure to language, socio-economic background, and cognitive abilities.

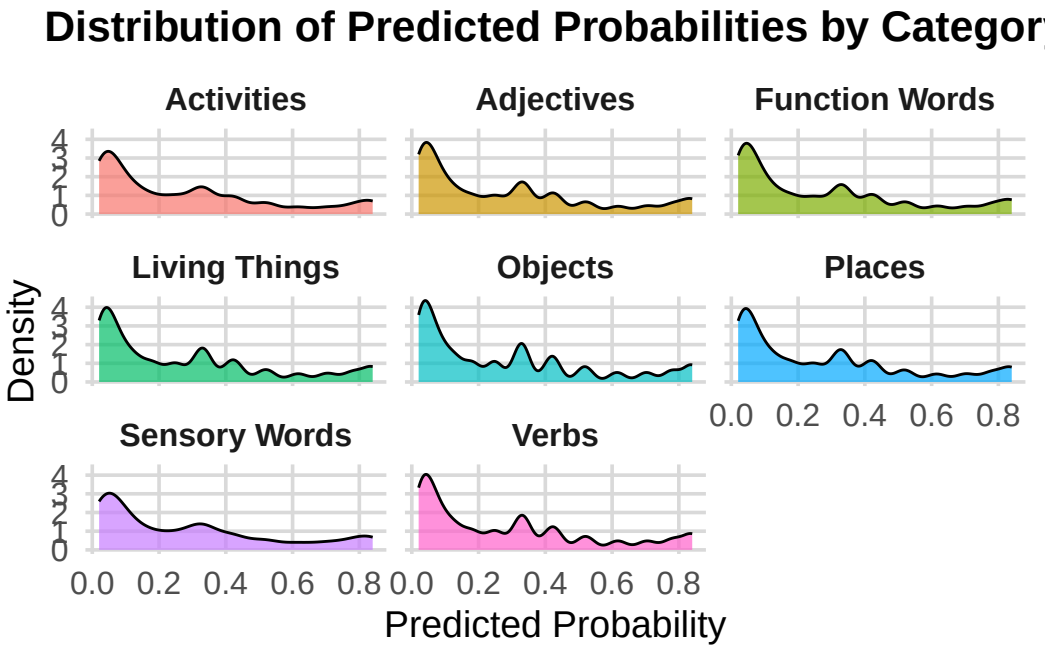


Figure 7: This density plot illustrates the predicted probabilities of having a high vocabulary across different broad lexical categories. Each curve represents the density of predicted probabilities within a category, showcasing the variation in predicted outcomes for categories such as Activities, 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.

The analysis of comprehension data revealed a similar trend, with a marked rise in vocabulary comprehension after 18 months. Median comprehension scores provided a robust central tendency, effectively avoiding skewness caused by outliers. The plateau observed around 30 months suggests a stabilization in vocabulary growth, likely reflecting the limitations of the CDI dataset rather than a developmental plateau.

5.2 Role of Predictors in Vocabulary Growth

Key predictors, including age, norming status, and broad lexical categories, significantly influenced vocabulary acquisition. Age, as expected, showed the strongest correlation, underlining its role as a primary determinant of language development. Children from the norming group exhibited higher vocabulary scores, possibly due to more structured linguistic environments or assessment settings.

Lexical categories revealed nuanced contributions to vocabulary development. Categories such as objects and verbs were associated with higher predicted probabilities of achieving advanced vocabulary levels, reflecting their functional importance in early communication. Conversely, categories like sensory words and adjectives showed less consistent trends, indicating their later emergence in the linguistic repertoire.

5.3 Third discussion point

5.4 Limitations and Future Directions

While the findings offer valuable insights, 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.

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. Figure 3.5 shows the number of administrations for particular languages that come from longitudinal datasets with a particular depth. 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 Additional data details

A.2 Data Sheet

B Model details

B.1 Model Summary

B.2 Diagnostics

B.2.1 Accuracy

```
analysis_data_train <- read_parquet(here::here("data/02-analysis_data/train_data.parquet"))
# Predicted probabilities and classes
predicted_classes <- ifelse(fitted(logistic_model) > 0.5, 1, 0)
confusion_matrix <- table(Predicted = predicted_classes, Actual = analysis_data_train$high_v

# Print confusion matrix and accuracy
print(confusion_matrix)
```

```
         Actual
Predicted      0      1
        0 540092 117529
        1  59450 112529
```

```
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("Accuracy:", round(accuracy, 3)))
```

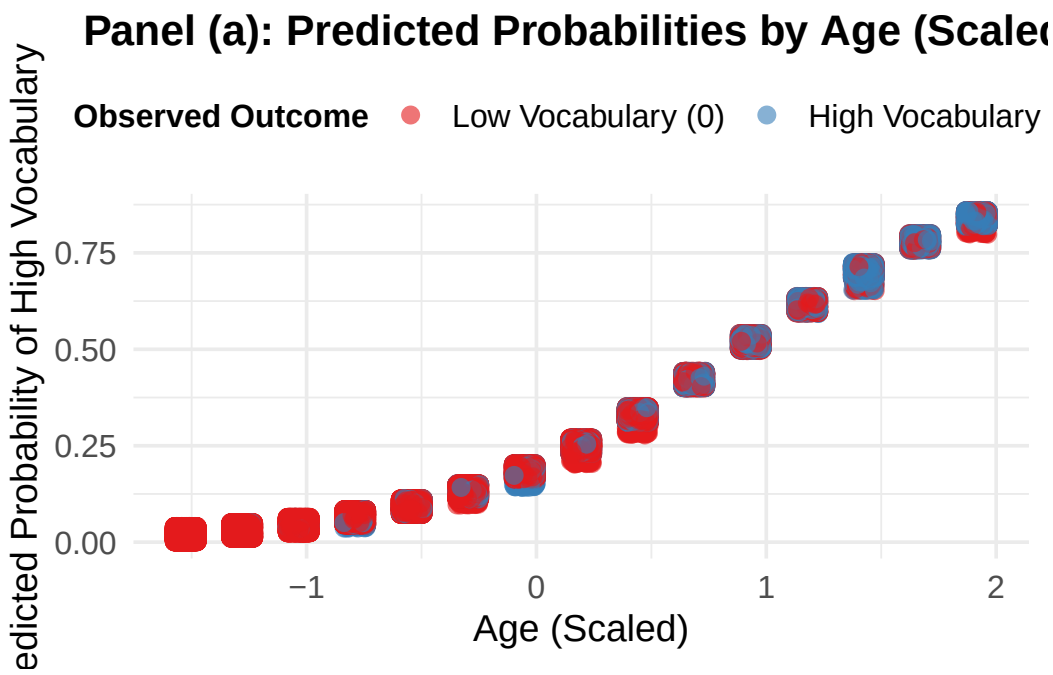
```
[1] "Accuracy: 0.787"
```

```
# Predict probabilities using the logistic model
analysis_data_test <- analysis_data_test %>%
  mutate(predicted_prob = predict(logistic_model, newdata = ., type = "response"))

# Prepare dataset for visualization
analysis_data_test <- analysis_data_test %>%
```

Table 3: model summary table

	(1)
(Intercept)	−1.417
	(0.016)
age_scaled	1.601
	(0.004)
is_normingTRUE	−0.144
	(0.038)
broad_categoryAdjectives	0.009
	(0.018)
broad_categoryFunction Words	0.005
	(0.019)
broad_categoryLiving Things	0.005
	(0.017)
broad_categoryObjects	−0.003
	(0.016)
broad_categoryPlaces	−0.003
	(0.018)
broad_categorySensory Words	−0.010
	(0.027)
broad_categoryVerbs	0.006
	(0.017)
Num.Obs.	829 600
AIC	703 694.5
BIC	703 810.8
Log.Lik.	−351 837.240
RMSE	0.37

```r
  mutate(high_vocabulary = factor(high_vocabulary, labels = c("0", "1")))

# Improved Panel (a): Scatterplot of Predicted Probabilities
ggplot(analysis_data_test, aes(x = age_scaled, y = predicted_prob, color = high_vocabulary)) +
  geom_point(alpha = 0.6, size = 3, shape = 16, position = position_jitter(width = 0.05, hei
  labs(
    title = "Panel (a): Predicted Probabilities by Age (Scaled)",
    x = "Age (Scaled)",
    y = "Predicted Probability of High Vocabulary",
    color = "Observed Outcome"
  ) +
  scale_color_manual(values = c("0" = "#E41A1C", "1" = "#377EB8"),
                     labels = c("Low Vocabulary (0)", "High Vocabulary (1)")) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    legend.position = "top",
    legend.title = element_text(size = 12, face = "bold"),
    legend.text = element_text(size = 12)
  )
```



B.2.2 Binned Residual Plot

```
#install.packages("arm")
library(arm)
```

```
Loading required package: MASS
```

```
Attaching package: 'MASS'
```

```
The following object is masked from 'package:dplyr':

    select
```

```
Loading required package: Matrix
```

```
Attaching package: 'Matrix'
```

```
The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```
Loading required package: lme4
```

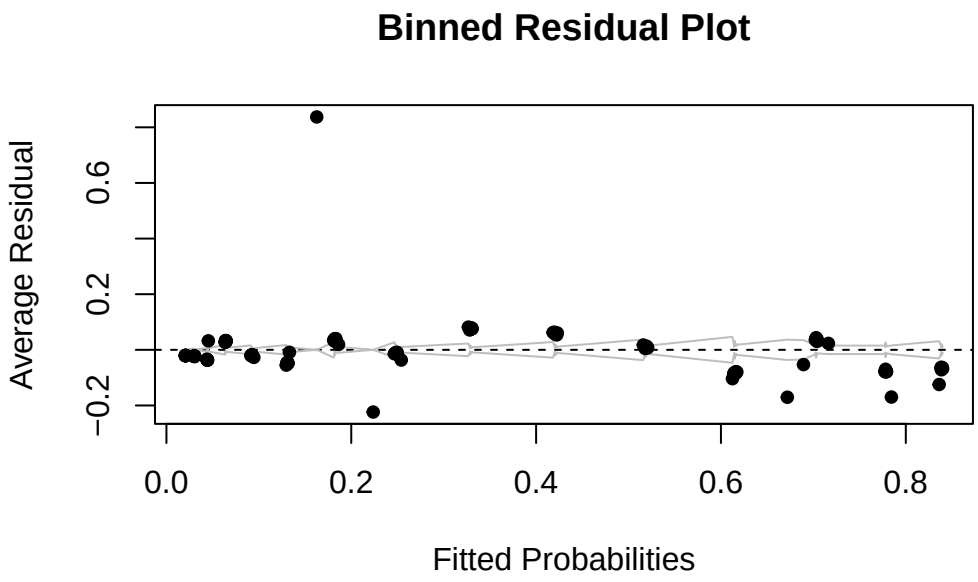
```
arm (Version 1.14-4, built: 2024-4-1)
```

```
Working directory is /Users/cassieliu/Desktop/Vocabulary_Learning_Pattern/paper
```

```
Attaching package: 'arm'
```

```
The following object is masked from 'package:scales':

    rescale
```

```r
binnedplot(fitted(logistic_model), residuals(logistic_model, type = "response"),
           xlab = "Fitted Probabilities", ylab = "Average Residual",
           main = "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), 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.

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.

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.

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.

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