Factors Influencing Early Childhood Comprehension of American English with a Bayesian Model*

Word production, child age, norming status and caregiver education have key effects on early language comprehension ability

Ziyuan Shen

December 3, 2024

In this paper, we develop a Bayesian linear model to analyze the factors that influence early childhood comprehension of American English from age 8 to 30 months. Using a dataset from the Wordbank database, our model examines predictors such as age, word production, caregiver education, and birth order to determine their impact on early childhood language comprehension skills. The analysis shows that children with more highly educated caregivers tend to have better language comprehension, while first-born children generally have higher comprehension compared to their siblings. This approach provides a detailed understanding of how developmental and environmental factors shape early childhood language abilities, offering a basis for designing targeted interventions to promote language development.

Table of contents

*Code and data are available at: Factors Influencing Early Childhood Comprehension of American English.

1 Introduction

Early childhood language comprehension is a fundamental area of cognitive development that influences future literacy and communication skills. Research has identified several key predictors of language comprehension in young children. For example, a study by Taylor, found that socioeconomic factors, family background, and personal characteristics significantly influence early childhood receptive vocabulary development (Zubrick, Taylor, and Christensen 2015). As highlighted by Lawton, the quality and quantity of language exposure in infancy are associated with brain development and long-term language achievement(Lawton, Araujo, and Kufaishi 2023). In addition, Bergelson's study showed that babies begin to understand common nouns as early as 6 to 7 months of age, which suggests that comprehension begins earlier than previously thought(Zonarich 2024). Despite these observations, we believe there is still a need for robust research that considers the range of demographic, developmental, and environmental factors that influence the early comprehension of American English learners.

In this paper, my estimand is the number of words comprehended by children learning American English. Our focus is on quantifying how developmental factors such as age, birth order, caregiver influences such as education level, and language context such as monolingual versus bilingual status affect comprehension outcomes. To analyze these relationships, we used Bayesian linear regression models. We initially built a full model and a reduced model to assess the contribution of these predictors. Upon comparison, we found that the reduced model retained the most important predictors, providing a more concise and easier-to-interpret result while maintaining a high level of predictive accuracy. Therefore, we chose the reduced model for our final analysis because it effectively captured the key factors that influence early childhood comprehension.

My findings show that developmental factors such as age, production ability of language, and caregiver education are important predictors of young children's level of comprehension. Children with more educated caregivers demonstrated better comprehension, which highlights the role of the home environment in language development. These findings are important because they provide a basis for carefully targeted early childhood interventions, which highlight the areas that need to be supported in order to promote language comprehension. By identifying the most influential factors, this study contributes to a better understanding of early language learning, which is essential for the development of effective educational practices and policies.

The structure of this paper is organized as follows: after this introduction, Section 2 details the dataset used, including the data collection and cleaning processes, and provides an overview of the key variables. Section 3 introduces the regression models employed in the analysis and discusses why these models are suitable for estimating comprehension outcomes. Section 4 presents the results, emphasizing the relationships between different predictors and comprehension. Finally, Section 5 concludes the paper by evaluating the implications of the findings, discussing their importance for educational policy, and outlining potential directions for future research.

2 Data

2.1 Data Overview

We used the statistical programming language R (R Core Team 2023) to retrieve, simulate, clean, analyze, and test early childhood language comprehension data. The dataset used for this analysis is from the Wordbank database of children's vocabulary growth (M. et al. Frank Braginsky, n.d.), specifically focusing on American English children. It contains comprehension and a variety of predictors, which include developmental, demographic, and environmental factors. Following the methodology discussed in "Telling Stories with Data" (Alexander 2023), we discover how different predictors affect language comprehension in early childhood.

The dataset used in this study provides different data on the language comprehension skills of young children among American English language learners. The dataset consists of 14,826 rows and 22 columns covering a range of early childhood developmental attributes such as comprehension, output and parental education. To ensure the validity of the analysis, we filtered the data to include children whose comprehension assessments and caregiver information were complete, which would ensure a general, unbiased representation of the target population. This cleaning allowed us to focus on high-quality, well-integrated cases, which provided reliable results in young children's comprehension skills. For key operations, please refer to Appendix B.1 for detailed data processing steps.

2.2 Data Measurement

The Early Childhood Language Development Assessment Measurement Data Set was collected using a standardized instrument such as the Communicative Development Inventory (CDI). Specifically, comprehension and vocabulary data are obtained through parent reports. These CDIs are survey instruments provided to parents or caregivers that allow them to document their child's language skills at specific developmental stages.

For example, let your child learn the meaning of different objects at home. This process naturally occurs through interaction with caregivers and is translated into quantifiable data

when parents marked items on the CDI checklist that their child understood or said. The resulting scores become items in our dataset, which reflect each child's language comprehension skills in numerical form.

In the data cleaning process, we prepared the raw election data for analysis by applying transformations, filtering, and restructuring using several R packages, including `tidyverse`(Wickham et al. 2019), `dplyr` (Wickham et al. 2023), `arrow` (Richardson et al. 2024), and `rsample` (Frick et al. 2024). Also, the following packages were used in this study:

- `here`(Müller 2020): Helps manage file paths in a project directory, which make it easier to locate and load data files.
- `ggplot2` (*ggplot2: Elegant Graphics for Data Analysi*, n.d.): Used for creating data visualizations like charts and graphs.
- `modelsummary`(Arel-Bundock 2022): Summarizes model outputs in tables that are easy to read.
- `rstanarm`(Goodrich et al. 2022): Fits Bayesian linear regression models using the `Stan` framework.
- `knitr`(Xie 2014): Converts R code and markdown text into reports, documents, or presentations.
- `kableExtra`(Zhu 2024): Beautify tables made with `knitr::kable` by adding formatting options.
- `ggridges`(Wilke 2024): Creates ridgeline plots, which are useful for visualizing distributions.
- `car`(Fox and Weisberg 2019): Provides tools for regression diagnostics and more accuracy linear model analysis.
- `broom.mixed`(Bolker and Robinson 2024): Provides functions to convert statistical model outputs, particularly from mixed-effects models into tidy data frames.

2.3 Outcome variables

2.3.1 Comprehension(Number of Words Understand By Children)

The primary outcome variable was the child's comprehension of American English, measured by the number of words the child understood. These data include responses from children of different ages (months) and allow for age-specific and general measures of language comprehension in early development.

Using the package ggplot2 (*ggplot2: Elegant Graphics for Data Analysi*, n.d.), Figure 1 shows the distribution of comprehension, which is the number of words children understand in early childhood. Each bar in the histogram represents the count of children who comprehend a specific range of words.

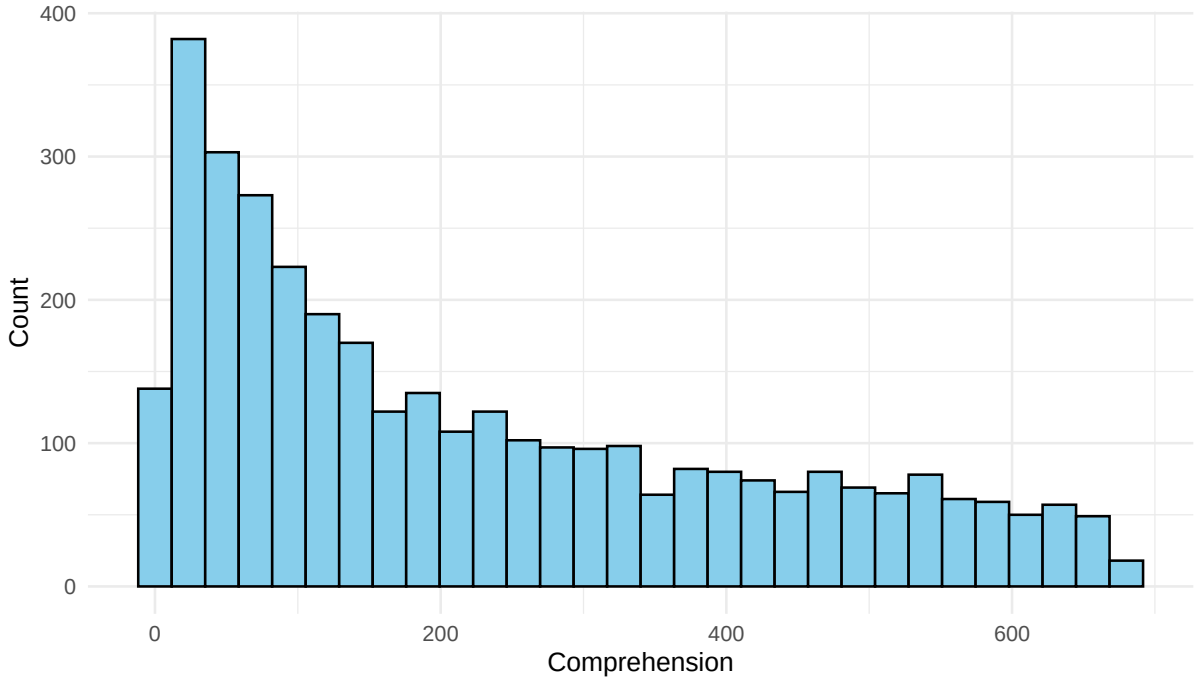


Figure 1: Distribution of comprehension among children. Each bar represents the count of children who understand a specific range of words, illustrating the varying levels of vocabulary comprehension in the dataset.

The histogram shows a right-skewed distribution, with a higher frequency of children having fewer understood words. The count gradually decreases as the number of comprehended words increases, suggesting that most children in the dataset have relatively lower comprehension levels, while fewer children understand a larger vocabulary.

2.4 Predictor variables

There are eight predictor variables for my study after cleaning. The Descriptions of Each Predictor Variable is in Table 3 shows at Appendix B.2.

2.4.1 Age

The `age` variable indicates the age (in months) of each child. This variable is important for understanding how comprehension develops over time in early childhood. By using age in months, we can capture growth on a more granular level rather than grouping children into broader age categories, which may overlook subtle developmental differences. Age is an important factor because it is directly related to a child's cognitive and language development, which can affect their ability to understand and produce words.

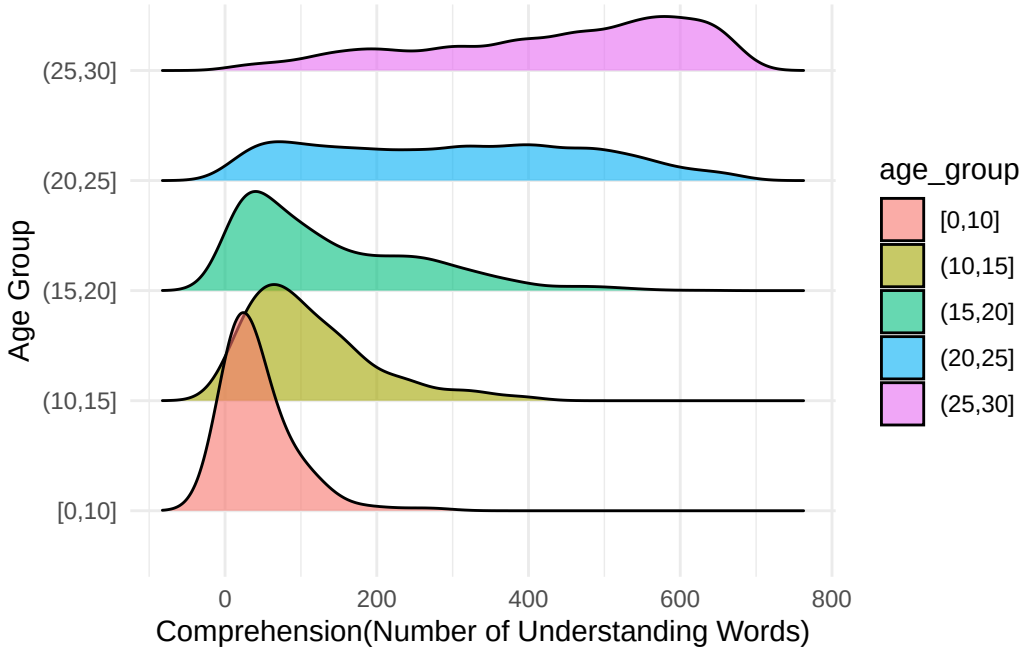


Figure 2: Ridgeline Plot of Comprehension Scores Across Different Age Groups. This plot illustrates the distribution of comprehension scores for each age group, highlighting how comprehension changes across age intervals.

Figure 2 presents a ridgeline plot that illustrates the distribution of comprehension across different age groups. The data is segmented into different age groups to analyze how comprehension levels vary across different stages of early childhood. Each ridge represents an age group, with the x-axis showing the number of words understood (comprehension) and the y-axis indicating the age group category. The density of each ridge shows how comprehension varies within each age group.

As we can see from the plot, comprehension is generally lower for children from 0 to 10 months of age at earlier ages. As children grow up, the distribution of comprehension skills is skewed toward higher levels. It is interesting to note that 15 to 20 month old have a wider distribution of comprehension skills, which indicates greater variability.

2.4.2 Birth Order and Caregiver Education

The `birthorder` variable indicates the position of the child in terms of the order of births within their family. It ranges from 1 (first-born) to 8, which allow us to understand whether being a first-born, middle, or later-born child influences comprehension. Birth order can play an important role in a child's language development due to varying attention levels or interactions they receive from caregivers and siblings.

The `caregivereducation` variable reflects the highest level of education attained by the child's caregiver. This variable includes levels such as "Primary," "Secondary," "College," "Some College," "Graduate". Caregiver education is also important in the analysis because it often serves as a proxy for socioeconomic status and access to language resources, both of which can significantly affect a child's language environment and comprehension ability.

Figure 3 shown above visualizes the distribution of birth order along with caregiver education level. Each bar represents the number of children with a specific birth order, ranging from 1 (first born) to 8. The bars are stacked by caregiver's education level, with different colors indicating categories such as "College", "Graduate", "Primary", and others.

As can be seen from the graph, the majority of children are first or second births, while the number of children in higher birth orders is lower. Caregivers with a "college" or "graduate" level of education make up a large percentage of first and second births, while lower birth order caregivers with "some college" or "secondary" level of education also make up a large proportion of caregivers. The predominance of lower birth orders suggests that most families in the data set tend to have only one or two children.

Figure 4 shows the mean comprehension of children grouped by the caregiver's educational level. This plot indicates that children whose caregivers have a higher level of education, such as `Some Graduate` and `Some Secondary`, tend to have higher average comprehension ability compared to children whose caregivers have only a primary education.

The `Primary` category has the lowest mean comprehension, suggesting that lower caregiver education is associated with lower comprehension in children. There is also less variation

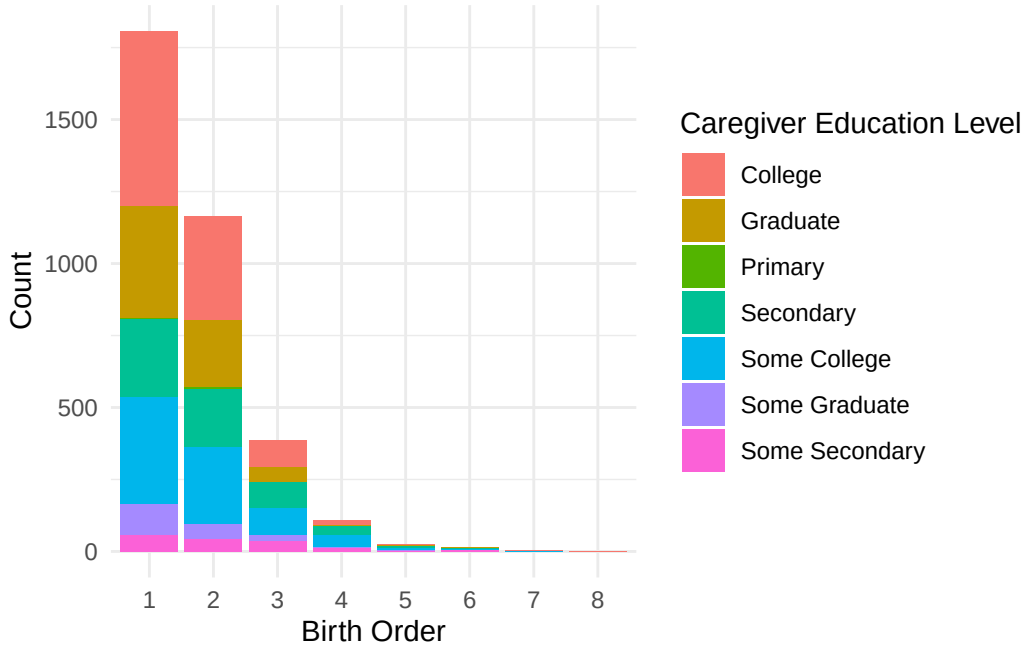


Figure 3: Distribution of Caregiver Education Levels by Birth Order. The stacked bar plot shows the count of children for each birth order, colored by the education level of their primary caregiver.

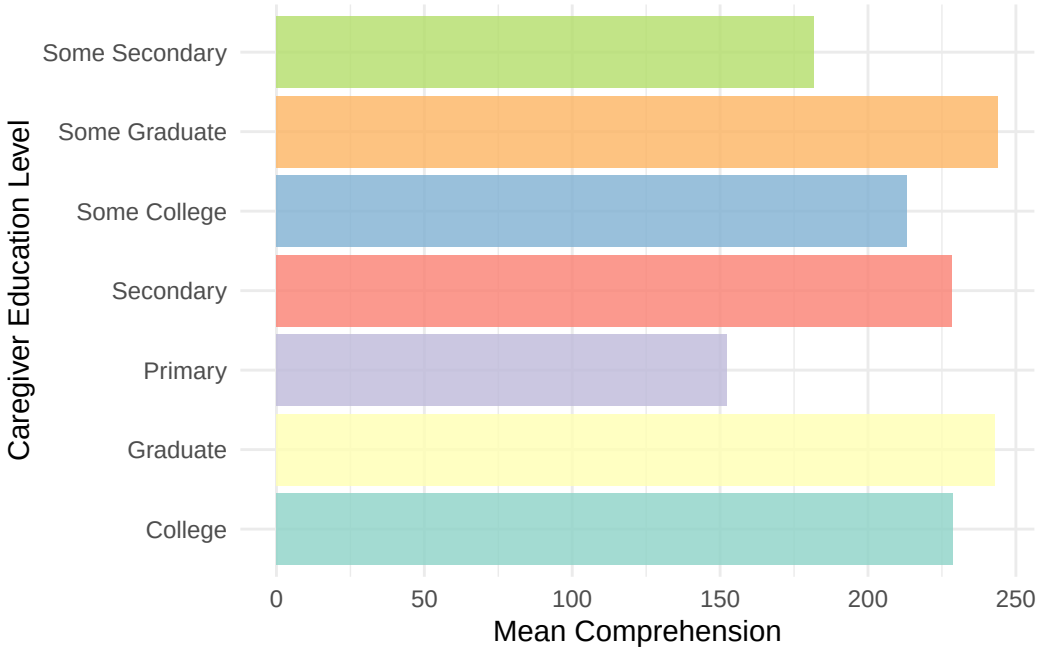


Figure 4: Corrected Colored Horizontal Bar Plot of Mean Comprehension by Caregiver Education Level. Each bar is colored differently to clearly distinguish between education levels.

across the categories, with the mean value for all groups being relatively close. This may indicate that while caregiver education level has an impact, it is not drastically different across these categories. The bars for `Some Graduate` and `Graduate` are among the highest, showing that higher education levels are correlated with increased comprehension. However, there is no indication of median, range, or outliers, as this plot represents only the average values for each group.

2.4.3 Race

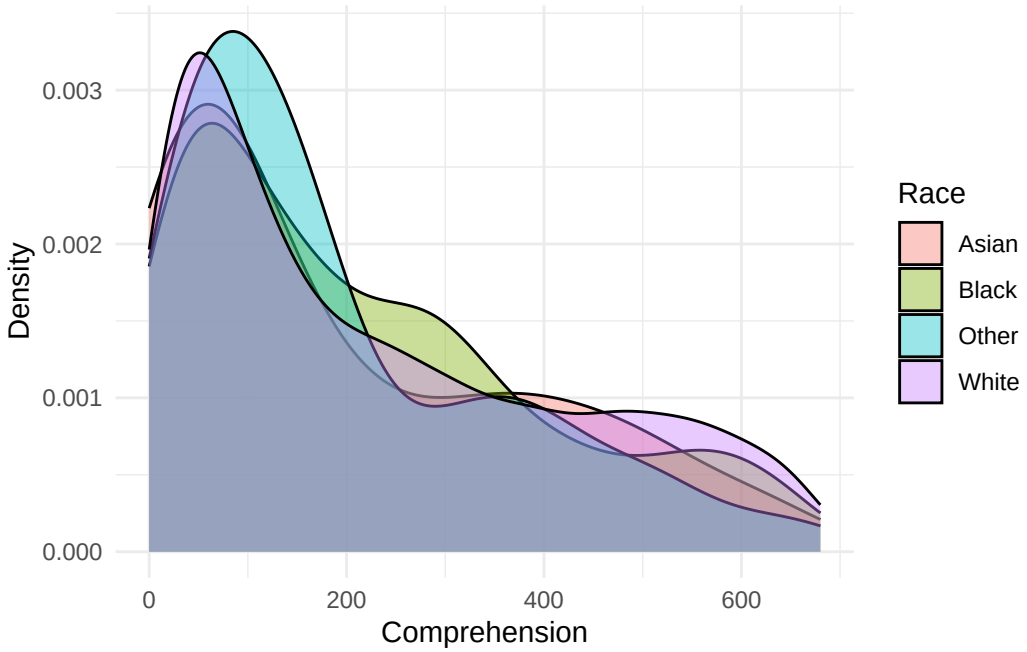


Figure 5: Density Plot of Comprehension by Race

Figure 5 visualizes the distribution of comprehension across different race groups, including Asian, Black, Other, and White. Each coloured density curve represents the distribution of comprehension scores for children from each race group.

From the plot, we can see that `White` children have a wider distribution of comprehension compared to other groups, with a peak density occurring at a higher number of words compared to other groups. `Asian`, `Black`, and `Other` race groups appear to have slightly lower comprehension distributions compared to White children, with peaks occurring at a similar lower range.

The overlap in the density curves shows comprehension across race groups share some similarities in their distributions, although the White group tends to show a broader range and higher scores.

3 Model

My modelling approach aims to quantify the relationship between various child and caregiver characteristics and comprehension in early childhood. For this analysis, we use a Bayesian linear model to examine how factors such as age, production (number of words produced) influence comprehension. The model is implemented using the `stan_glm` function, with a Gaussian distribution to capture the variability in comprehension.

The model assumes that the distribution of comprehension follows a normal distribution when these factors are considered. This Gaussian assumption favours parameter estimation, which is a standard method for linear regression. To prevent overfitting and maintain interpretability, we assume a modest before ensuring that uncertainty is balanced between predictors. This approach allowed us to assess the effects of child and caregiver characteristics on comprehension while maintaining the stability of the findings. See in Appendix C for more background details and diagnostics.

3.1 Alternative model

Alternative models, including a Bayesian logistic regression was considered. Logistic regression was rejected since the outcome (comprehension) is continuous, not binary. The Bayesian linear model was ultimately chosen for its balance between interpretability and the ability to handle uncertainty explicitly through priors.

To determine the best set of predictors, both a full and a reduced model of the Bayesian linear model were fitted. After comparing the models using Leave-One-Out Cross-Validation, the results Table 1 showed that the reduced model had an elpd difference of 0.0, while the full model had an elpd difference of -1.89 with a se difference of 0.17. This indicates that the reduced model performed better than the full model in terms of predictive accuracy with fewer predictors. Consequently, the reduced model was chosen for its simplicity and interpretability, without compromising on predictive performance. The detailed model selection process is in Appendix C.1

Table 1: Comparison of Full and Reduced Models Using LOO

	ELPD Diff	Sd Error Diff
reduced_model	0.00	0.00
full_model	-1.89	0.17

3.2 Model set-up

3.2.1 Reduced Model

The reduced model was developed by removing non-significant variables `sex` and `monolingual` from the full model. This reduced model simplifies interpretation and potentially improves model performance without losing significant explanatory power.

The model predicts the comprehension of children using the following predictor variables:

- Age(`age`): Represents the child's age in months.
- Production Words(`production`): The number of words the child can produce.
- Norming Status (`isnorming`): A binary variable, shoews if the child is being used for norming purposes (1 if true, 0 false).
- Birth Order (`birthorder`): Represents the child's birth order, ranging from 1 to 8.
- Caregiver Education (`caregivereducation`): The highest education level attained by the child's caregiver.
- Race (`race`): Represents the race of the child.

The model takes the form:

$$
\begin{aligned}
y_i \mid \mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{production}_i + \beta_3 \cdot \text{isnorming}_i + \beta_4 \cdot \text{birthorder}_i \\
&\quad + \beta_5 \cdot \text{caregivereducation}_i + \beta_6 \cdot \text{race}_i + \epsilon_i \\
\epsilon_i &\sim \text{Normal}(0, \sigma^2)
\end{aligned}
$$

Where:

- $y_i \mid \mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$ represents the outcome variable y_i (comprehension: number of words understood by child i), which is modeled as a normal distribution with mean μ_i and standard deviation σ.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ are the coefficients for each predictor.
- σ^2 is the variance of the error term.

The model is executed in `R` (R Core Team 2023) using the `rstanarm` package (Goodrich et al. 2022). Default priors from `rstanarm` (Goodrich et al. 2022) are used, with the priors set to have a mean of zero and a moderate standard deviation to ensure a reasonable level of regularization.

3.3 Model justification

Existing research in developmental psychology and linguistics suggests that factors such as age, production of words, birth order, caregiver's level of education, and other demographic variables can have a great impact on children's comprehension. For example, older children generally comprehend more words due to longer exposure to language and cognitive development. Since these two aspects of language acquisition are closely related, higher language expression tends to correlate with better comprehension. Caregiver's education is associated with a child's vocabulary since more educated caregivers provide a richer language environment. Birth order may also play a role, with research suggesting that first-born children typically receive more one-on-one attention, which may lead to differences in comprehension levels.

Bayesian linear regression modeling was chosen to predict comprehension because the outcome variables are continuous and can reasonably be assumed to follow a normal distribution. Linear regression is suitable for quantifying the relationship between multiple predictors and continuous outcomes, which provide interpretable coefficients for each independent variable. Bayesian methods allow us to incorporate a priori knowledge into the model and are particularly useful in situations where existing research provides information on expected effects. In addition, the bayesian inference method provides a way to quantify the uncertainty of parameter estimates, which provides us with a deeper understanding of the modeled relationships.

3.4 Create API for the Reduced Model

To make the reduced model more practical and usable, we created an API that provides real-time predictions of comprehension based on user-provided inputs. Researchers can input variables like a child's age, production score and caregiver's education to receive predictions with credible intervals. By using the `plumber` (Schloerke and Allen 2024), the API ensures usability and integration into various workflows.

4 Results

Section 4 examines the relationship between age, production, norming status, birth order, caregiver's education and race with respect to early childhood word comprehension. Using a dataset containing observations of children's linguistic abilities and various influencing factors, we apply a bayesian linear regression model to assess which key predictors have the most key impact on comprehension. Below, we present the results of our model and discuss the implications of our findings.

We predicted comprehension scores using our test dataset, but due to missing data for certain groups such as some levels of caregiver education or race categories, our model could not make predictions for all children. To address these gaps, we examined demographic trends from similar children in the dataset; if a group showed consistent patterns in linguistic development,

we assumed that trend would continue and used it to estimate missing comprehension values. This limitation arises because groups with missing values are not represented in the model, preventing accurate forecasts for those individuals. Further discussion on this limitation can be found in Section 5.

4.1 Model results and interpretation

The bayesian linear regression model built using our training dataset, which consists of 2,457 data points, estimated the factors that influence comprehension levels in children. For brevity, Table 2 shows only the first ten rows of the reduced model's coefficients, while the full reduced model summary is provided in Appendix C. The intercept is estimated at 77.564, representing the baseline comprehension level when all predictors are at their reference levels. The model achieved an R^2 value of 0.936, indicating that 93.6% of the variance in comprehension outcomes is explained by the predictors included. The adjusted R^2 value of 0.936 further confirms that the model captures the relationships between variables effectively without overfitting.

Table 2: Summary of key coefficients from the Bayesian linear regression model predict comprehension

	coefficient
(Intercept)	77.564
age	-3.476
production	0.983
isnorming	26.034
birthorder	-1.919
caregivereducationGraduate	1.531
caregivereducationPrimary	10.881
caregivereducationSecondary	6.342
caregivereducationSome College	-2.234
caregivereducationSome Graduate	-0.144

Certain predictors have a more pronounced impact than others. For instance, `production` (0.983) shows that a one-unit increase in the number of words produced by a child corresponds to a 0.983 increase in predicted comprehension outcomes. The variable `isnorming` (26.034) also has a positive effect, indicating that children in the norming status tend to have higher comprehension levels. Other predictors, such as `age` (-3.476) and `birthorder` (-1.919), have negative coefficients, suggesting that as age increases (within the studied age range), there may be a slight decrease in comprehension, and children born later in the birth order tend to have lower comprehension scores.

The caregiver's education level also contributes to variations in comprehension. For example, `caregivereducationPrimary` (10.881) indicates that children whose caregivers have a

primary education level tend to have higher comprehension compared to those whose caregivers have no formal education. However, `caregivereducationSomeCollege` (-2.234) and `caregivereducationSomeGraduate` (-0.144) show negative coefficients, suggesting a potential decrease in comprehension associated with these categories. A complete overview of all predictors is available in Appendix C.2.

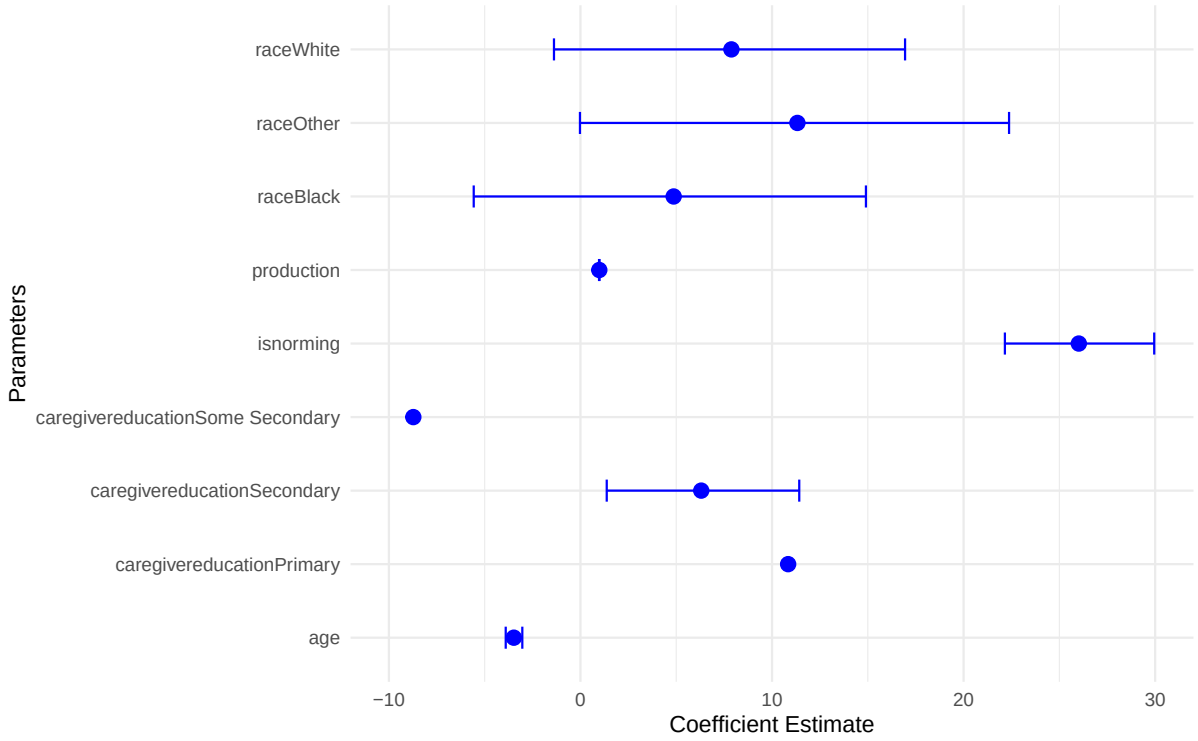


Figure 6: Coefficient Estimates with Confidence Intervals from Bayesian linear regression model

Based on Table 5 in Appendix C, Figure 6 shows the 90% credible intervals for selected coefficients from the Bayesian linear regression model used to predict comprehension by using package `broom.mixed`(Bolker and Robinson 2024). The variables in Figure 6 are chosen based on their coefficient magnitude in the regression results table, specifically the caregiver's education levels are chosen for those coefficients that have an absolute value greater than 3, which indicate a significant influence on the model.

The plot visualizes the estimates of key parameter and their confidence intervals, which provide a deeper understanding of the direction and uncertainty associated with these estimates. The intervals provide an understanding of the possible range of where the impact of each variable.

- `age`, `production`, `isnorming` and `caregivereducation` all have 95% credible intervals that do not include 0. Their CIs do not contain 0, indicating that these predictors have meaningful impacts on comprehension.

- `birthorder` and `race` have 95% credible intervals that include 0. Since their CIs include 0, this suggests they do not have key affect comprehension.

As shown above, the predictor variables of `age`, `production`, `isnorming` and `caregivereducation` had the greatest impact on `comprehension`. `production` and `isnorming` showed a strong positive correlation, indicating that the higher the production of words and norming status to the normative group, the better the comprehension of words. `age` and `caregivereducation` also play a major role, with different age and caregiver education having different effects.

5 Discussion

This study provides an in-depth study of early childhood language comprehension by examining various factors, such as age, caregiver education, word production, and birth order, through a Bayesian linear regression model. The findings underscore the significant influence of these predictors, contributing to a broader understanding of the mechanisms that underpin language acquisition during early childhood development.

5.1 The Role of Caregiver Education in Language Development

An important finding of this study is that caregiver education plays an important role in early childhood language comprehension. Children show better comprehension with caregivers who are more educated, which highlights the great importance of a rich home environment in early language development. This suggests that caregiver education can provide an important representation of language exposure, socioeconomic status, and access to resources that work together to create a favourable learning environment for children.

5.2 Effect of birth order on language comprehension

Another finding was the subtle effect of birth order on comprehension. The results showed that first-born children generally had better comprehension skills compared to their siblings. This may be because they receive more one-on-one attention from their caregivers before subsequent siblings are born. The effect of birth order suggests that parental dynamics and family structure play a key role in shaping early language skills, which suggests that attention needs to be more equitably distributed among children in a multi-child family.

5.3 Model Limitations and Uncertainties

While this study provides meaningful observations, there are many limitations. A major limitation is the high degree of uncertainty in some of the model's coefficient estimates, particularly those related to race and the educational level of some caregivers. The wide confidence intervals indicate variability and inconsistency in the data, which may prevent the generalization of the findings. In addition, using samples may introduce bias and reduce the representativeness of the results across populations. In addition, missing data for some population groups limits the ability of the model to make predictions for all groups, which may affect the overall robustness of the study results.

5.4 Recommendations for Future Research

Future research should address these limitations by collecting more representative data across different demographic groups. This would reduce uncertainty in model estimates and provide a better understanding of the relationships between different predictors and language comprehension. Longitudinal studies would also be useful for assessing these relationships' stability over time, providing a deeper understanding of the long-term effects of early intervention. In addition, further study of the role of caregiver education, particularly in families with multiple children, could provide more effective strategies to support equitable language development. The expansion of the dataset to include more vocabulary categories and contextual factors such as socioeconomic status, bilingual ability, and parental language practices also enhances our understanding of the mechanisms of early language learning and supports the development of more detailed educational policies.

Appendix

A CDI methodology overview and assessment

A.1 Overview

The Communicative Development Inventory (CDI) is a standardized assessment tool used to study early language development in children aged 8 to 30 months. Originally developed by (M. C. Frank et al. 2021), the CDI provides a reliable and general method for evaluating both vocabulary comprehension and production based on parent reports. This chapter revisits the strengths and limitations of CDI, its psychometric properties, and the reliability of the instrument in longitudinal studies.

A.2 Target population, frame, and sample

- Target Population: Infants and toddlers aged 8 to 36 months whose parents are willing to report on their child's language development.
- Sample Frame: The target sample includes children from different demographic backgrounds to capture a diverse representation of language development. Recruitment was conducted across pediatric clinics, childcare centers, and online forums targeting parents, ensuring participants come from various socioeconomic, educational, and geographic contexts.
- Sample Size: The study collected data from approximately 2,000 parents. A sample size of this scale ensures generalizability while maintaining a margin of error within $\pm 3\%$ at a 95% confidence level, thereby enhancing the robustness of findings.

A.2.1 Variables

- Age Groups: 8-12 months, 13-24 months, 25-36 months
- Gender: Boy, Girl
- Caregiver Education Level: Primary, Secondary, Some College, Bachelor's, Graduate, Other
- Household Income Level: <$30,000, $30,000-$59,999, $60,000-$99,999, >$100,000
- Caregiver Employment Status: Employed, Unemployed, Student, Homemaker
- Language at Home: Monolingual, Bilingual, Multilingual
- Geographic Region: Urban, Suburban, Rural
- Race: White, Black/African American, Hispanic/Latino, Asian, Native, Other

A.3 Sample recruitment

Recruitment Methods:

- Parents were recruited via two primary methods: direct outreach at pediatric clinics and digital advertisement on online parenting platforms such as forums and social media. This double approach ensured coverage of typically underrepresented groups such as rural families. The recruitment was conducted with bilingual support (English and Spanish) to improve inclusivity and response rates.
- Online Panel Use: An additional recruitment method involved using an online panel to increase the diversity of the sample. This panel was instrumental in reaching parents from areas with limited in-person outreach capabilities.

A.4 Sampling approach and trade-offs

Sampling Approach:

- The approach involves collecting CDI (Communicative Development Inventory) data across various languages and dialects, using both cross-sectional and longitudinal sampling. This data is sourced from multiple studies conducted over the years, sometimes gathered through paper forms, mail-in surveys, in-person sessions, or electronically. Samples are typically convenience samples, with limited control over representativeness.
- Types of Data: The data includes both "Words & Gestures" (WG, for infants) and "Words & Sentences" (WS, for toddlers), gathered from children ranging from infancy to around 36 months. Instruments include checklists and questionnaires aimed at assessing language comprehension and production.

Trade-offs:

- Advantages: The flexibility of using both WG and WS instruments allows for capturing a broad spectrum of early language development stages. The data provides meaningful ideas about different aspects of language development, including gestures, morphology, and grammar.
- Disadvantages: The sampling method is not necessarily representative due to reliance on convenience samples, which leads to biases depending on how researchers choose or access participants. Moreover, inconsistent administration methods such as electronic, paper and in-person lead to potential variability in data quality, affecting comparability between studies.

A.5 Non-response handling

Non-response Issues: The variability of the data is mentioned in the CDI methodology and is sometimes described as a "difficult dataset" where inconsistent or surprising results occur. This may indicate that non-response issues or participant misunderstanding of the questions affected data quality.

Strategies for Handling Non-response:

- Multiple Attempts: The research team attempted to handle potential non-response issues by improving the clarity of instructions such as simplifying written instructions for electronic forms, thereby reducing misinterpretations that could lead to "floor and ceiling" effects.
- Data Inclusion: Although recognizing the variability caused by nonresponse, the research team decided not to exclude any data sets. They instead accept the confusion and variability inherent in it, and recognize that exclusion of these data may lead to biased conclusions. This method of inclusion was intended to avoid circular arguments and preserve the natural variability in the dataset.

A.6 Longitudinal Stability of CDI Measurements

Stability Across Time: Longitudinal data from CDI demonstrate high stability in early language scores. Studies like De Houwer (HOUWER, BORNSTEIN, and LEACH 2005), showing major correlations in CDI measurements across different ages, indicating that early language acquisition remains relatively stable over time.

A.7 Psychometric Analysis and Item Response Theory (IRT)

- Measurement Properties of CDI Items: Using IRT, each CDI item was evaluated for difficulty and discrimination. Items like common nouns showed high discrimination, indicating their effectiveness in distinguishing between different levels of vocabulary ability.
- Psychometric Weaknesses: Items such as "mommy" and "daddy" have low discrimination because they are almost universally known by children, providing limited information about individual differences in vocabulary. However, these items are retained due to their cultural and developmental relevance.

IRT Evaluation Results:

- Difficulty: Items varied in difficulty, with abstract words proving harder for parents to report on than concrete items.

- Discrimination: High discrimination scores were associated with verbs and adjectives, suggesting these word types are better indicators of a child's linguistic development level compared to function words.

A.8 Simulation of CDI Sampling Approach

To further improve the reliability and representativeness of CDI findings, Bayesian inference can be used to assess different sampling strategies. Bayesian methods can integrate a priori information about sampling bias and provide a probability framework to assess the uncertainty of sampling outcomes. By incorporating a priori information related to demographic characteristics such as socioeconomic status and bilingual families, researchers can better model the effects of sampling bias and improve the robustness of their conclusions.

Simulations can also be extended to a variety of sampling methods using Bayesian modelling to estimate the impact of sampling bias, such as over-representation of certain demographic characteristics or under-sampling of key groups. Generating synthetic datasets containing these underrepresented groups would illustrate how these populations might affect language development patterns, and updates to the Bayesian model would adjust the model to reflect observed trends in the data over time.

A.9 Survey implementation and Structure

To reach a diverse audience of caregivers, multiple distribution channels will be used. The link to the survey will be distributed via email invitations, community events, and targeted social media advertisements. The link to the survey is here.

A.9.1 Budget Allocation

- Survey Panel Recruitment: $20,000
- Community Outreach: $10,000
- Incentives (Gift Cards and Lotteries): $10,000
- Data Cleaning and Validation: $5,000
- Advertising on Social Media: $5,000

A.9.2 Survey structure

Survey introduction:

Welcome!

We are conducting a survey to better understand the language development of children aged 8 to 36 months. Your participation is crucial in helping us gain valuable insights into early language acquisition and the factors that influence it.

Please note:

- All responses are confidential and will be used for research purposes only.
- This survey will take approximately 10 minutes to complete.
- Your participation is entirely voluntary, and you may withdraw at any time.
- All questions marked with an asterisk (*) are required.
- There are no right or wrong answers, we appreciate your honest opinions.

If you have any questions or concerns, please feel free to contact our research team at winnie keai23@gmail.com (Ziyuan Shen).

As a thank you for your participation, you will be entered into a raffle to win a grand prize of a gift card. We are deeply grateful for your participation and appreciate the time and effort that you put in.

Survey question:

Part 1: Eligibility and Consent

1. Is your child between the ages of 8 and 36 months?

- Yes / No

2. Are you the primary caregiver of the child?

- Yes / No

3. Are you willing to provide information about your child's language development?

- Yes / No

Part 2: Demographics

4. How old is your child?

- 8-12 months / 13-24 months / 25-36 months

5. What is your child's gender?

- Boy / Girl / Prefer not to say

6. What is your child's race or ethnicity? (Other, please specify)

- White / Black / African American / Hispanic or Latino / Asian / Other

7. What is your highest level of education?(Other, please specify)

- Primary / Secondary / Some College / Bachelor's Degree / Graduate Degree / Other

8. What is your current employment status? (Other, please specify)

- Employed / Unemployed / Student / Homemaker/ Other

9. What is your household income level?

- Less than $30,000 / $30,000-$59,999 / $60,000-$99,999 / Greater than $100,000 / Prefer not to say

10. How many language(s) do you speak at home?

- Monolingual (One language) / Bilingual (Two languages) / Multilingual (Three or more languages)

Part 3: Child's Language Development

11. At what age did your child start saying their first word?

- 8-12 months / 13-24 months / 25-36 months

12. How often does your child use words or phrases (like “mama”, “more”)?

- Frequently (several times a day) / Occasionally (once or twice a day) / Rarely / Not yet

13. Does your child use gestures (like waving, pointing) to communicate?

- Yes, often / Occasionally / No

14. How would you rate your child’s vocabulary comprehension?

- Understands basic words (like names of family members) / Understands common household objects (like cup, ball) / Understands more complex phrases (like “Give me the red ball”) / Does not yet understand any words

Part 4: Child’s Language Environment

15. How often do you engage in activities that promote language development (like reading, or singing songs)?

- Daily / Several times a week / Once a week or less / Never

16. Do you use any of the following methods to support your child’s language development? (Check all that apply)

- Reading books aloud / Singing songs / Playing interactive games (like peek-a-boo) / Speaking to the child in multiple languages / None of the above

17. How often does your child interact with other children?

- Frequently / Occasionally / Rarely / Does not interact with other children

Part 5: Additional Information

18. Do you believe there are any factors that have particularly influenced your child’s language development? (Such as: socioeconomic status, exposure to multiple languages, developmental disorders)

- Yes / No

19. Do you have any concerns about your child's language development?

- Yes / No

Part 6: Verify

20. Please select 'Agree' to verify that you are paying attention.

- Agree / Disagree

21. Do you agree to participate in this survey? Your responses will be kept confidential and used only for research purposes.

- Yes / No

End Part

Thank you for participating!

Your responses are valuable to our research on early language development. If you have any questions or would like more information about our research, please feel free to contact us at winniekeai23@gmail.com(Ziyuan Shen).

A.10 Conclusion

The CDI method, while not without its challenges, provides a generalized approach to understanding early language development through parental reports. The stratified sampling methodology used in CDI studies allows for a more representative understanding of language development across demographic contexts, thus mitigating potential biases in traditional parent-report measures. The CDI has been used in some studies to assess language development in children and adolescents. Although there are some limitations, such as potential bias in parent reports and subjectivity in assessing comprehension, the CDI has demonstrated strong reliability and utility in psychometric assessments, making it an important tool for researchers of child language development.

B Additional data details

B.1 Data manipulation and cleaning

In the data cleaning process, we prepared the raw data for analysis by applying transformations, filtering, and restructuring using several R packages, including `dplyr` (Wickham et al. 2023), `arrow` (Richardson et al. 2024), and `rsample` (Frick et al. 2024).

1. **Removing columns with missing values**: Using the `select(where(~ !all(is.na(.))))` function from `dplyr` to remove any columns that were entirely NA, thus retaining only those columns that contain meaningful data.
2. **Rename columns by removing underscores and updating names**: Using `gsub()` function to replace all underscores with an empty string (" ").
3. **Removing unnecessary columns**: We further removed columns deemed unnecessary for analysis, they were `downloaded`, `language`, `form`, `datasetname`, `childid`, `ethnicity`, `languageexposures`, `healthconditions`, `typicallydeveloping` by using `dplyr select()` function. This helped to streamline the dataset by eliminating redundant information.
4. **Handling missing values**: Removing rows with any missing values by using the `na.omit()` function, ensuring that our dataset contained only complete cases. This step was important to maintain data integrity and ensure accurate model training
5. **Transforming birth order**: The `birthorder` column was converted to a numeric sequence using `mutate()` and `factor()`. We mapped levels like "First", "Second", "Third" to a corresponding numeric value to facilitate easier use in statistical models.
6. **Converting categorical variables to binary**: Several categorical variables were converted into binary values to prepare the data for modeling. We used `mutate()` and `ifelse()` to recode variables:

- `isnorming` was converted to 1 for "TRUE" and 0 for "False".
- `sex` was coded as 1 for "Male" and 0 for "False".
- `monolingual` was similarly converted to 1 for "TRUE" and 0 for "False".

7. **Splitting the data**: We used the `rsample` package to perform a stratified split of the cleaned dataset based on the `race` variable. This ensured that all levels of `race` were well-represented in both the training (70%) and testing (30%) sets. We set a seed (`set.seed(123)`) for reproducibility.

8. **Saving the cleaned data**: Finally, the cleaned dataset was saved in different formats (CSV and Parquet) using `arrow` to facilitate efficient storage and accessibility. We saved both the full cleaned dataset and the resulting training/testing splits, enabling their use in further analyses and modeling.

B.2 Descriptions of Each Predictor Variable

Table 3

Variable	Description
age	The age of the child in months.
production	The number of words a child can produce.
isnorming	A binary variable indicating if the child is included in the norming sample (1 = Yes, 0 = No).
birthorder	The birth order of the child (e.g., 1 = first-born, 2 = second-born).
caregivereducation	The education level of the primary caregiver (e.g., Graduate, Some College).
race	The race of the child (e.g., White, Asian).
sex	The sex of the child (0 = Female, 1 = Male).
monolingual	A binary variable indicating if the child is monolingual (1 = Yes, 0 = No).

Note: This table provides descriptions of each predictor variable used in the study, explaining their relevance and values.

Descriptions of Each Predictor Variable

C Model details

C.1 Model Selection Process

1. **Fitting the Full Model**: By using Bayesian regression with a Gaussian distribution. The priors for both the intercept and the coefficients are set to normal distributions with a location of 0 and a scale of 2.5, allowing for flexibility in the regression parameters.
2. **Full Model Evaluation**:

In Table 4, it generates a summary table for a Bayesian regression model, using the `tidy()` function from the broom(Bolker and Robinson 2024) package and displaying it with `kable()`.

This summary table can help us understand the significance and uncertainty around each of the model's predictors. The 90% credible intervals can be particularly helpful in assessing the strength of each predictor's relationship with the outcome variable, `comprehension`.

From Table 4, we would like to remove both `sex` and `monolingual` to build a reduced model. Because both of their estimated effects were close to zero and credible intervals included zero,

indicating no significant effect on comprehension. Thus, removing them could simplify the full model while maintaining interpretability and predictive accuracy.

Table 4: Model Summary: Standard Deviations and 90% Credible Intervals for Predictors

term	estimate	std.error	conf.low	conf.high
(Intercept)	78.522	11.316	60.221	96.529
age	-3.481	0.275	-3.928	-3.042
production	0.983	0.008	0.971	0.996
isnorming	26.095	2.418	22.263	30.077
birthorder	-1.921	1.083	-3.744	-0.155
caregivereducationGraduate	1.516	2.841	-3.038	6.371
caregivereducationPrimary	10.214	18.487	-19.460	39.762
caregivereducationSecondary	6.486	3.037	1.532	11.176
caregivereducationSome College	-2.216	2.696	-6.693	2.242
caregivereducationSome Graduate	-0.135	4.509	-7.690	7.557
caregivereducationSome Secondary	-8.677	5.035	-17.102	-0.594
raceBlack	4.699	6.143	-5.881	15.044
raceOther	10.851	7.483	-0.762	22.904
raceWhite	7.697	5.493	-1.436	16.881
sex	0.034	1.999	-3.192	3.315
monolingual	-1.091	8.615	-14.817	13.088

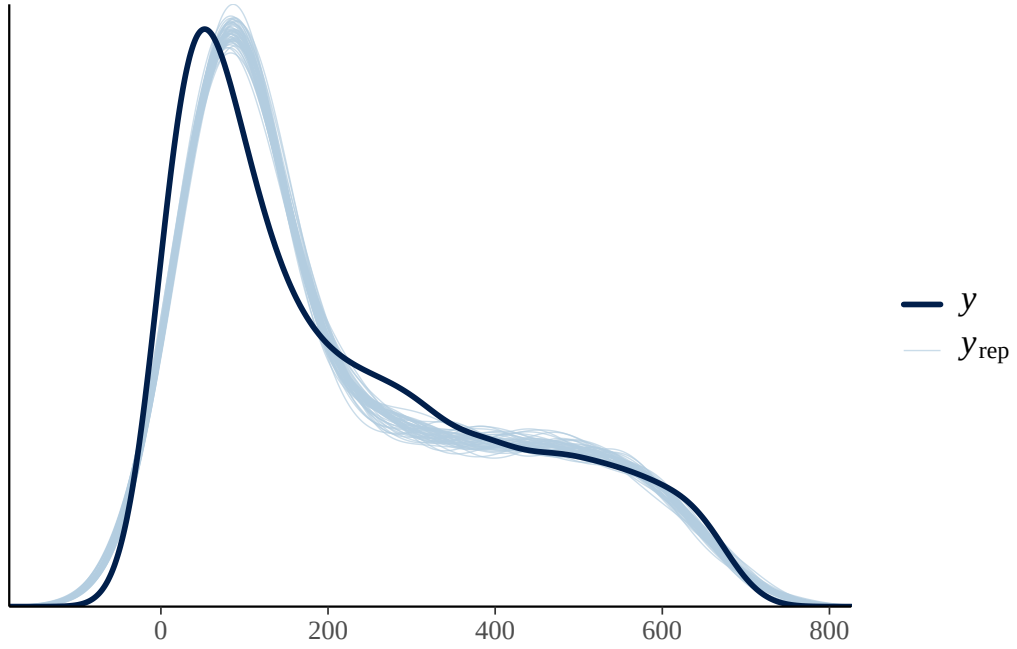
C.2 Model summary

Table 5 presents the coefficients from our model analyzing factors that influence comprehension in early childhood. The `intercept` value is 78.522, represents the baseline level of comprehension when all other predictors are at their reference levels. Key variables include `isnorming` with a positive coefficient indicating that children included in the norming sample tend to have higher comprehension scores. Different caregiver education levels show varying effects. For example, "caregivereducationPrimary`has a positive coefficient of 10.214, suggesting that having a caregiver with primary education is associated with increased comprehension. On the other hand,`caregivereducationSome Secondary‘ has a negative coefficient, indicating a potential decrease in comprehension scores for children whose caregivers have some secondary education.

Table 5: Coefficients from a regression model examining factors influencing comprehension

	coefficient
(Intercept)	77.573
age	−3.478
production	0.983
isnorming	26.014
birthorder	−1.921
caregivereducationGraduate	1.618
caregivereducationPrimary	10.841
caregivereducationSecondary	6.305
caregivereducationSome College	−2.210
caregivereducationSome Graduate	−0.163
caregivereducationSome Secondary	−8.725
raceBlack	4.865
raceOther	11.325
raceWhite	7.879
Num.Obs.	2457
R2	0.936
R2 Adj.	0.936
Log.Lik.	−13 004.522
ELPD	−13 018.3
ELPD s.e.	75.0
LOOIC	26 036.6
LOOIC s.e.	150.0
WAIC	26 036.5
RMSE	48.09

Table 6



C.3 Diagnostics

C.3.1 Posterior predictive check

In Table 6, we present a posterior predictive check by comparing the distribution of comprehension scores (denoted by y) across different age groups. The ridgeline plot illustrates the distribution of the observed comprehension scores for each age group and their corresponding replicated scores generated by the model (denoted by y_{rep}). In a well-fitting model, the replicated data should closely overlap with the observed data.

From the plot, we observe that the distribution of the replicated scores generally follows the same shape as the observed data's density, especially in the central regions of the age groups. This indicates that the model is capturing the main structure of the comprehension across different age intervals. AAny discrepancies between the replicated and observed lines would highlight areas where the model fails to adequately capture the observed data characteristics, such as specific patterns or variations in comprehension across age groups. However, in this case, the overall fit seems reasonable, as the replicated lines match the observed data's density well.

C.3.2 Variance Inflation Factors for Each Predictor in the Model

In Table 7 presents the Variance Inflation Factors (VIF) for each predictor in the model. VIF values are used to assess multicollinearity among the predictors, with values greater than 10 typically indicating a high level of multicollinearity. In this model, all predictors have VIF values below 3, suggesting that multicollinearity is not a key concern. For example, `age` and `production` have VIF values of 2.571 and 2.478, respectively, indicating a moderate relationship with other predictors. Other variables, such as `birthorder` and `race`, show lower VIF values, which indicates minimal multicollinearity.

Table 7: VIF for Each Predictor in the Model

	GVIF	Df	GVIF..1..2.Df..
age	2.571	1	1.603
production	2.478	1	1.574
isnorming	1.242	1	1.114
birthorder	1.064	1	1.031
caregivereducation	1.283	6	1.021
race	1.161	3	1.025

C.3.3 Gelman-Rubin diagnosis

Figure 7 is a visualization of the Gelman-Rubin diagnosis (often denoted as R-hat) for evaluating the convergence of Markov chain Monte Carlo (MCMC) samples in Bayesian models. The plot illustrates the R-hat values for various parameters. An R-hat value near 1 indicates that the MCMC chains have mixed well and are effectively sampling from the same posterior distribution. In Figure 7, all parameters exhibit an R-hat value at or below 1.05, suggesting that the model has reached adequate convergence, and the resulting parameter estimates are reliable for interpretation.

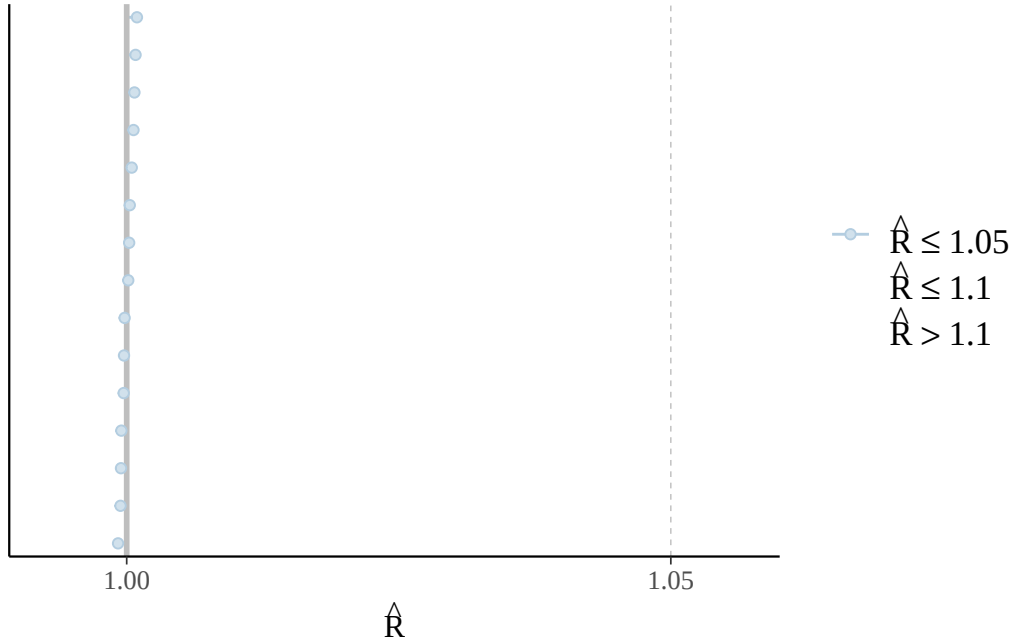


Figure 7: All parameters have R-hat values at or below 1.05, which indicates strong convergence of the MCMC chains and reliable parameter estimates.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

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.

Bolker, Ben, and David Robinson. 2024. *broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

Frank, Michael C., Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2021. "Variability and Consistency in Early Language Learning." *Variability and Consistency in Early Language Learning*. https://langcog.github.io/wordbank-book/.

Frank, Mika et al., Braginsky. n.d. "Wordbank Database." *Wordbank*. https://wordbank.stanford.edu/data/?name=admin_data.

Frick, Hannah, Fanny Chow, Max Kuhn, Michael Mahoney, Julia Silge, and Hadley Wickham. 2024. *rsample: General Resampling Infrastructure*. https://rsample.tidymodels.org.

ggplot2: Elegant Graphics for Data Analysi. n.d.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

HOUWER, ANNICK DE, MARC H. BORNSTEIN, and DIANE B. LEACH. 2005. "Assessing Early Communicative Ability: A Cross-Reporter Cumulative Score for the MacArthur CDI." *Journal of Child Language* 32 (4): 735–58. https://doi.org/10.1017/s0305000905007026.

Lawton, Will, Ozzy Araujo, and Yousif Kufaishi. 2023. "Language Environment and Infants' Brain Structure." *Journal of Neuroscience* 43 (28): 5129–31. https://doi.org/10.1523/JNEUROSCI.0787-23.2023.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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://github.com/apache/arrow/.

Schloerke, Barret, and Jeff Allen. 2024. *plumber: An API Generator for r*. https://CRAN.R-project.org/package=plumber.

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://dplyr.tidyverse.org.

Wilke, Claus O. 2024. *ggridges: Ridgeline Plots in 'Ggplot2'*. https://CRAN.R-project.org/

package=ggridges.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

Zonarich, Elizabeth. 2024. "Why Do Some Kids Learn to Talk Earlier Than Others?" *Harvard Gazette*, October. https://news.harvard.edu/gazette/story/2024/01/why-do-some-kids-learn-to-talk-earlier-than-others-childhood-development-linguistics/?utm_source=chatgpt.com.

Zubrick, Stephen R., Catherine L. Taylor, and Daniel Christensen. 2015. "Patterns and Predictors of Language and Literacy Abilities 4-10 Years in the Longitudinal Study of Australian Children." *PLOS ONE*, September. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0135612&utm_source=chatgpt.com#abstract0.