Factors Influencing Early Childhood Comprehension of American English*

Analyzing the impact of early childhood on English comprehension

Ziyuan Shen

November 26, 2024

This paper investigates the factors that influence language comprehension in young children by using a Bayesian linear regression model. We analyzed predictors including age, word production, caregiver education, and birth order. The analysis showed that factors such as caregiver education and word production have a significant association with comprehension levels. This study highlights the importance of both environmental and developmental factors in shaping early language abilities, providing a clearer understanding of how these elements contribute to a child's comprehension.

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 et al.(Zubrick, Taylor, and Christensen (2015)) found that socioeconomic factors, family background, and personal characteristics significantly influence early childhood receptive vocabulary development. As highlighted by Lawton et al.(Lawton, Araujo, and Kufaishi (2023)), the quality and quantity of language exposure in infancy are associated with brain development and long-term language achievement. In addition, Bergelson's study(Zonarich (2024)) 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. Despite these insights, I 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.

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

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. Using Bayesian regression modelling, I assessed the contributions of these predictors and identified the factors that have the greatest impact on 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

I 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 (Frank, 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), I explore how different predictors affect language comprehension in early childhood. Also, the following packages were used in this study:(Please refer to **?@sec-data-cleaning** for detailed data processing steps.)

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, I filtered the data to include children whose comprehension assessments and caregiver information were complete, which would ensure a comprehensive, 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.

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 `dplyr` (Wickham et al. 2023), `tidyverse` (Wickham et al. 2019), `arrow` (Richardson et al. 2024), and `rsample` (Frick et al. 2024). The detailed cleaning processes are in **?@sec-data-cleaning**.

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 (Wickham 2016), 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.

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.

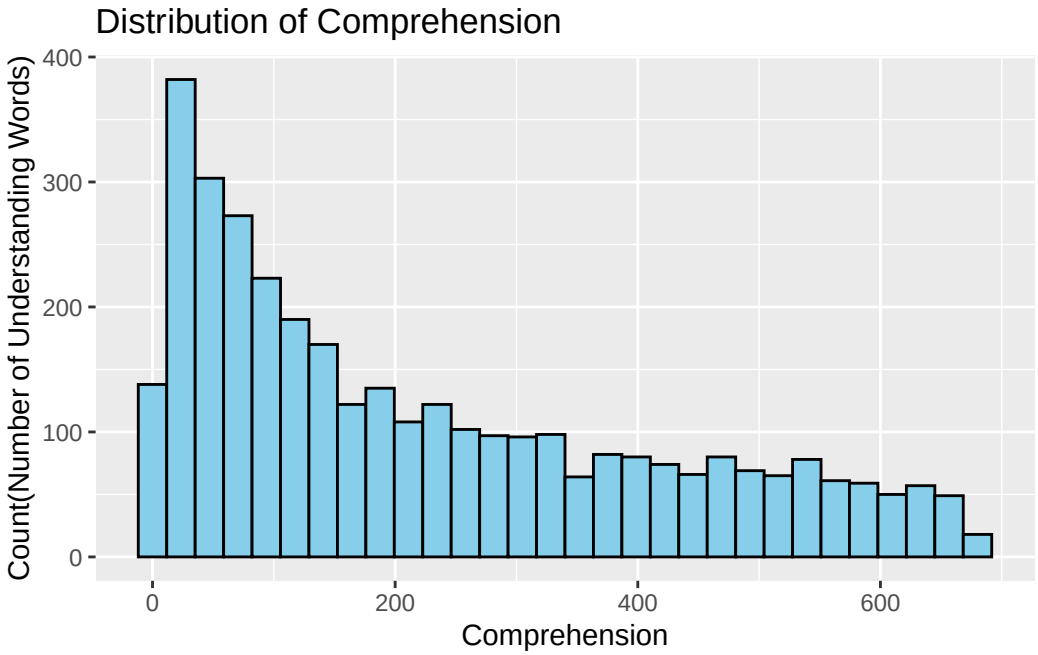


Figure 1: Distribution of comprehension among children. Each bar represents the count of children who understand a specific range of words, illustrating the varying levels of vocabulary comprehension in the dataset.

2.4 Predictor variables

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

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.

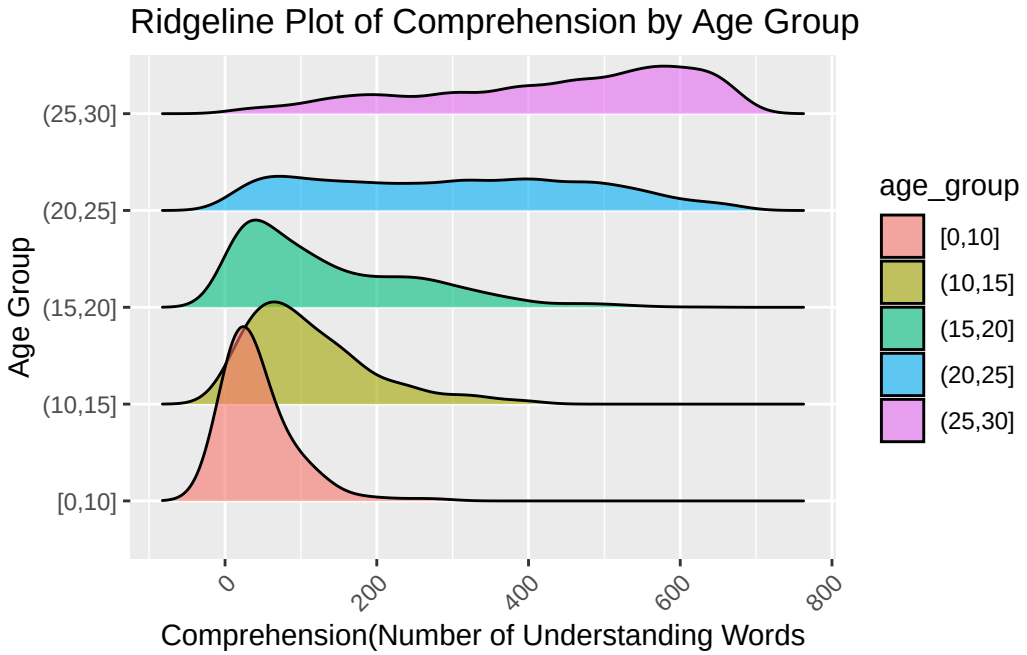


Figure 2: Ridgeline Plot of Comprehension Scores Across Different Age Groups. This plot illustrates the distribution of comprehension scores for each age group, highlighting how comprehension changes across age intervals.

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 `birth_order` 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 `caregiver_education` 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.

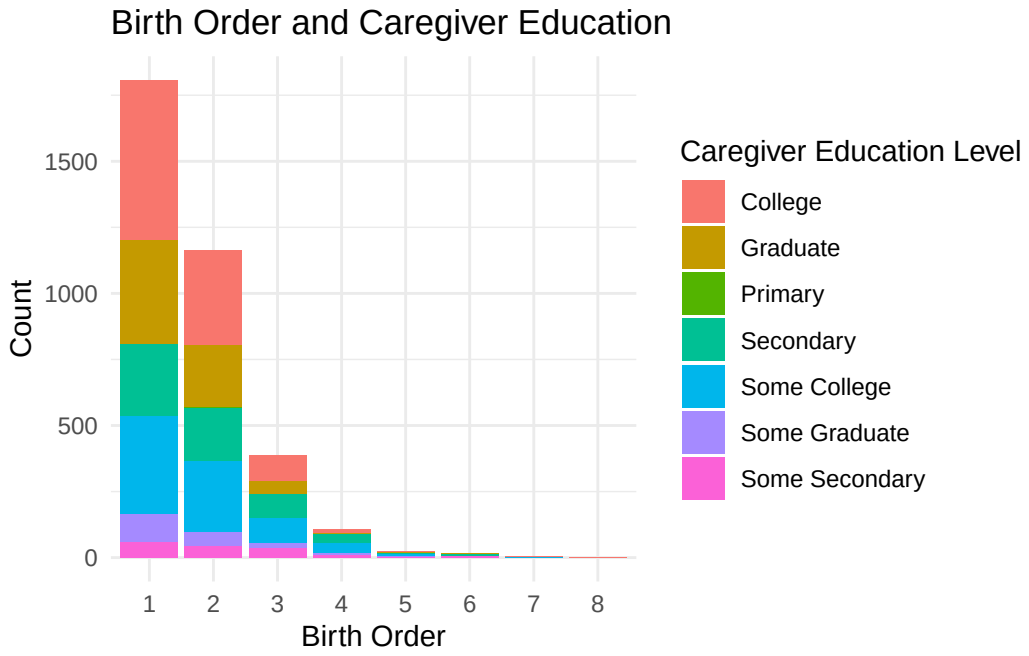


Figure 3: Distribution of Caregiver Education Levels by Birth Order. The stacked bar plot shows the count of children for each birth order, colored by the education level of their primary caregiver.

Figure 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 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.

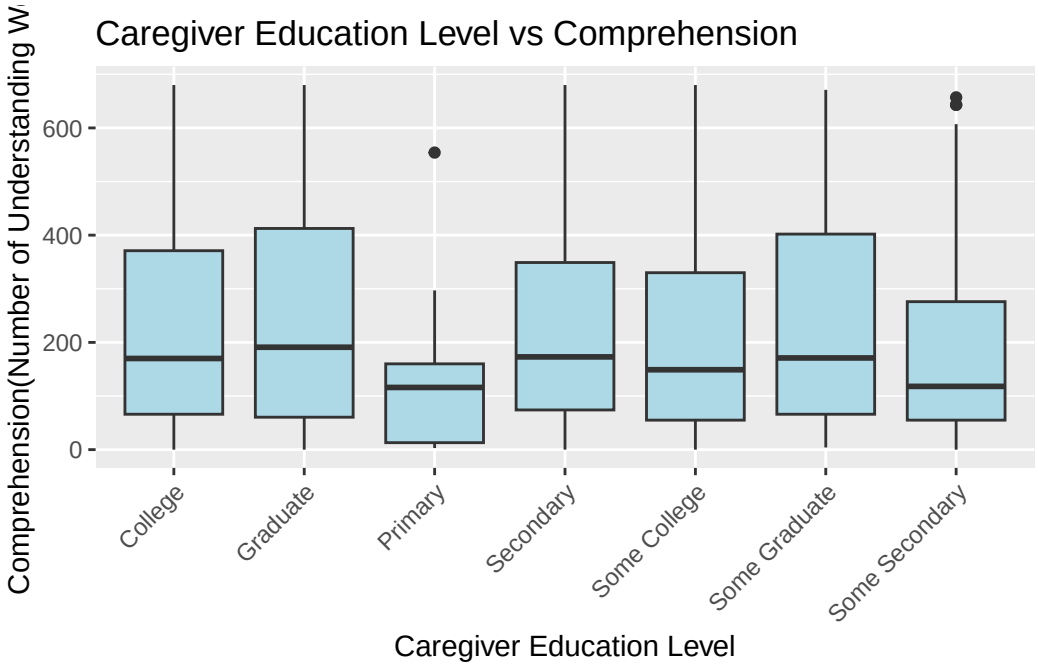


Figure 4: Boxplot of Comprehension by Caregiver Education Level. This plot shows the distribution of comprehension scores for children based on the education level of their caregiver, which highlights differences in comprehension across different educational backgrounds.

Figure 4 shows the distribution of children's number of comprehension words grouped by the caregiver's educational level. Children whose caregivers had higher education, such as `Graduate` and `Secondary` tended to have higher median comprehension scores compared to children with primary education. Median comprehension scores are lowest in the primary school category, with a narrower range compared to the other categories, which suggests that there is less variation in comprehension for children with caregivers at this level of education. `Graduate` and `Some College` show larger IQRs, indicating more variability in comprehension scores among children.

There are several outliers present, particularly for `Some Graduate` and `Graduate` education

levels, suggesting that some children scored significantly higher or lower compared to the majority.

2.4.3 Race

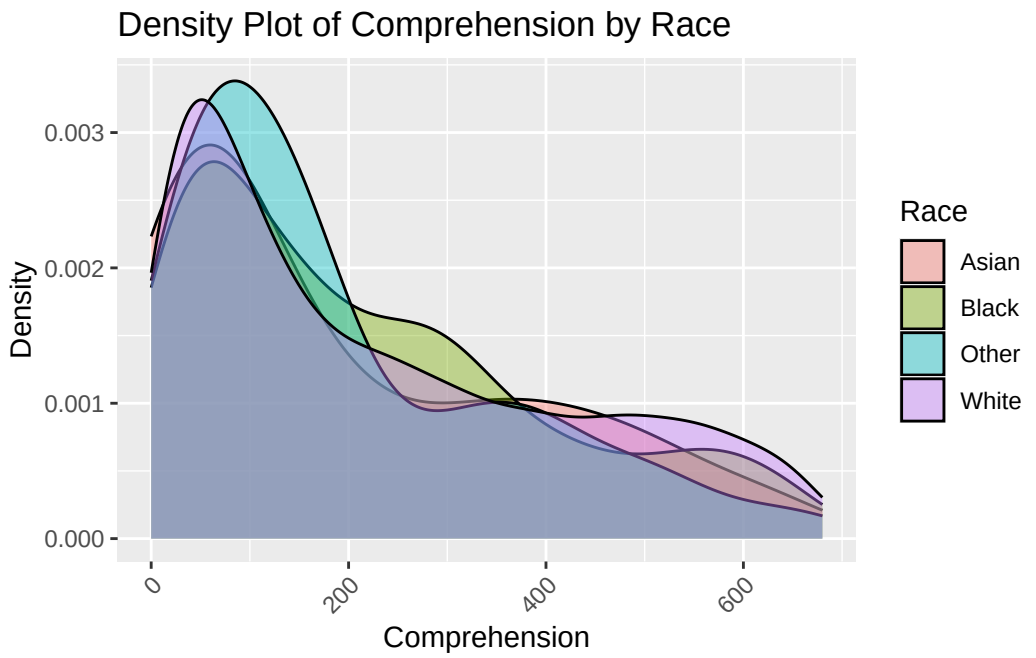


Figure 5: Comprehension by Race

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, I use a Bayesian linear model to examine how factors such as age, production (number of words produced), whether the child is in a norming status, birth order, caregiver education level, race, sex, and whether the child is monolingual 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 Models

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.

3.2 Model set-up

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 (`is_norming`): A binary variable, shoews if the child is being used for norming purposes (1 if true, 0 false).
- Birth Order (`birth_order`): Represents the child's birth order, ranging from 1 to 8.
- Caregiver Education (`caregiver_education`): The highest education level attained by the child's caregiver.
- Race (`race`): Represents the race of the child.
- Gender (`sex`): A binary variable, the gender of children (1 for female, 0 for male).

- Monolingual (`monolingual`): A binary variable, which indicates if the child is monolingual (1 if true, 0 false).

The model takes the form:

$$
\begin{aligned}
y_i \mid \mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{production}_i + \beta_3 \cdot \text{is\_norming}_i + \beta_4 \cdot \text{birth\_order}_i \\
&\quad + \beta_5 \cdot \text{caregiver\_education}_i + \beta_6 \cdot \text{race}_i + \beta_7 \cdot \text{sex}_i + \beta_8 \cdot \text{monolingual}_i + \epsilon_i \\
\epsilon_i &\sim \text{Normal}(0, \sigma^2)
\end{aligned}
$$

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.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8$ 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.

3.2.1 Model justification

Existing research in developmental psychology and linguistics suggests that factors such as age, production (vocabulary), 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 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, providing 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.

4 Results

Section 4 examines the relationship between age, production, is_norming, birth order, caregiver education, race, sex, and monolingual status 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 significant 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.

4.0.1 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 1 shows only the first ten rows of the model's coefficients, while the full model summary is provided in Appendix C. The intercept is estimated at 78.515, 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 1: Summary of key coefficients from the Bayesian linear regression model predict comprehension

	coefficient
(Intercept)	78.515

Table 1: Summary of key coefficients from the Bayesian linear regression model predict comprehension

	coefficient
age	-3.479
production	0.983
is_norming	26.115
birth_order	-1.922
caregiver_educationGraduate	1.577
caregiver_educationPrimary	10.381
caregiver_educationSecondary	6.403
caregiver_educationSome College	-2.204
caregiver_educationSome Graduate	-0.086

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 `is_norming` (26.115) also has a positive effect, indicating that children in the norming status tend to have higher comprehension levels. Other predictors, such as `age` (-3.481) and birth_order (-1.922), 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, `caregiver_educationPrimary` (10.381) 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, `caregiver_educationSome College` (-2.204) and `caregiver_educationSome Graduate` (-0.086) show negative coefficients, suggesting a potential decrease in comprehension associated with these categories. A complete overview of all predictors is available in Appendix C.

5 Discussion

This study provides insights into 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 Influence of Caregiver Education and Environmental Context

One of the prominent findings is the positive relationship between caregiver education levels and children's comprehension scores. This aligns with previous research that emphasizes the importance of an enriched home environment in supporting early learning. The implications suggest that interventions targeted at enhancing caregiver awareness and providing educational resources could significantly bolster children's language development.

5.2 Nuanced Relationships Revealed

Our analysis also uncovered intricate relationships that were not immediately evident. For instance, while caregiver education is generally a strong predictor, its effects appear more pronounced in first-born children compared to their younger siblings. Such a pattern suggests that first-born children might receive more undivided attention, which amplifies the effects of caregiver education. These nuanced findings call for targeted strategies, especially for later-born children, to ensure equitable developmental support across siblings.

5.3 Model Limitations and Uncertainties

Despite these findings, several limitations constrain the conclusions of this study. The linear model used in the analysis assumes a simple relationship between predictors and language comprehension. However, early language acquisition is influenced by a complex web of factors that may interact in non-linear ways. Additionally, the reliance on caregiver-reported data for education level introduces a risk of response bias, which could affect the reliability of the estimates. The dataset also had missing values in key demographic categories, potentially limiting the representativeness of our findings.

5.4 Recommendations for Future Research

Future studies could address these limitations by applying non-linear modeling techniques, such as Generalized Additive Models (GAMs), to better capture the complexities inherent in early language development. Longitudinal datasets that follow children over time would provide a richer context for understanding how developmental trajectories unfold in different family environments. Moreover, expanding the dataset to include more diverse demographics and additional variables, such as socioeconomic status or access to early education resources, could provide a more comprehensive picture of the factors influencing language comprehension.

Appendix

A Additional data details

Table 2: Variable Descriptions

Variable	Description
age	The age of the child in months.
production	The number of words a child can produce.
is_norming	A binary variable indicating if the child is included in the norming sample (1 = Yes, 0
birth_order	The birth order of the child (e.g., 1 = first-born, 2 = second-born).
caregiver_education	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).

Descriptions of Each Predictor Variable

B Model details

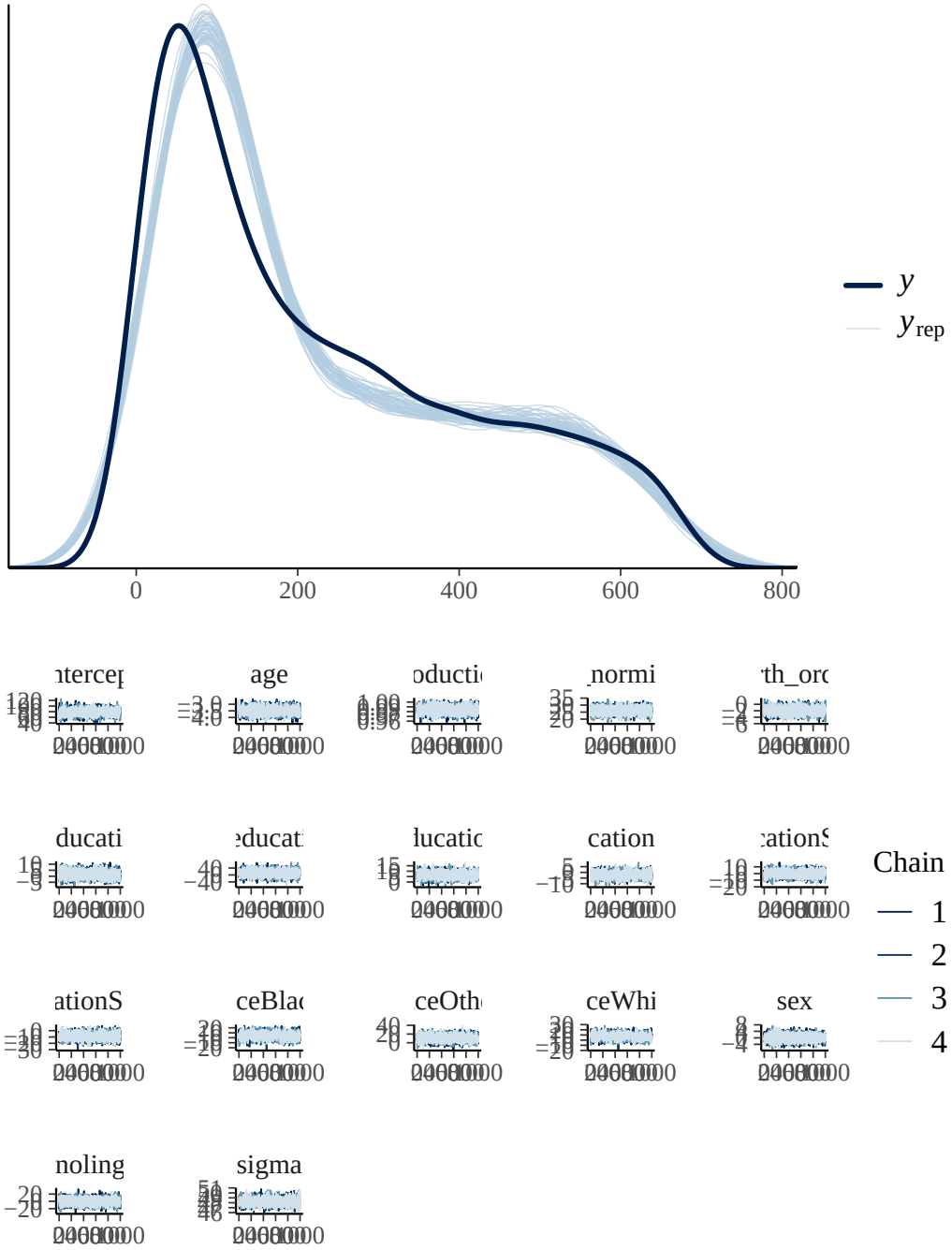
B.1 Posterior predictive check

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Table 3



(a) Trace plot

Figure 6: Checking the convergence of the MCMC algorithm

C Model details

C.1 Model summary

References

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

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.

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

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.

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/.

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://dplyr.tidyverse.org.

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.

Table 4: Coefficients from a regression model examining factors influencing comprehension

		coefficient
(Intercept)		78.522
		(11.316)
age		−3.481
		(0.275)
production		0.983
		(0.008)
is_norming		26.095
		(2.418)
birth_order		−1.921
		(1.083)
caregiver_educationGraduate		1.516
		(2.841)
caregiver_educationPrimary		10.214
		(18.487)
caregiver_educationSecondary		6.486
		(3.037)
caregiver_educationSome College		−2.216
		(2.696)
caregiver_educationSome Graduate		−0.135
		(4.509)
caregiver_educationSome Secondary		−8.677
		(5.035)
raceBlack		4.699
		(6.143)
raceOther		10.851
		(7.483)
raceWhite		7.697
		(5.493)
sex		0.034
		(1.999)
monolingual		−1.091
		(8.615)
Num.Obs.	17	2457
R2		0.936
R2 Adj.		0.935
Log.Lik.		−13 004.701
ELPD		−13 020.2
ELPD s.e.		75.0
LOOIC		26 040.4