Understanding the Ratings of Children's Books: A Data-Driven Analysis*

Exploring the Impact of Cover Type, Page Count, and Publication Period on Children's Book Ratings

Zien Gao

December 3, 2024

This study investigates the factors influencing children's book ratings, focusing on attributes such as cover type, page count, and publication period. Using simulated and cleaned datasets from various sources, we applied a Bayesian hierarchical linear model to explore these relationships comprehensively. We estimate that hardcover books, books in the 150-299 and 300-499 page ranges, and books published after 2000 have notably higher ratings, providing a clear indication of preferences in children's literature. These insights are critical for publishers, authors, and parents in making informed decisions about book design and selection. The findings highlight important characteristics that contribute to well-received children's literature, aiming to enhance both quality and reader satisfaction.

Table of contents

[*] Code and data supporting this analysis are available at: https://github.com/lauragao75/Understanding-the-Ratings-of-Children-s-Books-A-Data-Driven-Analysis

1 Introduction

Children's books play a vital role in early childhood development, shaping language skills, creativity, and understanding of the world. With the vast number of children's books published each year, understanding what makes a book popular and well-rated can provide valuable insights to authors, publishers, and parents. However, despite the abundance of children's books, there is a lack of empirical understanding about which specific attributes make a book highly rated by readers. Filling this gap can provide practical recommendations for those involved in creating and selecting children's literature.

In this study, we explore the factors that contribute to the popularity of children's books, specifically focusing on attributes like cover type, page count, and publication period. By analyzing these elements, we aim to understand which characteristics are linked to higher ratings for children's literature. Specifically, the primary estimand of this study is the relationship between various book characteristics and the ratings they receive. We aim to estimate how attributes such as cover type, page range, and publication period are associated with overall book ratings, providing insights into what features contribute to the perceived quality of children's literature.

The results of our analysis suggest that certain characteristics of children's books are associated with variations in ratings. The Bayesian model indicates that hardcover books tend to have higher ratings compared to paperback editions, which could be due to their perceived durability and quality. Regarding page count, books in the "150-299" and "300-499" ranges appear to have higher average ratings, implying that more substantial stories may be favored. Additionally, books published after 2000 tend to receive higher ratings, possibly reflecting evolving preferences or improved production standards in recent years. These findings reveal associations between book attributes and ratings, though they do not imply causation.

Understanding the factors associated with higher ratings for children's books is valuable for multiple stakeholders. For authors and publishers, these insights can inform decisions about book design, such as choosing a hardcover format or determining an optimal page length, to potentially increase a book's appeal. Parents and educators can also benefit by identifying books that align with popular trends and preferences, thereby selecting literature that children are more likely to enjoy. Ultimately, this research contributes to a better understanding of the characteristics that readers value in children's literature, helping to improve book quality and ensure that stories have a positive impact on young audiences.

The paper is organized into four main sections. Section 2 describes the dataset, its origin, and the cleaning process. Section 3 details the modeling approach used for analysis. Section 4 presents the findings from the model, and Section 5 interprets these findings, including their practical implications and future research directions.

2 Data

2.1 Data Overview

We use the statistical programming language R (R Core Team 2023) to generate, clean, and analyze the dataset. Our data, sourced from the children's book ratings dataset hosted on GitHub by Timothy A. Cookson (Cookson 2023), includes metadata such as cover type, publisher, and ratings. The dataset contains information like ISBN, title, author, cover type, pages, and rating, which form the basis for our analysis. We focus on attributes that may influence book ratings, such as the type of cover, number of pages, and the original year of publishing.

The libraries used in our analysis include `tidyverse` (Wickham 2023b), `arrow` (Richardson, Korn, et al. 2023), `broom` (Robinson and Hayes 2023), `ggplot2` (Wickham 2023a), `knitr` (Xie 2023), `kableExtra` (Zhu 2023), `here` (Müller and Bryan 2023), `bayesplot` (Gabry and Kay 2023), `broom.mixed` (Bolker and Vaughan 2023), `coda` (Plummer et al. 2006) and `modelsummary` (Arel-Bundock 2023). These libraries enable data manipulation, model building, and visualization, ensuring a comprehensive examination of the dataset to explore relationships between book characteristics and ratings.

2.2 Measurement

To understand the factors influencing the popularity of children's books, we translated various real-world phenomena into measurable attributes for our dataset. The dataset used in this analysis originates from a collection of children's books that includes detailed metadata about each book, such as cover type, page count, original publication year, and ratings given by readers. Each of these variables represents specific features that could influence a book's popularity, and they were selected to capture different aspects of what makes a book appealing to young readers and their families.

Cover Type is one of the key attributes measured. It represents the physical form of the book, categorized into "Hardcover" and "Paperback." This feature is often associated with a book's durability, visual appeal, and market positioning. Hardcover books are typically seen as premium products, potentially giving them an edge in ratings due to perceived quality.

Page Count is used to indicate the length of a book and is categorized into five different ranges: "Under 50," "50-149," "150-299," "300-499," and "500 and above." This transformation from a continuous variable to categorical ranges allows us to understand whether readers prefer shorter or longer books, which could affect the overall enjoyment and rating of a book.

Publication Year is another important attribute, reflecting the era during which the book was published. This variable was transformed into distinct periods: "Before 1950," "1950-1979," "1980-1999," and "2000 and after." Such categorization helps capture trends over time, as preferences in children's literature may shift due to changing societal norms, advances in education, or evolving cultural values.

Finally, **Ratings** are collected as numeric values ranging from 0 to 5, representing the perceived quality of each book according to its readers. These ratings serve as the core outcome variable in our analysis, allowing us to determine which attributes of children's books are most closely associated with higher reader satisfaction.

Together, these measured attributes offer a comprehensive framework to analyze which factors most influence the popularity of children's books. By taking real-world characteristics, such as the physical form and publication timing of a book, and translating them into well-defined, structured variables, we ensure that our dataset accurately represents elements that are believed to shape the appeal and success of children's literature.

2.3 Outcome Variables

Our analysis focuses on understanding the factors that contribute to higher ratings for children's books. The core outcome variable in this study is the **rating** of each book, which represents the overall popularity and perceived quality. In this section, we explore how ratings are distributed and visualize the relationships between ratings and different book attributes.

2.3.1 Ratings Distribution

The **ratings** variable represents the average score given to each book, ranging from 0 to 5. Ratings are a key indicator of a book's success and can reflect various dimensions, such as story quality, reader engagement, and production value.

Figure 1 shows the distribution of ratings across all books in the dataset. We observe that most books tend to have relatively high ratings, clustering around values between 3.5 and 4.5. This skew towards higher ratings may indicate a general satisfaction with the books or potential bias in how ratings are assigned.

2.4 Predictor Variables

In this section, we describe the predictor variables used in the analysis of children's book ratings. The predictor variables we considered include cover type, page range, and publication period. Each of these variables was analyzed to determine its association with the overall rating of children's books. Below, we present detailed descriptions, graphs, and tables for each predictor.

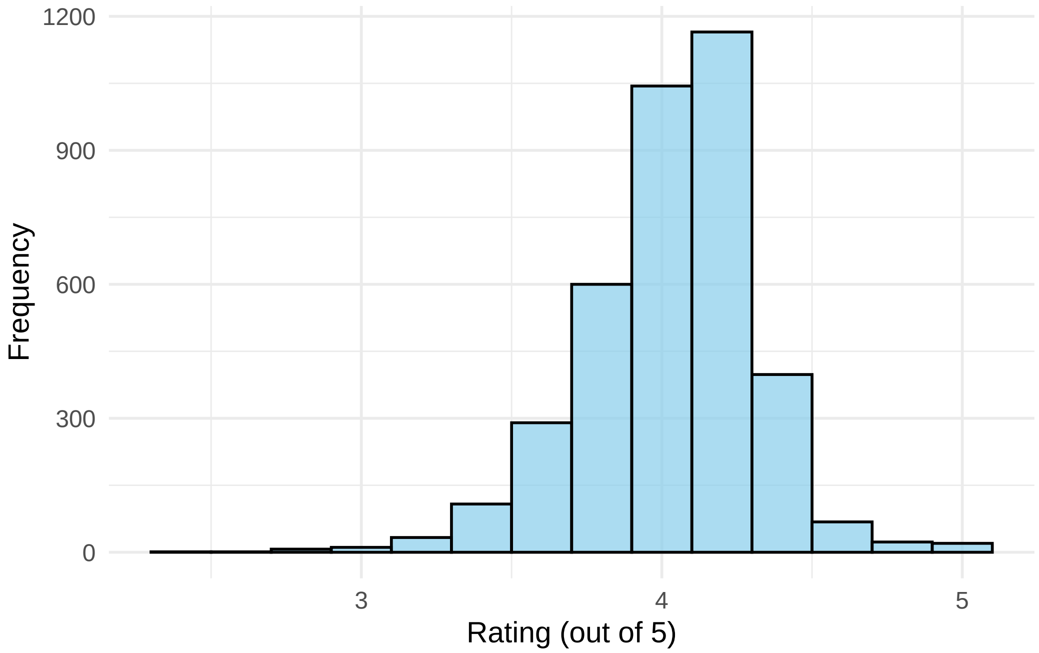

Figure 1: Distribution of Children's Book Ratings. Shows how the ratings of children's books are generally skewed towards higher values, indicating a general satisfaction with the literature available.

2.4.1 Rating by Cover Type

The **cover type** variable represents whether a book is in “Hardcover” or “Paperback.” It’s an important feature as the physical presentation of a book can influence reader perception and preferences. Here, we visualize the relationship between cover type and book ratings.

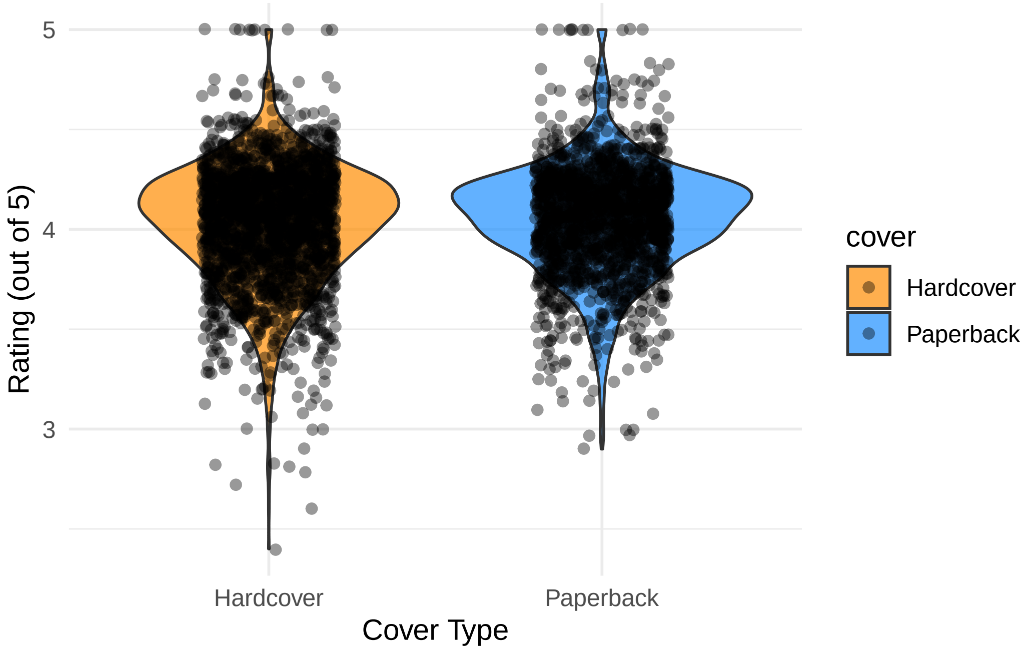

Figure 2: Relationship Between Book Ratings and Cover Type. Illustrates that hardcover books tend to receive slightly higher ratings compared to paperback editions, potentially due to their perceived durability and premium quality.

Figure 2 shows the distribution of ratings by cover type. Hardcover books generally tend to receive slightly higher ratings than paperback books, suggesting that readers may associate the physical sturdiness of hardcover books with higher quality. This could also reflect the fact that hardcovers are often positioned as premium products.

2.4.2 Rating by Page Range

Page range was categorized into five levels: “Under 50”, “50-149”, “150-299”, “300-499”, and “500 and above”. The goal was to determine if the length of a book influences its rating, as longer books might be perceived as offering a more immersive story.

Figure 3 illustrates the average ratings for different page ranges. It appears that books in the “150-299” range receive higher ratings on average, followed by books in the “300-499”

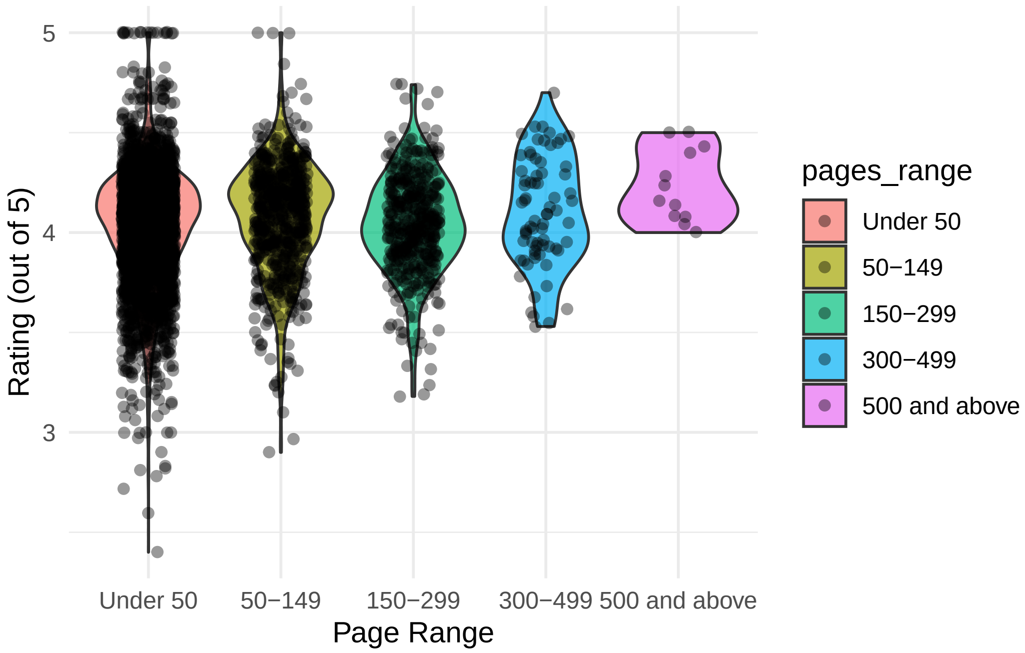

Figure 3: Average Ratings by Page Range. Indicates that books with 150-299 pages are more likely to receive higher ratings, potentially reflecting a balance between story depth and accessibility for children.

range. This could imply that readers prefer books of intermediate length, as they often strike a balance between offering enough content for engagement without being too lengthy for younger audiences.

2.4.3 Rating by Publication Period

Books published in different time periods may reflect changing styles, themes, and cultural relevance. The **publication period** variable categorizes books based on their original publication year.

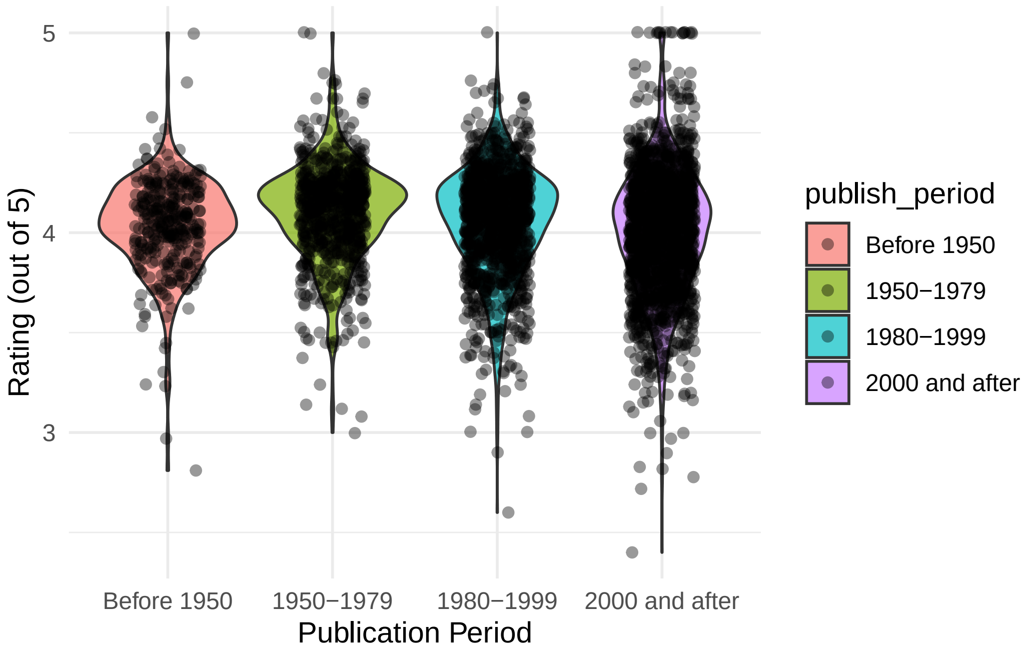

Figure 4: Average Ratings by Publication Period. Highlights that books published after 2000 tend to have higher ratings, possibly reflecting evolving preferences and improved production standards.

Figure 4 shows the average ratings for books from different publication periods. Books published in more recent years, particularly after 2000, tend to receive higher ratings. This could reflect improvements in the quality of children's literature, evolving reader preferences, or better production standards. Alternatively, it may indicate that recent books are simply more relevant to today's audiences.

2.5 Summary

Overall, we observe interesting patterns between book attributes and ratings. The physical characteristics of books, such as cover type and length, appear to play a role in how well they are rated. Similarly, the time of publication also seems to impact the perceived quality of a book.

3 Model

The goal of our modeling strategy is twofold. Firstly, we aim to quantify the relationship between book characteristics (such as cover type, page range, and publication period) and their respective ratings. Secondly, we seek to identify which of these characteristics are the most influential in driving higher ratings for children's books. This will help inform authors, publishers, and readers about what features may contribute to a book's popularity and perceived quality.

Here, we briefly describe the Bayesian analysis model used to investigate these relationships. We employed a Bayesian hierarchical linear model, which is particularly suitable for our study because it allows us to incorporate prior knowledge and account for potential variability between different book categories. This model provides more nuanced insights compared to classical statistical models, particularly when dealing with smaller datasets or datasets that might have underlying group structures.

The dependent variable in our model is the book rating (on a scale of 0 to 5), while the independent variables are the predictor variables discussed earlier: cover type, page range, and publication period. By using Bayesian modeling, we incorporate prior distributions for each of the coefficients, allowing for uncertainty in the estimates to be captured directly.

Background details and model diagnostics, including convergence checks and posterior predictive checks, are provided in Appendix - A. These diagnostics help ensure the robustness and reliability of our results, verifying that the model appropriately fits the data and that inferences drawn are trustworthy.

3.1 Model Assumptions and Limitations

The Bayesian hierarchical linear model makes several assumptions that should be considered when interpreting the results:

1. **Normality of Errors**: The model assumes that the residuals (errors) are normally distributed. This means that the difference between the predicted and observed values follows a normal distribution, which may not always hold, particularly if the data contains outliers or non-linear patterns.
2. **Linearity**: The relationship between the predictor variables (cover type, page range, and publication period) and the outcome variable (ratings) is assumed to be linear. If the true relationship is non-linear, the model may not accurately capture it.
3. **Homogeneity of Variance**: The model assumes that the variance of errors is constant across all levels of the predictor variables. Any deviation from this assumption (e.g., heteroscedasticity) may affect the reliability of the estimates.

4. **Independence of Observations**: The observations are assumed to be independent of one another. This means that each book rating is assumed to be independent, which might not always be the case if, for example, multiple ratings are influenced by a common author or publisher.
5. **Priors**: The priors used for the model are weakly informative, meaning that they have little influence unless the data strongly supports them. This helps to regularize the estimates but could still introduce some bias if the prior assumptions are incorrect.

3.1.1 Limitations

1. **Data Quality**: The quality of the model's results depends heavily on the quality of the data. Missing data, inaccuracies in reported ratings, or inconsistencies in book characteristics could introduce bias into the analysis.
2. **Limited Scope**: The model only considers a limited set of book attributes (cover type, page range, publication period). Other factors, such as author reputation, book awards, or illustrations, may also influence ratings but are not included here.
3. **Causality**: The model identifies associations between book characteristics and ratings but does not establish causation. Higher ratings may be linked to certain characteristics, but it is not necessarily true that changing these characteristics will directly improve a book's rating.
4. **Posterior Convergence**: The reliability of Bayesian inference depends on the convergence of the Markov chains. If the chains do not converge well, the inferences drawn may be unreliable.

3.2 Model Set-up

Define y_i as the average rating for book i. Then η_1 represents the effect of cover type, η_2 represents the effect of page range, and η_3 represents the effect of publication period.

$$\begin{aligned}
y_i \mid \mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha + \eta_1(\text{CoverType}_i) + \eta_2(\text{PageRange}_i) + \eta_3(\text{PublicationPeriod}_i) && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\eta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\eta_2 &\sim \text{Normal}(0, 2.5) && (5)\\
\eta_3 &\sim \text{Normal}(0, 2.5) && (6)\\
\sigma &\sim \text{Exponential}(1) && (7)
\end{aligned}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package (Team 2023). We use the default priors from `rstanarm`.

3.2.1 Model Justification

We expect a positive relationship between certain book characteristics and higher ratings. Specifically, hardcover books may receive higher ratings due to their perceived durability and premium quality. Medium-length books may be preferred for their balance between story depth and accessibility. Books published more recently may receive higher ratings, potentially reflecting changes in reader preferences or improved production standards.

4 Results

The primary goal of our analysis was to determine the factors that influence the ratings of children's books. Below, we summarize the findings from the Bayesian hierarchical linear model and interpret their practical implications for understanding the popularity of children's books.

4.1 Summary of Model Results

The summary statistics for our model parameters, presented in the appendix (Appendix - Section A), highlight key insights about the influence of book attributes on ratings.

4.1.1 Summary Statistics

To provide context for the model results, we present some key summary statistics of the dataset used in the analysis. Table 1 includes the average rating, distribution of book types, and the breakdown of page ranges and publication periods.

Table 1: Summary statistics of the children's book dataset, showing the distribution of key metrics such as average rating, book cover types, page ranges, and publication periods.

Metric	Value
Average Rating	4.04
Number of Hardcover Books	2053.00
Number of Paperback Books	1716.00
Page Range: 150-299	413.00
Page Range: 300-499	71.00
Page Range: 50-149	608.00
Page Range: 500 and above	12.00
Page Range: Under 50	2665.00
Publication Period: 1950-1979	697.00
Publication Period: 1980-1999	1122.00
Publication Period: 2000 and after	1643.00
Publication Period: Before 1950	307.00

4.1.2 Model Summary

The model included predictors such as **cover type** (hardcover vs. paperback), **page range**, and **publication period**. From the posterior summaries, we can infer several important relationships:

1. **Cover Type**: Books with a hardcover format tend to have slightly higher ratings than paperback editions. The estimated effect size is positive, indicating a preference for hardcover, which may be related to perceived durability and quality.
2. **Page Range**: Books with more pages, specifically in the range of **150-299 pages**, show a positive association with higher ratings. This finding implies that books with more substantial content are generally favored, though extremely lengthy books (over 500 pages) do not necessarily receive better ratings.
3. **Publication Period**: Books published in more recent periods, particularly after **2000**, are associated with higher ratings. This suggests a shift in consumer preferences towards modern content or improvements in book quality in recent decades.

These results are illustrated in Figure 8, where each plot represents the posterior distributions of the effects of cover type, page range, and publication period on ratings.

4.1.3 Regression Results Table

The regression table Table 2 provides the estimated effects of each predictor variable on children's book ratings.

4.2 Interpretation of Results

These findings suggest that certain characteristics of children's books are associated with varying levels of consumer appreciation, as reflected by their ratings. For instance, the preference for **hardcover editions** may indicate that buyers value quality and durability in children's books, possibly because these books are meant to be handled repeatedly by young readers.

The association with **page range** suggests that books providing moderate-length stories (150-299 pages) are highly valued. This range may offer enough depth for engaging storytelling without overwhelming younger readers.

Lastly, the preference for books published **after 2000** may indicate evolving preferences in children's literature. This could be due to the integration of more diverse themes and characters, better illustrations, or more modern narratives that resonate well with today's audiences.

These findings could help **authors**, **publishers**, and **parents** understand what characteristics are linked to better reception in children's literature, ultimately guiding decisions on book production, marketing, and selection.

5 Discussion

5.1 Factors Influencing Children's Book Ratings

In our study, we identified several key factors that influence children's book ratings. The **cover type**, **page range**, and **publication period** were found to significantly impact how books are rated. This section will delve deeper into these factors, examining why they may play a crucial role in determining a book's popularity.

The preference for **hardcover books** can be attributed to their perceived durability and quality. Since children's books are often read repeatedly, a hardcover format might offer more resilience, making it a preferred choice among parents and educators. This preference highlights an important aspect of consumer behavior when it comes to children's literature—practicality and longevity of the product are crucial.

The association between **page range** and ratings is another significant finding. Books in the **150-299 pages** range received higher ratings, suggesting that consumers value content that strikes a balance between being engaging and not overly lengthy. This page range might provide just the right amount of depth for storytelling, making the books suitable for young readers without overwhelming them. The implication here is that content length plays a critical role in the perceived quality and enjoyment of children's books.

Our analysis also revealed that books published **after 2000** tend to have higher ratings. This finding may be indicative of evolving tastes and preferences in children's literature, as well as improvements in book quality over the years. It is possible that newer books incorporate more diverse themes, modern storytelling techniques, and better visual appeal, all of which resonate well with today's readers. The shift in consumer preferences towards newer books could also be related to changes in societal values and expectations. For instance, children's literature has seen a growing emphasis on inclusivity and representation in recent years. Books that reflect diverse experiences and characters may be more likely to receive higher ratings, as they provide young readers with stories they can relate to and learn from. This trend underscores the importance of staying relevant and adapting to the changing landscape of children's literature.

5.2 Practical Implications for Authors and Publishers

The insights gained from this study have practical implications for **authors**, **publishers**, and **booksellers**. Understanding that **hardcover formats** are generally preferred can help publishers decide on the most marketable format for their books. Additionally, targeting the **150-299 pages** range may increase the likelihood of positive reception, as it appears to be the sweet spot for engaging yet manageable content.

For authors, being mindful of the themes and topics that resonate with modern audiences is crucial. Books that reflect contemporary issues, promote inclusivity, and feature diverse characters are more likely to receive favorable ratings. Publishers can also use this information to guide their acquisition and production decisions, ensuring that the books they bring to market align with current consumer preferences.

Ultimately, the findings of this study contribute to a better understanding of what makes children's books popular and well-rated. By considering factors such as **cover type**, **page range**, and **publication period**, stakeholders in the children's literature market can make informed decisions to enhance the quality and appeal of their offerings.

5.3 Weaknesses and Next Steps

While our study provides valuable insights, there are some limitations that should be acknowledged. One of the main weaknesses is that our dataset may not fully represent all children's books, especially those published outside the mainstream market or those that are self-published. The analysis is based on a subset of available books, which could introduce biases in our findings.

Another limitation is the lack of detailed reader demographics. Ratings can be influenced by the age, preferences, and reading habits of the reviewers, which were not accounted for in our model. Future research could incorporate demographic information to better understand how different groups perceive children's books.

In terms of next steps, expanding the dataset to include more books, especially those from diverse publishers and formats, would help improve the robustness of the analysis. Additionally, exploring other features, such as the presence of illustrations, themes, or author reputation, could provide a more comprehensive understanding of the factors influencing book ratings.

Finally, incorporating qualitative data, such as reader reviews, could help capture aspects of children's books that are not easily quantifiable but still contribute to their popularity and success.

A Appendix: Model Details

A.1 Regression Model Summary

Table 2 summarizes the regression results, showing the estimated coefficients, median absolute deviations, and credible intervals for each predictor.

A.2 Convergence Diagnostics

To ensure that the Bayesian hierarchical linear model has converged properly, we used multiple convergence diagnostics. The primary convergence diagnostic used is the **Gelman-Rubin statistic** (R-hat). The R-hat value should ideally be close to 1, indicating that the Markov chains have mixed well and reached convergence. Any R-hat value significantly above 1 suggests non-convergence and needs further investigation. In Figure 5 , we present the R-hat values for each model parameter.

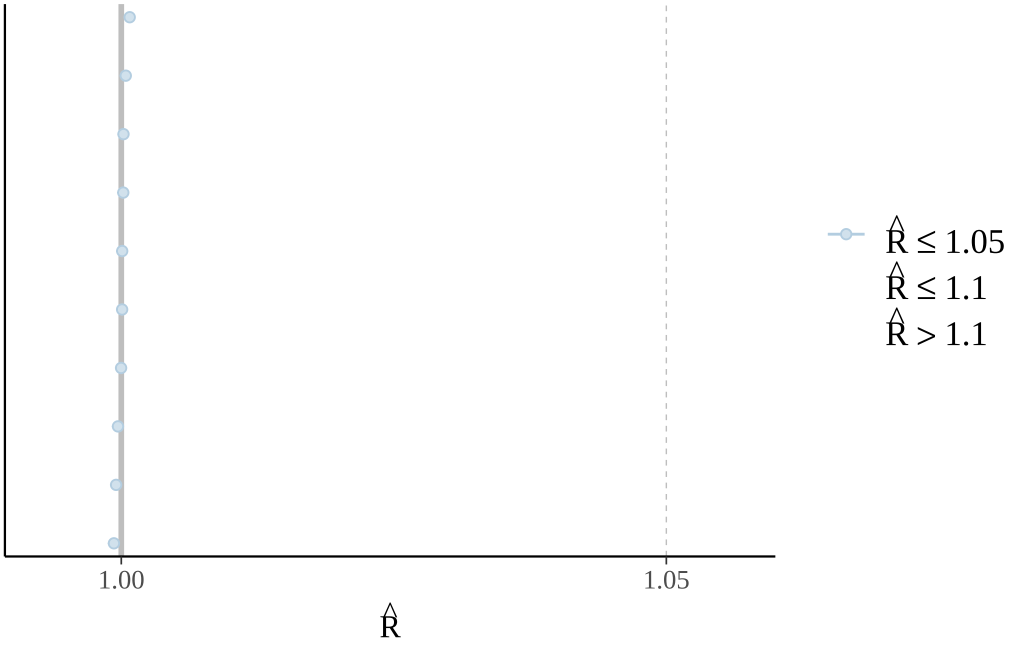

Figure 5: Gelman-Rubin convergence diagnostic (R-hat) values for each model parameter. All values are close to 1, indicating that the Markov chains have mixed well and achieved convergence.

Table 2: Regression model summary for predicting children's book ratings, including estimated coefficients and credible intervals.

	Children's Book Model
(Intercept)	4.09
	(0.02)
coverPaperback	0.00
	(0.01)
pages_range300-499	0.07
	(0.04)
pages_range50-149	0.03
	(0.02)
pages_range500 and above	0.20
	(0.08)
pages_rangeUnder 50	−0.01
	(0.02)
publish_period1980-1999	−0.04
	(0.01)
publish_period2000 and after	−0.09
	(0.01)
publish_periodBefore 1950	−0.07
	(0.02)
Num.Obs.	3769
R2	0.022
R2 Adj.	0.016
Log.Lik.	−607.569
ELPD	−616.0
ELPD s.e.	59.8
LOOIC	1232.0
LOOIC s.e.	119.6
WAIC	1232.0
RMSE	0.28

A.2.1 Trace Plots

Another convergence diagnostic involves examining trace plots. Trace plots show the sampled values of each parameter across iterations for each Markov chain. Well-mixed chains with no clear trends suggest that the model has converged. In Figure 6 , we observe the trace plots for key parameters in the model.

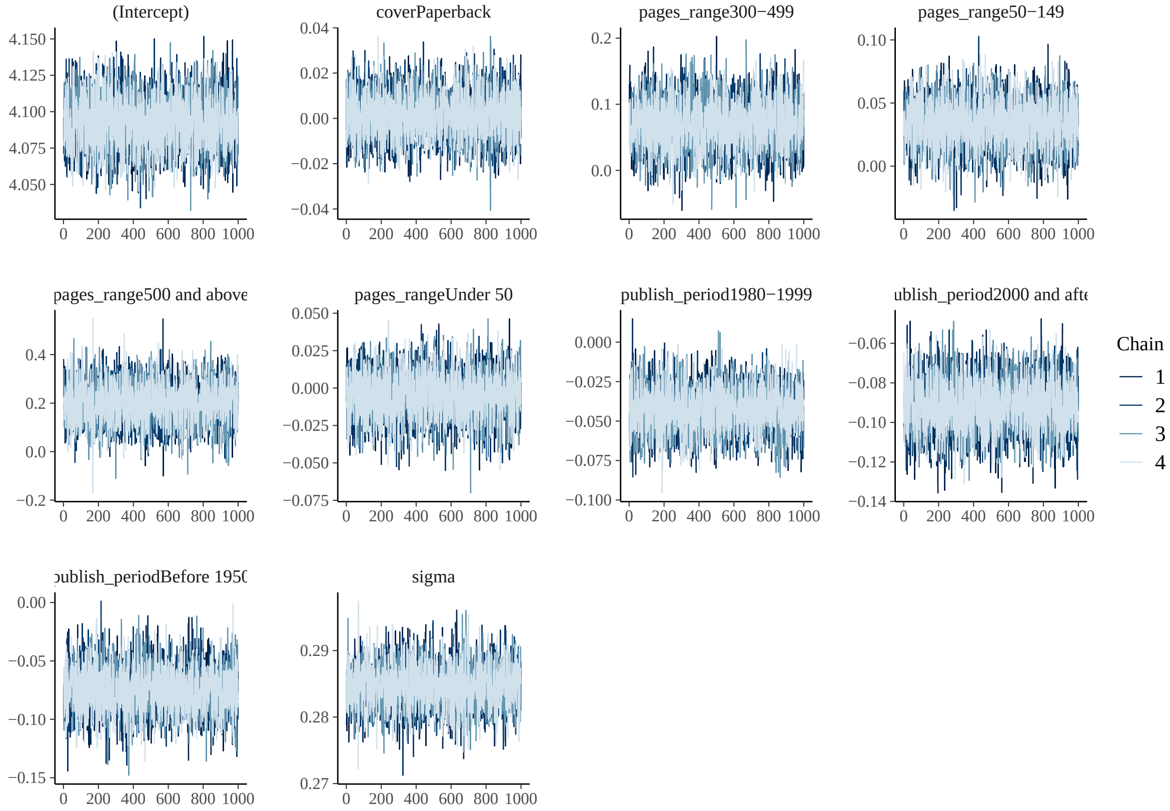

Figure 6: Trace plots for key parameters across the Markov chains, showing well-mixed chains with stable values. The lack of discernible patterns or trends confirms that convergence has been reached.

A.3 Posterior Predictive Checks

Posterior predictive checks are used to assess the goodness-of-fit of the Bayesian model. By simulating data from the posterior distribution, we can compare the simulated data to the observed data to determine if the model adequately represents the data. Figure 7 illustrates the results of posterior predictive checks for the Bayesian model.

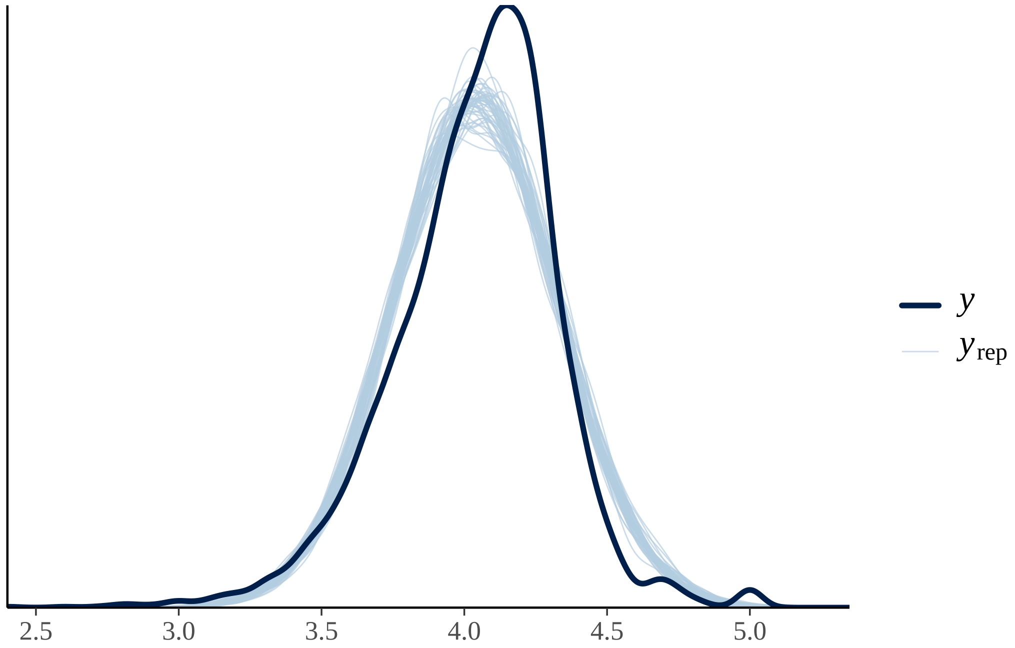

Figure 7: Posterior predictive checks showing a comparison between the observed data and simulated data generated from the model's posterior distribution. This helps evaluate whether the model adequately captures the key features of the observed data.

A.4 Summary Statistics for Posterior

Table 3 is a summary table of the key parameters in the model, including the estimated means, standard deviations, and 95% credible intervals.

Table 3: Summary statistics of the posterior distributions for key model parameters, including posterior means, standard deviations, and 95% credible intervals, which provide insights into the central tendencies and uncertainties of the parameter estimates.

term	estimate	std.error	conf.low	conf.high
(Intercept)	4.0930831	0.0182305	4.0577700	4.1289811
coverPaperback	0.0013643	0.0101316	-0.0180799	0.0212842
pages_range300-499	0.0696590	0.0366950	-0.0036223	0.1410696
pages_range50-149	0.0335828	0.0184475	-0.0027091	0.0692657
pages_range500 and above	0.1999264	0.0847749	0.0350992	0.3628462
pages_rangeUnder 50	-0.0056287	0.0155791	-0.0370709	0.0249475
publish_period1980-1999	-0.0422176	0.0139742	-0.0696154	-0.0144070
publish_period2000 and after	-0.0910799	0.0130185	-0.1168425	-0.0652273
publish_periodBefore 1950	-0.0747640	0.0195160	-0.1117509	-0.0359537
sigma	0.2845396	0.0032558	0.2782479	0.2909165

A.4.1 Posterior Density Plots

Posterior density plots provide a visual representation of the uncertainty in the model parameters. Figure 8 show the distribution of possible values for each parameter, which helps in understanding the spread and central tendency.

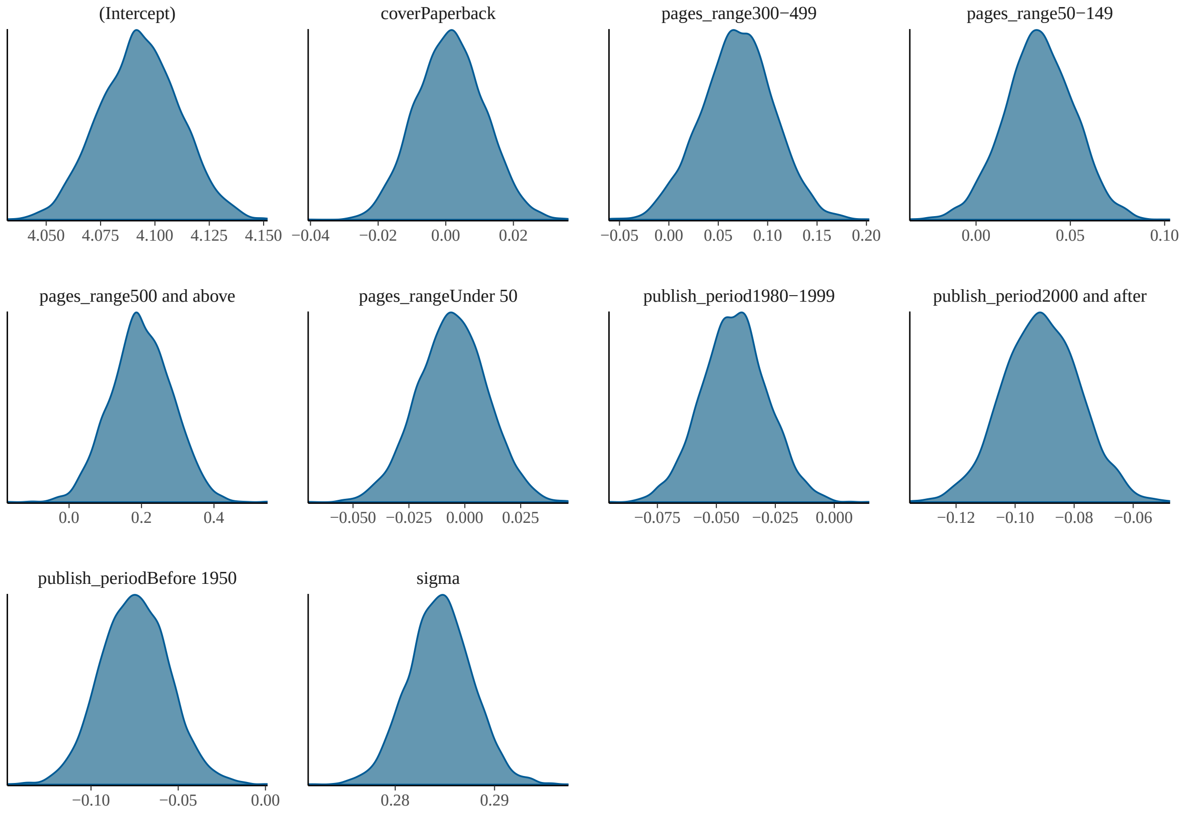

Figure 8: Posterior density plots for each model parameter, providing a visual representation of the uncertainty in the parameter estimates. The density curves illustrate the range and central tendency of plausible values derived from the data and priors.

B Appendix: Survey and Sampling Methodology in Children's Book Ratings Analysis

B.1 Overview of Children's Book Ratings Dataset and Its Significance

The dataset(Cookson 2023) used in this study, focused on children's book ratings, plays an important role in understanding the elements that make children's literature more appealing. The data, collected from online ratings and reviews, provides valuable insights into what readers—often parents or educators—find most engaging or high-quality in children's books. However, being sourced from an open online platform, the dataset carries inherent limitations, particularly concerning the sampling methodology and data collection biases.

B.2 Data Collection Challenges and Bias

The ratings data are observational in nature and were gathered without a structured survey framework. Instead, data collection relied on voluntary participation by readers rating and reviewing books online. This voluntary nature of participation presents several challenges, primarily due to the self-selection bias—those who choose to leave reviews may differ systematically from those who do not. For instance, individuals who provide ratings might be particularly passionate about children's books, leading to an overrepresentation of extreme opinions (either very positive or negative). This limits the ability to generalize findings to the broader population of book readers.

Non-response bias is another critical issue in observational datasets like ours. To address this, future studies could implement weighting adjustments to correct for underrepresented groups. For instance, if younger readers or readers from specific socioeconomic backgrounds are less likely to provide ratings, their responses could be weighted more heavily to reflect their actual proportion in the reader population. This technique would help adjust for disparities in participation rates and lead to more accurate estimations of the factors driving book ratings.

B.3 Stratified Sampling as an Ideal Approach

In an ideal scenario, a stratified sampling methodology would be employed to ensure the diversity of respondents, capturing the full spectrum of reader demographics. Stratified sampling involves dividing the population into subgroups (or strata) based on specific characteristics, such as age, reading frequency, and geographic location, and then sampling from each subgroup. This would ensure that the sample is representative of all reader groups, not just those who are the most vocal online. Such an approach would mitigate the risk of over-representing highly engaged readers and help provide a more balanced view of the factors influencing book ratings.

B.4 Limitations of Observational Data

The challenges associated with observational data are well-documented in the literature. Rubin discusses the difficulty of drawing causal conclusions from observational studies without experimental control (Rubin 1974). Cochran also highlights the limitations of survey data, noting the importance of understanding biases when interpreting results (Cochran 1977). Given these challenges, it is crucial to interpret the findings from our study with caution, acknowledging that correlations between book attributes and ratings do not imply causation.

B.5 Recommendations for Future Research

For future research, we recommend the development of a more structured survey methodology, incorporating stratified sampling and weighting adjustments to improve the representativeness of the dataset. Additionally, including more detailed questions regarding reader motivations, reading habits, and book preferences could provide deeper insights into the factors influencing ratings. Such improvements would not only enhance the quality of the data but also make the findings more actionable for publishers, authors, and educators interested in understanding and improving children's literature.

C Appendix: Datasheet for the Children's Books Ratings Dataset

C.1 Dataset Overview

The Children's Books Ratings Dataset contains metadata about children's books, focusing on attributes such as cover type, page count, publication period, and ratings. This dataset was derived from a raw dataset(Cookson 2023) originally created by Alex Cookson and subsequently cleaned and processed by Zien Gao for analysis. The main goal of the dataset is to provide insights into the factors that contribute to higher book ratings, which can help stakeholders like publishers, authors, and parents in selecting and promoting well-received children's books.

C.2 Motivation

The dataset was created to fill a gap in the understanding of which specific features of children's books are most strongly associated with higher ratings. Despite the vast amount of children's literature available, there has been limited empirical research on which attributes, such as cover type or page length, contribute to a book's popularity. By analyzing this data, stakeholders can make informed decisions that align with reader preferences.

C.3 Intended Use

This dataset can be used for:

- **Publishers and Authors**: To understand market preferences and make data-driven decisions to enhance the appeal of children's books.
- **Parents and Educators**: To identify books that are more likely to be well-received by children.
- **Researchers**: For academic purposes and to explore patterns in children's book ratings.

C.4 Composition

The dataset consists of individual entries, with each row representing a children's book. Key features include:

- **Cover Type**: Indicates whether the book is a hardcover or paperback.
- **Pages Range**: Number of pages, categorized into: "Under 50", "50-149", "150-299", "300-499", and "500 and above".
- **Publication Period**: Original publication year categorized as: "Before 1950", "1950-1979", "1980-1999", "2000 and after".

- **Rating**: A numerical value between 0 and 5, representing the average reader rating.

C.5 Collection Process

The original dataset(Cookson 2023) was compiled by Alex Cookson and made available on GitHub. The data was manually curated and then processed to focus on key features that influence book ratings. The cleaned dataset retains all relevant information needed for effective analysis while discarding unnecessary or redundant fields.

C.6 Preprocessing and Cleaning

The data was cleaned and preprocessed by Zien Gao using R(R Core Team 2023), utilizing libraries such as `tidyverse`(Wickham 2023b) and `arrow`(Richardson, Korn, et al. 2023). The cleaning process involved categorizing variables, handling missing values, and ensuring consistency. The dataset underwent several preprocessing steps to ensure data quality:

- **Handling Missing Values**: Entries with missing or invalid values were removed.
- **Categorization**: The **Pages Range** and **Publication Period** attributes were categorized for easier analysis.
- **Data Storage**: The cleaned dataset is stored in `.parquet` format for efficient storage and use.

C.7 Potential Biases and Limitations

- **Sampling Bias**: The dataset may reflect biases inherent in the original data collection, such as an overrepresentation of popular or well-marketed books.
- **Limited Attributes**: Only a limited set of book attributes are included, which may not capture all factors influencing book ratings. Features like genre or author popularity are absent.
- **Subjectivity of Ratings**: Ratings are subjective and may vary widely among different readers, making it challenging to draw definitive conclusions.

C.8 Ethical Considerations

- **Privacy**: The dataset does not contain any personal or identifiable information, ensuring that there are no privacy concerns.
- **Responsible Use**: Users should avoid making broad generalizations from the dataset, as ratings are subjective and may not represent all reader demographics.

C.9 Maintenance

- **Updates**: There are currently no plans for updates, but future versions could include more attributes or expanded datasets to improve the analysis.

C.10 Structure and Access

- **File Format**: The dataset(`clean_books_data.parquet`) is saved in `.parquet` format.
- **Access**: Available in the GitHub repository associated with this project, under the folder `data/analysis_data/`.

C.11 Future Improvements

- **Expanded Features**: Future iterations could include additional attributes like book genre, marketing data, or author information to provide deeper insights.
- **Comprehensive Data Collection**: Collecting a broader set of data from diverse sources would provide a more complete understanding of factors that drive book ratings.

References

Arel-Bundock, Vincent. 2023. *Modelsummary: Summary and Table of Statistical Models and Data*. https://CRAN.R-project.org/package=modelsummary.

Bolker, Ben, and Davis Vaughan. 2023. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Cochran, William G. 1977. *Sampling Techniques*. John Wiley & Sons.

Cookson, Timothy A. 2023. "Children's Book Ratings Dataset." https://github.com/tacookson/data/tree/master/childrens-book-ratings.

Gabry, Jonah, and Matthew Kay. 2023. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot/.

Müller, Kirill, and Jennifer Bryan. 2023. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Plummer, Martyn, Nicky Best, Kate Cowles, and Karen Vines. 2006. *Coda: Output Analysis and Diagnostics for MCMC*. https://CRAN.R-project.org/package=coda.

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, Uwe Korn, et al. 2023. *Arrow: Integration to Apache Arrow*. https://CRAN.R-project.org/package=arrow.

Robinson, David, and Alex Hayes. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Rubin, Donald B. 1974. *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*. Journal of Educational Psychology.

Team, Stan Development. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Wickham, Hadley. 2023a. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

———. 2023b. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'Kable()' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.