Fighting Fire with Fire

Using Fahrenheit 451 Logic to Analyze Literary Differences in Goodreads Ratings Across School-Banned Books

Nikita Jayaprakash

2025-12-10

1 Introduction

To quote Ray Bradbury's "Fahrenheit 451", "A book is a loaded gun in the house next door" (Bradbury (1951)). Ironically, it seems that an increasing number of schools across the United States has taken this literary advice to heart. PEN America keeps track of school book bans, and they reported 4,231 unique titles that were banned between July 1, 2023 to June 30, 2024; note that this encapsulates 10,046 instances of individual books banned across different schools in the United States (PENAmerica (2024)). We will use only a subset of these books during the same time period. This analysis is interested in comparing the Goodreads descriptions for books that are "more banned" (i.e. banned in more than 1 state within a single year) and books that are "less banned" (i.e. books banned in 1 state within a single year). In particular, this analysis explored the following questions:

- How do the literary features differ between the Goodreads descriptions of more-banned and less-banned books?
- What are the keywords that differ between the Goodreads descriptions of more-banned and less-banned books?
- Do the frequencies of these keywords and/or the occurrence of these features differ based on when they were banned within the year?

Using Goodreads descriptions rather than the full texts of the novels allows for easier access (especially with more recent books that aren't freely available). In addition, the use of Goodreads descriptions rather than the actual books allows for a focus on a larger breadth of banned books and is not restricted by copyright laws. Admittedly, this unfortunately partakes in another unfortunate logic highlighted in *Fahrenheit 451*, which is using summaries of books rather than the full text of the book (Bradbury (1951)). However, it is easier to assess topics of a book from the Goodreads description since these descriptions will mention general narrative details more explicitly rather than the subtlety in the actual novel. For this reason, we hypothesize that there will be literary differences and differences in keywords across more-banned and less-banned books.

2 Data

This analysis uses web-scraped Goodreads descriptions (and other variables like publication date, average rating, number of ratings) of a subset of the books that were banned between July 1, 2023 to June 30, 2024, based on the list published by PEN America (PENAmerica (2024)). As shown in Table 1, only 2713 descriptions – 2149 are "less banned" descriptions and 564 are "more banned" descriptions – out of 4239 listed on PEN America were successfully scraped, so this is still a convenience sample. This limits the possible generalizations of the data because we are focusing on the books banned within a year. This analysis uses web-scraped Goodreads descriptions (and other variables like publication date, average rating, number of ratings) of a subset of the books that were banned between July 1, 2023 to June 30, 2024, based on the list published by PEN America (PENAmerica (2024)).

Table 1: Here's a summary of the corpus. Note that there are fewer more-banned books.

Text_Type	Texts	Tokens
LessBanned	2149	186583
MoreBanned	564	51309

2.1 Florida Man Bans Books

From Table 2, it's evident that Florida has banned a vast majority of the books in this dataset. For this reason, this analysis will also focus on the dates of those book bans in order to see if there is any potential association between the dates of those book bans and any newsworthy events at that time in the United States. In addition to the statistical computations that are explained in the *Methods* section, this analysis will also conduct a qualitative investigation of the political atmosphere and legislation that could be related to the trends in banned books in Florida, both based on quantity – such as the spike in October 2023 in Figure 1 – and in linguistic trends identified in the *Results* section.

3 Methods

This analysis will use multi-dimensional factor analysis of the Biber features in order to assess if there are any distinct literary differences between more-banned and less-banned books. This will be accomplished by analyzing the factor loadings of the MDA model and seeing how those factors are distributed for more-banned and less-banned books (Biber (1992)). For comparison, this analysis will also view the distribution of these factors across the average Goodreads ratings for these books to see if there are certain features that indicate higher or lower public opinion of these descriptions. Note that the people who rate books on Goodreads are likely to also read the Goodreads descriptions, which indicates the relevance of comparing the MDA factor distribution across different average ratings. Finally, two F-tests are used to determine if the two MDA factors are independent across the more-banned and less-banned

Table 2: Top five states that have banned the most books in this dataset

State	# Banned Books
Florida	2099
Iowa	727
Wisconsin	275
Texas	238
Virginia	77

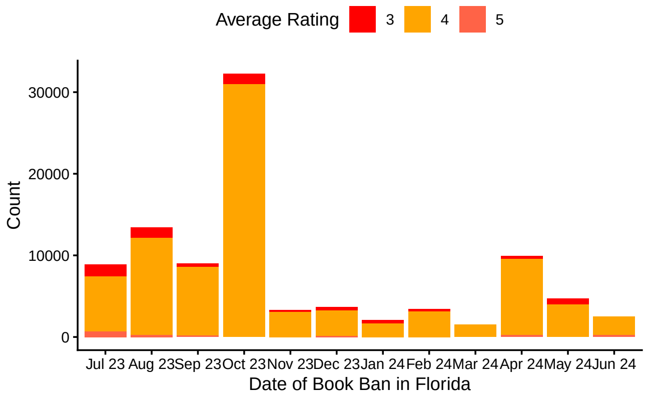


Figure 1: Florida had the most book bans completed in October 2023.

books at a significance level of $\alpha = 0.05$ with a Bonferroni correction for the two tests. Two additional F-tests are used to determine if the two MDA factor are independent across the Goodreads ratings at a significance level of $\alpha = 0.05$ with a Bonferroni correction for the two tests.

In addition, this analysis will use keyness – after dropping stopwords – in order to determine which words are keywords with the more-banned books as the target corpora and the less-banned books as the reference corpora (Brezina (2018)). One of the main benefits of using Goodreads descriptions is that these keywords may indicate controversial topics – such as "queer", "gay", "sex", "violence", etc. – that may have led to the book bans. According to Smith (2010), a "key keyword" is a keyword that is "key" in more than one of the related texts: after calculating the keyness values of the target corpus against the reference corpus, we only keep the texts that reach a significance threshold of 5%. We arrange these words in descending order based on their key range, which is the percent of texts in the target corpus for which keyness reaches the specified threshold. These are our key-keywords. Of the top 10 keywords, we will explore the frequency of the words with an effect mean over 3.8. Note that the effect means are the mean effect size by log ratio, which would yield key-keywords that are largely different between the target and reference corpora. Then, this analysis will compare the frequency of these words across the timeline of the book bans in Florida. We are using Florida as a case study because Florida accounts for a majority of the book bans in this dataset. This can guide further research into why these book bans were implemented at that specific time by investigating the social and political events at the time.

4 Results

4.1 Multi-Dimensional Factor Analysis of Biber Features

The screeplot determines that two factors are sufficient to account for a large portion of the variation in these description (see Figure 4). We reject the F-test (F(1, 2697) = 13.84, p < 0.001) for the null hypothesis that the groups of more-banned vs. less-banned books are independent of their dimension scores for factor 1 at a significance of 2.5%. We reject the F-test (F(1, 2697) = 8.48, p < 0.001) for the null hypothesis that the groups of more-banned vs. less-banned books are independent of their dimension scores for factor 2 at a significance of 2.5%. We reject the F-test (F(3, 2695) = 6.2, p < 0.001) for the null hypothesis that the groups of different Goodreads ratings are independent of their dimension scores for factor 1 at a significance of 2.5%. We reject the F-test (F(3, 2695) = 4.81, p < 0.001) for the null hypothesis that the groups of different Goodreads ratings are independent of their dimension scores for factor 2 at a significance of 2.5%. In response to the first research question, this shows that there are literary differences between more-banned and less-banned books.

Figure 2 shows that factor 1 is positively correlated with usage of third-person pronouns, the present tense and private verbs. Factor 1 is also negatively correlated with the usage of larger words, and Factor 1 yields high scores for more-banned books and low scores for less-banned books. In addition, moderate average Goodreads ratings – scores of 3 or 4 – have high dimension scores, like more-banned books, and more extreme Goodreads ratings – scores of 2 or 5 – have lower dimension scores, like the less-banned books.

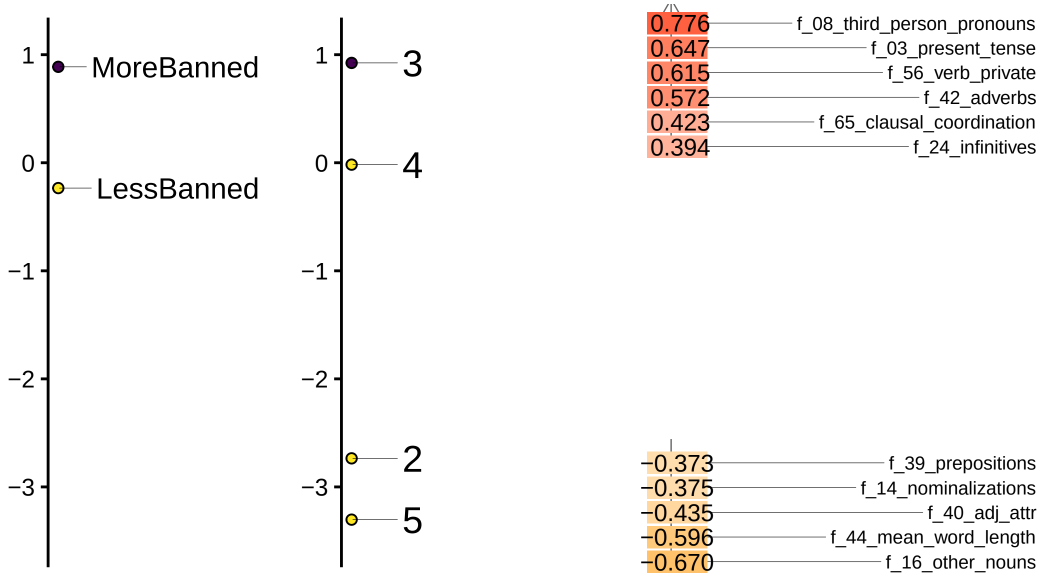


Figure 2: While both types of factors are statistically significant, the factor for Goodreads ratings has a larger spread than the factor for more-banned vs. less-banned.

Figure 3 shows that Factor 2 is positively correlated with usage of *be* as a main verb, predicative adjectives, and *it* as a pronoun. Factor 1 yields high scores for more-banned books and low scores for less-banned books. In addition, books with lower Goodreads ratings have higher dimension scores for factor 2, and books with higher Goodreads ratings have lower dimension scores.

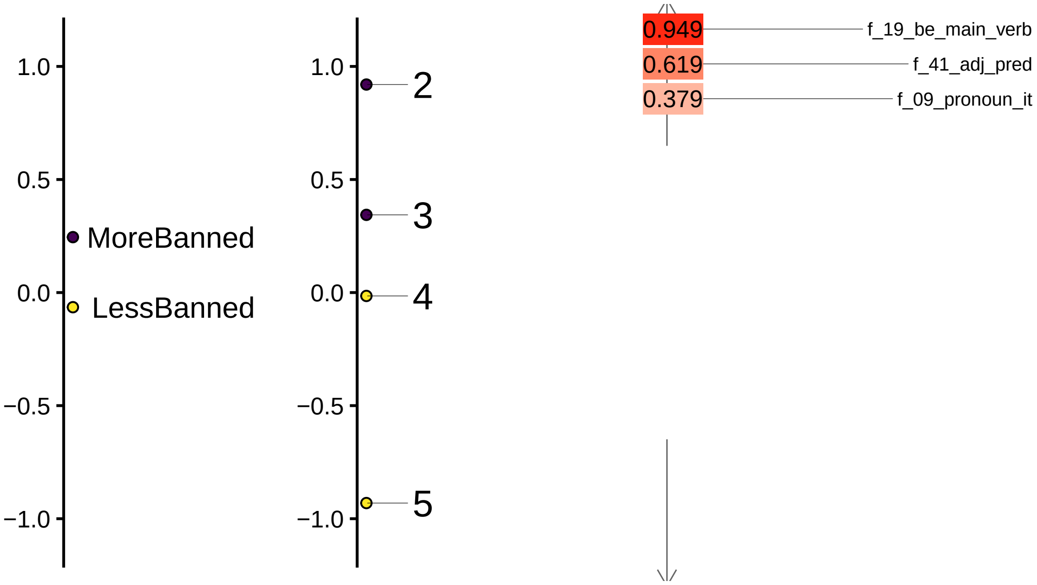


Figure 3: There aren't any Biber features with which Factor 2 is strongly negatively correlated.

4.2 Keyness

The descriptions for more-banned books are used as the target corpora, and the descriptions for less-banned books are used as the reference corpora. Both corpora are thresholded by a minimum frequency of 1. We sort by the top 10 key keywords based on the percent of texts in the target corpus for which keyness reaches the specified threshold of 0.05 (Table 4). In response to the second research question, these are the keywords that differ between more-banned and less-banned books. Note that we removed tokens that were incorrectly parsed, such as *s*, *n't* and *isbn*. From this subset, the highlighted words have an effect mean, which is the mean effect size by log ratio, greater than 3.8 (Table 3). The frequencies of these words will be explored in the descriptions of banned books in Florida.

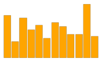
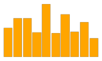
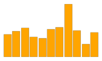
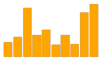
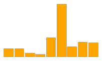
4.3 Florida Frequency Distributions

Now, from the subset of words determined above, we will analyze the distributions for lemmas in Florida banned books across the months between July 2023 until June 2024. These distributions are scaled by the total number of tokens in each month in order to account for the non-uniform distribution of banned books in Florida (see Figure 1).

Table 3

Token	Key Range	Effect Mean	Token	Key Range	Effect Mean
cover	6.560284	4.077670	**leave**	5.141844	3.823049
school	6.205674	1.349202	name	4.964539	3.791507
edition	5.673759	3.648773	maybe	4.787234	3.757955
question	5.319149	4.464849	alone	4.787234	3.766504
unforgettable	5.319149	3.933075	**gay**	4.609929	3.874688

All of the effect means are positive, which indicates that these key words have higher frequency in the target corpora (i.e. the more-banned books).

The distribution of the lemmas related to *cover* –  – and the lemmas related to *unforgettable*–  – seem fairly uniform across the different months. The distribution for the lemmas of *leave* –  – have a slight mode around March 2024. The distribution for the lemmas of *question* –  – seems bimodal, with a spike in August 2023 and June 2024. The frequency distribution for the lemmas of *gay* –  – is the most extreme with a large spike in January 2024. This addresses the third research question regarding how the frequency of these keywords differ throughout the year.

5 Discussion

More-banned books have higher dimension scores on Factor 1 of the MDA analysis, which are positively correlated with third-person pronouns and private verbs. These linguistic features are likely used to describe characters and their action because third-person pronouns refer to people and private verbs indicate how people feel. It's possible that more-banned books tend to be about the experiences of people, such as memoirs and fictional stories about marginalized groups, which leads to their banned status. Books with moderate Goodreads ratings also have higher dimension scores for Factor 1. Perhaps, this means that these character-centric books get more varied responses from the public because they indulge in niche, identity-related stories. More-banned books and books with low Goodreads ratings also have higher dimension scores on Factor 2, which is positively correlated with using *be* as the main verb and predicative adjectives. This kind of language is used to describe nouns. Perhaps this is present in the descriptions of banned books because it explicitly describes a character, which makes it easier to identify the topics of the book (and consequentially easier to find reasons to ban the book). In addition, this could explain the lower Goodreads ratings because it's too blunt without the nuisance of less explicitly descriptive language.

From the keyness analysis, there are keywords that differ between the Goodreads descriptions of more-banned and less-banned books, such as *cover*, *leave*, *unforgettable*, *question* and *gay*. It's easier to intuit why certain words may have increased prevalence as keywords with higher frequency in the target corpora of more-banned books. For example, while not one of the keywords used for the Florida frequency analysis, the prevalence of *edition* may be because certain banned books are older and have had many iterations published, such as *Handmaid's Tale*. In addition, the frequency distributions for *gay* across the Florida banned books reflects the general controversy around books about sexuality, particularly in schools. This answers the third research question that some of the frequencies of these keywords do differ based on when they were banned within the year. An interesting next step would be to use clustering to see if some of these words have a higher prevalence to a statistically significant degree within certain parts of the year.

Since this focuses on the banned books within a single year, it would be harder to generalize the results to additional years. In addition, this is only a subset of the total books that were banned within the time-period of July 2023 and June 2024 because we could not successfully scrape the Goodreads descriptions for all books in PEN America's list (PENAmerica (2024)). We cannot determine if there is any confounding between the included and excluded banned books. In addition, the dates for when certain book bans went into effect are missing from the original PEN America dataset, which limits the Florida frequency analysis.

5.1 Florida vs. Queer Literature

The spike in the lemma *gay* in January 2024 for Florida banned books is accounted for by a single novel, "Brave Face: A Memoir", by Shaun David Hutchinson. The excerpt from the Goodreads description below is provided below:

> …"I wasn't depressed because I was **gay**. I was depressed and **gay**." Shaun David Hutchinson was nineteen. Confused. Struggling to find the vocabulary to understand and accept who he was and how he fit into a community in which he couldn't see himself. The voice of depression told him that he would never be loved or wanted, while powerful and hurtful messages from society told him that being **gay** meant love and happiness weren't for him… (Hutchinson (2019))

According to a KUOW article, which is Seattle's NPR news station, the author currently lives in Seattle but grew up in Florida (Campbell (2024)). Hutchinson wrote this memoir so "that this book existed, that didn't exist when [he] was that age". This book, along with his other book, "We Are The Ants", are both banned in Florida schools. This article also mentions two other books that are both banned in Florida and have the lemma *gay* in this descriptions: "Tricks", by Ellen Hopkins and "Gender Queer: A Memoir", by Maia Kobabe. The Goodreads descriptions for both are provided below:

Excerpt from "Tricks", by Ellen Hopkins (Hopkins (2009)):

> Five teenagers from different parts of the country. Three girls. Two guys. Four straight. One **gay**…

Excerpt from “Gender Queer: A Memoir”, by Maia Kobabe (Kobabe (2019)):

> …Maia’s intensely cathartic autobiography charts eir journey of self-identity, … bonding with friends over erotic **gay** fanfiction, and facing the trauma of pap smears…

It’s evident that Goodreads descriptions can be indicative of which books certain states would like to ban, especially with Florida’s bans against queer literature. By focusing on books banned within a single year, we can investigate relevant legislation during this time period. Notable Florida legislation related to banning books, especially in schools, are as follows. HB 1467, which went into effect on July 2022, requires schools and educators to be transparent about what materials are used in class and making sure that those instructional materials are vetted (Senate (2022)). HB 1069, which went into effect on July 2023, provided requirements about teaching topics including pronouns, reproductive health and human sexuality as well as providing that the district school boards are responsible for materials used in classroom libraries (Senate (2023)). HB 1285, which went into effect on July 2024, states that residents who do not have a child with access to the district materials can only object to one material per month (Senate (2024))

The transparency mandated by HB 1467 likely gave parents and other community members more awareness of what books their students were learning in class. HB 1069 went into effect right at the beginning of the time period covered by this dataset and could be responsible for the higher counts for Florida book bans in July 2023 - October 2024 (see Figure 1). This legislation also targeted queer literature, like Hutchinson highlighted. While the effects of HB 1285 are out of scope of this dataset, it does show potential for repeating this analysis for the next year’s data – July 2024 to June 2025 – to see if these limits reduced the number of book bans in Florida or changed the linguistic features in the Goodreads descriptions.

5.2 There’s Still Hope

While such excessive banning of books feels as dystopian as “Fahrenheit 451”, there is still pushback from the public. For example, HB 1285 shows some legislative limits to how many books people are banning (Senate (2024)). In addition, there are community efforts to push back against these book bans, including from the authors themselves such as Hutchinson (Campbell (2024)). There are also organizational efforts, such as the Little Free Library program. This nonprofit works to provide 24/7 open access to books and have an interactive map that highlights both book ban hotspots – including in Florida – and nearby Little Free Library locations, which include access to banned books specifically (Aldrich (2024)). These efforts give hope that there are plenty of people fighting against these book bans and fighting for access to literature, regardless of the topic. After all, “the books are to remind us what asses and fools we are” (Bradbury (1951)).

6 Appendix

Table 4

Token	Key Range	Key Mean	Key SD	Effect Mean
s	9.397163	-0.3893798	3.447040	-0.2921697
cover	6.560284	0.2976709	1.382571	4.0776704
school	6.205674	0.2761773	2.175860	1.3492016
edition	5.673759	0.2934432	1.514876	3.6487733
question	5.319149	0.2675114	1.470470	4.4648486
unforgettable	5.319149	0.2296386	1.205174	3.9330747
leave	5.141844	0.2103758	1.163098	3.8230487
name	4.964539	0.2559177	1.609739	3.7915074
n't	4.964539	0.1831221	2.191650	1.2161340
maybe	4.787234	0.2450640	1.605248	3.7579549
alone	4.787234	0.1873932	1.171850	3.7665043
gay	4.609929	0.2636090	1.801329	3.8746883
tell	4.609929	0.2493503	1.608768	3.6948768
loss	4.609929	0.2352149	1.422577	4.4043171
sometimes	4.609929	0.2259163	1.358348	3.9091398
important	4.609929	0.1900824	1.226535	4.1188517
feelings	4.609929	0.1753117	1.148455	3.9550851
room	4.432624	0.3098551	2.331225	4.8635836
say	4.432624	0.1967047	1.349692	4.0563938
wrong	4.432624	0.1727494	1.221873	3.6528743

Top 20 key keywords, ordered by descending key range

7 Acknowledgments

Generative AI was used to implement the algorithm for the web scraper that was used to create this dataset and to create the sparkline graphics in the Results section.

Works Cited

Aldrich, Margret. 2024. "Book Ban Hotspots and Nearby Little Free Library Locations Shown on New Interactive Map." https://littlefreelibrary.org/2024/09/book-ban-hotspots-and-nearby-little-free-libraries-shown-on-new-interactive-map/.

Biber, Douglas. 1992. "The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings."

Bradbury, Ray. 1951. *Fahrenheit 451*. Simon; Schuster.

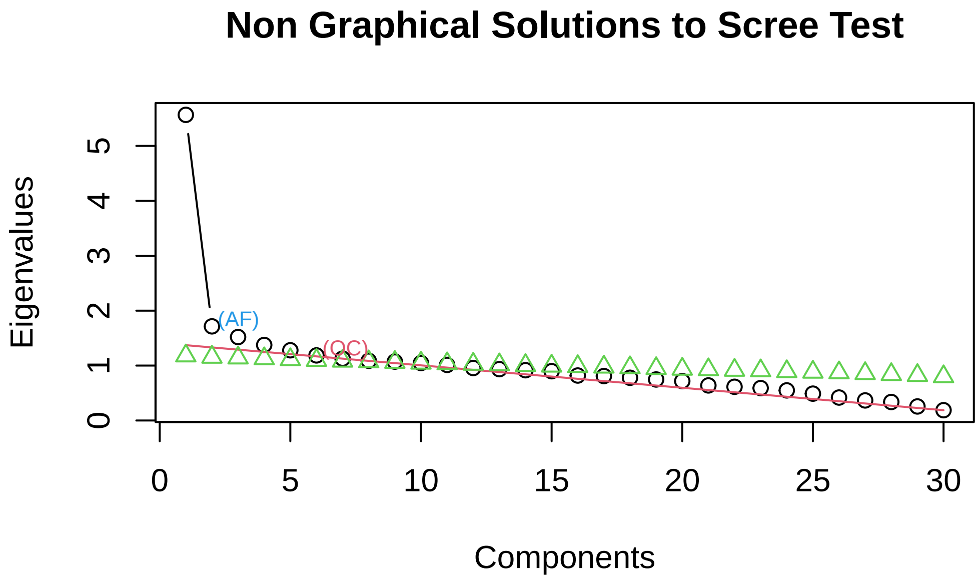


Figure 4: Elbow is at 2 factors

Brezina, Vaclav. 2018. "Statistics in Corpus Linguistics." In. Cambridge University Press.

Campbell, Katie. 2024. "This Seattle Author Wrote a Memoir for LGBTQ Youth. Now It's Being Banned." *KUOW*. https://www.kuow.org/stories/this-seattle-author-wrote-a-memoir-for-lgbtq-youth-now-its-being-banned.

Hopkins, Ellen. 2009. "Tricks." https://www.goodreads.com/book/show/5510384-tricks?from_search=true&from_srp=true&qid=44q3Sxwk5x&rank=1#CommunityReviews.

Hutchinson, Shaun David. 2019. "Brave Face: A Memoir." https://www.goodreads.com/book/show/42202041-brave-face?from_search=true&from_srp=true&qid=kaAsLtdNV4&rank=1.

Kobabe, Maia. 2019. "Gender Queer: A Memoir." https://www.goodreads.com/book/show/42837514-gender-queer?from_search=true&from_srp=true&qid=YZgZaMSqU6&rank=1.

PENAmerica. 2024. "PEN America Index of School Book Bans – 2023-2024." https://pen.org/book-bans/pen-america-index-of-school-book-bans-2023-2024/.

Senate, Florida. 2022. "CS/HB 1467: K–12 Education." https://www.flsenate.gov/Session/Bill/2022/1467/.

———. 2023. "CS/CS/HB 1069: Education." https://www.flsenate.gov/Session/Bill/2023/1069.

———. 2024. "CS/CS/HB 1285: Education." https://flsenate.gov/Session/Bill/2024/1285.

Smith, Mike. 2010. "Definition of a Key Key-Word." https://lexically.net/downloads/version5/HTML/index.html?keykeyness_definition.htm.