Analyzing the Reading Levels of *Fifty Shades of Grey* and *The DaVinci Code:* Learning More About Blockbuster Books

Sophie Clipson with Professor Chun and Professor Elkins
IPHS 200, Fall 2021

Abstract

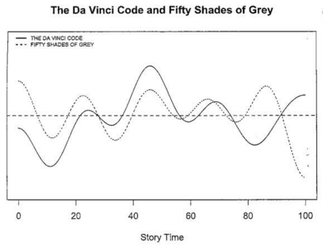
Through data analysis we have the opportunity to explore why certain books become blockbusters, even if at first glance it appears random. Using the Flesch-Kincaid Grade Level Test on the bestselling novels *Fifty Shades of Grey* and *The DaVinci Code*, I was able to produce a 'readability arc' for each book. These arcs began to tell the story of the purpose behind authorial choices in reading comprehension throughout the book, and, if readability arcs can serve as another pattern or criteria for a blockbuster novel.

Introduction

Conventionally, bestselling, and prizewinning books are assumed to fulfill certain criteria: a compelling plot, complex and relatable characters, and mastery of grammar. Past this, there are a handful of books that go above and beyond, collecting the attention and fanfare of millions of readers worldwide. At first glance, the books that achieve this seem random; however, there is evidence to suggest that it isn't.

Fifty Shades of Grey by E.L. James, and *The DaVinci Code* by Dan Brown attained unparalleled success, marking themselves as the highest-selling adult books of the past couple of decades. This begs the question: is there something about these books that makes their blockbusting sales possible?

Research by Jodie Archer and Matthew L. Jockers shows that *Fifty Shades of Grey* and *The DaVinci Code* are the only two books (in their corpus) that share a unique, perfectly rhythmic emotional curve. They argue that this symmetrically rhythmic curve, and the emotional experience it invokes in readers, is what allowed these two books to achieve the success they did.



[1]

Are there other qualities that help books achieve this success? Is there something else that *Fifty Shades of Grey* and *The DaVinci Code* share that allowed them to attain their unique success?

Ben Blatt in his book *Nabokov's Favorite Word is Mauve* suggests that a book's readability level could have influence over its success, quoting Dr. Seuss' belief that "simplicity brought success." Rudolf Flesch in the 1950s created the Flesch-Kincaid Grade Level Test, a mathematical formula for calculating the simplicity or complexity of any text, producing a score that indicates the grade-level required to read the book. Blatt's research shows that over the years, the bestseller's list has become full of fiction that is much simpler, with ultrapopular bestsellers having much lower readability scores than they have historically been.

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

[2]

Do *Fifty Shades of Grey* and *The DaVinci Code* have similar grade-level reading scores? More than their overall Flesch-Kincaid score, do these books have similar curves in their readability throughout the plot? Is there a specific readability arc that makes a book a bestseller?

[1]The emotional arcs of *Fifty Shades of Grey* and *The DaVinci Code* as calculated by Archer and Jockers in *The Bestseller Code.*

[2]Formula for the Flesch-Kincaid Grade Level Score.

Methodology

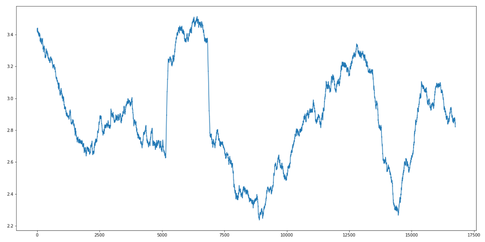
To get a preliminary understanding of the readability scores of each novel, I ran the '.txt' versions of *Fifty Shades of Grey* and *The DaVinci Code* through an online Flesch-Kincaid Grade Level calculator from Text Compare. I first ran the entirety of the *Fifty Shades of Grey* and *The DaVinci Code* text files through this program to get the overall Flesh-Kincaid Grade Level Scores of the texts, which were 3.52 and 6.21, respectively.

Next, with the help of Professor Chun, I used Google Colab and Python to parse the individual sentences from *Fifty Shades of Grey* and *The DaVinci Code.* This program parsed individual sentences, converted them into strings, and cleaned the text (specifically of short sentence lengths). After the data was cleaned, the program calculated the Flesch-Kincaid scores for each sentence, and produced a graph visualization of the readability scores over the course of the novels (by sentence).

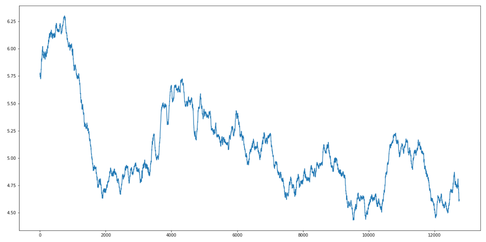
Using another Google Colab and Python notebook, I collected and visualized the sentiment arcs of *Fifty Shades of Grey* and *The DaVinci Code* in order to analyze the granular data Archer and Jockers highlight in *The Bestseller Code*. Similar to the previous Google Colab and Python notebook, this program imported the text versions of *Fifty Shades of Grey* and *The DaVinci Code*, parsed the paragraphs and sentences and cleaned those that had no letters in them, or did not meet a certain length. The program next used two sentiment analysis libraries, VADER and TextBlob, to calculate and visualize the sentiment arcs of the two novels.

With these graphs from Google Colab and Python, I used Microsoft Excel to compare and analyze the readability arcs of *Fifty Shades of Grey* and *The DaVinci Code*, as well as the readability arcs and sentiment arcs for each novel. To do so, I transposed the graphs of: (a) the readability arc of *Fifty Shades of Grey* and the readability arc of *The DaVinci Code*; (b) the readability arc and sentiment arc of *Fifty Shades of Grey;* and (c) the readability arc and sentiment arc of *The DaVinci Code.*

Using my background as an English major and general external knowledge I was able to notice patterns in the arcs, gain insights, and draw potential conclusions from the data. Each of the visualizations tell a story about the authorial choices in novels, and what makes a blockbuster book.



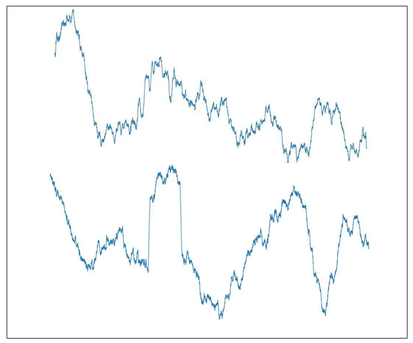
The Flesch-Kincaid Grade Level readability arc for *Fifty Shades of Grey* by E.L. James. The x-axis indicates the sentence number, and the y-axis indicated the Flesch-Kincaid grade-level score.



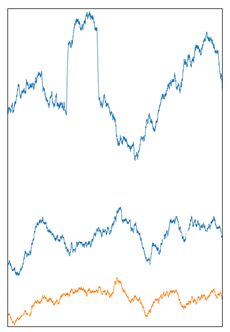
The Flesch-Kincaid Grade Level readability arc for *The DaVinci Code* by Dan Brown The x-axis indicates the sentence number, and the y-axis indicated the Flesch-Kincaid grade-level score.

Results

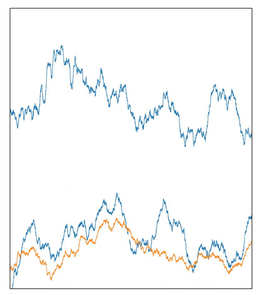
Below are the results of the readability and sentiment analysis of *Fifty Shades of Grey* and *The DaVinci Code*. For the various visualizations I transposed the graphs of the readability and sentiment arcs of the two novels in order to gain insights about the authorial choices for changing the reading level throughout the book.



This first visualization shows the compared readability arcs of *Fifty Shades of Grey* (top arc) and *The Davinci Code* (bottom arc). This visualization suggests that there is a degree of similarity between the readability curves of the two novels, indicating that an oscillation in reading level throughout a novel contributes to its success. As you can see, both *Fifty Shades of Grey* and *The DaVinci Code* have peaks and valleys in their readability as the novel progresses, and in several places these peaks and valleys align.



[3]



[4]

[3]The emotional arc (top arc) and sentiment arc (bottom two arcs – VADER and TextBlob) of *Fifty Shades of Grey.*
[4]The emotional arc (top arc) and sentiment arc (bottom two arcs – VADER and TextBlob) of *The Davinci Code.*

The two visualizations above suggest there is a potential relationship between sentiment and readability in these two novels. These parsed segments of the readability and sentiment arcs for *Fifty Shades of Grey* and *The DaVinci Code* show a correlated pattern in the level of readability and emotion in the novels. For both novels, it appears that in moments when there is high emotionality (or sentiment), as indicated by the bottom two arcs in each graph, there is also a peak in the Flesch-Kincaid grade-level, and vice versa.

In both visualizations, there are a couple of places where the readability and sentiment arcs diverge indicating that there is not a perfect correlation between reading level and emotionality level throughout the novels. As will be discussed in the conclusion, further analysis of the contexts of these cruxes as well as greater hands-on investigation of these arcs by humans could lend insights into these discrepancies.

Conclusion

Based on the calculations and subsequent visualizations of the readability and sentiment arcs of *Fifty Shades of Grey* and *The DaVinci Code* it is evident that there is a similarity in the readability patterns of the two novels, as well as a relationship between readability and emotionality as a book progress.

As discussed in the results section, a Flesch-Kincaid Grade Level analysis of *Fifty Shades of Grey* and *The DaVinci Code* produce connate readability arcs. Similar to Archer and Jockers research on the sentiment analysis of these two novels, the near-mirrored readability arcs of *Fifty Shades of Grey* and *The DaVinci Code* could indicate that the stratospheric success of these two extremely different novels may not be totally random. As Archer and Jockers stipulate in *The Bestseller Code*, although we cannot rely on these insights to determine bestsellers, we can use a Flesch-Kincaid analysis on manuscripts to help decide whether or not we pursue the book as a publishing house.

Looking closely at the comparison between the readability and sentiment arcs for both *Fifty Shades of Grey* and *The DaVinci Code*, it becomes clear that there is a potential correlation between the authorial choices on reading grade level and emotional events in books. In several locations we can see that as the readability arc peaks, such does the sentiment, and vice versa. Although we cannot draw a complete or direct conclusion about this relationship, this suggests that authors may manipulate the reading level of a book to complement the emotionality of an event in the novel, and to keep the reader engaged.

It is important to note that this study was limited in several ways, and that further research and development would be needed in order to confidently use these analyses in the publishing world. The first limitation to delineate of this study is the scope: only two books were analyzed in this study, and although they were chosen for specific reasons, research on this topic would benefit greatly if a larger corpus of books were used. It would be interesting to not only analyze other bestselling books, but to additionally survey prize-winning books, and, further, to compare the similarities and differences between their readability arcs. This study could also be improved by increased human-directed investigation. This includes examining the book context around the cruxes of both the readability and sentiment arc both in order to gauge the accuracy of the analyses and analyzing how the emotionality and readability in a book works and effects readers. This would entail identify each crux in the physical book, actively reading the parses, and then coding and topic modeling each.

Through these improvements to the study and more research on the interconnection between reading level and reader experience, as well as between readability and sentimentality, I believe this research could drastically change not only the publishing industry, but also expand the fields of psychology, English and AI.

Acknowledgements

Archer, Jodie and Matthew L. Jockers. *The Bestseller Code.* St. Martin's Press, 2016, New York, NY.

Blatt, Ben. *Nabokov's Favorite Word is Mauve.* Simon & Schuster, 2017, New York, NY.

Porter, JD. "Popularity/Prestige: A New Canon." *Stanford Literary Lab.* Stanford Literary Lab, 2018. https://litlab.stanford.edu/pop-pres-ny/. Accessed 1 December 2021.