5-15-2018

Beyond BookScan: How Publishers Can Use Alternative Data To Supplement Traditional Sales Metrics for Genre Fiction

Grace Evans
Portland State University

Beyond BookScan

How Publishers Can Use Alternative Data To Supplement Traditional Sales Metrics for Genre Fiction

Grace Evans

Paper submitted in partial fulfillment of the requirements for
the degree of Master of Arts in Writing: Book Publishing.
Department of English, Portland State University
Portland, Oregon
May 15, 2018

Abstract

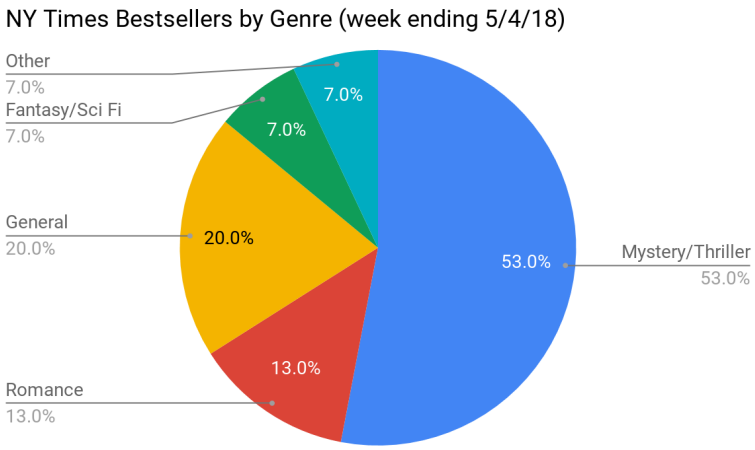
This paper explores available sales, rankings, and readership data for best-selling genre fiction in search of data correlations and relationships between trends and rankings from industry standard BookScan, market giant Amazon, and reader-sourced Goodreads. With these tools in hand, can publishers understand and assess the readership (and ultimately, the sales potential) of top-selling and midlist genre fiction? Given the inconsistent nature of available data, there is no global conclusion to be made; however, this paper points to the expansion and development of alternative data sources as a necessity for the continued evolution of publishing in the digital age.

Introduction

In a 2017 article published on *The Writing Cooperative* website community, author Erica Verrillo asks the question, "What are the most popular literary genres?" She goes on to explain the complexities of this seemingly simple question: Does it refer to the number of books consumed by readers? The number of books bought? Top genres requested by literary agents? Genres that make the most money? Verrillo, addressing a community of writers, ends her discussion with a shrug, advising her fellow authors to write what they love, not what they think is most popular.[1]

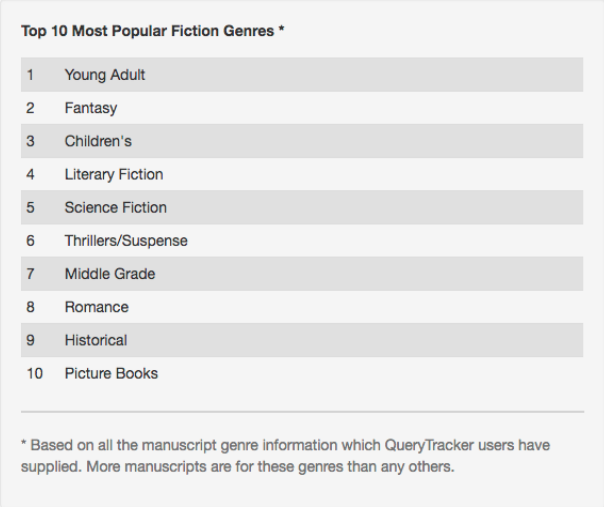
Verrillo's conclusion is a tempting one. The data available to answer such questions is, at best, incomplete, and at worst, hopelessly muddled. This paper attempts to take a sample survey of some of the data available to writers and publishers and to assess its utility to those in search of publishing success. In order to do this, the scope of the discussion has been restricted to adult fiction, and the focus of the conversation excludes the nuances of nonfiction markets as well as the infamous "Harry Potter phenomenon" of children's and young adult markets.

Even with the scope narrowed exclusively to adult fiction, is there data available to confirm or refute the industry reports? Sampling a week of the *New York Times* Combined Print and Ebook Fiction Best Seller list, a clear story emerges: mystery/thrillers reign supreme,



[1] Erica Verrillo, "What are the most popular literary genres?" *The Writing Cooperative* (blog), November 15, 2017, https://writingcooperative.com/what-are-the-most-popular-literary-genres-6db5c69928cc.

with romance taking up the next largest slice of genre pie after the amorphous "general" fiction.



Top 10 Most Popular Fiction Genres *

1	Young Adult
2	Fantasy
3	Children's
4	Literary Fiction
5	Science Fiction
6	Thrillers/Suspense
7	Middle Grade
8	Romance
9	Historical
10	Picture Books

* Based on all the manuscript genre information which QueryTracker users have supplied. More manuscripts are for these genres than any others.

Perusing the lists available on QueryTracker.net, a popular online network where agents and authors can find each other, fantasy tops the genre list, right under young adult, as most popular among users of that site. Thriller and romance do not even appear in the top five.[2]

Then there's the popular wisdom: romance makes the most money. The website of the Romance Writers of America (RWA) reports an estimated annual total sales value of $1.08 billion in 2013, and quotes a 2015 Nielsen report that gives romance novels a 34 percent share of the US fiction market.[3] And don't just take the RWA's word for it—read through Nielsen's subsequent well known and often-quoted 2016 report on the state of the "billion-dollar romance book industry" that outlines readership and market share for the year, as well as outlining where and in what format readers discover romance novels.[4] A more comprehensive report on romance readers is also available through the RWA website, which cites 94 percent of readers using print, 64 percent reading ebooks, and 38 percent engaging through audiobooks.[5] Because of the reader overlap inherent in the report, it is still difficult to isolate which readers and titles may be coming

[2] "Top 10 Genres," QueryTracker, https://querytracker.net/top-10-genres.php, accessed April 2018.
[3] "Romance Fiction Statistics," Romance Writers of America, https://www.rwa.org/p/cm/ld/fid=580, accessed May 2018.
[4] Nielsen, "Romance Readers by Numbers," *Nielsen Insights* (blog), May 26, 2016, http://www.nielsen.com/us/en/insights/news/2016/romance-readers-by-the-numbers.html.
[5] "Romance Readers," Romance Writers of America website, https://www.rwa.org/p/cm/ld/fid=582, accessed May 2018.

together in print spaces versus digital, and to understand how ebooks sales impact that particular market.

So, faced with the deceptively simple question, "What do people read most?" those looking for answers based in data are faced with the following conundrum.

The Trouble With BookScan

It is common industry knowledge that BookScan is the publishing industry's most significant metric, and also provides sales numbers with an often unknown margin of error. As reported by NPD, which acquired the US portion of the market information system from Nielsen in 2017:

> *NPD BookScan™ is the gold standard in POS tracking for the publishing market. It covers approximately 85 percent of trade print books sold in the U.S., through direct reporting from all major retailers, including Amazon, Barnes & Noble, Walmart, Target, independent bookstores, and many others.*[6]

Two of the claims above point to problematic data, particularly as it relates to the tracking of genre sales.

First, the missing 15 percent of data would be a workable data discrepancy, if only that 15 percent were reliably consistent. A 2013 *Forbes* article characterized those missing sales as coming primarily from gift stores, big box stores, library, export, corporate, and other special sales.[7] The problem with BookScan's underreporting is that it varies greatly by genre and market. Comic books, an often-quoted exception to BookScan's reliability, suffer from much greater

---

[6] "Books Market Research and Business Solutions," NPD, https://www.npd.com/wps/portal/npd/us/industry-expertise/books/, accessed May 2018.
[7] Suw Charman-Anderson, "Can Nielsen BookScan Stay Relevant In The Digital Age?" *Forbes*, January 7, 2013, https://www.forbes.com/sites/suwcharmananderson/2013/01/07/can-nielsen-bookscan-stay-relevant-in-the-digital-age/#2109cbaf761e.

underreporting than other categories, because of the much higher number of non-traditional retail outlets where they are sold. Increasingly, those non-traditional points of sale are not found in small stores or community meet-ups, but on the internet, which brings us to the second and likely more significant flaw in BookScan's numbers: ebook sales are not included in their reporting.

In the last decade, Nielsen added PubTrack Digital, "the only resource offering a comprehensive view of today's digital book market," in the US and UK; however, this reporting feature only claims "access to the top 80 percent of the traditionally published ebook market."[8] PubTrack is far from ubiquitous, and its scope, which automatically excludes non-traditional digital publishing, makes it a deeply flawed metric for a market that arguably has its deepest roots in non-traditional distribution. With the proliferation of opportunities for self-publishing, and ever-increasing channels for authors to transition between self-publishing and traditional publishing, this data gap can have significant implications on reporting for authors, titles, and markets.

Alternative Data

If BookScan is such a flawed tool, why do publishers continue to rely on it as a metric, particularly in acquisitions discussions regarding the success of upcoming projects? A very simple answer is that no ideal alternative exists. If you are looking for precisely quantifiable sales numbers, unless you have access to the carefully guarded vault of Amazon data, you're not going to find anything more comprehensive than BookScan. In search of reader statistics, most data is proprietary—publishers occasionally have access to their own data, but not to anything that spans a genre, or the industry at large.

---

[8] "Books Market Research and Business Solutions."

However, other data sources can provide publishers with different information. Various series of metrics, ratings, and rankings tell stories about book markets that reach beyond the flaws in BookScan numbers. Suw Charman-Anderson, author of the 2013 *Forbes* article titled "Can Nielsen BookScan Stay Relevant In The Digital Age?" interviewed a British publisher who posits that the industry as a whole is moving away from relying solely on BookScan numbers; instead, publishers are looking for ways to assess, engage, and expand an author's audience through other kinds of market research, particularly digital engagement.[9] Can these alternative data sources be analyzed, whether independently of or in tandem with sales numbers, to create a more complete picture?

Comparing Data

To explore the question above, we must first ask how the alternative data sources compare to BookScan and each other when analyzing successful genre titles. After just a quick overview, very different stories emerge about recent best-selling genre fiction when read through the lenses of the *New York Times* Best Seller list, Amazon, and Goodreads—all of which, it should be noted, pass the first test that BookScan fails by including ebooks in their metrics.

Methodology

Data was collected from four sources: BookScan, Amazon, Goodreads, and the *New York Times* Best Seller list. For the comparisons of data sources through rankings by genre, ranked titles from one sample week were pulled from each source, including title, author, and publisher information as well as any quantification of the title's popularity, where available (e.g. BookScan

[9] Charman-Anderson.

sales for the week or number of "read" tags assigned to the title on Goodreads). Titles were coded by genre, and the percentage of titles in each genre per ranking was calculated.

The title-based comparison took one top-selling title from three prominent genres, mystery/thriller, romance, and fantasy/sci fi. For each title, BookScan YTD sales were compared to average Amazon rating (0–5 stars), number of Amazon reviews, and number of users who rated the title on Goodreads. All three titles appeared on the *NYT* Best Seller lists.

Considerations

To compare and contrast the numbers as accurately as possible, this study has made every effort to collect data samples of corresponding sizes and time periods. Where manual classification of genre was necessary, categories were assigned with great attention to consistency.

However, a few notes about the quality of the comparisons—because several of the sources listed children and young adult titles in their fiction rankings, they have not been removed from the data, but rather segmented out as their own genre in an attempt to keep the comparisons as consistent as possible.

The classification of genre should also be acknowledged here as having its subjective moments. Each of these sources treats genre a little differently, with some categorizations separating mystery from suspense/thrillers, and others crowding many titles under the umbrella of general fiction rather than assigning it a genre label. In instances where the category was uncertain, every effort was made to assign the book according to where it would appear in a bookstore or library. In instances where a book could be labeled as multiple genres (as is often the case in the romantic suspense novels of Nora Roberts or Danielle Steel), titles were categorized in accordance with the author's primary market.

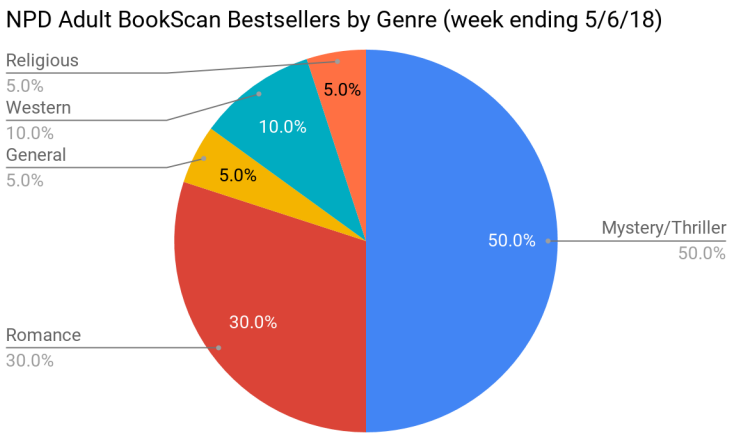
Examining Data Sources through Rankings by Genre

What happens when you put these data sources side by side? A snapshot comparison of a week's title rankings according to available data sources illuminates discrepancies; side by side and broken down by genre, each source tells a different story. Why is romance represented in BookScan and Amazon sales rankings, but not in Goodreads or Amazon's Most Read chart? Why are fantasy and science fiction titles represented on every ranking except BookScan?

The charts below provide a quick look at the genre breakdowns for each alternative source, provoking questions and discussion as each source reveals something of its data bias, intended audience, and usefulness to writers and publishers.

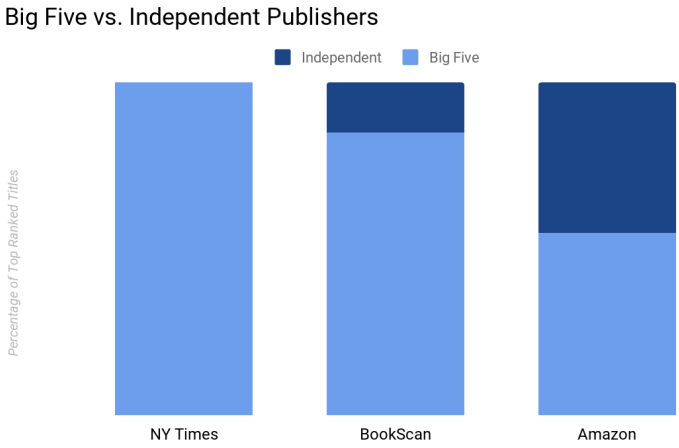
BookScan

Beginning with BookScan's data gives a helpful framework for the alternative data sources that follow. It is also important to acknowledge here that BookScan's data differs from the other rankings because its original data set does not include any children or young adult titles, and as such, that group is not represented on the chart. This data discrepancy seems to amplify the prominence of genre fiction in this ranking, with a full 50 percent of titles classified as mystery/thrillers, and the presence of religious fiction and western titles, both genres that appear in no other rankings.



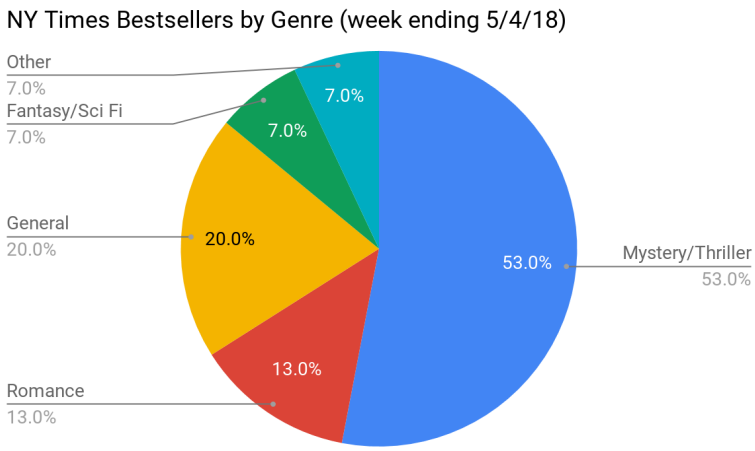
The New York Times

This paper referred earlier to the *New York Times* Best Seller list genre breakdown as a jumping off point to discuss some of the problems inherent in establishment data, sources that ignore grassroots initiatives, smaller actors, and sales and business practices that fall outside of a certain structure. While arguments about the accuracy or harmful nature of such skewed data continue, the *New York Times* indisputably bears the reputation of a brand centered on this establishment bias. In addition to genre breakdown, it is interesting in this case to see which publishers garner representation on this particular ranking. Of the books on the Combined Print and Ebook Fiction Best Seller list during the week ending May 7, 2018, 100 percent of them were published by Big Five houses or imprints. As shown by the chart on the right, even BookScan rankings include some smaller publishers, while Amazon charts further diversify from the Big Five establishment by including their own Amazon imprints and other Amazon-exclusive titles.



Without any diving any deeper than this, a key point can be extrapolated here about the nature of the data story told by the *New York Times*; it is biased toward established houses, authors, and market channels. It is the list of household names that appear in larger type on their book jackets than the titles of the books themselves: James Patterson, Danielle Steel, John

Grisham, and Nora Roberts (all these but one appeared on the list during the week of May 7). As such, it is unsurprising that the *NYT* genre breakdown appears quite middle-of-the-road when compared to other rankings. Its greater percentage of the ever-popular mystery/thriller genre replaces the religious fiction and westerns found in BookScan's ranking, while its outliers are fantasy/sci fi (which seems an inaccurate assessment of that genre based on its anecdotal popularity) and a humor title from best-selling satirist Christopher Moore. (It should be noted that while the Moore title is categorized as "other" in the chart here, it could also potentially be tagged as fantasy/sci fi, doubling that genre's prominence in this ranking.)



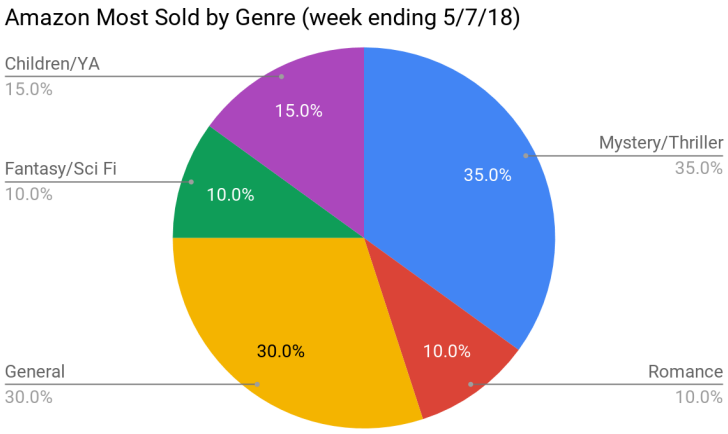
Amazon

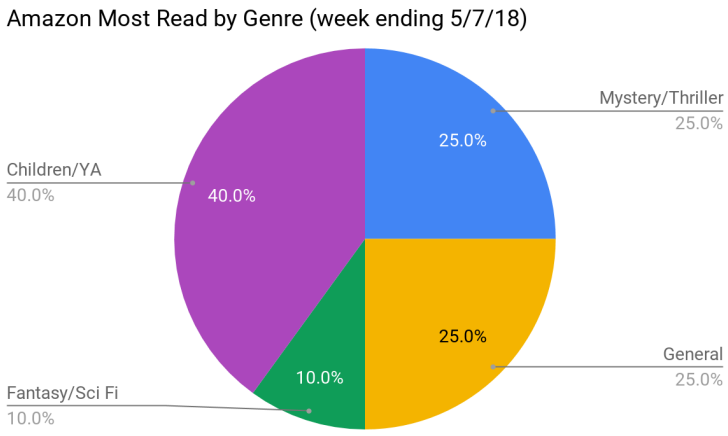
A very different brand of establishment bias emerges when looking at information available from Amazon. The company has a well known reputation as a data vault, containing much and sharing nothing. Frustratingly for publishers, this ostensibly includes comprehensive market data on the vast majority of ebook sales.

In their publicly available Amazon Charts, weekly rankings show top titles based on two separate metrics, most read and most sold, rankings that mimic the structure of the *NYT* while including titles published through Amazon imprints or, in one instance from the Most Sold data set, a self-published title. In reader-targeted data, Amazon displays user reviews and ratings,

prominently highlighting a book's 0–5 star rating and the number of customer reviews in search results and book pages. This information is said to have an impact on searchability and positioning of the book within the Amazon network, but the details and extent of that impact are largely unknown, guarded within the data vault.

From the accessible data, it is notable that Amazon confirms the trend (later increased by Goodreads), of prominent percentages of genre-defying fiction, with 30 percent of the Most Sold titles and 25 percent of the Most Read titles categorized as "general." These general titles have lessened the prominence of the mystery/thriller category, while the Most Read data set has an unparalleled 40 percent of children/YA titles, due entirely to a recent release of enhanced ebook editions of the Harry Potter series.



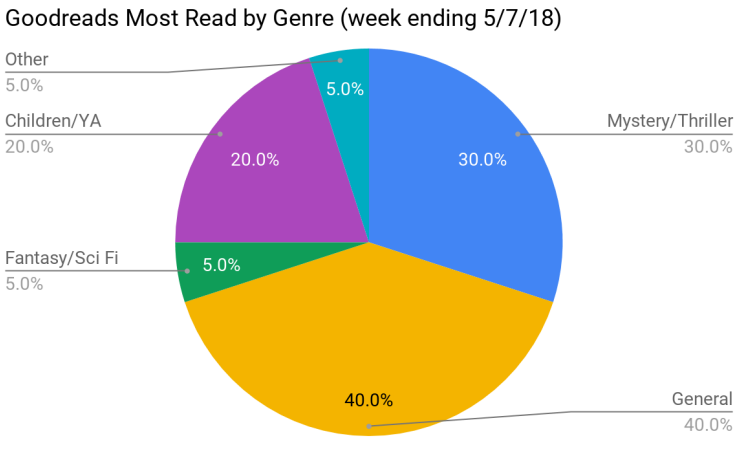


Goodreads

Self-identified as a "social cataloging site," the Goodreads argument for relevance to its audience is in a different kind of data. "Our recommendation engine analyzes 20 billion data points to give

suggestions tailored to your literary tastes," the website claims on an About Us page that also boasts 65 million members and 68 million book reviews.[10] Positioned as the anti-bestseller list, Goodreads markets itself as an online window into your friends' bookshelves, a deeply personalized, egalitarian, and unpretentious experience for readers. Whether in spite of or because of this positioning, Goodreads does contain mountains of accessible data, including reader reviews (which contain keywords galore), reader numbers and ratings, and lists of crowdsourced rankings for every genre, audience, and literary yen imaginable. However, Goodreads data is purposefully packaged for readers, not writers or publishers, and wading through it for purposes of market research can be a disorganized and fragmentary experience.

Given the similarity in data queries and the corporate relationship between Amazon and Goodreads (the former acquired the latter in 2013), one would expect the most positive data correlation between Amazon Most Read and Goodreads Most Read; this seems to be true, although the numbers are not identical, making it clear that while Amazon's metrics may involve data input from Goodreads, there are other factors involved.



The Goodreads Most Read ranking does not include the Harry Potter editions that skew the Amazon numbers, and also contains the largest percentage of titles categorized as general fiction. Mystery/thrillers remain prominent

[10] "About Us," Goodreads, https://www.goodreads.com/about/us, accessed May 2018.

among genres, while the outlying "other" found here is Michelle McNamara's *I'll Be Gone in the Dark*, a true crime title about the Golden State Killer, which is likely experiencing an uptick in popularity due to recent current events. Notably and puzzlingly, romance is not represented on either Amazon or Goodreads Most Read rankings. The high number of "general" titles likely comes from the lack of tight categorization present in the Goodreads structure; by its crowdsourced nature, the genre demarcations in its own data sets are more loosely organized and less consistently applied.

Examining Genre through Ratings by Title

While the genre breakdown of data source rankings offers insight into the biases and quirks characteristic of each particular alternative data source, a deeper dive into a few best-selling titles further illustrates the distinct ranges of those different data sets. Using case studies to more narrowly define quantitative analyses creates enlightening comparisons across data sources for a single title and across titles within a single source. Looking at Amazon, Goodreads, and BookScan numbers for one romance title, one mystery/thriller title, and one science fiction title—all of which have appeared on the *NYT* Best Seller lists—insights emerge about the different behaviors of genre readers, and thus the different positioning and markets open to publishers of those genres.

The table below offers a sampling of quantitative analyses, collected from some of the sources discussed in the section above in order to comparatively examine three genre fiction best sellers by sales and readership response data points. The three books are:

1. David Baldacci's *The Fallen*, the latest in his mystery/thriller Memory Man series, which appeared in first or second position in every ranking discussed above except Goodreads,

2. Romance giant Nora Roberts' latest romantic suspense, *Come Sundown*, and
3. Current science fiction darling *Ready Player One*, a debut from author Ernest Cline originally published in 2011 and experiencing a spike in popularity due to the recent film adaptation.

Title & Author	***Genre***	**BookScan YTD**	**Amazon Rating**	**Amazon No. of Reviews**	**Goodreads No. of Ratings**
The Fallen *David Baldacci*	*Mystery/Thriller*	119,278	4.6	582	5,442
Come Sundown *Nora Roberts*	*Romance*	27,360	4.4	2,230	17,753
Ready Player One *Ernest Cline*	*Science Fiction*	265,669	4.6	17,260	575,261

This chart offers two directions for comparison, horizontally (across each row) and vertically (down each column). Examining the data in columns, the most apparent discrepancy is in the BookScan data for *Come Sundown*. For an incredibly prolific best-selling author, the BookScan reported sales seem very low, especially in comparison to *The Fallen*. Given the expansive and diverse landscape of the romance genre, this could be due to shifts in demographic and engagement among romance readers, but the available data is inconclusive. Conversely, the high sales numbers for *Ready Player One* are easily explained by the recent release of the movie adaptation. Down the column, the BookScan figures are incongruent with the data indicating genre popularity; based on the ranking breakdowns, one would expect *The Fallen* to lead in sales, followed by *Come Sundown*, then *Ready Player One*.

One and two columns over, the Amazon ratings show little variation among the three titles, while the number of reviews posted to Amazon vary wildly. Again, *Ready Player One*

takes the lead, likely because of the movie attention, while the mystery/thriller title's review numbers are underwhelming. This trend continues into the Goodreads ratings, where *Ready Player One* has garnered over one hundred times more reader ratings than *The Fallen*. From a purely genre-driven perspective, according to the landscapes of top rankings laid out in the previous section of this paper, this narrower quantitative data sampling seems to make little sense.

However, in the case of both Goodreads ratings and Amazon reviews, the differences between all three titles could be explained to some extent by their publication dates—*Ready Player One* has been out since 2011 and *Come Sundown* was published in January 2017, while *The Fallen* arrived on the scene just this year.

In fact, with those publication dates in mind, a horizontal analysis of the data could be interpreted as a fairly straightforward representation, not of genre, but of timelines—again with the gaping exception of the BookScan figures. Reading the rows, it seems reasonable to interpret the trends in the figures regarding reader engagement (Amazon and Goodreads) as factors of the time reader have had to engage with the titles: several months for *The Fallen*, a year and a half for *Come Sundown*, and a whopping seven years for *Ready Player One*.

Perhaps predictably, the *Ready Player One* row contains the most positive data correlation, while the *Come Sundown* data suggests that BookScan data does not match the trends in the engagement data; however, it cannot be conclusively determined from this sample whether those data relationships are ultimately affected by factors of genre or time.

Lessons Learned

The comparisons presented above clearly demonstrate that BookScan sales do not positively correlate with reader engagement. They also shows that reader engagement varies across platforms. Variations between richly engaging data sources (like Amazon and Goodreads) likely emerge from differences in audience demographics and preferences for discovery platforms among readers of different genres. While sometimes poorly organized for purposes of market research, such platforms nevertheless offer a wealth of data to be mined for information about readership and market potential.

But is this alternative data is useful to publishers? Trends in reader engagement by genre tell a story about where the readers (and purchasers) of those genres are acquiring those books, and thus where publishers can go to find them. And if Suw Charman-Anderson's 2013 interviewee was correctly identifying a shift in publisher focus from sales to audience engagement when deliberating acquisitions decisions, it seems clear that a more sophisticated analysis of data sources for reader engagement, particularly in digital spheres, can only add value in strategic decision-making.

Without delving too deeply into individual social media metrics or online self-publishing platforms, there are overlaps and intersections to be further explored, even within the limitations of the nuclear world of traditionally successful genre publishing. Clearly, those books that top BookScan rankings and the *New York Times* Best Seller lists need little expansion of their market research or engagement; however, there are lessons to be learned from their success that could

make a significant difference in the ways that publishers explore market opportunities for midlist authors and titles.

Obviously, further research is needed. A next step (though one that far exceeds the scope of this paper) would be to refine the usefulness of reader engagement data sources by tracking the data they provide over a time period to see if it is a reasonable predictor of success over that period, or whether, at some point, the trending line levels out or begins to decline. Understanding time as a factor in reader engagement would be a crucial step toward understanding how to maximize that engagement in support of frontlist titles, and take advantage of its role in sustaining or bolstering a backlist. By investing in more sophisticated systems for quantitative analysis of alternative data sources, publishers could develop new, comprehensive strategies driven by data beyond BookScan, broadening industry understandings of genres' readership potentials and opening doors to new markets.

Bibliography

"About Us." Goodreads. https://www.goodreads.com/about/us. Accessed May 2018.

"Books Market Research and Business Solutions." NPD. https://www.npd.com/wps/portal/npd/us/industry-expertise/books/. Accessed May 2018.

Charman-Anderson, Suw. "Can Nielsen BookScan Stay Relevant In The Digital Age?" *Forbes*. January 7, 2013. https://www.forbes.com/sites/suwcharmananderson/2013/01/07/can-nielsen-bookscan-stay-relevant-in-the-digital-age/#2109cbaf761e.

Nielsen. "Romance Readers by Numbers." *Nielsen Insights* (blog). May 26, 2016. http://www.nielsen.com/us/en/insights/news/2016/romance-readers-by-the-numbers.html.

"Romance Fiction Statistics." Romance Writers of America. https://www.rwa.org/p/cm/ld/fid=580. Accessed May 2018.

"Romance Readers." Romance Writers of America. https://www.rwa.org/p/cm/ld/fid=582. Accessed May 2018.

"Top 10 Genres." QueryTracker. https://querytracker.net/top-10-genres.php. Accessed April 2018.

Verrillo, Erica. "What are the most popular literary genres?" *The Writing Cooperative* (blog). November 15, 2017. https://writingcooperative.com/what-are-the-most-popular-literary-genres-6db5c69928cc.