Don't Judge a Book by Its Cover: Differential Learning on Amazon

Omar Abdel Haq *

December 15, 2025

Latest, Complete Version Here.

Abstract

Online marketplaces now present consumers with dense bundles of ratings and reviews, raising new questions about how people learn from such information-rich environments. We study Amazon's book marketplace using a daily title–country panel of sales ranks for over 10,000 books in 2018, matched to star ratings, review counts, genres, and review texts. First, we relate within-title changes in sales ranks to lagged average star ratings, and show that higher lagged ratings predict improved next-day ranks, but with striking heterogeneity across genres. We summarize this heterogeneity in a continuous genre-level "star-rating effect index," which is large for commercial fiction and narrative non-fiction and close to zero for more specialized or scholarly categories. Second, after scoring one-star reviews along vertical (broad quality) and horizontal (audience fit) dimensions, we show that both types of content predict helpful-vote counts, and that their relative importance varies systematically with the index: in high-index genres, helpfulness is especially sensitive to vertical quality content, whereas in low-index genres it loads more on horizontal fit. These patterns are difficult to reconcile with a purely algorithmic exposure story and instead suggest that readers in different genres make different trade-offs between simple rating heuristics and more detailed text-based evaluation, with implications for the design of rating and review systems.

*Abdel Haq: Harvard Business School (email: oabdelhaq@hbs.edu)

1 Introduction

The explosion of e-commerce has transformed not only where we shop but how we learn about what we buy. A single Amazon page now displays more product information than a day of mall-hopping once could: star averages, rating distributions, verified-purchase badges, "most helpful" votes, and thousands of review texts. Search costs have fallen toward zero, but information selection and interpretation costs have risen sharply. Consumers who were once constrained by scarcity of information now face an over-abundance of noisy signals and must decide, in a matter of minutes, how much weight to give each cue and, ultimately, whether or not to click "Add to Cart".

This paper examines how learning unfolds in such information-dense environments. The setting is Amazon's book marketplace, where purchases are largely discretionary and heavily shaped by social cues. A panel of daily sales ranks matched to all ratings and reviews posted over the course of a year makes it possible to connect crowd-generated signals to a clear performance outcome: each book's position in the sales hierarchy. Because the analysis is restricted to information and platform signals that are determined before shoppers decide to buy, movements in sales rank can be read as the revealed consequence of two possible mechanisms: one is how consumers learn from and act on the information displayed on a product page, and the other is how Amazon's own ranking and recommendation algorithms translate those same signals and recent demand into changes in a book's visibility.

The first part of the analysis measures how strongly Amazon's aggregate rating signal co-moves with demand. Using panel regressions on a daily title–country dataset, the paper relates changes in each book's sales rank to its previously displayed average star rating, while conditioning on past rank, review volume, price, and time since publication. In the pooled sample, higher lagged ratings are associated with statistically significant, though not necessarily sizable, improvements in next-day rank. Estimating these relationships separately by genre reveals striking heterogeneity. In commercial genre fiction and narrative non-fiction (for example Romance, Mystery, Women's Fiction, and Biographies & Memoirs), small movements in the displayed rating are tightly linked to changes in sales ranks, whereas in more scholarly or specialized categories (such as History & Criticism and various academic subfields) the same variation in ratings has little detectable relationship with subsequent sales. This genre-specific responsiveness is summarized in a continuous "star-rating effect index" that records how strongly sales in each category respond to shifts in the average rating.

The second part of the analysis links this genre-level variation to how users process the richer information contained in review text. Focusing on one-star reviews, the paper uses a large language model (OpenAI's o3) to construct two scores for each review: a vertical score that captures broadly relevant quality (clarity, coherence, factual accuracy, craftsmanship) and a horizontal score that captures match quality (the fit between the book and the reader's tastes, beliefs, or needs, such as "too technical," "very partisan," or "good for teens"). These vertical and horizontal scores are then related to the number of "helpful" votes a review receives. Across both high- and low-index genres, higher

vertical and horizontal scores are systematically associated with more helpful votes, and in pooled specifications the strength of these associations varies with the genre's star-rating effect index: the impact of vertical content on helpfulness grows with the index, while the impact of horizontal content diminishes.

Altogether, these patterns provide a descriptive backbone for the paper's interpretation of how learning operates in information-dense environments. The fact that ratings and sales move closely together in some genres but hardly at all in others, and that the mapping from review text to helpfulness depends on the same genre-level index, suggests that readers do not use Amazon's information architecture in a uniform way. In categories where the star-rating effect index is high, one natural interpretation is that many shoppers can rely on the displayed average rating as a sufficiently informative summary of others' experiences, and detailed critiques are most valued when they speak to widely relevant quality concerns. In low-index categories, where aggregate ratings appear less predictive of sales dynamics, the greater explanatory power of review content for helpfulness votes points toward a more text-based mode of evaluation in which audience fit and match-specific details play a larger role.

The remainder of the paper develops this consumer-learning interpretation, using the genre-level index and the review-content measures to separate simple rating-based shortcuts from more elaborate forms of learning on Amazon. A central alternative explanation, however, is that Amazon's own recommendation and ranking algorithms respond mechanically to ratings, review features, and recent demand, so that part of the observed rating–sales co-movement could arise from algorithmic changes in exposure rather than from consumers' direct processing of on-page information. We therefore also provide institutional background on Amazon's recommendation system and spell out whether such algorithmic responses could generate patterns similar to those documented here. Throughout the paper, these institutional details are used alongside the genre-level and review-text variation to assess how far the main results can be attributed to algorithmic exposure alone and how much is left for consumer learning.

This study relates to several strands of prior research. The first concerns online reviews, ratings, and demand. Reimers and Waldfogel (2021) show that both professional reviews and Amazon crowd star ratings have sizable causal effects on book sales, with crowd ratings generating substantially larger aggregate welfare gains because they cover far more products. Acemoglu et al. (2022) develop a Bayesian model of learning from online reviews that highlights how selection into reviewing affects whether and how quickly consumers can learn true product quality under different rating-system designs. Earlier empirical work, including Chevalier and Mayzlin (2006), demonstrates that variation in online book reviews causally affects relative sales across retailers, and Luca (2016) shows that Yelp ratings causally affect restaurant revenue using a regression discontinuity design based on Yelp's rounding rules. This paper adds to this literature by showing that, even within a single marketplace and uniform information environment, the extent to which consumers learn from and act on star ratings differs across genres, with some relying heavily on the heuristic and others turning instead to richer text signals.

The second strand draws on research on attention, learning, and persuasion. Models

of limited attention and context-dependent weighting, including Bordalo et al. (2013)'s salience framework, show that people tend to latch onto simple, prominent cues and give less weight to more informative but harder-to-process details. Work on rational inattention Sims (2003) makes a related point: when information processing is costly, individuals conserve effort, relying on coarse signals rather than analyzing everything in depth. Extending these ideas, the theory of Bayesian persuasion Kamenica and Gentzkow (2011) and the model-based approach of Schwartzstein and Sunderam (2021) examine how senders shape receivers' beliefs by deciding what information to highlight or suppress. Research on herd behavior and informational cascades, beginning with Banerjee (1992) and Bikhchandani et al. (1992), similarly shows how agents may set aside their own private signals and instead follow visible public cues when inference becomes challenging. The pattern documented in this paper, that consumers in some genres may lean more heavily on the mean star rating while others downplay it and instead read more meaning into the review text, fits naturally within this broader view. Star ratings offer a quick, attention-saving heuristic, while reviews provide richer but more demanding signals; which one consumers rely on shifts with the decision environment.

The third strand of the literature draws on works in information design, which study how the structure and precision of signals shape behavior. Bergemann et al. (2018) show that an optimal information seller screens heterogeneous users by offering menus that range from fully informative to deliberately coarse experiments. Taneva (2019) demonstrates that, even in strategic environments with multiple agents, providing only partial detail can be optimal, underscoring that coarse signals may be sufficient for some decision makers. Bergemann and Morris (2019) synthesize this broader literature and clarify when finer information partitions enhance choices and when coarser ones are preferable, highlighting the trade-offs audiences face between simplicity and richness. Complementing these theoretical insights, Haaland et al. (2023) show experimentally that the format and depth of information systematically influence individual belief updating.

The remainder of the paper is organized as follows. Section 2 summarizes past economic studies on the book market, describes the information environment on Amazon (including sales ranks, star ratings, and review text), explains the Amazon recommendation procedure, and outlines the construction of the dataset. Section 3 explains the empirical strategy and the construction of the main measures used in the analysis. Section 4 presents the main empirical findings. Section 5 discusses their implications, considers whether they could be attributed to non-learning mechanisms, and examines how the results can be used to improve the rating and review system in the book market. Section 6 concludes.

2 Background

2.1 Economic Studies of the Book Market

Economic work on the book market has expanded sharply in recent years, but it still lags behind research on other cultural industries. The survey in Cameron (2022) emphasizes that, relative to film and recorded music, economists have devoted comparatively little attention to the production, pricing, and consumption of books, even though they are a long-standing and central cultural medium. The emerging literature it synthesizes is organized around three themes that are directly relevant for this paper: how responsive book demand is to prices and income; how digital technologies and online platforms are transforming formats and distribution; and how the information-goods aspect of books, especially reviews and recommendations, shapes reading decisions.

Within this literature, Crosby (2022) treat books not as a single homogeneous good but as a bundle of competing formats. Using a discrete choice experiment with Australian readers, they estimate a latent-class model in which individuals choose among hardback, paperback, ebook, audiobook, and a no-purchase option. Results show strong heterogeneity in preferences and price sensitivity across reader types and formats, suggesting that digital disruption has not simply replaced print. This persistent heterogeneity across formats and genres motivates the genre-level perspective adopted in the present paper, in which responsiveness to ratings and reviews is allowed to vary across categories rather than being constrained to be uniform.

Reimers and Waldfogel (2021) shift the focus from prices and formats to the role of pre-purchase information in online book markets. They estimate how both expert reviews and crowd ratings causally affect demand for individual titles on Amazon, translating movements in sales rank into quantity responses. By translating changes in sales rank into quantities, they obtain elasticities of quantity with respect to Amazon's average star rating that range from roughly 0.4 to 0.8 depending on the number of underlying ratings. They also estimate a relatively small book-level price elasticity of around -0.17, consistent with money price being only one component of the full cost of reading. Their results establish that online ratings are an important driver of book sales; this paper builds on this insight by showing that the strength of the rating–demand linkage is itself heterogeneous across genres and systematically related to how readers evaluate review text.

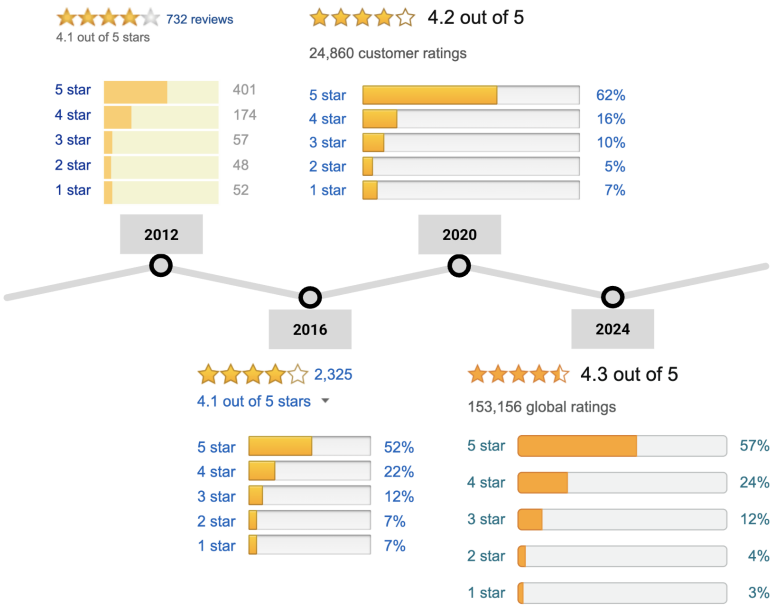
Finally, books are classic experience goods: their quality can be assessed only after reading. The theoretical analysis in Chen et al. (2022) highlights that when quality is not directly observable before purchase, reductions in search costs can have non-monotonic effects on welfare and firms' incentives to invest in quality. In environments where search is very cheap but product quality remains hard to evaluate ex ante, such as large online marketplaces, some form of pre-purchase information or friction is needed to sustain good matches and high quality. This perspective underscores the central role of Amazon's rating and review system in the book market and motivates the paper's focus on how readers use these signals when making purchase decisions.

2.2 Amazon's Book Marketplace

Amazon is a natural setting for studying how readers use ratings and reviews. The broader U.S. book market is large, around $40 billion in annual revenue with continued growth projected (Grand View Research (2023)). Within this market, Amazon has grown, since opening as an online bookstore in 1994, into the dominant force in U.S. book retail, now responsible for more than half of all book sales and roughly three quarters of online purchases (Chow and Gutterman (2020)). Its 2013 acquisition of Goodreads, a large social platform where millions of users record what they read, rate titles, and write reviews (Vinjamuri (2013); Olanoff (2013)), further underscores how central Amazon has become to the discovery and evaluation of books.

On each Amazon book page, shoppers encounter a dense bundle of crowd-generated information. Customers rate titles on a five-point scale, and each page displays an aggregate star rating (rounded to one decimal place) alongside a histogram of the distribution of 1–5 star reviews. Amazon reports that these star ratings are not simple averages but are produced by models that place greater weight on recent reviews and on feedback from verified purchasers (Amazon (2025b)). Individual reviews may include text, photos, or videos, and other users can vote on whether a review is "helpful," with the most helpful reviews surfaced more prominently. Amazon also applies authenticity checks and review-quality standards, removing content it deems inauthentic or unhelpful (Amazon (2025c)). As a result, shoppers see both a coarse summary signal (the average star rating) and a curated set of richer evaluations. Figure 1 illustrates examples of Amazon's rating layout across the years.

Figure 1: Amazon's Star Rating User Interface



The main outcome variable in this paper is Amazon's internal sales rank. For each book, Amazon reports a rank that summarizes its relative sales performance within a given category and country, with lower numbers indicating higher sales. Sales ranks are updated frequently and incorporate both recent and historical sales activity, although the precise algorithm is proprietary (Amazon (2025a)). The high temporal resolution of this measure makes it well suited for analyzing how changes in the information displayed on a product page correspond to subsequent shifts in relative demand.

2.3 Amazon's Recommendation System

Because algorithmic exposure is a central alternative mechanism linking ratings and reviews to sales ranks, this subsection summarizes the main features of Amazon's recommendation system. Early on, Amazon's recommendation architecture, described by Linden et al. (2003), replaced user-based collaborative filtering with an item-based approach. Rather than searching for customers with similar purchase histories, which scales poorly, Amazon precomputes a table of item-to-item similarities by identifying products that are purchased together more often than expected by chance. This offline computation enables fast recommendations at runtime: when a customer views or purchases an item, the system retrieves related items from the table and aggregates them into a ranked list. In this way, changes in the demand for a title can feed back into the set of products that are co-purchased with it and, through those co-purchase links, into how prominently it is displayed to other shoppers.

Over time, Amazon has refined how it measures relationships between items. Early methods relied on simple co-purchase counts, which could be skewed by heavy buyers or very popular products. Smith and Linden (2017) describe improvements that account for uneven customer purchasing patterns by estimating how often items would co-occur by chance and then comparing that baseline with actual data to highlight meaningful correlations rather than noise. The system also incorporates timing (items bought close together versus far apart) and purchase order (such as accessories bought after electronics) to interpret item relationships more accurately. For books, this implies that titles with similar purchase and browsing histories are more likely to be recommended together, and that recent shifts in a book's demand can quickly alter its platform exposure. As a result, a shock to ratings or reviews that raises sales for a given title may also change how often that title is recommended.

More recently, Amazon has incorporated LLMs to adjust the presentation of recommendations and product information. According to Amazon's 2024 announcement on generative-AI-based personalization, these models rewrite product descriptions and generate contextual recommendation labels based on recent customer search and browsing activity (Levine, 2024). These additions do not replace the underlying collaborative filtering system and were implemented after the sales window studied here (the 2018 calendar year). Consequently, while they illustrate the broader trend toward algorithmic curation of information, they do not play a direct role in the empirical patterns analyzed in this paper.

2.4 Dataset Construction

The empirical work in this study uses the dataset assembled by Reimers and Waldfogel (2021). Their construction begins with a list of books that were either actively selling in 2018 or received professional critical attention. To build this list, they merge several sources: all titles that appeared in the weekly USA Today top-150 bestseller lists during 2018; all books reviewed that year by the New York Times and five major U.S. newspapers (the Boston Globe, Chicago Tribune, Los Angeles Times, Wall Street Journal, and Washington Post); all books reviewed in 2018 in Publishers Weekly that were available in hardback or paperback before 2019; and books reviewed in 2018 by a group of prominent Goodreads reviewers.

The next step links these titles and their editions to Amazon. Using keepa.com, they collect daily information for each edition, including Amazon sales rank, price, number of user ratings, and average star rating for the 2018 calendar year, with separate coverage for the U.S., Canadian, and U.K. Amazon sites. Keepa reports sales ranks at relatively high frequency, which makes it possible to observe how pre-purchase information and professional reviews correspond to changes in sales rank over time. This process yields Amazon sales-rank data for about 94 percent of the title list, or 10,641 of 11,324 titles.

For the analyses in this paper, the Reimers and Waldfogel dataset is supplemented with Amazon genre classifications obtained from separate datasets of individual Amazon reviews (Ni et al. (2019), Hou et al. (2024)). Books in the Reimers and Waldfogel dataset are matched either on ISBNs or exact title matches to books in the larger review datasets, which are enriched with genre classifications. This addition allows each title to be linked to Amazon's internal genre categories and provides detailed information for each review, including verified-purchase status, helpfulness votes, review text, review time, and star rating. This additional data is found for approximately 83% of the corpus. Figure 1 lists the 15 largest genres based on this augmentation.

Table 1: Top 15 Genres by Book Count

Genre	Book Count
Children's Books	2,428
Thrillers & Suspense	1,281
Teen & Young Adult	961
Growing Up & Facts of Life	885
Biographies & Memoirs	736
Mystery	708
Science Fiction & Fantasy	647
Romance	612
Contemporary	577
Comics & Graphic Novels	462
Politics & Social Sciences	453
Graphic Novels	390
History	384
Action & Adventure	376
Historical Fiction	364

3 Methods

3.1 Panel Construction and Baseline Specification

We construct a panel of Amazon books observed at the title–country–day level. For each title, we have its sales rank, price, displayed star rating, total number of customer reviews, publication date, and Amazon's genre classification. The panel covers Amazon's U.S., Canadian, and U.K. marketplaces.

The dependent variable in our regressions is the logarithm of the daily sales rank. Within each title–country series, we construct one-day lags of the log sales rank, the star rating, and the log number of reviews. To ensure that these lags truly represent consecutive days, we treat a lag as missing whenever the prior observation for that title–country pair is not exactly one calendar day earlier. We also compute third-degree polynomials in days until and since publication, following Reimers and Waldfogel (2021). We restrict the panel to days with complete data for the log sales rank and its lag, the lagged log star rating, the log price, the lagged log number of reviews, and the publication-age polynomials. Observations without a valid previous-day record for the lagged variables (including the first observed day for each title and days with gaps in the daily sequence) are dropped. These criteria reduce the working sample from 8,705,192 title–country–day observations to approximately 3,213,344 daily observations.

Our baseline specification relates today's log sales rank to its one-day lag, today's log price, the lagged log number of reviews, and the lagged log star rating. Following Reimers and Waldfogel (2021), we absorb time-invariant differences in demand across titles and countries using title–country fixed effects; we also include publication-age polynomials and indicators for reviews and recommendations published in major outlets. Formally, for title j in country c on day t, we estimate:

$$\begin{aligned}\log(\text{Rank}_{jct}) = \alpha_{jc} + \rho\, \log(\text{Rank}_{jc,t-1}) + \beta_p \log(\text{Price}_{jct}) + \beta_r \log(\text{Reviews}_{jc,t-1}) \\ + \beta_s \log(\text{Stars}_{jc,t-1}) + f(\text{Age}_{jct}) \\ + \sum_k \gamma_k\, \text{Recommendation Indicator}^{(k)}_{jc,t-1} + \varepsilon_{jct}.\end{aligned}$$

where α_{jc} denotes title–country fixed effects, Age_{jct} is the number of days until or since publication, and $f(\cdot)$ is a polynomial in age of up to third degree. We estimate the model by ordinary least squares (OLS) with heteroskedasticity-robust standard errors. An augmented specification includes an interaction between the lagged log number of reviews and the lagged log star rating, allowing the relationship between additional reviews and changes in sales rank to vary with the rating level.

3.2 Genre-Level Star-Rating Effects

To study how the association between star ratings and sales ranks varies across book categories, we use the canonical Amazon genre tags associated with each title. For each genre

with sufficient coverage, we re-estimate the sales-rank specification described above on the subsample of titles in that genre, using the same within-title demeaning, controls, and robust standard errors. From each genre-specific regression, we extract the coefficient on the lagged log star rating as a summary measure of how strongly changes in the average star rating are associated with subsequent changes in sales rank within that genre.

These coefficients are used in two ways. First, we rank genres by the magnitude of this star-rating effect and define high-effect and low-effect sets, consisting of the genres with the strongest and weakest estimated associations between lagged ratings and subsequent sales ranks. Second, we map the full set of genre-specific coefficients into a continuous genre star-rating effect index, normalized to lie between 0 and 1, with higher values indicating a stronger estimated impact of the lagged log star rating on relative sales performance.

3.3 Vertical & Horizontal Review Scores

To characterize the content of individual reviews, we use a large language model, namely OpenAI's o3, to assign two scores to each review. A vertical score captures absolute or objective quality, including clarity of writing, coherence of argument, factual accuracy, and the quality of editing or research. The horizontal score captures content related to specific audiences or use cases (for example, whether the book is described as suitable for children, highly technical, or oriented toward a particular viewpoint). Each review receives a vertical and horizontal score on a 0 to 5 scale. For regression analysis, we standardize these scores within the review sample, yielding z-scored vertical and horizontal measures that are comparable across reviews and genres. The prompt used to extract these scores can be found in Appendix Figure A1. We note that the LLM use in this scenario is largely exploratory and is yet be validated in line with Ludwig et al. (2025).

3.4 Helpfulness Regressions for One-Star Reviews

To examine how review content is related to helpfulness evaluations, we analyze individual one-star reviews associated with titles in the main sample and assigned to the genres studied above. For each review, we observe the number of "helpful" votes on Amazon and define the dependent variable as $\log(1 + \text{Helpful Votes})$.

We estimate three OLS specifications with robust standard errors. In the first two, we run separate regressions for reviews in the high-effect and low-effect genre groups (the top five genres from each end of the star-rating effect index). In each case, $\log(1 + \text{Helpful Votes})$ is regressed on the standardized vertical and horizontal scores. In the third specification, we pool all one-star reviews across genres and include the continuous genre star-rating effect index and its interactions with both standardized scores. This specification allows the association between review content and helpfulness votes to vary systematically with the genre's estimated star-rating effect.