

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

Omar Abdel Haq \*

December 15, 2025

## 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.

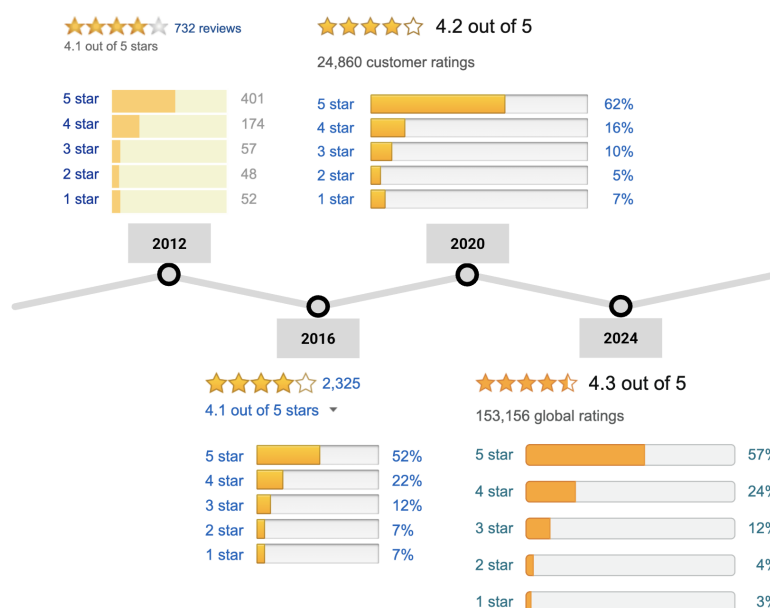
Finally, books are classic experience goods: their quality can be assessed only after reading. The theoretical analysis in Chen et al. (2022) highlights that when quality is not directly observable before purchase, reductions in search costs can have non-monotonic effects on welfare and firms’ incentives 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:

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

where  $\alpha_{jc}$  denotes title–country fixed effects,  $\text{Age}_{jct}$  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 to 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.

## 4 Results

Table 2 summarizes how Amazon’s variables relate to within-title changes in sales performance. The coefficients capture how day-to-day movements in ratings, reviews, and prices are associated with movements in a title’s log sales rank, holding fixed all time-invariant differences across books.

The estimates show substantial persistence in sales. The coefficient on lagged log sales rank is about 0.78, indicating that most of today’s rank is explained by yesterday’s position. Prices matter in the expected direction: higher Amazon list prices are associated with worse sales ranks. A one percent increase in price corresponds to roughly a 0.08 percent increase in log sales rank.

Table 2: Amazon Sales Rank Predictors

	(1)	(2)
Lagged Log Sales Rank	0.781*** (0.000)	0.781*** (0.000)
Log Amazon Price	0.081*** (0.002)	0.082*** (0.002)
Lagged Log Number of Reviews	0.061*** (0.001)	0.151*** (0.005)
Lagged Log Star Rating	−0.078*** (0.008)	−0.014 (0.009)
Number of Reviews $\times$ Star Rating <i>Interaction of Lagged Log Terms</i>		−0.061*** (0.004)
Demeaned within Book	Yes	Yes
Polynomials of Time Until & Since Publication	Yes	Yes
Observations	3,213,344	3,213,344
$R^2$	0.670	0.670

The table reports the results of OLS regressions estimated on a daily panel of Amazon book titles, where each title–country combination is treated as a distinct book. The dependent variable is the log of the daily sales rank. All specifications are estimated after demeaning the outcome and regressors within book, so the coefficients are identified from within-title variation over time rather than from cross-sectional differences across titles. Both columns also include flexible polynomial controls in days until and since publication.

Column (1) reports a baseline specification that relates today’s log sales rank to the previous day’s log sales rank, today’s log Amazon price, the previous day’s log number of reviews, and the previous day’s star rating. Column (2) augments this specification by adding an interaction between the previous day’s log number of reviews and the previous day’s star rating, allowing the effect of additional reviews to depend on the rating level. The sample is restricted to days for which all variables are observed and a previous-day observation exists for the lagged covariates. Robust standard errors are reported in parentheses, and statistical significance at the 10%, 5%, and 1% levels is denoted by \*, \*\*, and \*\*\*, respectively.

The variables of primary interest are the star rating and the count of customer reviews. In the baseline specification without interactions (column (1)), a higher lagged log star rating is associated with a subsequent improvement in sales. A one-unit increase in the previous day’s log star rating corresponds to a reduction in log sales rank of about 0.08, even after conditioning on lagged rank, price, and review volume. Under the identifying assumptions in Section 3, this pattern is consistent with the interpretation that changes in the displayed average rating have economically meaningful short-run effects on relative sales performance.

Column (2) adds an interaction between the lagged log review count and the lagged log star rating. The negative, highly significant interaction shows that the association between additional reviews and subsequent sales performance depends on the current rating. Over the observed range, more reviews are associated with higher (worse) ranks, but this association is much weaker for well-rated books than for poorly rated ones. In this sense, the marginal effect of new user feedback is state dependent: extra reviews hurt low-rated titles most and high-rated titles least, reinforcing existing demand differences.

Figure 2 shows that the impact of the star rating varies considerably across genres. The figure plots, for each genre, the estimated coefficient on the lagged average rating from regressions analogous to those in Table 2. All specifications include the same controls and title fixed effects, and all use log sales rank as the dependent variable. The coefficients are generally negative, indicating that higher lagged ratings are associated with improved subsequent sales, but their magnitudes differ substantially. Some genres exhibit large effects of rating changes on sales ranks, while others show effects closer to zero.

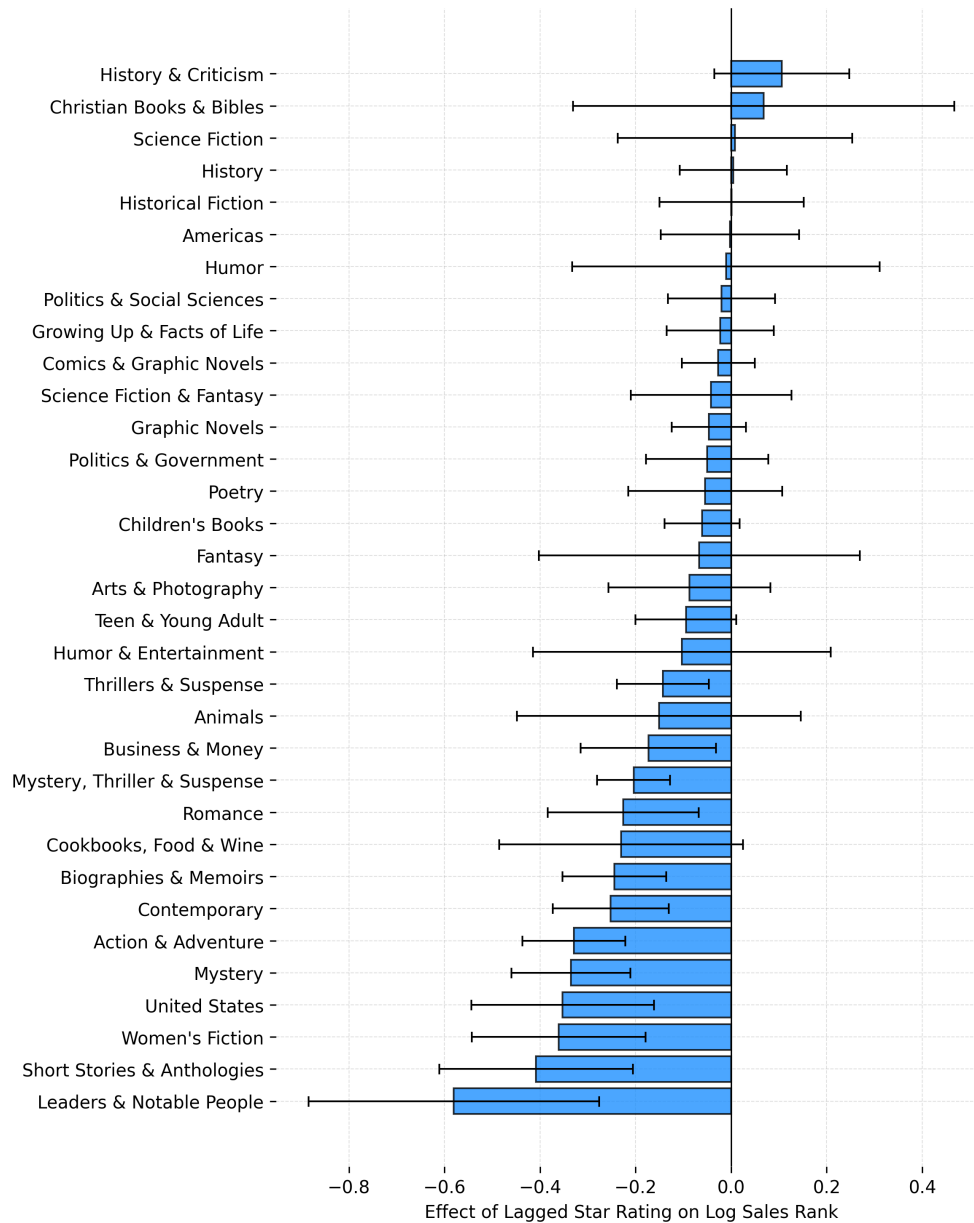
We use these coefficients to construct a genre-specific star-rating effect index, which summarizes the strength of the association between lagged ratings and subsequent sales within each category. Under the identifying assumptions in Section 3, this index can be interpreted as capturing heterogeneity in the effect of star ratings on relative sales performance across genres. We use it to define the high- and low-effect genre groups in the next analysis.

Table 3 turns from purchases to the evaluation of individual one-star reviews. In our data, one-star reviews receive the highest number of helpful votes, making them a natural setting for studying how review content relates to helpfulness assessments. The dependent variable is the log of one plus the number of helpful votes received. The key regressors are standardized vertical and horizontal review scores generated by OpenAI’s o3 model. The vertical score summarizes content related to product quality, while the horizontal score summarizes content related to match-specific or taste-based issues.

In both the high- and low-effect genre subsets (columns (1) and (2)), the vertical and horizontal scores are positively and significantly associated with helpful votes. In the high-effect genres, a one standard deviation increase in the vertical score is associated with an increase in log helpful votes of about 0.31, and a similar increase in the horizontal score is associated with an increase of about 0.15. The estimated effects are somewhat larger in the low-effect genres, but the qualitative pattern is similar: even among one-star reviews, higher vertical and horizontal scores are systematically linked to more helpful

votes. The  $R^2$  values indicate that these two variables account for a larger share of the variation in helpfulness within the low-effect genres than within the high-effect ones.

Figure 2: Lagged Log Star Rating Effect by Genre



Each bar shows the estimated effect of the previous day's star rating on today's log sales rank for the genre listed on the vertical axis. Estimates come from separate OLS regressions on daily title-country panels with book fixed effects and standard controls for lagged sales rank, price, reviews, and time until/since publication; standard errors are robust. Whiskers denote 95%-confidence intervals, and negative coefficients indicate that higher lagged log star ratings are associated with subsequent improvements in a book's sales rank within the corresponding genre.

The pooled specification in column (3) incorporates the genre star-rating effect index. The negative coefficient on the index indicates that, holding review content constant, one-star reviews in genres with larger estimated star-rating effects tend to receive fewer helpful votes on average. The interaction terms show that the association between review content and helpfulness varies with the index: as the index increases, the coefficient on the vertical score becomes more positive, while the coefficient on the horizontal score becomes less positive. Thus, the way vertical and horizontal content relate to helpful votes differs systematically across genres depending on the estimated strength of the star-rating effect.

Table 3: 1-Star Review Helpfulness Estimates

Genres	Dependent Variable: Log (1 + Helpful Votes)		
	High Rating Effect (1)	Low Rating Effect (2)	All (3)
Vertical Score	0.306***	0.414***	0.143***
<i>Standardized within Sample</i>	(0.020)	(0.026)	(0.024)
Horizontal Score	0.146***	0.207***	0.261***
<i>Standardized within Sample</i>	(0.020)	(0.027)	(0.022)
Genre Star-Rating Index			-0.823***
<i>Normalized</i>			(0.055)
Vertical Score $\times$ Star-Rating Index			0.359***
			(0.057)
Horizontal Score $\times$ Star-Rating Index			-0.168***
			(0.054)
Observations	4,197	1,924	26,284
$R^2$	0.090	0.162	0.109

Each column reports coefficients from an OLS regression where the dependent variable is Log (1 + Helpful Votes) for 1-star reviews. The key regressors are standardized vertical and horizontal review scores, which summarize how a language model rates the review text along an “objective quality” dimension (vertical) and a “match quality” dimension (horizontal). Columns (1) and (2) estimate separate specifications for reviews in the five genres with the highest and lowest star rating effects on sales rank, respectively. Column (3) pools all examined genres and adds a genre-level star-rating index (normalized between 0 and 1) as well as its interactions with the vertical and horizontal scores, allowing the relationship between these scores and helpfulness to vary systematically with the extent of the star-rating effect on sales. Standard errors are reported in parentheses. Statistical significance at the 10%, 5%, and 1% levels is denoted by \*, \*\*, and \*\*\*, respectively.

Taken together with the sales-rank results, these patterns show that Amazon’s aggregated rating is closely linked both to relative sales performance and to how one-star reviews are evaluated. The mean star rating is strongly associated with sales, and the vertical and horizontal content of one-star reviews are strongly associated with helpfulness assessments, with both sets of relationships varying across genres.

## 5 Discussion

This section interprets the sales-rank and helpfulness estimates, with an emphasis on what they imply about how ratings and reviews matter in practice. We first discuss when the lagged star-rating coefficients in Table 2 can be read as causal, and through which channels such effects might operate. We then use the cross-genre heterogeneity and the helpfulness regressions to argue that differences in how readers use star ratings and review text are likely to be an important part of the story.

### 5.1 Interpreting the Star-Rating Effect

The panel specification in Table 2 compares, within each title-country, days that follow relatively high lagged ratings and review counts to days that follow relatively low ones, controlling for lagged rank, price, publication age, and observed professional-review events. Under the identifying assumptions in Section 3, that there are no remaining time-varying shocks that simultaneously move lagged ratings, lagged review counts, and next-day sales ranks conditional on these controls, the negative coefficient on the lagged log star rating can be interpreted as a short-run causal effect of changes in the displayed rating on relative sales performance.

Even under this interpretation, several channels could generate the observed relationship. One possibility is that the star rating functions primarily as an information signal for shoppers: when the average rating ticks up (down), some marginal viewers become more (less) willing to buy. A second possibility is that the rating affects sales indirectly, by altering Amazon’s internal allocation of visibility: for example, through its search or recommendation algorithms. This would mean that higher-rated books receive more impressions and thus more sales. A third possibility is that the coefficients reflect residual confounding from unobserved shocks, such as marketing campaigns or media mentions that are not fully captured by the included professional-review indicators and that both increase sales and change the composition of reviewers.

The genre-level evidence in Figure 2 helps to assess these explanations. The estimated effects of lagged ratings on sales ranks differ sharply across categories, with large negative coefficients in commercial fiction and narrative non-fiction genres and much smaller or statistically insignificant coefficients in more scholarly or specialized categories. Because the sales-rank algorithm itself does not depend on genre labels, explaining this pattern purely through internal visibility rules would require Amazon to use ratings in a strongly genre-contingent way: for example, allowing star ratings to affect search or recommendation rankings in Romance but not in History & Criticism. While such heterogeneity in platform algorithms cannot be ruled out, it is not an obvious baseline.

A purely non-behavioral interpretation in which the genre star-rating effect index simply reflects how strongly Amazon’s collaborative-filtering and ranking systems amplify ratings would also predict tight links between the index and variables that directly feed into those systems, such as sales performance and review volume. Empirically, these

links are weak. At the genre level, the correlation between the star-rating effect index and average log sales rank is only  $r = -0.029$  ( $p = 0.004$ ), and the correlation with average log review count is  $r = -0.168$  ( $p < 0.001$ ). The first is essentially zero in magnitude, and the second, while precisely estimated, is modest and negative. Genres where ratings appear to have the largest estimated impact on sales do not systematically correspond to those with the highest sales or the largest review bases, as one would expect if the index were mainly picking up algorithmic amplification of already popular or heavily reviewed titles. Instead, the index is nearly orthogonal to these basic platform signals, making it unlikely that it is driven solely by mechanical features of Amazon’s recommendation and ranking algorithms.

By contrast, the observed pattern fits naturally with the idea that the same rating signal has different importance across categories. In genres where purchases are likelier to be discretionary and prior knowledge is limited, an aggregate rating is likely to be a convenient summary of others’ experiences, which makes it a more influential cue. In genres where purchases are likelier to be guided by outside information or detailed content needs, the same coarse signal may be less informative and so less heavily used. On this interpretation, the genre-specific star-rating effect index summarizes how strongly demand in each category responds to movements in the rating signal, whether that response is mediated directly through individual choice, through changes in platform visibility, or both, and the cross-genre variation in the index reflects genuine differences in how readers use the rating heuristic rather than just differences in how the platform’s algorithms treat ratings.

## 5.2 What Matters When Ratings Matter Less?

The helpfulness regressions in Table 3 speak to what aspects of review content are associated with attention and approval, and how that association varies with the estimated star-rating effects across genres. Across all specifications, higher vertical and horizontal content scores are linked to more helpful votes on one-star reviews, indicating that review text that contains more information about either broadly relevant quality or match-specific issues tends to be rated as more helpful.

The split-sample estimates for high-effect and low-effect genres show that the magnitudes of these associations, and the fraction of variation they explain, differ by category. In the low-effect genres, where lagged ratings have weaker estimated associations with subsequent sales, vertical and horizontal content scores are somewhat more predictive of helpfulness, and the  $R^2$  values are higher. One natural reading is that in categories where the average rating is a weaker predictor of sales dynamics, users’ evaluations place relatively greater weight on the text of the reviews.

The pooled specification with the continuous genre star-rating effect index provides a more systematic view. Holding text content constant, one-star reviews in genres with larger estimated star-rating effects tend to receive fewer helpful votes on average, and the interactions between the index and the content scores indicate that the way vertical and



horizontal content relate to helpfulness shifts with the strength of the star-rating effect. As the index increases, the association between vertical content and helpfulness becomes stronger, while the association between horizontal content becomes weaker. This suggests that in genres where the rating signal is more tightly linked to sales, one-star reviews that speak to widely relevant quality issues (clarity, coherence, accuracy) are especially likely to be rated as helpful, whereas in genres where the rating signal is weaker, match-specific information plays a relatively larger role in helpfulness assessments.

Viewed through the lens of heuristic decision making, these patterns suggest that differences in how heavily readers rely on star ratings can be used to anticipate what kind of information they are looking for in reviews. In high-index genres, where many shoppers seem comfortable using the average rating as a heuristic for “is this good enough?” review text is most valued when it clarifies broadly relevant quality concerns. In low-index genres, where the rating heuristic appears less informative, readers instead turn to the text to assess horizontal fit—whether the book matches their technical background, beliefs, or intended use. From a design perspective, this implies that platforms could use observed reliance on ratings to tailor which reviews are surfaced. For example, in low-index genres, it may be especially useful to filter or highlight reviews that contain strong horizontal “fit” information, whereas in high-index genres it may be more valuable to surface reviews that emphasize vertical quality. Such tailoring could reduce information overload and improve matching efficiency by bringing the most decision-relevant textual signals to the top of the page in the categories where they matter most.

Tailoring which reviews are surfaced is particularly relevant given Amazon’s introduction of review-organizing modules on product pages. Figure 3 illustrates these organizers for two books that lie at opposite ends of the star-rating effect index. For the programming textbook in Panel A (a low-index genre), the organizer aggregates mixed feedback on clarity and readability: some readers describe the writing as accessible, while others find it exhausting or opaque. This pattern is consistent with our interpretation that audience fit is central in such categories and that the coarse average star rating is a relatively weak guide. Our results suggest that organizer designs for low-index genres could be made substantially more informative by explicitly eliciting and using simple background information from reviewers, such as prior experience, intended use, or technical goals, and then grouping or filtering reviews by these audience segments and their horizontal fit assessments. By contrast, the historical fiction novel in Panel B (a high-index genre) is summarized almost entirely in terms of vertical quality themes (for example, plot and writing), which aligns with the idea that the average rating is already a good heuristic for overall quality and that broadly relevant quality information is what additional review text mainly needs to clarify. The evidence implies that platforms could use observed reliance on ratings to adapt their review-organization tools: emphasizing audience-segmented, match-specific information where the rating heuristic is weak, and emphasizing vertical quality summaries where it is strong.

Figure 3: Differences between Review Organizers by Book Type

*Panel A: Amazon Review Organizer for a Python Programming Textbook*

**Customers say**

Customers find this Python programming book comprehensive and ideal for learning, taking them from basics to advanced concepts. The book's readability and writing style receive mixed feedback - while some say it reads like a novel and is well-written, others find it exhausting to read and unclearly written. The book's length is also a point of contention, with some praising its extensive content while others find it overly long.

 Generated from the text of customer reviews

**Select to learn more**

✓ Detail (156) | ✓ Learning material (118) | ✓ Python knowledge (60) | ✓ Ease of use (41) |  
— Readability (75) | — Book length (64) | — Language knowledge (51) | — Writing style (44)

*Panel B: Amazon Review Organizer for a Historical Fiction Novel*

**Customers say**

Customers find this book to be one of their favorites, praising its riveting prose and unique plot structure. The writing is beautifully poetic, and customers appreciate how it helps describe the horrors of war while being more real than non-fiction. The stories are heart-wrenching and emotionally believable, with one customer noting how it delves deeply into the psyche of the writer.

 Generated from the text of customer reviews

**Select to learn more**

✓ Readability (896) | ✓ Story telling (328) | ✓ Writing quality (318) | ✓ Thought provoking (201) |  
✓ Pacing (197) | ✓ Emotional content (160) | ✓ Realistic (106) | ✓ Detail (73)

These patterns do not rule out alternative explanations: differences across genres in typical review length, age, or readership could also shape helpfulness votes, and the vertical and horizontal scores themselves are noisy, model-based proxies. Nonetheless, the systematic way in which review-content coefficients and helpfulness  $R^2$  values vary with the genre star-rating effect index indicates that the mapping from review text to perceived helpfulness is not homogeneous across categories and is related to the role that ratings appear to play in sales dynamics. Together with the weak correlation between the index and basic platform signals such as sales performance and review volume, this evidence is more naturally explained by heterogeneous use of rating and text heuristics across genres than by a purely non-behavioral, algorithmic account.

## 6 Conclusion

This paper has studied how readers use ratings and reviews in Amazon’s book marketplace. Using a daily title–country panel, we related changes in sales ranks to lagged star ratings, and documented substantial heterogeneity across genres in the strength of the rating–sales linkage. We summarized this heterogeneity in a genre-level star-rating effect index and linked it to text-based measures of review content, constructed with a large language model. One-star reviews with higher vertical (quality-related) and horizontal (match-related) scores receive more helpful votes, and the relative importance of these dimensions in predicting helpfulness varies systematically with the genre’s star-rating effect index.

The results suggest that readers in different genres make different trade-offs between coarse rating heuristics and richer textual information. In high-index genres, where the rating signal is strongly associated with sales dynamics, readers appear to rely more on the average rating and value one-star reviews that clarify broadly relevant quality concerns. In low-index genres, where the rating–sales link is weaker, helpfulness votes place relatively greater weight on horizontal fit content. The weak correlations between the genre star-rating effect index and basic platform signals such as average sales performance and review volume make it unlikely that the index simply reflects mechanical features of Amazon’s recommendation algorithms, and instead point toward genuine heterogeneity in how ratings and reviews are used. From a design perspective, this suggests that platforms could tailor which reviews they surface by genre: for example, highlighting vertically informative reviews in high-index categories and horizontally informative “fit” reviews in low-index categories, to reduce information overload and improve matching.

The analysis has several limitations that point to directions for future work. First, the estimates are based on observational data and a dynamic panel specification, so they are best interpreted as descriptive associations: they are consistent with, but do not conclusively establish, causal learning from ratings and reviews. Second, additional evidence is needed to assess the role of the recommendation system in generating the observed results. One avenue for exploration is to examine the books each title is co-recommended with, and test whether movement in one title is systematically associated with movement in its co-recommended neighbors, or whether the behavioral explanation is the more dominant factor at play. Third, the vertical and horizontal content scores are model-based proxies, and selection into reviewing and voting may differ across genres in ways that are not fully observed. Finally, the LLM-based construction of the vertical and horizontal scores needs to be further validated, and additional aspects of the reviews, such as length, Lexile level, and other linguistic features, should be controlled for in the relevant regressions.

Future research could combine similar genre-level and text-based measures with exogenous shocks to ratings or exposure, or with randomized experiments on information display, to sharpen identification. Extending the approach to other product categories, time periods, and platforms, and incorporating richer measures of user heterogeneity,

would help to assess how general these patterns are and how information design can best accommodate heuristic decision making while still supporting high-quality matches.

## References

- Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2022). Learning from Reviews: The Selection Effect and the Speed of Learning. *Econometrica* 90(6), 2857–2899.
- Amazon (2025a). About Amazon Best Sellers Rank.
- Amazon (2025b). Everything You Need to Know About Amazon Reviews.
- Amazon (2025c). Reviews: Amazon Customer Service.
- Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics* 107(3), 797–817.
- Bergemann, D., A. Bonatti, and A. Smolin (2018). The Design and Price of Information. *American Economic Review* 108(1), 1–48.
- Bergemann, D. and S. Morris (2019). Information Design: A Unified Perspective. *Journal of Economic Literature* 57(1), 44–95.
- Bikhchandani, S., D. Hirshleifer, and I. Welch (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy* 100(5), 992–1026.
- Bordalo, P., N. Gennaioli, and A. Shleifer (2013). Salience and Consumer Choice. *Journal of Political Economy* 121(5), 803–843.
- Cameron, S. (2022). Cultural Economics, Books and Reading. pp. 1–10.
- Chen, Y., Z. Li, and T. Zhang (2022). Experience Goods and Consumer Search. *American Economic Journal: Microeconomics* 14(3), 591–621.
- Chevalier, J. A. and D. Mayzlin (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research* 43(3), 345–354.
- Chow, A. R. and A. Gutterman (2020, April). How Coronavirus Is Affecting Independent Bookstores.
- Crosby, P. (2022). Don’t Judge a Book by its Cover: Examining Digital Disruption in the Book Industry Using a Stated Preference Approach. In *The Economics of Books and Reading*, pp. 91–121.
- Grand View Research (2023). The United States Books Market Size & Outlook, 2033.
- Haaland, I., C. Roth, and J. Wohlfart (2023). Designing Information Provision Experiments. *Journal of Economic Literature* 61(1), 3–40.
- Hou, Y., J. Li, Z. He, A. Yan, X. Chen, and J. McAuley (2024). Bridging Language and Items for Retrieval and Recommendation.

- Kamenica, E. and M. Gentzkow (2011). Bayesian Persuasion. *American Economic Review* 101(6), 2590–2615.
- Levine, I. (2024, Sep).
- Linden, G., B. Smith, and J. York (2003). Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing* 7(1), 76–80.
- Luca, M. (2016). Reviews, Reputation, and Revenue: The Case of Yelp.com. *Harvard Business School Working Paper* (12-016).
- Ludwig, J., S. Mullainathan, and A. Rambachan (2025). Large Language Models: An Applied Econometric Framework.
- Ni, J., J. Li, and J. McAuley (2019). Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 188–197.
- Olanoff, D. (2013, March). Amazon Acquires Social Reading Site Goodreads.
- Reimers, I. and J. Waldfogel (2021). Digitization and Pre-Purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings. *American Economic Review* 111(6), 1944–1971.
- Schwartzstein, J. and A. Sunderam (2021). Using Models to Persuade. *American Economic Review* 111(1), 276–323.
- Sims, C. (2003). Implications of Rational Inattention. *Journal of Monetary Economics* 50(3), 665–690.
- Smith, B. and G. Linden (2017). Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Computing* 21(3), 12–18.
- Taneva, I. (2019). Information Design. *American Economic Journal: Microeconomics* 11(4), 151–185.
- Vinjamuri, D. (2013, March). Three Hidden Benefits of the Amazon Acquisition of Goodreads.

## Appendix A: Supplementary Materials

Figure A1: Prompt Used to Elicit Vertical and Horizontal Review Scores

SYSTEM: You are an assistant analyzing Amazon book reviews. Your task is to score, on a 0–5 scale, how much vertical quality evaluation and horizontal taste/audience matching information each review contains. Here are some useful definitions:

- Vertical content focuses on the book’s overall or absolute quality, mostly independent of who is reading it. Examples include writing quality, editing, structure, pacing, factual accuracy, research quality, credibility, translation quality, production value.
- Horizontal content focuses on WHO the book is for and whether it is a good fit for particular tastes, preferences, or audiences. Examples include whether the book is too dark/light, too biased/neutral, too technical/basic, or aimed at specific groups.

Respond with exactly two integers between 0 and 5 inclusive, in this order: Vertical Score, Horizontal Score. Do not include any other text in your response.

USER: How would you score the following review on its vertical and horizontal components?

*[Review Placeholder]*