

Digitization and Product Discovery: The Causal and Welfare Impacts of Reviews and Crowd Ratings

Imke Reimers
Northeastern University

Joel Waldfogel
University of Minnesota, NBER, ZEW

November 30, 2019

Abstract

Digitization has led to product proliferation, straining traditional institutions for product discovery; but digitization has also spawned crowd-based rating systems providing information on all products. Using the book market as our context, we assemble data on daily Amazon sales ranks, star ratings, and prices for thousands of the top-selling books of 2018, along with information on their professional reviews in the New York Times and other major outlets. Using various fixed effects and discontinuity-based empirical strategies, we estimate that a New York Times review raises estimated sales by 78 percent during the first five days following a review and by 5.2 percent overall; and the elasticity of sales with respect to an Amazon star is about 0.75. We use these causal estimates to calibrate structural models of demand for measuring the welfare impact of pre-purchase information. The effects of professional reviews, clearly evident in the descriptive results, also arise in the welfare analysis. While the existence of professional reviews raises consumer surplus by almost \$3 million, the effects of having a star rating system are much larger. The existence of star ratings raises consumer surplus from Amazon book purchases by over \$40 million or by roughly 15 times as much as the overall market effect of professional reviews. While crowd-based information now accounts for the vast majority of pre-purchase information, the absolute effects of professional reviews have not declined over time.

When choosing among experience goods, consumers benefit from guidance prior to purchase. Traditionally, professional critics – such as product reviewers in prominent media outlets – played important roles in providing this guidance.¹ One of digitization’s many impacts has been a sharp increase in the number of new creative products, exacerbating the product discovery problem while also taxing the capacity of professional critics to review all of the offerings.² The possibility of realizing the welfare gains from a plethora of new products is diminished by the difficulty that consumers might have in discovering which products to consume. While the number of new products has always exceeded the capacity of professionals to review them, this gap has only grown with digitization. Crowd-based ratings – such as Amazon stars based on user ratings – on the other hand, are available for all products, raising the possibility that another facet of digitization, ubiquitous crowd ratings, can provide information that allows the realization of welfare gains from discovery of new products.

These considerations lead to the question of how the new crowd-based pre-purchase product information made available by digitization affects purchase behavior and, by extension, welfare, along with the related question of whether professional critics or the crowd will coordinate the matching of products to consumers. To these ends, we ask the following questions. First, do professional reviewers and crowd ratings have causal impacts on demand; and if so, how large are these impacts? Second, how does the growing availability of crowd-based ratings alongside the reviews of professional critics affect which, how many, and which sorts of, products are consumed? Third, in addition to the specific causal impacts of reviews and ratings on sales, how does the existence of these pre-purchase information institutions – professional reviews and crowd ratings – affect the welfare of consumers? And fourth, has the growth of crowd-based ratings reduced the influence of professional reviews.

This paper explores these questions in the market for books. Books provide an auspicious

¹See [Deutschman \(2004\)](#); [Pompeo \(2017\)](#), or [Martin \(2011\)](#) for descriptions of various professional critics and their influence on product markets.

²See [Waldfoegel \(2017\)](#) for evidence on the growth in new products. In 2014 New York Times film critic Manohla Dargis implored the film industry to make fewer movies. See [Dargis \(2014\)](#).

context for study for a few reasons. First, books are experience goods, so that pre-purchase information is potentially useful. Second, the number of professional reviews, and particularly the number appearing in highly visible outlets, is relatively small and therefore feasible to observe and quantify. Third, and perhaps most important, we have high frequency data on book demand at Amazon – which accounts for about 45 percent of the physical books market – that helps identify causal relationships.³ We have daily measures of Amazon sales ranks and their crowd-based star ratings, for 4,283 titles (appearing in 9,146 editions) during 2018, for three English-language Amazon sales domains (the US, Canada, and the UK). Reviews and star ratings are inherently endogenous, as raters and reviewers decide whether and when to give feedback, in addition to what they say. More appealing books sell more and receive more positive feedback. Our high-frequency data from multiple platforms allow us to deal with this endogeneity using three strategies, one for reviews and two for star ratings. We treat the appearance of a professional review as a discontinuous jump in attention delivered to the title, and we look for a corresponding jump in our daily sales measure. We measure the impacts of star ratings with a cross-platform longitudinal comparison for measuring causal impacts of star ratings. We also provide evidence for a causal mechanism using a discontinuity approach based on Amazon’s visual display of ratings in half star increments.

Our descriptive analysis gives us credible causal evidence on the links between pre-purchase information – reviews and ratings – and sales ranks; but we also seek to perform welfare analyses, which requires two translational steps. First, we transform effects of pre-purchase information on sales ranks into effects on quantities, allowing the calculation of the elasticities of quantity sold with respect to the Amazon price, the star rating, and the percentage impact of a professional review on sales. Second, we use those elasticities to calibrate nested logit models of demand that facilitate welfare analysis. Our measured welfare effects of ratings and reviews allow the consumers’ ex ante choice utility to differ from their ex post consumption utility.

³See <https://www.publishersweekly.com/pw/by-topic/industry-news/financial-reporting/article/78929-print-unit-sales-increased-1-3-in-2018.html>.

We have four broad findings. First, professional review outlets, notably the New York Times and to a lesser extent other US newspapers, have clear impacts on sales. In the five days following a New York Times review, a book's estimated sales improve by 78 percent, on average, with slightly larger effects for more positive reviews. Over the entire year, a New York Times review raises sales by 5.2 percent. Second, the crowd also has clear effects on sales: using a variety of measurement approaches including title fixed effects, cross-platform intertemporal comparisons, and discontinuity approaches, the elasticity of sales with respect to Amazon stars averages about 0.75, and it is larger when the stars are based on more underlying ratings. Third, while both professional critics and crowd ratings affect consumer welfare, the effects of crowd ratings are much larger. While professional reviews raise consumer surplus by just under \$3 million, the effect of star ratings, at \$41 million of additional consumer surplus for Amazon sales alone, is roughly 15 times larger. Fourth, while the welfare benefit of the crowd adds substantially to the influence of professionals, a supplementary analysis of weekly sales data on books reviewed by the New York Times, 2004-2018, shows that impact of professional reviews has not waned. We conclude that digitization has delivered not only a proliferation of new products but also new information mechanisms that add substantially to the value of the pre-purchase information available to consumers from traditional review sources. These crowd-based reviews provide pre-purchase information on all products, including those neglected by professional critics, and do so without undermining effects of professional critics on the books and genres they do cover.

The paper proceeds in six sections. Section 1 provides background on the book market, the evolution of the information environment with digitization, and a discussion of the existing literature. Section 2 presents a simple theory of choice with and without pre-purchase product information that organizes our descriptive and welfare analyses. Section 3 describes our data on Amazon sales ranks, star ratings, and prices, as well as reviews in major newspapers. Section 4 presents our empirical strategies for measuring causal impacts of professional reviews and star ratings, on sales ranks. We also present estimates, as well as translations

of the estimated effects on log sales ranks into effects on quantities sold. Section 5 then turns to welfare analysis. Using structural demand models calibrated to our causal quantity estimates, we measure the respective welfare gains arising from Amazon star ratings and professional reviews. Section 6 asks whether the effect of New York Times reviews on sales changed between 2004 and 2018. Section 7 concludes.

1 Background

1.1 The U.S. Product and Information Environment for Books

In 2000, roughly 80,000 fiction and non-fiction titles were released in the United States, and the number of new titles released annually has grown sharply since then. In 2012, when 100,000 new U.S. titles appeared in hardback form, the number of new U.S. ebook titles was 280,000.⁴ This figure, while impressive, only counts the titles with ISBNs (“international standard book number”), which many self-published titles lack. Clearly, there has been substantial growth in the number of new book titles released in the U.S. Large physical bookstores only carry roughly 200,000 titles, so only a small fraction of new titles have traditionally been marketed directly to consumers (Greenfield, 2012). Even before digitization, product discovery was a significant challenge; the challenge has grown substantially since.⁵

1.2 Professional Reviews

There is a two-part professional reviewing ecosystem that supports retailer, library, and consumer discovery of new products. One part consists of reviews targeted at libraries and bookstores, from outlets such as Publishers Weekly, Library Journal, and Kirkus. These “B2B” outlets review relatively large numbers of titles – although a small share of releases –

⁴These figures are based on queries of the Bowker Books in Print database for numbers of English-language hardback and ebook titles published in the U.S.

⁵See Waldfoegel and Reimers (2015) and Waldfoegel (2017) for additional data on the growth in new books since digitization.

but have rather limited audiences. In 2018, they each reviewed about 3,500 to 7,500 books, and the respective sites attracted no more than 2.15 million site visits in December 2018, according to Similarweb data.⁶

The other, consumer-facing part of the reviewing environment consists mainly of reviews in daily newspapers, including the New York Times, the Wall Street Journal, the Washington Post, the Los Angeles Times, the Boston Globe, and the Chicago Tribune. These major U.S. newspapers contain many fewer but still substantial numbers of book reviews. Most reviewed between 93 and 248 titles during 2018 (using the Bowker measure). By contrast the Bowker data include 1,800 hardcover editions published in 2018 that were reviewed by the New York Times. Of the titles reviewed in any of these major US newspapers, about 80% were reviewed by the New York Times. Newspapers have far more general traffic and visibility than B2B book review outlets. For example, the Washington Post had 120.5 million monthly visitors in December 2018, while the New York Times had 302.5, according to Similarweb. Measured by both volume of reviews and visibility to consumers, the New York Times is the preeminent US book review outlet.

1.3 Crowd-based Star Ratings at Amazon

Amazon allows users to review and rate books on a five-point scale, and Amazon aggregates users' ratings into star ratings for each book. A few features of the ratings system are noteworthy. First, in contrast to professional reviews, which are available for a small share of titles, crowd ratings are available for all of them. Second, as users leave ratings, Amazon aggregates these individuals' ratings into an overall rating, which they report to a tenth of a star, although the aggregation is not a simple averaging.⁷ Third, consumers can also

⁶We get a rough count of the number of titles reviewed during 2018 by querying Bowker's Books in Print, which contains indicators for whether a book was reviewed by each of a number of major outlets. An entry in Bowker is an edition rather than a title, so we restrict attention to hardcover editions to reduce duplication. Moreover, Bowker's list includes new editions of titles published in the past. Despite these sources of duplication, the Bowker data are useful for rough comparison of the volumes of reviews across professional sources.

⁷Rather, "Amazon calculates a product's star ratings based on a machine learned [sic] model instead of a raw data average. The model takes into account factors including the age of a rating, whether the ratings

prominently see the number of users who have thus far rated each book and therefore the number of underlying ratings on which the visible star rating is based. Leaving ratings is common. While all books start with no ratings, the average title in our sample, described in more detail below, has 326 underlying ratings by the end of 2018. Fourth, the star ratings for a particular book differ across Amazon’s country platforms. Fifth, Amazon visually depicts the rating using a star system with only half-star increments. Ratings of 4.8 and above are depicted visually as having 5 stars, while books with ratings between 4.3 and 4.7 are depicted as having 4.5 stars, and so on. A user who hovers over the stars can see the star rating displayed to a tenth of a star. Thus, users have access to both a “continuous” star measure in tenths of a star and a discontinuous visual measure based on half stars.⁸

1.4 Existing Literature

Our study is related to three existing literatures. First, our study is related to work measuring the impact of professional reviews on product sales. [Reinstein and Snyder \(2005\)](#); [Sorensen \(2007\)](#); [Berger et al. \(2010\)](#) and [Garthwaite \(2014\)](#) provide three examples of studies employing careful empirical strategies to document impacts of professional reviews on movie and book sales. Existing studies of reviews and book sales document causal impacts using weekly sales data. We are able to build on this work using higher frequency daily data for a large sample of books.

Second, our study is related to existing work on the impact of word of mouth reviews on sales. Prominent examples include [Chevalier and Mayzlin \(2006\)](#); [Luca \(2016\)](#); [Duan et al. \(2008\)](#); [Forman et al. \(2008\)](#); [Helmets et al. \(2019\)](#), and [Senecal and Nantel \(2004\)](#). [Chevalier and Mayzlin \(2006\)](#) makes use of a cross-platform comparison of books’ sales ranks and star ratings to measure impacts of crowd opinions, in the form of star ratings, on sales. [Luca \(2016\)](#) makes use of a reporting discontinuity – that crowd ratings are denominated in

are from verified purchasers, and factors that establish reviewer trustworthiness.”

⁸Digitization has also fostered growth in amateurs who distribute their book reviews at Goodreads. Because these reviewers appear not to have effects on book sales; we relegate their discussion to the Appendix section [A](#).

half stars – to measure causal impacts of Yelp ratings on restaurant sales. Our descriptive analysis below implements approaches that build on both of these. In addition to estimating causal effects of an additional star rating, we also embed the causal estimates in a structural model that allows us to quantify the value of the star rating system to consumers. Our structural welfare analysis makes the distinction between ex ante “decision utility” and ex post experienced utility, similar to [Jin and Sorensen \(2006\)](#); [Allcott \(2011\)](#), and [Train \(2015\)](#).

2 Theory: Information, Purchase, and Welfare

2.1 Information and Purchase

Reviews and ratings provide information that can affect consumers’ tendency to purchase products. To be concrete – and to put this in a framework that we return to below – suppose a consumer i has the following utility function for a product j when reviews exist:

$$u_{ij} = u(R_j, p_j; x_j)$$

In this setup, R_j is the pre-purchase product information (rating or review) on product j , p_j is the product’s price; and x_j contains other observables on product j . Because pre-purchase information exists, R_j is both a measure of quality and of the pre-purchase perception of quality.

If reviews and ratings did not exist, then consumers might instead form predictions of quality based on characteristics of the product which we summarize in this setup as a predicted rating, \hat{R}_j . Expected utility absent the reviews would then be

$$u_{ij} = u(\hat{R}_j, p_j; x_j).$$

A surprisingly positive review – when a product is better than expected so that $R_j > \hat{R}_j$ – could increase its consumption relative to its consumption in their absence, and vice versa.

Whether this would happen, of course, depends on the causal impact of review information on purchase (and therefore, we infer, utility). Hence, our main causal empirical task below is to measure the causal impact of reviews and ratings on purchase. We also need a measure of the quality that consumers would expect for each book in the absence of pre-purchase information (\hat{R}_j). We discuss this prediction problem in section 5.

2.2 Review Information and Welfare

Assessing effects of pre-purchase information on welfare requires a distinction between expected ex ante utility and experienced ex post utility. For this we follow studies such as [Jin and Sorensen \(2006\)](#), [Allcott \(2011\)](#), and [Train \(2015\)](#).

Suppose that consumers were uninformed prior to purchase and, in particular, that they believed the product’s quality to be lower than its true quality ($\hat{R}_j < R_j$ in the above example). Then their ex ante demand curve would be given by the dashed curve in Figure 1. They would choose Q_1 units, and at purchase they would expect consumer surplus equal to region A . Upon consumption, however, they would perceive the product’s true value, so that the ex post experienced consumer surplus would be regions $A + B$. Had they been informed prior to purchase, they would have chosen Q^* units and would have experienced their ex ante CS – regions $A + B + C$ – as ex post consumer surplus. Therefore, the value of access to this pre-purchase review information is region C .

There is an analogous case, in which consumers believe the product is better than it actually is ($\hat{R}_j > R_j$) and consume Q_2 units. While the consumers expected even more prior to purchase, their experienced consumer surplus is regions $A + B + C$ less region D . If the consumers had access to information prior to purchase, they would have consumed Q^* , generating consumer surplus of $A + B + C$. Hence, the value of information to these consumers is region D . Generically, the welfare gain from having pre-purchase information arises from a “triangle” associated with either consuming too much or too little of the product in the absence of having pre-purchase information. The base of this triangle is the amount

by which quantity deviates from the informed quantity, and its height is determined by the shape of the demand curve for the product.

Thus, the change in welfare from reviews and ratings is

$$\Delta CS = CS_{\text{ratings}} - [CS_{\text{no ratings}} + \textit{adjustment}].$$

In this formula, CS_{ratings} is the ex ante (and ex post) CS associated with the consumption decision made in light of ratings and reviews, $CS_{\text{no ratings}}$ is the CS associated with the consumption decision made without the benefit of ratings, and *adjustment* is the dollar value of the surprise in product quality for the units consumed. For example, when consumers choose Q_1 units in Figure 1 but should have chosen Q^* , then $CS_{\text{ratings}} = A + B + C$, $CS_{\text{no ratings}} = A$, and *adjustment* = B . In what follows, we first estimate the impacts of reviews and ratings on purchase. We then embed these estimates in a structural model that allows us to estimate the gain on CS arising from the existence of ratings and reviews.

3 Data

3.1 Data Set Construction

The ideal dataset for addressing our questions would be a high-frequency panel on prices, quantities sold by day, and ratings and review information for every book published over some period, or at least a representative sample of all titles, including those reviewed by major outlets. Our data resemble the ideal in some respects but also have some features that require adaptation. There are two broad challenges in assembling a dataset for our study, choosing the group of books to study and getting price and quantity data at sufficiently high frequency for identification. In what follows, we first explain which book titles are included in the study. Second, we describe how we obtain each professional outlet’s review timing. Third, we explain how we obtain lists of ISBNs for particular editions of each title, which

we use to get Amazon data on prices, sales ranks, and star ratings.

We create a list of books that reflects what sells by starting with the most comprehensive publicly available bestseller list. USA Today produces a weekly top 150 bestseller list. During 2018, this list includes 1,901 distinct titles (including 4,355 editions). We supplement this list with all books reviewed in the New York Times and the other major review outlets during 2018 – 1,076 titles (1,918 editions) – as well as a list of books of interest to lay readers outside of the right tail of the sales distribution. These are the 2,222 titles (4,920 editions) published in 2018 and reviewed in the same year by widely followed users of the site Goodreads.⁹ The grand list – the universe of books we study – thus includes 4,283 distinct titles (9,146 editions). Most of these are published during 2018, but some are published earlier, as some books published prior to 2018 still sell enough during 2018 to appear on the 2018 bestseller list; and some of the books reviewed by professional reviewers in 2018 were published in 2017.

We obtain review dates for books reviewed in the New York Times by directly searching for all book reviews on the newspaper’s website. For the other newspapers (the Boston Globe, the Chicago Tribune, the Los Angeles Times, the Wall Street Journal, and the Washington Post), we use the Bowker Books in Print directory. We obtain lists of hardcover editions published in 2018 and reviewed by the newspaper, and then find the reviews written in 2018 using Google searches of, say, “Chicago Tribune book review [author title].” For books reviewed by the New York Times, we also have a measure of whether the review was positive, based on whether the book was included on a New York Times “recommended” list in the weeks after its New York Times review appeared. Each week, the New York Times lists between about 8 and 12 books recently reviewed in their newspaper as recommended. Of the titles reviewed in the New York Times, roughly 40 percent are “recommended.”

For each of the 4,283 titles in the sample, we obtain a list of the books’ ISBNs by searching for the title and author on Bowker. A title can have multiple editions with separate sales

⁹We include all reviewers on Goodreads’ “most-popular reviewers” lists as of June 2019 who have more than 10,000 followers.

ranks, prices, and ratings. We use the ISBNs to retrieve daily Amazon data on the respective editions' sales ranks, prices, number of ratings, and Amazon stars from keepa.com, which provides Amazon data on physical book editions.

We obtain daily Amazon data for the US site as well as two other domains selling English-language books, the Canadian and UK sites. The benefits of these data are considerable for causal identification of rating and review effects. Because we have high-frequency data, we can look for high-frequency variation in the sales rank with the appearance of reviews. Moreover, we can make use of high-frequency changes in prices and crowd ratings, all of which can differ across domains as well as over time, to ascertain their impacts on sales.¹⁰ In addition, because we have data on the same edition at different national Amazon domains, we can also identify impacts of Amazon star ratings and prices using cross-platform variation in the changes in, say, ratings and the changes in sales ranks.

Along with these advantages come some disadvantages. First, our data cover only one retailer - Amazon - and not the entire market. Still, during 2018 Amazon accounted for 44.5 percent of the sales of physical books in the US in 2017 – the year before our sample – so our data cover a major part of the market.¹¹ Second, our sample includes only a subset of the titles sold at Amazon. Third, we observe the sales rank and not the sales quantity for each edition. We are thus in the position of following other authors faced with rank rather than quantity data (e.g. [Chevalier and Goolsbee, 2003](#); [Brynjolfsson et al., 2003](#); [Reimers, 2019](#)). Ranks are valuable measures of quantity, but many of our analyses below require a way to translate ranks into estimates of sales quantities. Amazon does not disclose how it calculates its sales rank, but a few things are clear.¹² First, many ranks are updated at least daily, often hourly. Second the ranks are not based only on the most recent day. [Figure 2](#) shows the time series of the Amazon sales rank for a book with modest sales. When a sale

¹⁰While Nielsen includes list prices, it does not provide information on the prices actually charged for books.

¹¹See <https://www.publishersweekly.com/pw/by-topic/industry-news/financial-reporting/article/78929-print-unit-sales-increased-1-3-in-2018.html>.

¹²See <https://www.amazon.com/gp/help/customer/display.html?nodeId=525376>.

occurs, the rank improves sharply, then drifts up for days. This clearly indicates that the sales rank is based on a moving average of sales that appears to have a long – multi-day – memory. This will be relevant to both their modelling and their interpretation.

3.2 Summary Statistics

Table 1 provides a description of the sample. The first column includes all of the editions and domains in the estimation sample. The overall sample, in column (1), includes 9,146 distinct editions and just over 1.6 million daily observations. Columns (2)-(4) report statistics separately for the US, Canada, and the UK. The US sample includes 8,631 editions, and the Canadian and UK samples include about 3,800 editions each. The US sample includes substantially more reviews. Columns (5)-(7) report statistics for three (overlapping) sets of titles, those reviewed in the professional review outlets, those reviewed by Goodreads top reviewers, and those in the USA Today bestseller sample.

Half of star ratings in the sample are between 4.1 and 4.7. The number of individual ratings underlying these star ratings varies across book and time. By construction, books enter the platform with no ratings, so many titles have star ratings based on few underlying ratings for a time. The median number of underlying ratings is 60. A quarter of the observations have star ratings based on 15 or fewer, and a tenth of the observations have star ratings based on ten or fewer ratings.

We also obtain genre information on sample titles from Bowker. As Figure 3 shows, the professionally reviewed, and crowd-rated, books have different genre distributions. The professionally reviewed subsample has higher proportions in genres such as biography, history, and social science and lower shares in genres such as self-help, romance, and juvenile fiction, compared to the books in our sample that were not reviewed professionally. Even without deeper analysis, Figure 3 shows an important fact. Professionals focus on serious genres, so the appearance of crowd reviews raises the amount of pre-purchase information available for the less serious genres.

3.3 Supplementary Weekly Nielsen Data

In addition to the main analysis sample consisting of daily sales ranks, we also make use of weekly sales data from the Nielsen Bookscan database for three ancillary analyses. We have weekly US sales quantities for the top 100-selling physical editions of each week. We use these data for 2018 for estimating the elasticity of sales quantities with respect to ranks, which we use for translating ranks into quantities. We later employ these data for 2015-2018 for estimating a nested logit substitution parameter, which we use in our welfare calculations, as described in Appendix section C.

We use a different extract from the Nielsen data for a third exercise, estimating the impact of a consistent subset of New York Times reviews – for the 100 New York Times annual “notable books” on sales over time. For each even-numbered year 2004-2018 we employ the weekly Nielsen sales data for nine weeks before, and nine weeks after, their original New York Times review dates.

4 Empirical Strategies and Descriptive Results

We have two goals in this section. First, we provide credibly causal evidence on the relationships between pre-purchase information (reviews and crowd ratings) and sales ranks. Second, we translate measured effects on sales ranks into effects on quantities – such as the elasticity of the quantity sold with respect to the Amazon star rank – that we can use to calibrate structural models for welfare analysis.

To accomplish the first goal, we run regressions of log sales ranks on three groups of variables, as well as various fixed effects. The three groups of variables are: a) indicators for whether a title has received a review from a professional outlet, for example a dummy for whether a title received a New York Times review in the past five days; b) platform-specific measures of Amazon crowd ratings, prices and the number of reviews; and c) a lagged value of the log sales rank for the title at the platform. We also include country-specific title fixed

effects as well as, in some specifications, fixed effects for the time until and since the book’s publication. To allow for the possibility that star effects vary with the number of underlying ratings, we also include the interaction of stars and the numbers of ratings.

Generically, the specifications can be described via the following equation:

$$\begin{aligned} \ln(r_{jct}) = & \theta \ln(r_{jc,t-1}) + h_{\tau c} + a \ln(p_{jct}) + g \ln(R_{jct}) + m \ln(\text{ratings}_{jct}) \\ & + n \ln(\text{ratings}_{jt}) \ln(R_{jt}) + \pi_{\tau'c} + \mu_{jc} + \epsilon_{jct} \end{aligned} \quad (1)$$

In this model, r_{jct} , p_{jct} , R_{jct} , and ratings_{jct} are the sales rank, price, star rating, and number of underlying ratings for title j on platform c on day t . The term $h_{\tau c}$ is a platform-specific coefficient for τ days relative to the appearance of a review. Initially, we allow for separate h terms for each of the days leading up to and following the appearance of professional reviews. We then aggregate across days to produce average review effects for, say, the first five days after a review appears, etc. Finally, the terms $\pi_{\tau'c}$ are fixed effects for the τ' days relative to the book’s publication to account flexibly for time patterns of sales around the publication date, and the term μ_{jc} is a platform-specific title fixed effect. We perform some estimates using only the data from the US platform; other estimates use all three platforms.

4.1 Effect Estimates

We first focus on the effect of professional reviews on rankings. To that effect, we estimate a version of the model with flexible review effects. In particular, we include the h terms for each of the 20 days before and 100 days after the appearance of a review with the last pre-review day as baseline. We include two sets of these terms, one for the New York Times and another set for the other professional review outlets, collectively. We estimate this model only on US data. Figure 4 reports the time patterns of New York Times and other professional reviews. As the left panel shows, a New York Times review delivers a large and immediate improvement in the sales rank at the appearance of the review. The log rank

improves by -0.4, then returns to its baseline trend a few weeks later. As the right panel shows, professional reviews at other outlets also have detectable effects, but they appear to be much smaller. The coefficients on the remaining variables are shown in column 1 of Table 2.

For books reviewed by the New York Times, we can distinguish the books recommended by the New York Times versus the remainder of those reviewed. Figure 5 compares coefficient estimates for recommended vs other books, summarizing the professional review effects with three indicators, for 0-5 days after a review, 6-10, and 11-20 days. We also include an indicator that is one from ten days before until 20 days after the appearance of a review so that the post-review effects are defined relative to the ten days before. The review effects for both groups of reviewed books are positive, although the effects are larger for the recommended books. The finding that all reviews have a positive effect on sales suggests that professional reviews might have market expanding effects beyond the information provided in them.

We proceed by focusing on the effects of star ratings, and we continue to control for professional reviews by aggregating the effects to the first five days after a review, days six through ten, and days eleven through 20. Credibly estimating the effect of Amazon star ratings is more challenging, as these ratings evolve less discontinuously, and potentially endogenously, over time. Still, features of the environment give us promising avenues of credible identification. First, we have panel data, and we observe each title's daily log sales rank, along with the evolution in each title's star rating. Second, Amazon's country-specific star ratings for each title evolve separately over time. This allows us to pursue identification approaches using both temporal and cross-platform variation that build on [Chevalier and Mayzlin \(2006\)](#).

A simple approach to measuring the impact of ratings and prices on sales would be to estimate the relationship between a title's sales rank and its rating, across titles within a platform at a point in time, but the obvious shortcoming of this approach is that titles that

are “worse” may have both lower ratings and higher (worse) sales ranks, entirely apart from the possible causal impact of ratings on sales. A possible solution to this problem would be to control for the unobserved quality of the title, using panel data on the editions on a particular platform and including an edition fixed effect. Then the effects of the log price (p) and star rating (R) on the log sales rank would be identified from the within-title changes.

We implement a variant of this approach in column (2) of Table 2, using only US data. The specification in column (2) includes both title fixed effects and the π terms (for time until and since publication). The coefficient on the log star rating is -0.107. This estimate also shows impacts of professional reviews that are consistent with those in Figure 4. The appearance of a New York Times review improves the log sales rank by 0.24 in the five days after the review; and impacts of other professional reviews are much smaller.

We explore whether the impact of star ratings depends on the number of underlying ratings on which the stars are based by adding an interaction of $\ln(R_{jct})$ and $\ln(\text{ratings}_{jct})$, in column (3). The interaction term is negative and significant, indicating that the star ratings have larger effects as they are based on more underlying ratings.

The US-data-only approach of columns (2) and (3) is vulnerable to a concern that some unobserved factor is changing both attitudes toward a title, and its sales, over time. A second alternative is to use multiple platforms, i.e. the Amazon sites for different countries, selling the same book. Then one could make use of the possible differences in ratings across platforms to ask whether the cross-platform rating differential gives rise to a cross-platform sales rank differential. We combine the cross-platform and time series approaches, using multiple points in time at multiple platforms. This allows the fixed effect for a title to differ across platforms. Moreover, it assumes that sales ranks for a title move together over time at different platforms, except for the impact of differential prices, rankings and reviews across the platforms’ environments. This is analogous to the approach that [Chevalier and Mayzlin \(2006\)](#) employ with two time observations. Our data allow us to implement this approach with hundreds of daily observations per title.

Column (4) implements the estimation with all three platforms and platform-specific title fixed effects as well as the π terms for time until and since publication. Column (5) adds the interaction of $\ln(R_{jct})$ and $\ln(ratings_{jct})$. In these specifications we interact the h terms for time since a professional review with the platform. This allows the effect of, say, the New York Times to differ between the US and the other platforms' countries. The US NYT effect for the first five days is about -0.25 in both specifications. While the coefficient on $\ln(R_{jct})$ is smaller in absolute value in column (4) than in column (2), the interaction specifications in columns (3) and (5) give very similar results.

4.1.1 Robustness Checks

While the interaction specifications appear to fit, it is worth exploring whether they are overly restrictive. To this end we replace the $\ln(ratings_{jct})$ term in column (5) with dummies for deciles of the number of underlying ratings and the interactions of these indicators with $\ln(R_{jct})$. Figure 6 reports results. The magnitude of the coefficient on star ratings ($\ln(R_{jct})$) rises essentially monotonically in the deciles for numbers of underlying ratings. While stars have essentially no effect on the sales of books with numbers of ratings in the lower decile, the effect grows to -0.50 for the top deciles. We conclude that our interaction specifications provide faithful representations of the underlying relationships.

The way that Amazon reports its star ratings gives rise to an additional identification strategy for measuring the impact of star ratings on sales. On a book's page, a customer sees an image of the number of stars that is denominated in half stars, but if one hovers over the star image, one sees a number of stars to a single decimal place. It is easy for a user to see the decimal star rating, but the visual, half-star image may have additional salience. This suggests an additional, discontinuity method for identifying the impact of stars that is reminiscent of Luca (2016). We look for jumps in log sales ranks at the decimal star ratings for which the visual half stars jump by one half. This occurs, for example, at 2.8, 3.3 and 3.8 stars, etc. To explore this we estimate variants of model (5) above where we include

a series of dummies for each of the possible decimal star ratings instead of the continuous measure $\ln(R_{jt})$. Figure 7 displays the pattern of coefficients on the decimal rating dummies, with vertical lines at the decimal ratings at which the visible star rating jumps by one half. Ninety percent of these ratings fall between 3.4 and 5, so we focus on this range. We see an overall trend of better ranks as the rating increases, with larger jumps at the discontinuities. We take this as additional evidence that star ratings have a causal impact.

4.2 Translating Ranks into Quantities

The evidence above indicates that reviews have an impact on sales ranks, but two steps are required to translate coefficients from our models into elasticities. First, we have no information on quantities of titles sold by rank at Amazon. We do, however, have information on the sales of the top-100 weekly physical bestsellers according to Nielsen. Assuming that sales follow a power distribution with respect to ranks as in $q_j = Ar_j^{-B}e^{\epsilon_j}$, we can summarize these data by a regression of log quantities on log ranks. This regression yields $B = 0.54$ (with a standard error of 0.004).

Second, equation (1) is a partial adjustment model. We find the full effect of a right hand side variable on the log rank by setting $\ln(r_{jt}) = \ln(r_{j,t-1})$. Then the derivative of a book's rank with respect to, say, the log price, from equation (1), is $\frac{a}{1-\theta}$. Combining the above, the reduced form elasticities of quantity with respect to price and the rating are, respectively:

$$\begin{aligned}\epsilon_p &= \frac{\partial \ln(q_j)}{\partial \ln(p_j)} = \frac{aB}{1-\theta}, \\ \epsilon_{R_j} &= \frac{\partial \ln(q_j)}{\partial \ln(R_j)} = \frac{(g + n \ln(\text{ratings}_{jt}))B}{1-\theta}.\end{aligned}$$

Analogously, the effect of a review on the log sales quantity is $\frac{hB}{1-\theta}$. We obtain standard errors for these estimates by taking 500 parametric bootstrap draws from the estimated joint distributions of the parameters from Table 2, as well as from the distribution of B from the estimation of quantities as a function of ranks.

Table 3 reports estimates of quantity effects from model (5) in Table 2. Rows 2-5 report the elasticity of the quantity sold with respect to the Amazon star rating. Because of the interaction of the star rating with the number of underlying ratings, the effect varies across the distribution of the number of ratings. At the 25th percentile, the elasticity is 0.511, while it is 0.769 at the median and 1.055 at the 75th percentile. At the mean, the elasticity is 0.783. The next rows report the effects of reviews, during particular time windows after their appearance, on log sales. For example, the 0.58 in the NYT 0-5 row indicates that sales increase by 78 percent during the 0-5 days after the appearance of a NYT review ($e^{0.579} - 1 = 0.78$).

The bottom panel of Table 3 reports percentage impacts of reviews on annual simulated sales. We estimate daily sales quantities for an edition j by assuming that q_{jt} is proportional to $\frac{1}{\exp(\ln(\text{rank}_{jt})^B)}$ using $B = 0.54$ as described above. We estimate the counterfactual sales absent professional reviews by substituting the following for the log-rank:

$$\ln(\text{rank}_{jt}) - \frac{h_k B}{1 - \theta} (\text{review indicator } k)_{jt}.$$

Here, the indicator k refers to, say, the first five days after the receipt of a New York Times review. We aggregate these estimated quantities across all days in the year, then compare the baseline to the calculated values corresponding to the absence of the respective sources of pre-purchase information to calculate the percentage impacts on sales. For example, according to our preferred specification, receiving a New York Times review (but not another professional review) raises sales by 5.17 percent during 2018.

Finally, Table 3 also reports a price elasticity of demand of -0.42. This title-level elasticity appears to be rather inelastic on its face. We offer two comments at this point. First, it is widely understood that Amazon prices below the static profit-maximizing level. In a 2013 60 Minutes interview, Amazon CEO Jeff Bezos stated, “We do price elasticity studies, and every time the math tells us to raise prices.”¹³ We find similarly inelastic estimates in Reimers

¹³See <https://www.cbsnews.com/news/amazons-jeff-bezos-looks-to-the-future/>.

and Waldfogel (2017). Second, as we will discuss further below, while the absolute size of the welfare effects of pre-purchase information depends on the price coefficient, the relative size of the welfare effects of professional vs crowd reviews is invariant to it.

5 Welfare Analysis

Our descriptive analysis above gives us the relationships between three important factors – star ratings, professional reviews, and prices – and the quantities of books sold. One of the shortcomings of the analysis above is that there is no obvious way to compare the size of the welfare benefit from the availability of professional reviews versus the existence of Amazon stars. As our theoretical model suggests, however, the counterfactual change in consumer surplus is a more natural basis for comparison. A structural demand model allows us to undertake this calculation; and we use our descriptive estimates from above to calibrate a nested logit model of demand. We then present estimates of the welfare impact of star ratings and professional reviews.

5.1 Preliminaries

In order to build a nested logit model of demand and to develop estimates of the welfare effects of reviews and ratings, we need a few components in addition to the descriptive quantity effects estimated above. These include the market size (M), the total 2018 US physical book sales, the number of unit sales accounted for by sample books, and a nested logit substitution parameter σ , which are summarized in Table 4. We discuss their derivations in Appendix section B.

In addition, we need measures of ex ante book quality, denominated in stars, that consumers would have expected absent the stars' existence. To this end, we model consumers' beliefs about book quality in the absence of stars (\hat{R}_j) via a regression of Amazon stars for books on the US platform on publisher fixed effects, genre fixed effects, and dummies for

authors' prior experience. For each edition, we use the average star rating across all days. The resulting regression explains 26.25 percent of the variation in log stars. We then treat the fitted value as a measure of the ex ante quality of each book that consumers would have expected absent the star rating system. We explore the sensitivity of our results to the explained share of the variation in star ratings in the robustness section below.

Rather than covering all retailers and all titles, our data cover one retailer, albeit a major one, and roughly 9,000 editions. We would like to make statements about market-wide welfare effects of ratings and reviews, and this requires some behavioral and scaling assumptions. First, we assume that effects of professional reviews operate equally at all retailers, not just at Amazon. Second, we assume that effects of Amazon stars operate only via Amazon sales, but we also assume that Amazon star effects operate on all books sold at Amazon, not just on those titles in the sample.

5.2 A Simple Structural Model

To perform our welfare analysis, we calibrate a nested logit model to the estimated elasticities. We begin by defining

$$\delta_j = \ln(s_j) - \sigma \ln(s_{j|g}) - \ln(s_0),$$

where $s_j = q_j/M$, $s_{j|g} = q_j/Q$, and $s_0 = 1 - Q/M$. Note that $Q = \sum q_j = 192$ million units sold market-wide for the sample titles, as shown in Table 4 and derived in the appendix. Each product's share is then

$$s_j = \frac{e^{\delta_j/(1-\sigma)} D^{1-\sigma}}{1 + \sum e^{\delta_{j'}/(1-\sigma)} 1 + D^{1-\sigma}},$$

where $D = \sum e^{\delta_j/(1-\sigma)}$.

Let δ_j be the utility in the status quo, when reviews and ratings are present. We can

write this as

$$\delta_j = \delta_j^0 + \alpha p_j + \gamma_j R_j + \psi_j,$$

where α , γ_j , and ψ_j are utility function parameters. While these parameters are unknown, they are related to estimated parameters a , g , h , and n , respectively. We can calculate the nested logit expressions for the derivatives of quantity with respect to price and rating, set these equal to the reduced form derivatives described above, then solve for the utility function parameters.

In particular, the nested logit model gives a simple expression for the price elasticity of demand:

$$\hat{\epsilon}_p = \alpha_j \frac{p_j}{1 - \sigma} (1 - \sigma s_{j|g} - (1 - \sigma)s_j).$$

Given $\hat{\epsilon}_p$ from the descriptive analysis above, σ (from the appendix), and $s_{j|g}$, s_j , and p_j (which are data), this formula gives us a parameter estimate of α_j for each j , which we average for our estimate of utility function parameter α . Analogously, we can infer γ_j by solving the following equation:

$$\hat{\epsilon}_{R_j} = \gamma_j \frac{R_j}{1 - \sigma} (1 - \sigma s_{j|g} - (1 - \sigma)s_j).$$

Rather than taking a simple average of the γ_j 's, we exploit the interaction of stars and the numbers of ratings by solving the structural model using a γ_j that is parameterized as a function of the number of reviews that each book has received by the end of 2018.

We can solve for the utility function parameters associated with reviews in a related way, although reviews are binary rather than continuous. Our descriptive analysis tells us how each sales quantity q_j would have been different in the absence of reviews, q'_j . But since the descriptive model includes no measure of the impact of reviews on unreviewed books, it implicitly identifies the effect of reviews on reviewed books from the difference between the change in sales for reviewed books relative to unreviewed books.

Hence, the model analog of our descriptive measure $\ln(q_j/q'_j)$ is the review-induced per-

centage change in sales for reviewed books, relative to the review-induced percentage change in sales for unreviewed books. Define s_j^r as sales of reviewed books in the presence of reviews, s_j^u as sales of unreviewed books in the presence of reviews, $s_j'^r$ as sales of reviewed books in the absence of reviews, and $s_j'^u$ as sales of unreviewed books in the absence of reviews. Then the equation of the descriptive fact and its model analogue is

$$\ln(q_j/q'_j) = \ln(s_j^r/s_j'^r) - \ln(s_k^u/s_k'^u),$$

where j indexes reviewed books and k indexes non-reviewed books. A few lines of algebra show that $\ln(q_j/q'_j) = \frac{\psi_j}{1-\sigma}$. Given σ , we therefore know ψ_j ; and we parameterize ψ_j by dividing books into four groups, those reviewed by the New York Times, those reviewed by other professional outlets, those reviewed by neither, and other.

Given values of the utility function parameters, we can compare two counterfactual scenarios – without Amazon star ratings and without professional reviews – to the baseline when both are present. We are interested in the effects of the two sorts of pre-purchase information on the consumer surplus achieved in the market. First we need expressions for the status quo utility level, as well as its analogues in the absence of crowd and professional reviews. Status quo utility of product j is given by the data:

$$\delta_j = \ln(s_j) - \sigma \ln(s_{j|g}) - \ln(s_0),$$

while counterfactual utility absent Amazon stars is given by $\delta_j^s = \delta_j - \gamma_j(R_j - \hat{R}_j)$; and counterfactual utility absent professional reviews is given by $\delta_j^p = \delta_j - \psi_j$.

The change in CS associated with star ratings is given by

$$\Delta CS = \frac{M}{\alpha} \left[\ln \left(1 + \sum \exp \left(\frac{\delta_j}{1-\sigma} \right)^{1-\sigma} \right) - \ln \left(1 + \sum \exp \left(\frac{\delta_j^s}{1-\sigma} \right)^{1-\sigma} \right) - \sum \gamma_j (R_j - \hat{R}_j) q_j^s \right],$$

where the term $\sum \frac{\gamma_j}{\alpha} (R_j - \hat{R}_j) q_j^s$ is the adjustment reflecting the possibility that what is

consumed has ex post utility that differs from the ex ante value, and q_j^s is the quantity of product j chosen in the absence of star ratings.

The respective changes in consumer surplus from the presence of professional reviews is given by the analogous equations, with δ_j^s replaced by δ_j^p , and with $\sum \gamma_j(R_j - \hat{R}_j)q_j^s$ replaced by $\sum \psi_j q_j^p$, where q_j^p is the quantity of product j chosen in the absence of professional reviews.

5.3 Results

In the demand model we scale the quantities of each book sold so that total inside sales are the total market-wide sales of the sample titles. This way, because all professionally reviewed titles are in the sample, results on reviews directly measure market-wide effects. To get Amazon star effects, we take our estimate of the change in CS from Amazon stars per dollar of model Amazon revenue. We then multiply this per-dollar measure of the change in surplus by Amazon’s revenue from sales of all titles (695 million unit sales industry-wide x \$17.54 per title x Amazon’s 44.5 percent market share).

Table 5 shows the welfare results. Both forms of pre-purchase information have impacts on sales and revenue. Professional reviews raise revenue by \$23.20 million. On net, the presence of star ratings raises book revenue by \$42.81 million across all books – about twice as much as do professional reviews. For the “unexpectedly good” titles with star ratings that beat expectations, pre-purchase information raises sales by \$185.95 million, while it decreases sales of “unexpectedly bad” titles by \$143.14 million.

Effects on net revenue mask effects on consumer welfare, as pre-purchase information raises consumer surplus both when it raises and when it decreases spending. Professional reviews, by shifting consumption to “good” titles, raise consumer surplus by \$2.54 million. While the existence of star ratings has an effect on net revenue similar to the net revenue impact of professional reviews, the impact of star ratings on CS is much larger. Compared with the counterfactual environment without star ratings, the status quo delivers \$41.6 million in additional consumer surplus from Amazon sales alone, which is roughly 15 times

the effect of professional reviews on CS through all retail channels. Thus, the vast majority of the value of pre-purchase information available to consumers following digitization is delivered by crowd-based rating systems.

5.4 Robustness

Our welfare analyses depend on estimated parameters. Here we explore the sensitivity of our basic results to different parameter values α, σ , and the consumers' ability to predict a book's true star rating. The parameter α determines the absolute size of welfare effects. It does not, however, affect the relative size of the respective effects of professional reviews and crowd ratings on consumer surplus; the term α is a factor of proportionality on our measure of ΔCS . Hence, our conclusions on the relative impacts of professional reviews and star ratings are unaffected by α . The impact of the substitution parameter σ on ΔCS is less obvious a priori. We have experimented with values of σ between 0 and 0.9 as well as with nests on the genre level, and results change only minimally.

The measured welfare benefits of Amazon star ratings also depend on the accuracy of consumers' predictions of product quality absent star ratings. Our baseline model of this is a regression of stars on observables, and the regression explains 26.25 percent of the variation. It is possible that the regression understates, or overstates, the ability of consumers to predict quality. We can explore the sensitivity of our Amazon stars welfare benefit measure to prediction accuracy using the approach of [Aguiar and Waldfogel \(2018\)](#). We add the following explanatory variable to the regression: $R_j + \kappa \cdot \epsilon_j$, where ϵ_j is a standard normal random error, and κ is a scale factor we vary to produce variation in the prediction accuracy, which we summarize by the R^2 of the regression. [Figure 8](#) shows how the change in CS from the presence of Amazon stars varies with prediction accuracy, with dots for our baseline estimate with an estimated R^2 of 0.2625 and for the zero-information case ($R^2 = 0$). If our model understates prediction accuracy, then the true welfare benefit is lower. For example, if prediction accuracy corresponded to an R^2 of 50 percent, then the welfare benefit would be

roughly \$20 million. If consumers could perfectly predict quality absent star ratings ($R^2 = 100$ percent), the star ratings would deliver no welfare benefit. If R^2 were 80 percent, then Amazon stars would add roughly as much consumer benefit as professional reviews.

6 Professional Review Effects over Time

The causal and welfare estimates above indicate substantial impacts of crowd ratings on sales. Because crowd reviews did not exist prior to digitization, it is reasonable to ask whether the newfound influence of the crowd is displacing the influence of professional reviewers. This is, broadly, a test of whether the crowd substitutes for professionals. Ideally, we would repeat the foregoing analyses for earlier years, prior to the diffusion of online retail. This is infeasible, though, because crowd reviews have become ubiquitous, and our daily ranking data do not reach back far enough to repeat our analysis for a time without crowd reviews. So instead of relying on daily ranking data, we employ the approach of [Berger et al. \(2010\)](#) and collect weekly physical book sales data to estimate the impact of New York Times book reviews on demand, from 2004 to 2018.

For this analysis, we first find the New York Times review dates for all books that made the 100 Notable Books of the Year list, for all even years from 2004 to 2018. We manually search for these books' ISBNs in the Nielsen Bookscan database to collect weekly unit sales.¹⁴ We then estimate regressions of the form

$$\ln \left(\frac{s_{jt}}{s_{j,t-1}} \right) = \lambda \text{review}_{jt} + \beta x_{jt} + u_{jt},$$

where s_{jt} denotes the sales of book i in week t , $\text{review}_{jt} = 1$ in the week immediately following the New York Times review, and x_{jt} includes controls for the number of weeks since the book's release, and a dummy variable that equals one in all weeks after publication

¹⁴We limit our analysis to the list of notable books because their reviews are likely most positive, and because manually searching for ISBNs is quite time consuming. We obtain these books' ISBNs from Goodreads.

(to account for pre-sales). Like [Berger et al. \(2010\)](#), we drop all observations more than nine weeks before or after the review. The form of the dependent variable means that our coefficient of interest, λ , measures the impact of a review on the rate of change of sales.

Figure [A.1](#) illustrates the impacts of a review on sales from our regressions for the New York Times Notable Books of each year, including only books that were reviewed more than one week after their publication. We find that professional reviews had a positive effect on the rate of change in sales in all years, with positive and statistically significant coefficients throughout. Interestingly, the coefficient is largest for books reviewed in 2018 – about twice as large that for reviews in other years. If anything, the effect of the New York Times book reviews has increased over the last 15 years. This may in fact be due to digitization: it is possible that the outlet was able to utilize digitization to improve the reach of its book reviews. For example, the number of digital-only subscriptions to the New York Times rose from about 100,000 in March 2011 (when a metered paywall was introduced) to over 2.5 million in the third quarter of 2018 ([Richter, 2018](#)).¹⁵

7 Conclusion

Digitization has delivered a challengingly large number of new products, straining the capacity of both critics and consumers to discover those meriting their attention. At the same time, digitization has delivered a potential solution in new mechanisms for aggregating user product ratings into potentially useful pre-purchase information for other consumers. Using Amazon daily data on sales ranks, prices, and star ratings for over 9,000 book editions, along with information on review timing in professional review outlets, we document causal impacts of reviews on sales ranks. We then transform these estimates into impacts on quantities, which we use to calibrate nested logit demand models for welfare analysis. We find that book reviews in the New York Times and other major newspapers have substantial impacts

¹⁵[Richter \(2018\)](#) also shows a large jump in subscriptions around the time of the 2018 presidential election, suggesting that other forces were also at play that may have increased the reach of the New York Times.

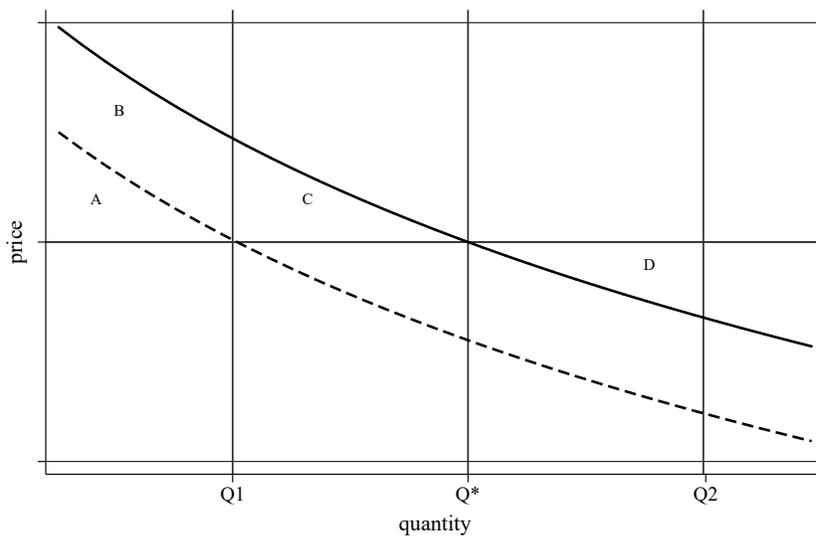
on book sales – NYT reviews raise sales by over 75 percent in the five days after a review and by 5.2 percent over the year. We also document that the causal elasticity of quantity sold with respect to Amazon stars averages about 0.75. Because these two forms of pre-purchase information have causal impacts on buying behavior, they also affect welfare. The clear effects of professional reviews in the descriptive results are also present in the welfare analysis: professional reviews raise revenue by \$24.06 million and raise consumer surplus by 12 percent as much. Crowd ratings have net impacts on revenue that are twice as large as the net revenue impacts of professional reviews, but the difference is even larger for consumer surplus. The existence of crowd ratings adds \$41 million to consumer surplus from Amazon book purchases alone, or 15 times the impact of professional reviews on surplus derived from purchases through all channels. We conclude that digitization, in addition to delivering a proliferation of new products, has also added substantially to the value of the pre-purchase information available to consumers. While the value of crowd-based pre-purchase information to consumers is now much larger than the consumer benefit derived from professional reviews, the absolute impacts of professional reviews have not declined over time. Crowd ratings are available for all products, not just books; particularly given the smaller role of professional reviewers for other products, crowd-based ratings may add substantial benefit for consumers of other products.

References

- AGUIAR, L. AND J. WALDFOGEL (2018): “Quality predictability and the welfare benefits from new products: Evidence from the digitization of recorded music,” *Journal of Political Economy*, 126, 492–524.
- ALLCOTT, H. (2011): “Consumers’ perceptions and misperceptions of energy costs,” *American Economic Review*, 101, 98–104.
- BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2014): “High-dimensional methods and inference on structural and treatment effects,” *Journal of Economic Perspectives*, 28, 29–50.
- BERGER, J., A. T. SORENSEN, AND S. J. RASMUSSEN (2010): “Positive effects of negative publicity: When negative reviews increase sales,” *Marketing Science*, 29, 815–827.
- BRYNJOLFSSON, E., Y. HU, AND M. D. SMITH (2003): “Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers,” *Management Science*, 49, 1580–1596.
- CHEVALIER, J. AND A. GOOLSBEE (2003): “Measuring prices and price competition online: Amazon.com and BarnesandNoble.com,” *Quantitative marketing and Economics*, 1, 203–222.
- CHEVALIER, J. A. AND D. MAYZLIN (2006): “The effect of word of mouth on sales: Online book reviews,” *Journal of marketing research*, 43, 345–354.
- DARGIS, M. (2014): “As indies explode, an appeal for sanity: flooding theaters isn’t good for filmmakers or filmgoers,” *New York Times*, 9.
- DEUTSCHMAN, A. (2004): “The Kingmaker Walt Mossberg makes or breaks products from his pundit perch at a little rag called The Wall Street Journal,” *WIRED-SAN FRANCISCO*-, 12, 120–132.
- DUAN, W., B. GU, AND A. B. WHINSTON (2008): “Do online reviews matter?—An empirical investigation of panel data,” *Decision support systems*, 45, 1007–1016.
- FORMAN, C., A. GHOSE, AND B. WIESENFELD (2008): “Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets,” *Information systems research*, 19, 291–313.
- GARTHWAITE, C. L. (2014): “Demand spillovers, combative advertising, and celebrity endorsements,” *American Economic Journal: Applied Economics*, 6, 76–104.
- GREENFIELD, J. (2012): “Seven Advantages Barnes & Noble has in the Bookseller Wars,” *Digital Book World*.
- HELMERS, C., P. KRISHNAN, AND M. PATNAM (2019): “Attention and saliency on the internet: Evidence from an online recommendation system,” *Journal of Economic Behavior & Organization*, 161, 216–242.
- JIN, G. Z. AND A. T. SORENSEN (2006): “Information and consumer choice: the value of publicized health plan ratings,” *Journal of health economics*, 25, 248–275.

- LUCA, M. (2016): “Reviews, reputation, and revenue: The case of Yelp. com,” *Com* (March 15, 2016). *Harvard Business School NOM Unit Working Paper*.
- MARTIN, A. (2011): “The End of the Career Food Critic,” <https://www.theatlantic.com/national/archive/2011/09/sam-sifton-departure-and-the-end-of-the-career-food-critic/337844/>.
- NARULA, S. K. (2014): “Millions of people reading alone, together: The rise of Goodreads,” *The Atlantic*.
- POMPEO, J. (2017): “Michiko Kakutani, the Legendary Book Critic and the Most Feared Woman in Publishing, is Stepping Down from the New York Times,” *Vanity Fair*.
- REIMERS, I. (2019): “Copyright and generic entry in book publishing,” *American Economic Journal: Microeconomics*, 11, 257–84.
- REIMERS, I. AND J. WALDFOGEL (2017): “Throwing the books at them: Amazon’s puzzling long run pricing strategy,” *Southern Economic Journal*, 83, 869–885.
- REINSTEIN, D. A. AND C. M. SNYDER (2005): “The influence of expert reviews on consumer demand for experience goods: A case study of movie critics,” *The journal of industrial economics*, 53, 27–51.
- RICHTER, F. (2018): “The “Failing” NY Times Passes 2.5 Million Digital Subscriptions,” .
- SENECAL, S. AND J. NANTEL (2004): “The influence of online product recommendations on consumers’ online choices,” *Journal of retailing*, 80, 159–169.
- SORENSEN, A. T. (2007): “Bestseller lists and product variety,” *The journal of industrial economics*, 55, 715–738.
- TRAIN, K. (2015): “Welfare calculations in discrete choice models when anticipated and experienced attributes differ: A guide with examples,” *Journal of choice modelling*, 16, 15–22.
- WALDFOGEL, J. (2017): “How digitization has created a golden age of music, movies, books, and television,” *Journal of Economic Perspectives*, 31, 195–214.
- WALDFOGEL, J. AND I. REIMERS (2015): “Storming the gatekeepers: Digital disintermediation in the market for books,” *Information economics and policy*, 31, 47–58.

Figure 1: Illustration – welfare analysis of pre-purchase information



Notes: This figure illustrates demand curves under full information about a product's quality (solid line) and with limited ex ante information about the quality when the expected quality is less than the true quality (dashed line). The corresponding consumer surplus under full information is areas $A + B + C$; under limited ex ante information, it is $A + B$.

Figure 2: Amazon sales rank evolution of a sample book

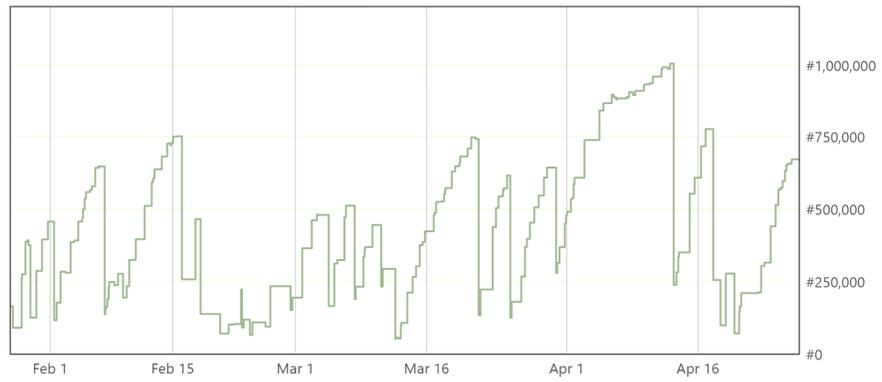
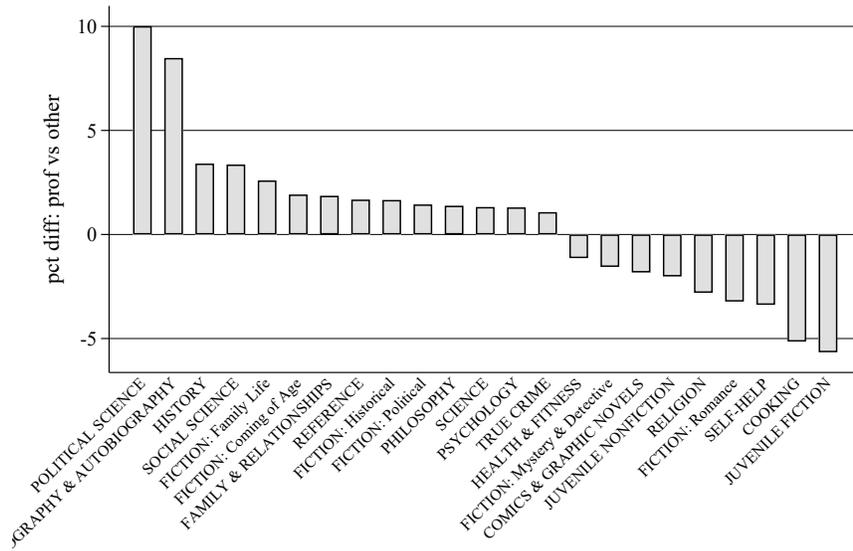
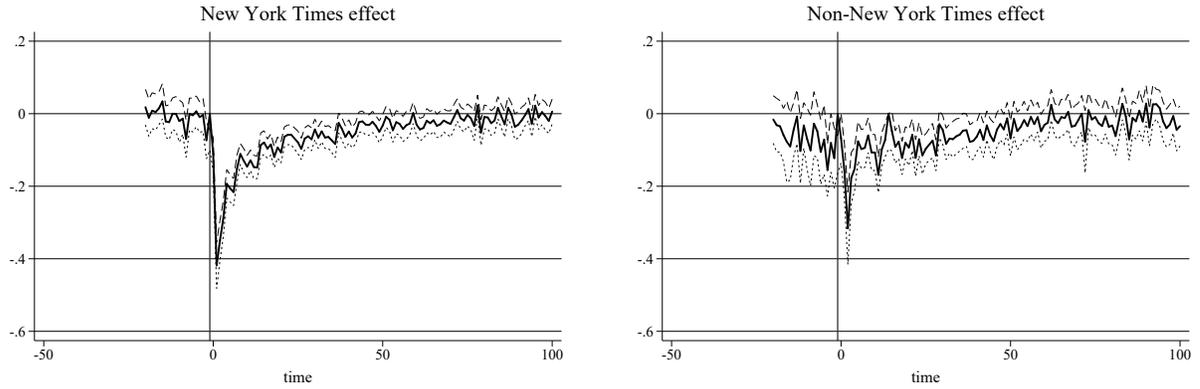


Figure 3: Composition of genres – reviewed vs. not



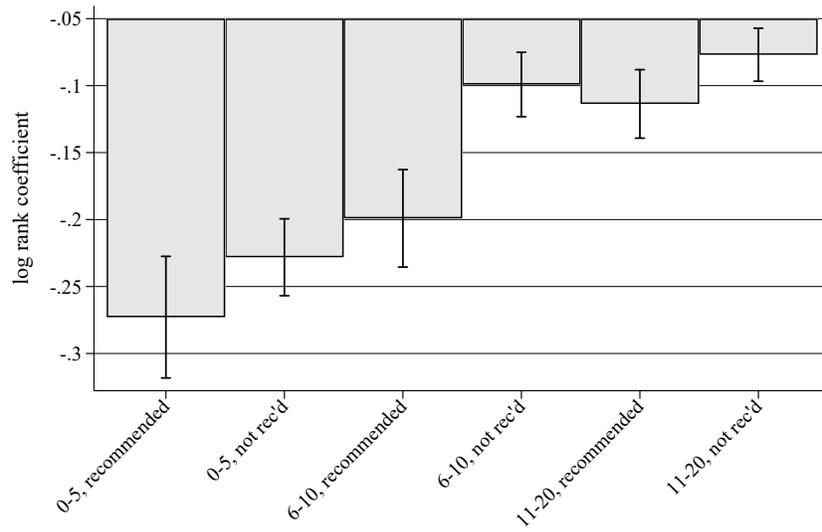
Notes: we calculate sample sales during 2018 by genre and by whether the books were reviewed in professional outlets. The figure reports the difference in the genre distributions between the professional outlets and others, including only the genres that differ by at least one percentage point.

Figure 4: Daily effects of professional reviews on sales ranks



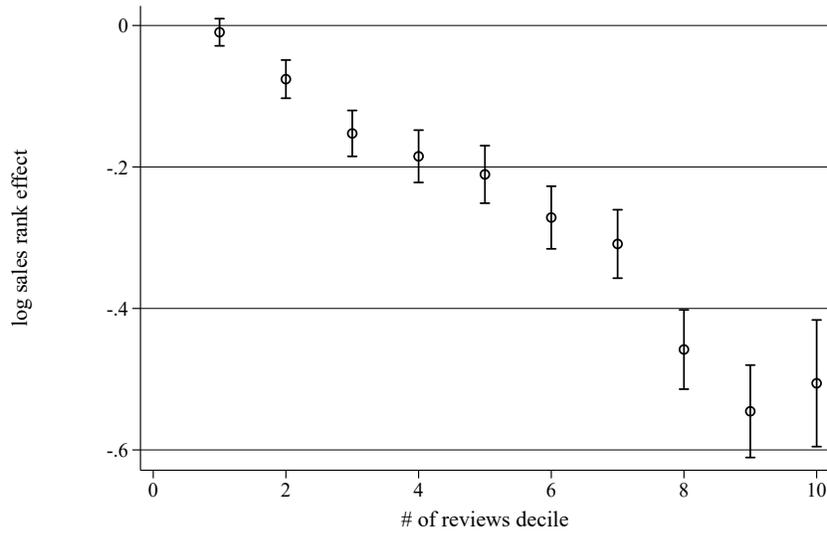
Notes: These figures show coefficients and 95% confidence intervals for each day before and after a New York Times (left panel) and other major review (right panel). The estimates are from a regression of log-rank on its lag, price, stars, and the number of underlying ratings, in addition to title fixed effects and days-since publication dummies, using Amazon US data.

Figure 5: Effect of New York Times reviews – recommended vs. not



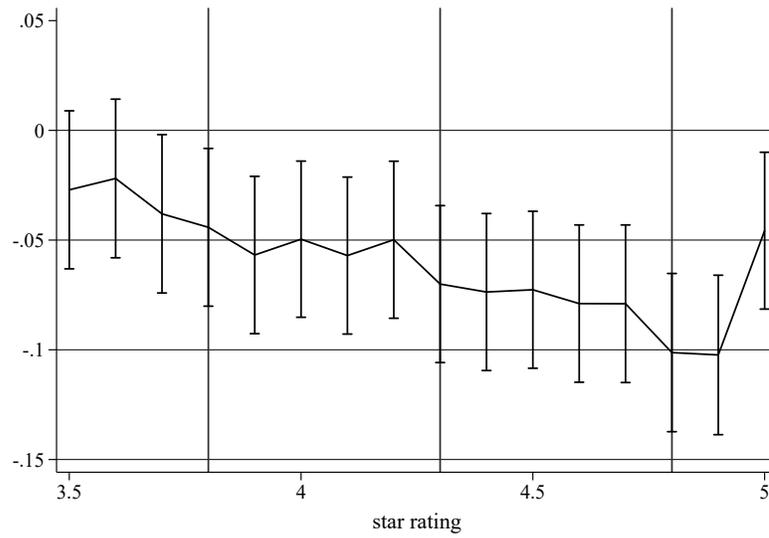
Notes: This figure shows coefficients and 95% confidence intervals for days after a New York Times review, separately for books that were recommended and those that were not. The estimates are from a regression of log-rank on its lag, price, stars, and the number of underlying ratings, in addition to title fixed effects and days-since publication dummies, using Amazon US data.

Figure 6: Effects of Amazon star ratings by ratings deciles



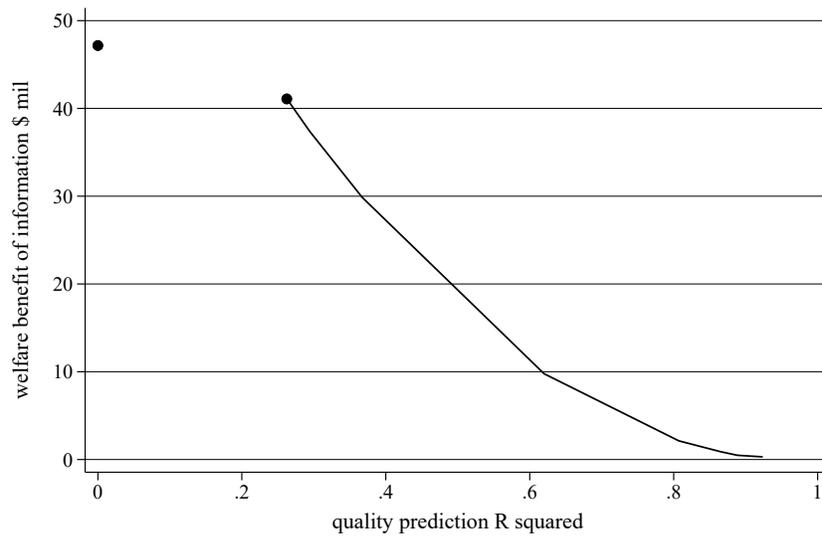
Notes: This figure shows coefficients and 95% confidence intervals for a book's star rating interacted with each decile of the number of underlying ratings, in a regression of log-rank on its lag, price, days since various professional reviews, and the number of underlying ratings, in addition to title-platform fixed effects and days-since publication dummies, using Amazon data from the US, Canada, and Great Britain.

Figure 7: Effects of Amazon star ratings on sales



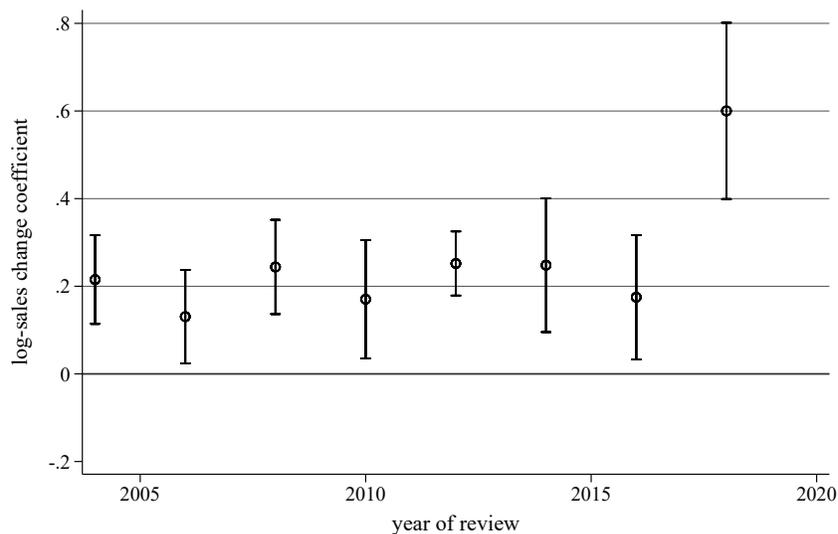
Notes: This figure displays coefficients and 95% confidence intervals for each star rating dummy (from 3.5 to 5 stars) in a regression of log-rank on its lag, price, days since various professional reviews, and the number of underlying ratings, in addition to title-platform fixed effects and days-since publication dummies, using Amazon data from the US, Canada, and Great Britain.

Figure 8: Prediction accuracy and the welfare benefit of crowd information



Notes: This figure depicts the estimated welfare gains from the existence of a star rating system for varying levels of explained variation in star ratings, as measured by the R^2 of a regression of the book's average true star rating on genre and publisher dummies as well as author experience controls. The baseline accuracy, $R^2 = 0.2625$, is denoted by the second dot. We increase the R^2 by adding the following explanatory variable to the regression: $R_j + \kappa \cdot \epsilon_j$, where ϵ_j is a standard normal random error and κ is a scale factor of varying size. The first dot ($R^2 = 0$) assigns the overall average star rating to all books.

Figure 9: Effects of NYT reviews from 2004 to 2018



Notes: This figure displays coefficients and 95% confidence intervals for the review dummy in regressions of $\ln\left(\frac{s_{jt}}{s_{j,t-1}}\right)$ on a dummy for the week after a New York Times review was given and controls for the number of weeks since the book's publication. The regressions were done separately for each even review year from 2004 to 2018, on all books on the New York Times Notable Books lists for their respective years.

Table 1: Sample characteristics

	all	Canada	Great Britain	US	Professionally reviewed	USA Today	Goodreads
price	18.12	24.73	14.60	17.54	19.98	18.38	17.52
star rating	4.38	4.39	4.35	4.39	4.27	4.39	4.40
sales rank	448,609	203,921	639,306	451,150	385,160	432,394	481,372
reviews	326.18	27.73	111.42	471.95	130.90	560.55	166.72
<u>star rating percentiles</u>							
10th	3.7	3.6	3.6	3.8	3.5	3.8	3.8
25th	4.1	4.1	4	4.2	4	4.2	4.1
50th	4.5	4.5	4.5	4.5	4.4	4.5	4.5
75th	4.7	4.9	4.8	4.7	4.7	4.7	4.8
90th	5	5	5	4.9	5	4.9	5
editions	9,146	3,891	3,860	8,631	1,918	4,355	4,920
observations	1,612,489	264,615	325,921	1,021,953	304,312	728,823	666,343

Notes: Average prices, star ratings, sales ranks, and number of reviews, across all days in 2018. The samples include all editions of books that entered the 2018 USA Today weekly bestseller lists, that were reviewed professionally in 2018, or that were reviewed by highly-followed Goodreads reviewers. Column (1) includes all books and all platforms (Amazon US, Canada, and Great Britain). Columns (2)-(4) include all books on each individual platform. Columns (5)-(7) include all platforms for each sample of book titles.

Table 2: Effects of crowd and professional reviews on log sales ranks

	(1)	(2)	(3)	(4)	(5)
lagged log sales rank	0.785*** (0.000766)	0.786*** (0.000764)	0.786*** (0.000764)	0.760*** (0.000656)	0.760*** (0.000656)
log Amazon price	0.194*** (0.00406)	0.193*** (0.00406)	0.195*** (0.00406)	0.188*** (0.00307)	0.190*** (0.00307)
log reviews	0.0484*** (0.00125)	0.0488*** (0.00125)	0.171*** (0.00846)	0.0376*** (0.00100)	0.160*** (0.00739)
log star rating	-0.108*** (0.0118)	-0.107*** (0.0118)	-0.00354 (0.0142)	-0.0680*** (0.00860)	-0.00323 (0.00971)
log reviews x log stars			-0.0829*** (0.00568)		-0.0827*** (0.00497)
US: NYT, 0-5 days		-0.240*** (0.0129)	-0.241*** (0.0129)	-0.257*** (0.0130)	-0.257*** (0.0130)
US: NYT, 6-10 days		-0.129*** (0.0110)	-0.130*** (0.0110)	-0.152*** (0.0111)	-0.153*** (0.0111)
US: NYT, 11-20 days		-0.0883*** (0.00931)	-0.0897*** (0.00930)	-0.101*** (0.00938)	-0.103*** (0.00937)
CA: NYT, 0-5 days				-0.112*** (0.0422)	-0.111*** (0.0422)
CA: NYT, 6-10 days				-0.0640 (0.0407)	-0.0633 (0.0407)
CA: NYT, 11-20 days				-0.00996 (0.0348)	-0.0107 (0.0348)
GB: NYT, 0-5 days				-0.0298 (0.0259)	-0.0294 (0.0259)
GB: NYT, 6-10 days				-0.000492 (0.0240)	-0.000872 (0.0240)
GB: NYT, 11-20 days				0.0130 (0.0207)	0.0129 (0.0207)
US: other, 1-10 days		-0.0300* (0.0173)	-0.0294* (0.0173)	-0.0370** (0.0175)	-0.0364** (0.0175)
US: other, 11-20 days		0.000674 (0.0165)	0.00115 (0.0165)	-0.00114 (0.0167)	-0.000516 (0.0167)
CA: other, 1-10 days				0.0414 (0.0706)	0.0454 (0.0706)
CA: other, 11-20 days				0.0711 (0.0680)	0.0759 (0.0680)
GB: other, 1-10 days				0.0601 (0.0612)	0.0584 (0.0612)
GB: other, 11-20 days				0.0742 (0.0616)	0.0759 (0.0615)
Observations	1,021,765	1,021,765	1,021,765	1,612,014	1,612,014
R-squared	0.968	0.968	0.968	0.961	0.961

Notes: regression of Amazon log daily sales rank on its one-day lag, as well as the log price, log number of reviews, the log of the star rating, and indicators for whether the title had recently been reviewed by the New York Times or another major US outlet. The sample includes titles on the USA Today bestseller list during 2018, as well as titles reviewed in the New York Times and other major US papers during 2018, and books from 2018 reviewed by highly followed Goodreads reviewers in 2018. The first three columns include only data from Amazon's US site. Columns (4) and (5) include data from Amazon's US, Canadian, and Great Britain sites. Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Causal quantity effects

	effect	se
Price elasticity	-0.427	0.007
Amazon stars elasticity (25th pctile)	0.511	0.028
Amazon stars elasticity (50th pctile)	0.769	0.040
Amazon stars elasticity (75th pctile)	1.055	0.056
Amazon stars elasticity (mean)	0.783	0.041
NYT 0-5	0.579	0.029
NYT 6-10	0.345	0.025
NYT 11-20	0.231	0.021
OTH 0-10	0.082	0.039
OTH 11-20	0.001	0.038
<u>% effect of review on annual q</u>		
Other only	0.579	(0.510)
NYT only	5.169	(0.272)
both	6.055	(0.695)

Notes: The price and Amazon star rows show estimated elasticities of quantity sold with respect to price and Amazon stars, respectively, based on column (5) of Table 2. The Amazon stars elasticities are divided into percentiles of the number of underlying ratings. The NYT and OTH rows show percentage impacts on reviews on sales during the relevant numbers of days after the reviews. The bottom panel shows the percentage impacts of being reviewed in the New York Times or other professional outlets on estimated sales over the year. Standard errors are based on 500 parametric bootstrap replications. We draw from the estimated joint distributions of the parameters from Table 2, as well as from the distribution of B from $q_j = Ar_j^{-B}$ using the Nielsen data.

Table 4: Model inputs

variable	value
<u>Market-wide unit sales</u>	
2018 US unit sales (mil)	695
2018 unit sales of Nielsen top 100 (mil)	67.4
Nielsen top 100 share of US sales	0.0984
top 100 share of sample sales	0.3557
US unit sales of sample titles (mil)	192.28
<u>Amazon unit sales</u>	
share of physical sales (2017)	0.455
all titles (mil)	316.2
sample titles (mil)	81.6
<u>Overall market size</u>	
US pop 2018 (mil)	327.2
market size (12*pop)	3,926.4

Notes: This table reports inputs for the nested logit model in Section 5. 2018 US unit sales are from <https://www.publishersweekly.com/pw/by-topic/industry-news/financial-reporting/article/78929-print-unit-sales-increased-1-3-in-2018.html>; 2018 unit sales of Nielsen top 100 are calculated from the Nielsen Bookscan database. US unit sales of the titles in our sample are $Q = \frac{1}{0.3557} \times 0.0984 \times 695$. Amazon's share of physical sales: <https://www.idealogy.com/blog/changing-book-business-seems-flowing-downhill-amazon/>. Amazon unit sales of all titles and our sample titles are market-wide sales times Amazon's share.

Table 5: Welfare impacts of professional reviews and Amazon star ratings

	Stars	Reviews
Δ Revenue (net)	42.81 (4.86)	23.21 (2.24)
$R > \hat{R}$	185.95 (9.57)	
$R < \hat{R}$	-143.14 (7.71)	
Δ CS (total)	41.58 (4.19)	2.54 (0.34)
Δ CS / Δ Revenue	0.971 (0.13)	0.109 (0.01)

Notes: all dollar figures in millions. Baseline revenue is \$3,057.74 million. The scaled dollar figures for star ratings are calculated by multiplying their impact on CS per dollar spent by our estimate of the spending on physical books at Amazon in 2018. This, in turn, is the 695 million volumes sold during 2018, times their average price (\$17.54), times Amazon's share of the market (45.5%). Because we include all of the books reviewed at the New York Times and the other major papers in the sample, the model's direct measure of the change in CS from these reviews requires no scaling. Figures are based on estimates in column 5 of Table 2. Standard errors are based on 100 parametric bootstrap draws.

Appendix

A Goodreads Amateur Review Impacts

In addition to making crowd opinion visible, digitization has also fostered growth in amateurs who distribute their book reviews. Goodreads is a site where readers rate and review books. Goodreads was founded in 2007 and has grown quickly. According to [Narula \(2014\)](#), Goodreads had 25 million users by 2014, and the number of registered users reached 80 million in November 2018. As of April 2016, Goodreads reported that 50 million reviews had been posted to the site. Leaving reviews, and having substantial numbers of followers, are relatively uncommon; but some users have large numbers of followers and leave substantial numbers of reviews. Hence, the appearance of reviews by users with substantial numbers of followers can provide crowd analogues to reviews in critical outlets.

It is difficult to say exactly how visible Goodreads reviews are in comparison with major elite review sources. Based on Similarweb traffic data for December 2018, Goodreads received 98.6 million visits, compared with 302.5 for the New York Times' entire website.

To explore possible impacts of Goodreads reviews on sales, we assemble a list of reviews from Goodreads "most popular" reviewer lists and include those with more than 10,000 followers. This produces a list of 1,742 reviews of titles reviewed and published during 2018.

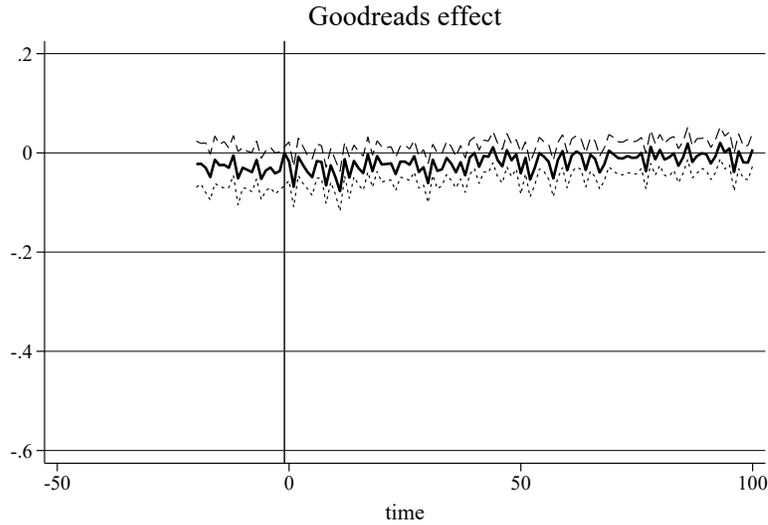
We measure their impacts on sales ranks using two approaches employed in the paper. First, we measure the daily impacts for 20 days before and 100 days after the appearance of highly followed reviewer's review. This is analogous to our approaches in [Figure 4](#) in the paper. [Appendix Figure ??](#) shows the result: there is no apparent effect of these amateur reviewers' reviews on Amazon US sales ranks.

We then measure the impacts for three discrete periods following the appearance of these Goodreads reviews: 0-5, 6-10, and 11-20, also analogous to our approaches in the US-data-only columns of [Table 2](#). We find insignificant effects (coefficients of -0.008 (se=0.008), -0.011 (0.008), and -0.009 (0.006)). We also checked whether effects were different for the books receiving 5-star ratings, but the coefficients were statistically indistinguishable.

B Logit Preliminaries

We obtain the market sizes and collective unit sales of the books in our sample as follows. First, for market size, we assume that each member of the US population is making a monthly decision of whether to purchase a book, so $M = 12 \times 327$ million. Second, we estimate the annual sales per sample title as follows. From data outside our sample, we know the total physical sales for the year (695 million units), and we know that the top 100 weekly titles account for 9.84 percent of

Figure A.1: Goodreads daily effects



Notes: The figure shows coefficients and 95% confidence intervals for each day before and after a Goodreads top reviewers' review. The estimates are from a regression of log-rank on its lag, price, stars, and the number of underlying ratings, in addition to title fixed effects and days-since publication dummies, using Amazon US data.

total physical sales, or 68.3 million units.¹⁶ Because our estimation sample includes only rank data, even with an estimated elasticity of the quantity sold with respect to the sales rank (B), we can only estimate daily sales of each title up to a scalar ρ , i.e. $q_{jt} = \rho / \ln(r_{jt}^B)$. But we can choose ρ so that the sum of weekly top 100 sales in the sample equals 68.3 million. This in turn indicates that sample titles collectively account for 192 million units. Fourth, we know that Amazon accounts for 44.5 percent of US book sales.

C Estimating the Nested Logit Parameter

We can infer the degree of substitutability using the Nielsen top 100 sales weekly data, which we have for 2015-2018. For this purpose, we need a few additional pieces of information, along with an instrumental variables strategy. We describe these in turn.

First, we obtain weekly data on total physical book sales from Publisher's Weekly, which reports this in most but not all weeks. We refer to this as Q_t . We have these data for 124 weeks during 2015-2018. Based on a Pew report indicating that one quarter of people have not read a book

¹⁶See <https://www.publishersweekly.com/pw/by-topic/industry-news/financial-reporting/article/78929-print-unit-sales-increased-1-3-in-2018.html>.

during the prior year, we set market size M equal to three quarters of the US population, implicitly assuming that people are making a weekly choice about whether to purchase a book.

We then define the following variables:

$$s_{jt} = q_{jt}/M, \quad s_{jt|g} = q_{jt}/Q_t, \quad s_{0t} = 1 - Q_t/M.$$

As in Berry (1994), we seek to obtain σ from a regression of $\ln(s_j) - \ln(s_0)$ on $\ln(s_{jt|g})$. Intuitively, identification comes from the relationship between the number of products available and whether the share of the population buying books increases.

There is seasonality in the book market, with a substantial increase in sales around Christmas. Publishers know this and may release more books around Christmas, raising a concern that book the number of books coming out as well as demand might rise around Christmas. This would look like an effect of product entry on market expansion, even if it were not. To address this, we include week-of-the-year dummies.

Second, we need an instrument for the books' inside shares $s_{jt|g}$. One natural idea would be the number of products available in each week. In our data it is by construction 100. More to the point, however, not all products are of equivalent importance. We can appeal to the logic of BLP instruments, which are terms involving the other products in the choice set. Here, for example, we can measure the number of products in the top 100 that were originally released in the past week, 2 weeks, and so on, up to ten weeks. Further, because we have the Nielsen weekly top 100 going back to 2015, we can construct measures of authors' past sales. We can then use measures of the past sales of authors whose new books are in the top 100 this week. We implement this with a series of measures: the number of authors in the current top 100 whose previous sales are in some interval, for 7 intervals.

This gives us 17 possible instruments. To avoid choosing among them arbitrarily, we use the variable selection approach of Belloni et al. (2014). We estimate IV regressions in which we use LASSO techniques for the choices of a) which week dummies to include in the main equation, and b) which instruments to include in the first stage. The procedure selects 4 of the 17 possible instruments and 16 of the possible week dummies. Not surprisingly, the weeks before Christmas are selected. The resulting estimate of σ is 0.471 (with a standard error of 0.0457).