Digitization and Pre-Purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings

By Imke Reimers and Joel Waldfogel*

Digitization has led to many new creative products, straining the capacity of professional critics and consumers. Yet, the digitization of retailing has also delivered new crowd-based sources of pre-purchase information. We compare the relative impacts of professional critics and crowd-based Amazon star ratings on consumer welfare in book publishing. Using various fixed effects and discontinuity-based empirical strategies, we estimate their causal impacts on sales. We use these causal estimates to calibrate a structural demand model. The aggregate effect of star ratings on consumer surplus is, in our baseline estimates, more than ten times the effect of traditional review outlets.
JEL: L15, L81

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. While the number of new products has always exceeded the capacity of professionals to review them, this gap has only grown with digitization.[2] Crowd-based ratings – such as Amazon stars – on the other hand, are available for essentially all products, raising the possibility that another facet of digitization, ubiquitous crowd ratings, can provide information that allows the realization of welfare gains from new products.

These considerations raise the question of how the new crowd-based pre-purchase information made available by digitization affects purchase behavior and, by extension, welfare. To address this, we ask the following specific questions. First, do professional reviews and crowd ratings have causal impacts on demand; and if so, how large are these impacts? Second, how do the two pre-purchase information institutions – professional reviews and crowd ratings – affect the welfare of

* Reimers: Department of Economics, Northeastern University (email: i.reimers@northeastern.edu). Waldfogel: Carlson School of Management and Department of Economics, University of Minnesota, NBER, and ZEW (email: jwaldfog@umn.edu). We are grateful to seminar and conference participants at Chapman University, Cornell University, the University of Georgia, the University of Gießen, the University of Luxembourg, the University of Minnesota, the NBER Summer Institute, the National Taiwan University, Northeastern University, the Pennsylvania State University, Queens University, Tel Aviv University, Toulouse School of Economics, ZEW, and the University of Zurich. We thank Judith Chevalier, Christian Helmers and Maria Ana Vitorino, as well as the referees and editor, for comments.

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

[2]See Waldfogel (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).

consumers? Finally, has the growth of crowd-based rating systems reduced the influence of professional reviews?

We address these questions in the book market, which provides an auspicious study context for a few reasons. First, books are experience goods, so that pre-purchase information is especially 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 US physical book market – that helps identify causal relationships.[3] We have daily measures of Amazon sales ranks and their crowd-based star ratings, for 21,546 editions (from 10,641 titles) selling 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 write. 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 two strategies, one for professional reviews and one 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 an approach in the spirit of Chevalier and Mayzlin (2006), employing both book fixed effects and cross-platform intertemporal comparisons.

Our descriptive analysis gives us causal evidence on the links between pre-purchase information – reviews and ratings – and sales ranks. Quantification of the importance of these two mechanisms requires a framework for welfare analysis. Our measured welfare effects of ratings and reviews allow the consumers' ex ante choice utilities – when consumers have limited pre-purchase information – to differ from their ex post consumption utilities. Calculating welfare effects requires two translational steps beyond the causal evidence. First, we transform effects of pre-purchase information on sales ranks into effects on quantities. This allows the calculation of the elasticities of quantity sold with respect to the Amazon price and the star rating, as well as the percentage impacts of professional reviews on sales. Second, we use our estimated elasticities to calibrate nested logit (and Marshallian) models of demand that facilitate welfare analysis. We also support our interpretation of pre-purchase information with evidence, from individuals' Amazon book ratings, that consumers who buy professionally reviewed titles enjoy them.

We have four broad findings. First, professional review outlets, notably the New York Times, have clear impacts on sales. In the five days following a New York Times review, a book's estimated sales improve by 55 percent. Sales for

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

this time interval improve by roughly 80 percent if the book is not only reviewed but also "recommended" by the New York Times. Over the entire year, a New York Times review raises sales by 2.8 percent. Second, the crowd also has clear effects on sales: The elasticity of sales with respect to Amazon stars averages 0.63, and it is larger when the stars are based on more underlying ratings. Third, per title, the effects of professional reviews are about twice as large as the effects of star ratings; but in the aggregate, professional reviews raise consumer surplus by \$3.18 million while the aggregate effect of Amazon star ratings (\$35.83 million using our baseline approach) is over ten times larger. Despite operating only on the Amazon platform, which is just under half of the book market, effects of the crowd dwarf professional review effects because star ratings are ubiquitous while only a small share – a few percent of the thousands of books published per year – have prominently visible professional reviews. While improved pre-purchase information can affect welfare either by inducing substitution among products or by expanding the market, the welfare gains we calculate arise almost entirely from consumer substitution of better for worse books. Finally, we do not find evidence of substitution between reviews and star ratings: The effect of a New York Times review, in a supplementary analysis covering weekly sales from 2004 to 2018, has not waned with the growth of digitization.

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 essentially all products, including those in categories neglected by professional critics, and they do so without undermining the effects of professional critics on the books and genres professionals do cover.

The paper proceeds in eight sections. Section I provides background on the book market, the evolution of the information environment with digitization, and a discussion of the existing literature. Section II presents a simple theory of choice with and without pre-purchase product information, which organizes our descriptive and welfare analyses. Section III describes our main data on Amazon sales ranks, star ratings, and prices, as well as reviews in major newspapers. Section IV presents our empirical strategies for measuring causal impacts of professional reviews and star ratings on sales, as well as empirical estimates and translations of these into effects on quantities sold. Section IV also provides evidence, from a supplementary dataset on individual user star ratings (Ni, Li and McAuley, 2019), that consumers who buy professionally reviewed titles enjoy them. Section V 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 VI demonstrates the robustness of our welfare results to an alternative, descriptive model of consumer surplus as well as an alternative model of the pre-purchase information consumers would have in the absence of digitization. Section VII examines the complementarity of professional reviews and crowd ratings. Section VIII concludes.

I. Background

A. The US Product and Information Environment for Books

In 2000, roughly 80,000 fiction and non-fiction books were released in the US, and the number of new works released annually has grown sharply since then. In 2012, when 100,000 new US titles appeared in hardback form, the number of new US ebook titles was 280,000.[4] This figure, while impressive, only counts the books with ISBNs ("international standard book numbers"), which many self-published titles lack. Clearly, there has been substantial growth in the number of new books released in the US. The largest physical bookstores, however, only carry roughly 200,000 new and old books, 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]

B. 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 but have rather limited audiences. In 2018, Publishers Weekly reviewed 5,693 books – compared with 6,808 in 2000 and 5,596 in 2010 – or just a few percent of new releases.[6] The Publishers Weekly site attracted 2.15 million visits in December 2018; and the other B2B sites had far fewer, according to Similarweb. Because of the limited reach of B2B sites among end-consumers, we focus on consumer-facing review outlets.

The consumer-facing part of the 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, all of which we include in our sample. We describe the sample in detail below, but here we note three things. First, the elapsed time between publication and review date varies substantially across books, with over three quarters of reviews happening more than one week after the book's publication. Second, books reviewed by these outlets vary widely in popularity: only 9.3 percent of the books professionally reviewed in 2018 appear on the USA Today weekly top 150 list during 2018. Third, the New York Times has far more reviews than the other outlets. In 2018, the New York Times published 1,700 reviews with information

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

[5] See Waldfogel and Reimers (2015) for additional data on the growth in new books since digitization.

[6] These are the numbers of book reviews returned from searches in Publishers Weekly.

on about 2,000 newly-released titles.[7] The New York Times is reported to review one percent of the titles they receive, and the number of titles reviewed in these outlets has not grown over time.[8] Newspaper websites have far more traffic and visibility than the sites for 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 million, according to Similarweb. Measured by both volume of reviews and visibility to consumers, the New York Times is the preeminent US book review outlet.[9]

C. *Crowd-based Star Ratings at Amazon*

Like other digital retailing platforms, Amazon allows users to review and rate books on a five-point scale, and Amazon aggregates this user feedback into star ratings for each book. A few features of the ratings system are noteworthy. First, in contrast to professional reviews, crowd ratings are available for the vast majority of titles. 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.[10] Third, consumers can easily observe the number of underlying ratings on which the visible star rating is based. While all books start with no ratings, 91 percent of the titles in our overall sample, described in more detail below, receive a star rating by the end of 2018. The average edition in our sample has 984 ratings on Amazon's US domain by yearend 2018. Fourth, the star rating for a particular book differs across Amazon's country platforms.

D. *Existing Literature*

Our study is related to four existing literatures. First, it is related to work measuring the impact of professional reviews on product sales. Reinstein and Snyder (2005), Sorensen (2007), Berger, Sorensen and Rasmussen (2010) and Garthwaite (2014) provide four examples of studies employing careful empirical strategies to document impacts of professional reviews on movie and book sales.

[7] The number of reviews is based on the number of New York Times urls containing "books/review." A review can cover more than one book.

[8] Lozada (2015) discusses how the New York Times selects books to review. Based on "book review" queries at `www.nytimes.com`, the New York Times published 10 percent fewer reviews in the five years between 2014 and 2018 than they had between 2004 and 2008.

[9] We ignore magazines because they have small reach or review few books. For example, the New Yorker, which typically provides one long and four short reviews per weekly issue, had 14.9 million site visitors according to Similarweb, about a tenth of the volume at the Washington Post, which we show below to have a negligible effect on sales. While Oprah's book club has been documented to have large effects (Garthwaite, 2014), the club has reviewed only four books per year on average (`https://www.oprahmag.com/entertainment/books/g23067476/oprah-book-club-list/`).

[10] Rather, "Amazon calculates a product's star ratings based on a machine-learned model instead of a simple average. ...These models take into account factors such as how recent the rating or review is and verified purchase status. They use multiple criteria that establish the authenticity of the feedback. The system continues to learn and improve over time" (`https://www.amazon.com/gp/help/customer/display.html`).

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. Examples include Chevalier and Mayzlin (2006); Luca (2016); Duan, Gu and Whinston (2008); Forman, Ghose and Wiesenfeld (2008); Helmers, Krishnan and Patnam (2019), and Senecal and Nantel (2004). Third, our study is related to an emerging literature documenting the consumer benefits of pre-purchase information from digital platforms (Lewis and Zervas, 2016; Farronato et al., 2020; Liu, Ranjan and Shiller, 2020). Finally, our structural welfare analysis makes the distinction between ex ante "decision utility" and ex post experienced utility (Jin and Sorensen, 2006; Allcott, 2011; Train, 2015).

II. Theory: Information, Purchase, and Welfare

A. *The Roles of Ratings and Reviews in Product Purchase*

While both reviews and ratings are pre-purchase information, consumers interact with them in different ways; and they may have different effects on purchase. Professional reviews, which exist for a subset of books, are delivered to consumers as newspaper articles. A review creates awareness of a book, as well as delivering information about its appeal to readers of the review. By conveying information about the book, a review can change the tendency for a consumer to purchase the product.

Consumers interact differently with Amazon star ratings than with newspaper book reviews. Consumers encounter star ratings when shopping for particular books. Rather than alerting consumers to a book's existence, ratings provide quality assessments for those consumers already considering those titles.

If reviews and ratings were not available, then consumers would still have some pre-purchase information; they would form estimates of the quality of book j, perhaps based on product characteristics or traditional word of mouth, which we summarize as a predicted rating, $\hat{R}_j$. Surprisingly positive pre-purchase information – when a product is rated better than consumers would have expected so that $R_j > \hat{R}_j$ – could increase its consumption relative to its consumption in their absence, and vice versa.

B. *Pre-Purchase Information and Welfare*

Assessing effects of pre-purchase information on welfare requires a distinction between expected ex ante utility and experienced ex post utility, as in Jin and Sorensen (2006), Allcott (2011), and Train (2015).

Suppose that consumers were poorly informed prior to purchase and, in particular, that they believed a product's quality to be lower than its true quality ($\hat{R}_j < R_j$). Then their ex ante aggregate demand curve for the product would be given by the dashed curve in Figure 1. They would choose Q_1 units, and at purchase they would expect consumer surplus (CS) equal to region A. Upon consumption, however, they would ascertain the product's true value, so that the ex post experienced CS 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 information prior to purchase is the difference between these surplus regions, or C.

There is an analogous case, in which consumers believe a 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 they 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 consuming either too much or too little of the product when lacking 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.

Our welfare framework in this section considers each product in isolation, but full welfare effects depend on substitution across products and possible market expansion. If the increases in consumption of the "good" products exactly offset the reductions in consumption of the "bad" products, then the welfare effects would arise purely from reallocation. But pre-purchase information can also change total consumption across all products. In our baseline welfare analyses below – a random utility model in which consumers choose among all products at once – we highlight the respective roles played by reallocation across books vs market expansion.

The empirical welfare analysis also requires characterizations of the consumers' product quality predictions absent star ratings or professional reviews. We model the absence of these types of pre-purchase information differently. In the absence of ratings, we assume consumers would expect title j to have some estimate of quality, which we term $\hat{R}_j$. We present two distinct behavioral and statistical approaches to modelling $\hat{R}_j$ in Sections V and VI. We handle professional reviews differently. While professional reviews can in principle be positive or negative, it turns out empirically – both in our data and in previous work (Berger, Sorensen and Rasmussen, 2010) – that reviews tend to have positive impacts on sales. We therefore treat professional reviews as information that directs their readers to products they would enjoy, and we provide evidence consistent with this interpretation in Section IV.C.

III. Data

A. Basic Data Set Construction

The ideal dataset for addressing our questions would be a high-frequency panel on prices and quantities, as well as ratings and review information, for every book sold during some period, along with all of the books professionally reviewed in major outlets. Our data resemble the ideal in some respects but also have some features that require adaptation.

There is no readily available data source with the information we seek, so we build one. First, we create a list of books selling during 2018 that also includes all of the professionally reviewed titles (as well as their review dates). Second, we get sales, price, and pre-purchase information data on as many titles as possible. To create our list of titles, we begin with the top-selling 150 titles of each week from the USA Today bestseller list.[11] During 2018, the USA Today list includes 2,116 distinct titles published either in 2018 or earlier. We then add 1,904 titles reviewed by the New York Times or by the five additional US newspapers (the Boston Globe, the Chicago Tribune, the Los Angeles Times, the Wall Street Journal, and the Washington Post) during 2018. Finally, to ensure that our sample includes titles outside of both bestsellers and books attracting mainstream critical attention, we include two additional sources. First, we include the 3,725 titles reviewed by Publishers Weekly during 2018 and available in hardcover or paperback before 2019.[12] Second, we include the 4,546 titles reviewed in 2018 (and published in 2018 or earlier) by widely- followed users of the site Goodreads.[13] For each of these titles, we obtain a list of its editions' ISBNs by searching for the title and author in the Bowker Books in Print directory.

We obtain review dates for all books reviewed in the New York Times by directly searching at the newspaper's website. For the other newspapers, we assemble lists of reviewed titles from the Bowker Books in Print directory.[14] For books reviewed by the New York Times, we also have a measure of whether the title was among those more favorably reviewed, based on whether the book was included on a New York Times "recommended" list in the weeks after its review appeared. Each week, the New York Times recommends about eight to twelve recently reviewed books, or roughly 40 percent.

The second step in the process is to obtain daily Amazon data. We get data on the respective editions' daily sales ranks, prices, number of ratings, and Amazon

[11]See `https://www.usatoday.com/entertainment/books/best-selling/` for the rankings. They are based on "data from booksellers representing a variety of outlets: bookstore chains, independent bookstores, mass merchandisers and online retailers." See `https://www.usatoday.com/story/life/books/2013/06/04/about-usa-todays-best-selling-book-list/2389075/`.

[12]We obtain these by searching all book reviews directly on the Publishers Weekly website.

[13]We include all reviewers on Goodreads' "most-popular reviewers" lists as of June 2019 who have more than 10,000 followers.

[14]We search for hardcover editions published in 2018 and reviewed by the newspaper, and we then find the dates of reviews appearing in 2018 using Google searches of, say, "Chicago Tribune book review [author title]."

stars for 2018 from keepa.com (Keepa GmbH, 2019), separately for the US site as well as two other domains selling English-language books, the Canadian and UK sites. These data allow us to 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.[15] We are able to obtain Amazon sales rank data for the vast majority (94 percent) of the titles and editions we seek. Of 11,324 titles in the overall list, we obtain sales data for 10,641 titles. We refer to the data encompassing these titles as the "sales data sample."

Unified econometric estimation of the impacts of various kinds of pre-purchase information on Amazon sales ranks requires not only the daily sales ranks but also information on the timing of professional reviews, as well as prices, average star ratings, and the number of reviews that users have left for each book. Some of the Amazon data – prices, number of ratings, or average star ratings – are unavailable for some titles, which reduces to 8,770 the number of titles that we include in our "estimation sample." The titles in this sample account for about 87 percent of the estimated sales in the sales data sample.

Along with the advantages of our high-frequency data come some disadvantages. First, our data cover only one retailer – Amazon – and not the entire market. Still, Amazon accounted for 44.5 percent of US sales of physical books in 2017 – the year before our sample – so our data cover a major part of the market.[16] Second, we observe the sales rank and not the sales quantity for each edition. We are thus in the position of other authors faced with rank rather than quantity data in book publishing (e.g. Chevalier and Goolsbee, 2003; Brynjolfsson, Hu and Smith, 2003; Reimers, 2019). Amazon does not disclose how it calculates its sales ranks, but a few things are clear. Many ranks are updated at least daily, often hourly; and ranks are not based only on the most recent day but are based on a moving average of sales that has a long – multi-day – memory.[17]

Third, we also seek to quantify the respective aggregate impacts of professional reviews and star ratings, but we do not observe all books. Hence, we scale our estimates of star rating effects to known total Amazon book sales to deliver a population estimate. We need not do this with the professional reviews because we observe virtually all of them. We explore how sample representativeness might affect our results in the online appendix.

[15] In addition to higher frequency, the availability of transaction prices is another advantage over weekly sales data from Nielsen, which include only list prices.

[16] See `https://www.publishersweekly.com/pw/by-topic/industry-news/financial-reporting/article/78929-print-unit-sales-increased-1-3-in-2018.html`. This is not a shortcoming of our star ratings analysis, as our estimated effect of star ratings necessarily covers only Amazon.

[17] See `https://www.amazon.com/gp/help/customer/display.html?nodeId=525376` for some detail. Figure **??** shows the time series of the Amazon sales rank for a book with modest sales. When a sale occurs, the rank improves sharply, then drifts up for days until the next sale occurs.

B. Summary Statistics

Table 1 describes the estimation sample. The overall estimation sample, in column (1), includes 8,770 titles (in 13,652 distinct editions) and 3.2 million daily observations across all domains. Columns (2)-(4) report statistics separately for the US, Canada, and the UK. The US sample includes 12,201 editions, and the Canadian and UK samples include 6,800 and 6,339 editions, respectively. The US sample includes substantially more underlying Amazon ratings.

The inter-quartile range of sample star ratings runs between 4.1 and 4.7. The numbers of individual ratings underlying these star ratings vary across books 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. In the US, the median number of underlying ratings is 113. A quarter of the observations have star ratings based on 21 or fewer, and a tenth of the observations have star ratings based on six or fewer ratings.

We also obtain genre information on sample titles from Bowker. As Figure **??** shows, the professionally reviewed subsample has higher proportions in "serious" genres such as social science, biography, and history and lower shares in genres such as cooking, romance, and juvenile fiction and nonfiction. The availability of crowd ratings raises the amount of pre-purchase information available for the genres not attracting the attention of professional review outlets.

C. Supplementary Data

We have weekly Nielsen Bookscan top 100 ranks and sales by title for 2015-2018 (Nielsen NPD, 2019), as well as weekly sales of the New York Times 100 Notable books for each of the even-numbered years 2002-2018. We use these data for three things: calculating the relationship between Amazon ranks and Nielsen sales quantities, estimating a substitution parameter for a nested logit model of demand, and measuring the impact of New York Times reviews on the sales of the Notable titles over time.

In addition, we have a dataset with individual Amazon book ratings during most of 2018 (Ni, Li and McAuley, 2019). We can match 265,747 of these individual star ratings left by 196,719 Amazon reviewers to 4,382 editions (accounting for about 30% of the titles) in our sample. We use these data for two exercises. First, we use them to examine how professionally-reviewed titles appeal to consumers who rely on these reviews. Second, we use these micro-level reviews to model a word of mouth approach to consumers' predictions of book quality ($\hat{R}_j$) in the absence of digitization in Section VI.A.

IV. Empirical Strategies and Descriptive Results

We have three goals in this section. First, we provide causal evidence on the relationships between pre-purchase information (reviews and crowd ratings) and sales ranks. Second, we translate the measured sales rank coefficients into effects on quantities that we use to calibrate structural models for welfare analysis. Finally, we provide evidence that consumers who buy professionally reviewed titles enjoy them more than other titles.

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) a lagged value of the log sales rank for the book at the platform; b) indicators for whether a title has recently received a review from a professional outlet; and c) book- and platform-specific measures of Amazon crowd ratings, prices, and the numbers of underlying ratings. To allow for the possibility that star effects vary with the number of underlying ratings, we also include their interaction.

The specifications can be described via the following equation:

$$\begin{aligned}(1)\quad \ln(r_{jct}) \; = \; & \theta \ln(r_{jc,t-1}) + a\ln(p_{jct}) + g\ln(R_{jct}) + m\ln(ratings_{jct}) \\ & + n\ln(ratings_{jct})\ln(R_{jct}) + h_{\tau c} + f(U_{jt}, S_{jt}) + \mu_{jc} + \epsilon_{jct}\end{aligned}$$

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 professional review. Finally, U_{jt} is days until, and S_{jt} days since the publication of title j, $f(U_{jt}, S_{jt})$ includes first through third order terms to account flexibly for time patterns of sales before and after the publication date; and the terms μ_{jc} are platform-specific edition fixed effects.

A. Effect Estimates

We first focus on the effect of professional reviews. To that end, we include the h terms for each of the days before and after the appearance of a review with the last pre-review day as a baseline. We estimate two sets of these terms, one for the New York Times and another for the other professional review outlets collectively. We estimate this model on only US data. Figure 2 reports the time patterns of New York Times and other professional review effects for 20 days before and 40 days after the review. As the left panel shows, a New York Times review delivers a large and immediate improvement in the sales rank when the review appears. The log rank improves by 0.4, then returns to its baseline about two weeks later. As the right panel shows, professional reviews at other outlets also have detectable effects, but they are much smaller. The coefficients on the remaining variables are shown in column (1) of Table 2.

In our remaining specifications we summarize the professional review effects with three indicators, for 0-5 days, 6-10 days, and 11-20 days after a review. 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. In addition, we include separate measures for the New York Times-recommended and other New York Times-reviewed books to account for differences in the reviews' positivity.

Estimating the effect of Amazon star ratings is more challenging than estimating review effects, as these ratings evolve less discontinuously, and potentially endogenously, over time. A simple approach to measuring the impact of ratings and prices on sales would be to estimate the cross-sectional relationship between an edition's sales rank and its rating, but the obvious shortcoming of this approach is that books 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 book's unobserved quality, following editions on a particular platform over time.

We implement a variant of this approach in column (2) of Table 2, using only US data. The specification includes both edition fixed effects and polynomial functions of time until and since publication, so that all coefficients are identified from within-title variation. The coefficient on the log star rating is -0.121, indicating that a one-percent increase in a book's star rating generates a 0.121 percent improvement in its ranking. This specification also shows impacts of professional reviews that are consistent with those in Figure 2. For a book that is not ultimately recommended, a New York Times review improves the log sales rank by 0.20 in the five days after the review. For a recommended title, the effect is about seven percentage points larger. New York Times reviews without (with) a subsequent recommendation improve the log sales rank by 0.06 (0.15) in days 6-10 following the review. For days 11-20 New York Times effects are smaller, as are the effects of other professional reviews.

We allow the impact of stars to depend on the number of underlying ratings by adding an interaction of $\ln(R_{jct})$ and $\ln(ratings_{jct})$, in column (3). The inclusion of the interaction term shrinks the star rating main effect, and the interaction term itself is negative and significant, indicating that star ratings have larger effects when they are based on more underlying ratings.

The US-data-only approach of columns (1) through (3) is vulnerable to a concern that some unobserved factor is changing both attitudes toward a title and its sales over time. An alternative is to make use of the differences in a book's ratings changes across platforms to ask whether the cross-platform rating change differential gives rise to a cross-platform sales rank change differential for the same book. This is analogous to an 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. This approach supports a causal interpretation unless there are title-specific shocks to both star ratings and demand that are also country-specific.

Column (4) implements the estimation with all three platforms and platform-specific edition fixed effects as well as the polynomial functions of time until and since publication. Column (5) adds the interaction of $\ln(R_{jct})$ and $\ln(ratings_{jct})$. In these specifications, we allow the coefficients reflecting effects of professional reviews to vary by platform, although we only report the US effects. The US New York Times no-recommendation effect for the first five days is about -0.21 in both specifications, while the analogous recommended coefficient is -0.29. 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 similar estimates of the derivative of the sales rank with respect to the star rating.

We also explore whether the interaction specifications are overly restrictive. We replace the $\ln(ratings_{jct})$ term in column (5) with dummies for 50 equal-sized groups according to the number of underlying ratings and the interactions of these indicators with $\ln(R_{jct})$.[18] Figure 3 reports the results. The absolute magnitude of the coefficient on star ratings rises fairly steadily in the log number of ratings until groups near the high end. While stars have essentially no effect on the sales of books with numbers of ratings in the lowest quantile, the coefficient grows in magnitude to -0.6 for the groups near the top.[19]

B. *Translating Ranks into Quantities*

A few steps are required to translate the coefficients from the log rank regressions into quantity elasticities.

Quantities

While Amazon does not disclose sales quantities, we do have information on market-wide sales of the top-100 weekly bestsellers (along with the New York Times Notable books) for 2018. We can match these with 874 of the editions in our sales data sample.

We follow other researchers in assuming that sales follow power laws in ranks (Chevalier and Goolsbee, 2003). Our context has the complication that the Amazon rank data are daily and are based on moving averages of sales while the Nielsen sales data are weekly. We model the relationship between the editions' daily ranks and weekly unit sales by minimizing the squared deviation between predicted weekly sales based on the daily Amazon ranks and actual weekly 2018 Nielsen sales data. That is, we use nonlinear least squares to estimate A and B in: $q_{jw} = \sum_{t \in w} A r_{jt}^{-B} + \upsilon_{jw}$, where t denotes day, w denotes week and υ_{jw} is an additive error. We take 500 boostrap draws on edition weeks to produce standard errors, in parentheses. This yields A=10,167.0 (2,359.3) and B=0.45 (0.058). Given estimates of A and B, we calculate sales of edition j during 2018 by summing across days of the year: $q_j = \sum_{w \in 2018} A r_{jt}^{-B}$. Note that we can calculate sales of all books in the sales data sample, regardless of whether we have data on the other Amazon variables (price, etc.).

[18]Because of bunching in the distribution of # ratings, there are only 46 distinct groups.

[19]The coefficient falls in magnitude for some of the top groups. Below we explore the sensitivity of our main welfare results to a flexible specification that allows the effect of stars to vary across the groups in Figure 3.

Elasticities

Equation (1) is a partial adjustment model, so we find the full effect of a right hand side variable by setting $\ln(r_{jt}) = \ln(r_{j,t-1})$. Then the derivative of a book's log rank with respect to, say, the log price, from equation (1), is $\frac{a}{1-\theta}$. Translating this into a quantity elasticity requires the derivative of log quantity with respect to the log sales rank. The parameter B from above provides an elasticity estimate based on daily data, albeit using Amazon's daily rank variable r_{jt} that reflects both contemporaneous and past sales.[20] Then the reduced form elasticities of quantity with respect to price and the star rating are, respectively:

$$\begin{aligned} \epsilon_p &= \frac{\partial \ln(q_j)}{\partial \ln(p_j)} = \frac{aB}{1-\theta} \qquad \text{and} \\ \epsilon_{Rj} &= \frac{\partial \ln(q_j)}{\partial \ln(R_j)} = \frac{(g + n\ln(ratings_{jt}))B}{1-\theta}. \end{aligned} \tag{2}$$

Table 3 reports estimates of quantity effects from model (5) in Table 2. Rows 2-5 report elasticities 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 with the number of ratings. At the 25$^{\text{th}}$ percentile, the elasticity is 0.392, while it is 0.621 at the median and 0.839 at the 75$^{\text{th}}$ percentile. At the mean, the elasticity is 0.616. The next rows report the effects of professional reviews, during particular time windows after their appearance, on log sales. For example, the 0.438 in the NYT non-recommended first five row indicates that sales increase by 55 percent during the first five days after the appearance of a New York Times review that is not eventually accompanied by a recommendation ($e^{0.438} - 1 = 0.55$).

The bottom panel of Table 3 reports percentage impacts of reviews on annual simulated sales estimating daily sales quantities for an edition j as above. We estimate the counterfactual sales absent professional reviews by substituting the following for the log rank:

$$\ln(rank_{jt}) - \frac{h_k B}{1-\theta} 1(k)_{jt} \tag{3}$$

Here, h_k denotes a coefficient measuring the impact of professional reviews on sales ranks, $1(k)$ is an indicator, and k refers to both the outlet and timing of the review, such as the first five days after the receipt of a recommended 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

[20] We could instead use contemporaneous weekly sales and rank data from Nielsen. Both approaches yield similar estimates: our estimate gives 0.47 using daily data, while we get 0.54 with weekly data. Larger absolute-value elasticities deliver larger welfare benefits of stars in relation to professional reviews, so for conservatism we use the daily estimate. We explore the sensitivity of our results to alternative measures of the quantity-rank elasticity in the online appendix.

absence of the respective sources of pre-purchase information. This allows us to calculate the percentage impacts on sales. For example, receiving a New York Times review without a recommendation (but not another professional review) raises sales by 2.19 percent during 2018. This effect is equivalent to that of a year-long increase of 0.16 in a book's star rating.[21]

Finally, Table 3 also reports a price elasticity of demand of -0.17. This book-level measure appears rather inelastic. We offer four 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."[22] Second, Reimers and Waldfogel (2017) also obtain inelastic estimates for books at Amazon. Third, the money price of a book is a small component of the full cost of consumption, which may explain the low money price elasticity. Finally, as we will discuss further below, while the absolute size of the welfare effects of pre-purchase information depends on the price coefficient, our main study finding – the relative size of the welfare effects of professional vs crowd reviews – is invariant to it.

C. Professional Reviews and Ex Post Satisfaction

The goal of our analysis is to compare the welfare effects of the two types of pre-purchase information. Pre-purchase information can only raise consumer surplus if it directs consumers to products they like. Establishing this is straightforward for star reviews: Higher star ratings indicate that consumers liked these products more. It is more complicated for professional reviews. It is not self-evident that professional reviews direct their readers to books they will enjoy. To explore this, we merge the books in our sample with underlying 2018 Amazon review data (Ni, Li and McAuley, 2019) for those titles. This gives us a sample of 265,747 individual star ratings on 4,382 editions in our sample left by 196,719 distinct users. We use these data to ask whether the users who read and rate professionally reviewed books – and may therefore rely on professional reviewers – tend to prefer the books that are professionally reviewed.

The average Amazon star rating for a book that is not professionally reviewed is 4.48. Users who have not rated a professionally reviewed book rate these critic-ignored books at 4.50, on average, whereas "highbrow" users, who have rated a professionally reviewed book, rate these critic-ignored books less favorably, at 4.30 on average. While highbrow users are less favorable than other users in their ratings of the critic-ignored books, they rate the professionally reviewed books more highly: 4.34 stars vs 4.30 (difference = 0.041, se = 0.009). This suggests that consumers who read professionally reviewed books tend to like them more

[21]The elasticity of sales with respect to stars is 0.616 at the mean. So, the increase in star rating needed to produce an equal effect is 3.6 percent (=0.0219/0.616). Because the mean star rating is about 4.40, this is an increase of 0.16 stars.

[22]See `https://www.cbsnews.com/news/amazons-jeff-bezos-looks-to-the-future/`.

than other books.

V. Welfare Analysis

While the foregoing analysis demonstrates causal impacts of professional reviews and star ratings on sales, it does not provide a theoretically consistent way to compare the welfare benefits from the availability of professional reviews and the existence of Amazon stars. As our theoretical model in Section II suggests, however, the counterfactual change in consumer surplus provides a natural basis for comparison, and a structural demand model facilitates this calculation. We calibrate our descriptive estimates to structural models of demand, allowing us to explicitly explore the roles of substitution toward better books vs market expansion in welfare effects. We then estimate the welfare impacts of star ratings and professional reviews based on a one-level nested logit model of demand, using the sales data sample for the US market.

A. A Simple Structural Model

To perform our welfare analysis, we calibrate a nested logit model to the estimated elasticities. We begin by defining a consumer i's utility from product j as $u_{ij} = \delta_j + \epsilon_{ij}$, where ϵ_{ij} follows a Type 1 extreme value distribution, and

$$\delta_j = \ln(s_j) - \sigma \ln(s_{j|g}) - \ln(s_0). \tag{4}$$

The s terms describe the product's shares: $s_j = q_j/M$, $s_{j|g} = q_j/Q$, and $s_0 = 1 - Q/M$, where M denotes the market size and $Q = \sum_{k \in \mathcal{J}} q_k$ is the sum of quantities for the books in our sales data sample $\mathcal{J}$.[23] Finally, σ reflects the substitutability of products and, by extension, the degree of market expansion arising from the presence of pre-purchase information. A σ of 1 implies full substitution and no expansion. We estimate this as 0.373 in the online appendix, but we also show that our estimates are little affected by σ. Each product's share is then

$$s_j = \frac{e^{\delta_j/(1-\sigma)}}{1 + \sum_{k \in \mathcal{J}} e^{\delta_k/(1-\sigma)}} \frac{D^{1-\sigma}}{1 + D^{1-\sigma}},$$

where $D = \sum_{k \in \mathcal{J}} e^{\delta_k/(1-\sigma)}$ (see Berry, 1994).

Let δ_j be the mean utility of product j 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

[23] To determine the market size, we assume that each US person makes a monthly choice of whether to purchase a book. Recall that we calculate annual q_j by translating sales ranks into quantities using A and B from Section IV.B.

unknown, they are, respectively, related to the estimated parameters a; g and n; and h from equation (1). We can calculate the nested logit expressions for the elasticities of quantity with respect to price and rating, set these equal to the reduced form elasticities described above, then solve for the utility function parameters.

For example, the nested logit model gives a simple expression for the price elasticity of demand:

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

Given $\hat{\epsilon}_p$ from the descriptive analysis above, σ (from the online appendix), and $s_{j|g}$, s_j, and p_j (which are data), we solve for a parameter estimate of α_j for each j, which we average to obtain our estimate of utility function parameter α. We infer γ_j, for each book, by solving the analogous equation for star rating elasticities. Unlike α, which does not meaningfully vary across titles, γ_j varies meaningfully because of the interaction between stars and the number of ratings.

We solve for the utility function parameters associated with professional reviews in a related way. Our descriptive analysis tells us how each sales quantity q_j would have been different in the absence of reviews, ${q'}_j$. The model analogue of our descriptive measure $\ln(q_j/{q'}_j)$ is the review-induced percentage change in sales for reviewed books, relative to the review-induced percentage change in sales for unreviewed books. A few lines of algebra show that in our nested logit model, $\ln(q_j/{q'}_j) = \frac{\psi_j}{1-\sigma}$. Given the degree of substitutability σ, we therefore know ψ_j for each book.

Given values of the utility function parameters, we can compare the consumer surplus achieved in the market in counterfactual scenarios – without Amazon star ratings or without professional reviews – to the baseline when both are present.[24] To do this, in addition to the status quo utility level δ_j from equation (4), we need its analogues in the absence of crowd ratings and professional reviews. Counterfactual utility absent Amazon stars is defined by $\delta_j^s = \delta_j - \gamma_j(R_j - \hat{R}_j)$, where we determine $\hat{R}_j$ as described below; and counterfactual utility absent professional reviews is defined by $\delta_j^p = \delta_j - \psi_j$.

The change in CS associated with star ratings is given by equation (5):

$$\Delta CS = \frac{M}{\alpha}\left[\ln\left(1 + \sum \exp\left(\frac{\delta_j}{1-\sigma}\right)^{1-\sigma}\right) - \ln\left(1 + \sum \exp\left(\frac{\delta_j^s}{1-\sigma}\right)^{1-\sigma}\right) - \sum \gamma_j(R_j - \hat{R}_j)s_j^s\right], \tag{5}$$

where the term $\sum \gamma_j(R_j - \hat{R}_j)s_j^s$ is an adjustment reflecting the possibility that what is consumed has ex post utility that differs from the ex ante value (such as area B in Figure 1), and s_j^s is the market share of product j in the absence of star ratings (so that Ms_j^s corresponds to Q_1 in Figure 1).

[24]We leave prices unchanged in these counterfactual scenarios because unreported estimation exercises show negligible effects of either type of pre-purchase information on prices.

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

B. *Empirical Implementation*

Book Sales Absent Pre-purchase Information

We assume that a professional review affects consumers' perceptions of the book; and the review effect estimated above reflects the revelation of the book's true quality and that it is better than expected. Modeling an environment without star ratings is more complicated. Star ratings exert a continuous effect on sales. In our baseline approach, we assume that consumers would rely on word of mouth or personal experience to develop predictions of book quality reflecting the historical relationship between book characteristics and quality. In particular, we model consumers' beliefs about book quality ($\hat{R}_j$) via a regression of Amazon stars on publisher fixed effects, genre fixed effects, and dummies for the author's prior experience. For each edition, we use the average star rating across all days. The resulting regression explains 33.94 percent of the variation in average book star ratings. 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 examine a different approach to modeling pre-purchase information based on word of mouth in Section VI.

From Sample to Population Estimates

Our counterfactual simulations apply the causal estimates from the estimation sample to the sales data sample, which encompasses many but not all books available to consumers. Total book sales in 2018 are 695 million units (Milliot, 2019), whereas our estimates of A and B mapping Amazon sales ranks into quantities indicate that our sales data sample accounts for 331.55 million sales, or a little under half of the total. To determine the industry-wide impacts of stars and reviews from our sales data sample, we therefore need to do some scaling.

Our coverage of professionally reviewed titles is essentially complete, so we need not scale the results. Star ratings, by contrast, are available for the vast majority of books at Amazon, so we need to scale our welfare estimates up for this part of the analysis. To create aggregate star rating estimates from our sales data sample, we scale sample quantities up to population sales (by multiplying by 695/331.55). We then scale down to Amazon's share of physical sales (by multiplying by 0.445). For titles in the sales data sample for which we do not observe star ratings, we assume that digitization delivers no welfare benefit. That is, we treat $(\hat{R}_j - R_j)$ as 0.[25]

[25]Our approach provides a conservative estimate, for two reasons. First, we over-estimate the benefit

C. Results

Table 4 shows the revenue and consumer surplus results using the structural model with baseline σ. We obtain standard errors by taking 500 bootstrap draws on A, B, σ, and the estimated parameters in column (5) of Table 2. Our first exercise asks how the presence of star ratings at Amazon, and professional reviews, respectively affect book industry revenue. On net, stars raise revenue by \$27.51 million (se = 11.79 million). Beneath this net change is a \$92.56 (17.00) million increase in spending on books whose star ratings are better than consumers would have expected, while the books whose star ratings are worse than expected experience a \$65.05 (6.29) million reduction in revenue. The presence of professional reviews, which only provide positive information about books, raises spending by \$19.98 (2.10) million.

Of greater interest to us is the impact of professional reviews and star ratings on consumer surplus. The existence of the professional reviews appearing in 2018 raises the consumer surplus from 2018 US book spending by \$3.18 (0.41) million. Our estimate of ΔCS from stars, based only on Amazon sales, is \$35.83 (6.98) million. This is 11.27 (2.28) times as large as the effect of professional reviews. This is our main finding.[26]

Our basic specifications include an interaction of the number of ratings and the average star rating. We also estimate a model with 50 groups of books according to the number of ratings they have received, along with interactions of indicators for these groups with the log star rating (see Figure 3). Using this flexible specification, the resulting ΔCS ratio is 11.13 (2.23). Hence, our main result is not driven by the specification of the interaction in the descriptive analysis.

The relative impacts of star ratings and professional reviews depend on the respective per-book impacts, as well as the prevalence of professional reviews. To compare per-book impacts, we perform a welfare calculation based on counterfactual removals of the star ratings on only the professionally reviewed books (leaving the star ratings in place on the others). As Table 4 shows, the presence of star ratings on the reviewed books adds \$1.68 (0.30) million, in comparison to the \$3.18 (0.41) million added by the reviews. On a per-book basis, then, the dollar value of the impact of professional reviews – which affect the entire market – is about twice as large as that of stars, which we assume to operate only at Amazon. Despite being smaller per book, aggregate effects of stars are ten times

of professional reviews by applying the estimation sample coefficients to the sales sample. Unreported regressions of log sales on professional review variables using the entire sales sample (without the other control variables) yield larger coefficients for observations that are in the sales but not the estimation sample. Second, we under-estimate the effect of stars by allowing no welfare benefit for titles that are in our sample but missing Amazon rating measures.

[26]The relative welfare effects of stars and reviews vary substantially across genre. While the ratio is under one for political science and social science, it is over 50 for romance and juvenile fiction, and over 300 for juvenile nonfiction, crime fiction, and cooking. These differences are driven largely by the rarity of professional reviews in those categories.

larger in the aggregate because of the broad incidence of star ratings.

While our direct evidence on star ratings covers only Amazon's 44.5 percent of the market, we note that non-Amazon sales may also be affected by digital word of mouth analogous to Amazon stars. First, other digital platforms, such as Barnes & Noble and Goodreads, feature user-based star ratings. Second, users might use Amazon as a source of information and shop elsewhere. In the extreme, if the effect of stars we document here informed all 2018 book spending, the ΔCS from stars would be \$80.52, or 25.32 times as large as the effect of professional reviews.

D. *Substitution vs Market Expansion*

Our baseline results use an estimated substitution parameter σ of 0.373, which is consistent with books being imperfect substitutes for one another, and therefore gives some scope for market expansion to affect the welfare estimates. We can calculate alternative results – ΔCS for stars and professional reviews – for a range of substitution parameters. As σ rises toward 1 – the case with full substitution and no market expansion – the change in total consumption approaches zero. As Table 4 shows, the change in CS from stars goes from \$36.09 (7.43) million (when $\sigma = 0$) to \$34.67 (6.29) million (when $\sigma = 0.95$), the change in CS from professional reviews goes from \$3.27 (0.41) to \$3.03 (0.39) million, and the ratio varies between 11.18 (2.34) and 11.42 (2.20). In other words, our basic result – in both absolute dollar and relative ΔCS terms – is insensitive to the substitution parameter. Our result arises almost entirely from consumers switching from "worse" to "better" books.

VI. Robustness

In this section we explore the sensitivity of our basic result to our modeling assumptions on consumers' beliefs about book quality absent the availability of digital star ratings, and to an alternative, Marshallian triangle approach to CS calculation. The online appendix explores the robustness of our results to our α and B parameter estimates, to the substitution patterns among books, as well as sample representativeness issues.

A. *Consumer Pre-purchase Information Absent Digital Stars*

Better Prediction Within the Baseline Approach

The measured welfare benefits of Amazon star ratings 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 33.94 percent of the variation. It is possible that the regression understates, or overstates, the ability of consumers to predict quality. We explore the sensitivity of our result to prediction accuracy using the approach of Aguiar and Waldfogel

(2018). We add the following explanatory variable to the star rating regression: $R_j + \kappa \cdot \nu_j$, where ν_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.

A larger R^2 corresponds to more pre-purchase information absent star ratings and therefore a lower welfare benefit of stars. At our baseline R^2 of 33.9 percent, the ratio of CS benefits from stars vs professional reviews is 11.27. When there is no pre-purchase information ($R^2 = 0$), the ratio is about 13; and if prediction accuracy corresponded to an R^2 of 50 percent, then the ratio would be roughly seven. If consumers could perfectly predict quality absent star ratings ($R^2 =$ 100 percent), star ratings would deliver no welfare benefit. Consumers would need to be very well informed in the absence of reviews and ratings – an R^2 of over 80 percent – in order for Amazon stars to add as little consumer benefit as professional reviews.

Word of Mouth Approach to Quality Prediction Absent Stars

We can take a different approach to pre-purchase information absent digitization, which corresponds more closely to word of mouth (WOM). Without Amazon, consumers might consult acquaintances offering their assessments as the equivalent of individuals' star ratings of books; and they might effectively take mental averages over their friends' assessments. We can construct a simple model to approximate this kind of information gathering using the micro data on Amazon star ratings from Ni, Li and McAuley (2019). Suppose, for illustration, that a consumer would consult one person familiar with a book. We can simulate this by taking a single random draw from the customer ratings for that book. The single consulted rating would play the role of $\hat{R}_{ij}$, where i denotes consumer i, and j denotes title j, which we would term $\hat{R}^1_{ij}$. Alternatively, consumer i might consult N "friends," in which case we could measure the consumer's counterfactual prediction of quality by the average of N underlying ratings, $\hat{R}^N_{ij}$. As N rises to the total number of reviews available in the digital era, $\hat{R}^N_{ij}$ approaches R_j, and the WOM information people obtain from others causes pre-purchase information to be as good without digitization as it is with digitization.

For any number of "friends" N, we compute $\hat{R}^N_{ij}$ for each title j and consumer i. We then calculate a consumer-specific deviation between the consumer's estimate and the true quality $(\hat{R}^N_{ij} - R_j)$. This allows us to calculate consumer i's change in CS across all books, ΔCS_i, via the single-consumer version of equation (5). Because every consumer has different friends, we repeat this exercise 100 times. We then scale up the average across these 100 ΔCS_i to the market size for an aggregate ΔCS effect.

We encounter some complications implementing this approach. We have underlying ratings for only 4,382 editions of the sample total; and because the underlying ratings data are available only through October 2018 (as well as because Amazon stars are not simple averages of underlying ratings), the averages from the micro data do not equal the reported Amazon stars. Still, we can compare the WOM approach on the subset of editions for which we have the micro data to our baseline approach by scaling the ΔCS for stars from this subset of editions to the industry total as above, and we compare these effects to our baseline effects of professional reviews, for varying numbers N of "friends."[27]

With N=1, $(\hat{R}_{ij}^N - R_j)$ tends to be very large, in part because individual ratings take only integer values, while the averages across many users (R_j) are continuous. As N rises, $\hat{R}_{ij}^N$ approaches R_j, and ΔCS for stars, and therefore the ratio, falls. We find that if consumer word of mouth is based on consulting five or fewer friends who have already read the book, then pre-purchase information made available by digitization raises CS more than in our baseline estimates. If consumers consulted ten friends, the ratio would be 6.2. Even if consumers consulted 20 friends to form their estimates, digitization would deliver three times as much ΔCS as professional reviews. We conclude that our basic result, that the increase in consumer surplus from digital star ratings is much larger than the increase in surplus from professional reviews, is robust to different models of the pre-purchase information consumers would have if Amazon did not report star ratings.

B. Marshallian Approach

One might be worried that the results are driven by the logit error. We offer an alternative, product-by-product approach to ΔCS calculation that is more transparent but does not reflect the quantities of each product chosen in market equilibrium. Following our theoretical framework in Section II, we can calculate changes in Marshallian consumer surplus as triangles. Our descriptive model gives us estimates of the change in consumption between the baseline and the counterfactual states without pre-purchase information, Δq_j. Table 3 gives us the price elasticity of demand, which in turn implies a demand curve slope of m. Then we approximate the change in CS from the availability of pre-purchase information – separately for each book – as $\Delta CS_j = (\Delta q_j^2)/2m$. When we aggregate these estimates across products, separately for stars and professional reviews, we get \$37.66 (8.01) million as the ΔCS from stars and \$4.76 (0.60) million as the ΔCS from reviews, for a ratio of 7.91 (1.80).

As with our baseline estimates, these estimates entail changes in total consumption with the availability of pre-purchase information. We can eliminate market expansion from this calculation by applying a scalar factor of proportionality x such that $x\Sigma_{\Delta q_j<0}\Delta q_j + (1/x)\Sigma_{\Delta q_j \geq 0}\Delta q_j = 0$ for each of stars and professional reviews. With this restriction, the change in CS from stars is \$40.52 (7.83), while ΔCS from reviews is \$3.89 (0.48); and the ratio is 10.41 (2.24). This is similar to both the unrestricted Marshallian approach and the baseline logit results. We conclude, first, that our basic result is driven almost exclusively by better

[27] In this approach, in contrast with the baseline regression approach, each book's quality is equally predictable. That is, the predictions are based on the same number of friend draws.

matching and, second, that it is not driven by the logit error.

VII. Substitutability of Crowd and Professional Reviews

Crowd ratings and professional reviews may function as substitutes or complements for one another. This section explores this both over time and across titles.

A. Professional Review Effects over Time

The growing availability of star ratings may have changed the effect of reviews. To explore this, we would ideally repeat the foregoing analyses for earlier years, prior to the diffusion of online retail and associated crowd ratings. This is infeasible, however, because our daily ranking data do not reach back far enough. Still, we can employ the approach of Berger, Sorensen and Rasmussen (2010) and use weekly physical book sales data to estimate the impact of New York Times book reviews on demand.

For this analysis, we assemble data on a comparable list of books over time, the 100 New York Times Notable Books for even-numbered years from 2004 to 2018. We manually search for these books' ISBNs in the Nielsen Bookscan database to collect weekly unit sales, and we obtain their review dates from the New York Times.[28] We then estimate regressions of the form:

$$\ln\left(\frac{q_{jw}}{q_{j,w-1}}\right) = \lambda \text{review}_{jw} + \beta X_{jw} + u_{jw},$$

where q_{jw} denotes the sales of book j in week w, $\text{review}_{jw} = 1$ in the week immediately following the New York Times review, and X_{jw} includes controls for the number of weeks since the book's release. We also include a dummy variable that equals one in all weeks after publication. Like Berger, Sorensen and Rasmussen (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 4 presents coefficient estimates. We find that professional reviews had positive and significant effects on sales in all years. The size is roughly constant through 2016, then more than doubles in 2018. If anything, the effect of the New York Times book reviews has increased over the last 15 years. This may be due to a different aspect of digitization: The number of digital-only subscriptions to the New York Times rose from about 100,000 in March 2011 to over 2.5 million in the third quarter of 2018 (Richter, 2018).

[28]We limit our analysis to the list of notable books because their reviews are likely most positive.

B. Title-level Substitution

Finally, we explore the possible substitution between star ratings and professional reviews at the micro level. Using variants on the regressions in Table 2 – and data from all three domains – we include only books that are professionally reviewed, and we interact a post-professional review indicator with the main variables in equation (1), the price, stars, number of ratings, and their interaction. The changes in the coefficients on star rating and the review interaction following the professional review appearance are such that the star effect is larger following the review appearance, reflecting complementarity between reviews and ratings.

VIII. Conclusion

Digitization has delivered 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, we find that book reviews in the New York Times and other major newspapers have substantial impacts on book sales. New York Times reviews raise sales by at least 55 percent in the five days after a review and by 2.8 percent over the year. We also show that Amazon star ratings significantly affect quantities sold.

Because these two forms of pre-purchase information have causal impacts on buying behavior, they also affect welfare. Using our baseline approach, the existence of Amazon crowd ratings adds more than ten times as much consumer surplus on the platform as professional reviews add to surplus derived from purchases through all channels. Pre-purchase information affects welfare mainly by leading consumers to those products that they would enjoy more – including many products for which information would be very limited without digitization – rather than by inducing consumers to buy more products overall. 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. Finally, the absolute impacts of professional reviews have not declined with the rise of crowd ratings.

Digitization allows aggregation of consumers' product experiences into measures that deliver substantial welfare benefits. Crowd ratings may provide an important source of welfare benefit in categories where professional reviews are uncommon. We expect that this will have a number of effects on both distribution platforms as well as underlying suppliers. For example, information that leads to useful product recommendations allows platforms to create positive user experiences, whether it is choosing the right cultural product or an appealing hotel or restaurant. Digitally-enabled crowd information also may affect underlying markets. First, by enriching the information environment, ubiquitous rating information may enable new or otherwise unknown products to compete with well-known existing products. Second, crowd information may serve as a substitute for other mechanisms by which products have traditionally become known to consumers, such as investments in advertising or brand, or third-party certification. Finally, while the star ratings at Amazon are common across users and are publicly disclosed, much of the information that platforms – such as Amazon and Netflix – collect on user behavior is proprietary; and platforms use this information to create personalized recommendations. It would be of substantial interest to know how the welfare benefits of such personalized recommendations compare with the impacts of the general stars we study.

REFERENCES

Aguiar, Luis, and Joel Waldfogel. 2018. "Quality predictability and the welfare benefits from new products: Evidence from the digitization of recorded music." *Journal of Political Economy*, 126(2): 492–524.

Allcott, Hunt. 2011. "Consumers' perceptions and misperceptions of energy costs." *American Economic Review*, 101(3): 98–104.

Berger, Jonah, Alan T Sorensen, and Scott J Rasmussen. 2010. "Positive effects of negative publicity: When negative reviews increase sales." *Marketing Science*, 29(5): 815–827.

Berry, Steven T. 1994. "Estimating discrete-choice models of product differentiation." *The RAND Journal of Economics*, 242–262.

Brynjolfsson, Erik, Yu Hu, and Michael D Smith. 2003. "Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers." *Management Science*, 49(11): 1580–1596.

Chevalier, Judith A, and Dina Mayzlin. 2006. "The effect of word of mouth on sales: Online book reviews." *Journal of marketing research*, 43(3): 345–354.

Chevalier, Judith, and Austan Goolsbee. 2003. "Measuring prices and price competition online: Amazon. com and BarnesandNoble. com." *Quantitative marketing and Economics*, 1(2): 203–222.

Dargis, Manohla. 2014. "As indies explode, an appeal for sanity: flooding theaters isn't good for fimmakers or filmgoers." *New York Times*, 9.

Deutschman, Alan. 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(5): 120–132.

Duan, Wenjing, Bin Gu, and Andrew B Whinston. 2008. "Do online reviews matter?—An empirical investigation of panel data." *Decision support systems*, 45(4): 1007–1016.

Farronato, Chiara, Andrey Fradkin, Bradley Larsen, and Erik Brynjolfsson. 2020. "Consumer Protection in an Online World: An Analysis of Occupational Licensing." National Bureau of Economic Research.

Forman, Chris, Anindya Ghose, and Batia Wiesenfeld. 2008. "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets." *Information systems research*, 19(3): 291–313.

Garthwaite, Craig L. 2014. "Demand spillovers, combative advertising, and celebrity endorsements." *American Economic Journal: Applied Economics*, 6(2): 76–104.

Greenfield, Jeremy. 2012. "Seven Advantages Barnes & Noble has in the Bookseller Wars." *Digital Book World.*

Helmers, Christian, Pramila Krishnan, and Manasa Patnam. 2019. "Attention and saliency on the internet: Evidence from an online recommendation system." *Journal of Economic Behavior & Organization*, 161: 216–242.

Jin, Ginger Zhe, and Alan T Sorensen. 2006. "Information and consumer choice: the value of publicized health plan ratings." *Journal of health economics*, 25(2): 248–275.

Keepa GmbH. 2019. "Keepa - Amazon Price Tracker." *https://keepa.com.*

Lewis, Gregory, and Georgios Zervas. 2016. "The welfare impact of consumer reviews: A case study of the hotel industry." *Unpublished manuscript.*

Liu, Siqi, Bhoomija Ranjan, and Benjamin Shiller. 2020. "Are coarse ratings fine? Application to crashworthiness ratings' format." *Unpublished manuscript.*

Lozada, Carlos. 2015. "How to get your book reviewed in the New York Times, if your name isn't David McCullough." *Washington Post.*

Luca, Michael. 2016. "Reviews, reputation, and revenue: The case of Yelp. com." *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, , (12-016).

Martin, Adam. 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/.*

Milliot, Jim. 2019. "Print Unit Sales Increased 1.3% in 2018." *Publishers Weekly.*

Nielsen NPD. 2019. "Bookscan." *https://bookscan.npd.com.*

Ni, Jianmo, Jiacheng Li, and Julian McAuley. 2019. "Justifying recommendations using distantly-labeled reviews and fine-grained aspects." 188–197.

Pompeo, Joe. 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, Imke. 2019. "Copyright and generic entry in book publishing." *American Economic Journal: Microeconomics*, 11(3): 257–84.

Reimers, Imke, and Joel Waldfogel. 2017. "Throwing the books at them: Amazon's puzzling long run pricing strategy." *Southern Economic Journal*, 83(4): 869–885.

Reinstein, David A, and Christopher 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(1): 27–51.

Richter, F. 2018. "The "Failing" NY Times Passes 2.5 Million Digital Subscriptions." *https://www.statista.com/chart/3755/digital-subscribers-of-the-new-york-times/*.

Senecal, Sylvain, and Jacques Nantel. 2004. "The influence of online product recommendations on consumers' online choices." *Journal of retailing*, 80(2): 159–169.

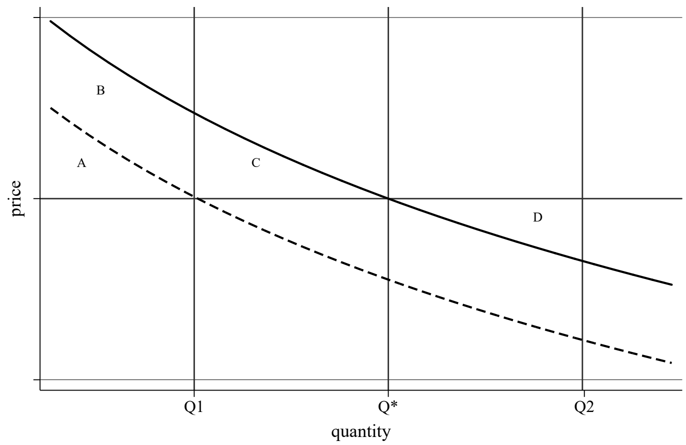
Sorensen, Alan T. 2007. "Bestseller lists and product variety." *The journal of industrial economics*, 55(4): 715–738.

Train, Kenneth. 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, Joel. 2017. "How digitization has created a golden age of music, movies, books, and television." *Journal of Economic Perspectives*, 31(3): 195–214.

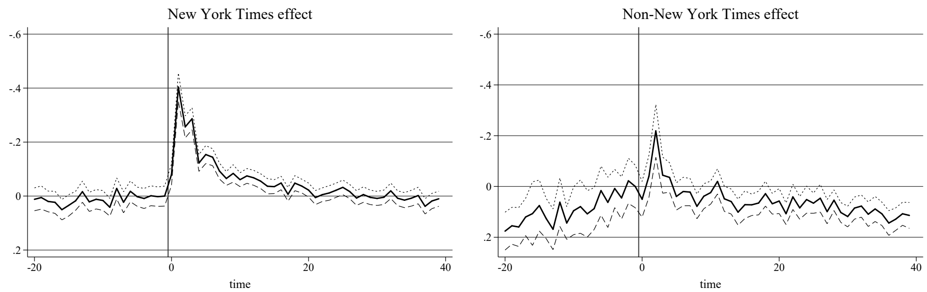
Waldfogel, Joel, and Imke Reimers. 2015. "Storming the gatekeepers: Digital disintermediation in the market for books." *Information economics and policy*, 31: 47–58.

FIGURE 1. ILLUSTRATION – WELFARE ANALYSIS OF PRE-PURCHASE INFORMATION



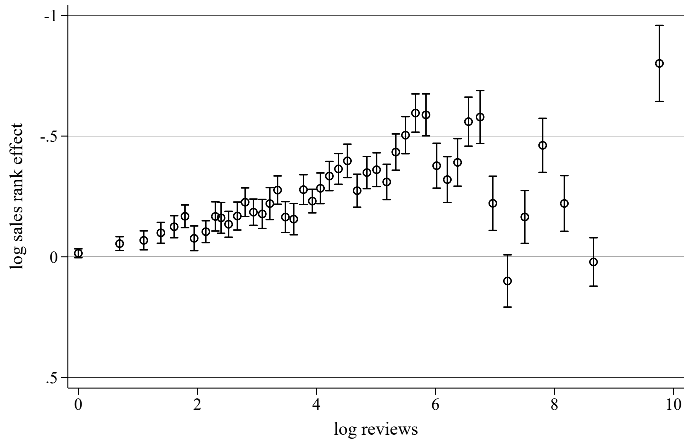
Notes: This figure illustrates demand curves under full information about a product's quality (solid line) and with limited ex ante information about the quality when the expected quality is less than the true quality (dashed line). The corresponding consumer surplus under full information is areas $A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

Figure 2. Daily effects of professional reviews on sales ranks



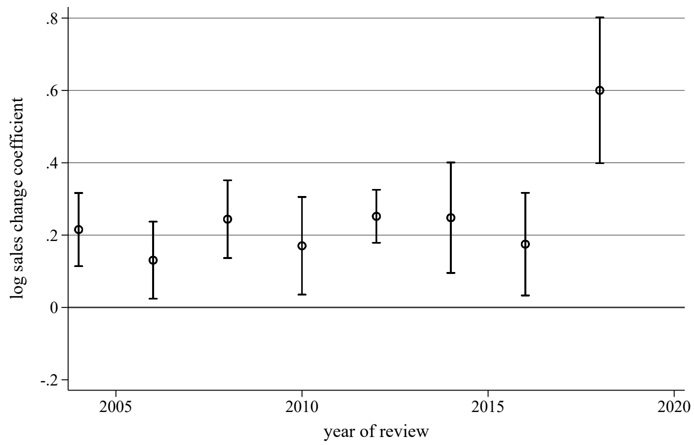
Notes: These figures show coefficients and 95% confidence intervals for each day before and after a New York Times (left panel) and other major review (right panel). The estimates are from a regression of log rank on its lag, price, stars, and the number of underlying ratings, in addition to title fixed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

Figure 3. Effects of Amazon star ratings for 50 ratings quantiles



Notes: This figure shows coefficients and 95% confidence intervals for a book's star rating interacted with each of 50 quantiles 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 polynomial functions of the days until and since publication, using Amazon data from the US, Canada, and Great Britain. The y-axis is reversed to reflect that sales move inversely with sales ranks.

FIGURE 4. EFFECTS OF NEW YORK TIMES REVIEWS FROM 2004 TO 2018



Notes: This figure displays coefficients and 95% confidence intervals for the review dummy in regressions of $\ln\left(\frac{q_{jw}}{q_{j,w-1}}\right)$ on a dummy for the week after a New York Times review 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
price	16.42	21.07	13.12	15.85
star rating	4.38	4.35	4.36	4.40
sales rank	562,085	347,955	781,661	562,232
# ratings	745.51	96.34	199.48	1,220.60
star rating percentiles				
10^{th}	3.8	3.7	3.7	3.9
25^{th}	4.1	4.1	4.1	4.2
50^{th}	4.5	4.4	4.5	4.5
75^{th}	4.7	4.7	4.7	4.7
90^{th}	5.0	5.0	5.0	4.9
titles	8,770	4,747	5,021	8,274
editions	13,652	6,800	6,339	12,201
observations	3,220,809	722,335	703,209	1,795,265

Notes: Average prices, star ratings, sales ranks, and number of reviews, across all days in 2018, for the estimation sample. Column (1) includes all books and all platforms (Amazon US, Canada, and Great Britain). Columns (2)-(4) include all books on each individual platform.

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

	(1)	(2)	(3)	(4)	(5)
lagged log sales rank	0.812	0.813	0.812	0.782	0.782
	(0.0006)	(0.0006)	(0.0006)	(0.0005)	(0.0005)
log Amazon price	0.0737	0.0725	0.0730	0.0804	0.0810
	(0.0023)	(0.0023)	(0.0023)	(0.0017)	(0.0017)
log # ratings	0.0772	0.0782	0.147	0.0618	0.150
	(0.0010)	(0.0010)	(0.0061)	(0.0008)	(0.0054)
log star rating	-0.117	-0.121	-0.0409	-0.0789	-0.0161
	(0.0108)	(0.0108)	(0.0138)	(0.0078)	(0.0091)
log # ratings × log stars			-0.0467		-0.0594
			(0.0041)		(0.0036)
NYT not recommended					
0-5 days		-0.199	-0.199	-0.213	-0.214
		(0.0128)	(0.0128)	(0.0129)	(0.0129)
6-10 days		-0.0618	-0.0621	-0.0800	-0.0805
		(0.0113)	(0.0113)	(0.0114)	(0.0114)
11-20 days		-0.0341	-0.0344	-0.0393	-0.0399
		(0.0094)	(0.0095)	(0.0096)	(0.0096)
NYT recommended					
0-5 days		-0.267	-0.267	-0.285	-0.285
		(0.0183)	(0.0183)	(0.0184)	(0.0184)
6-10 days		-0.153	-0.154	-0.180	-0.181
		(0.0155)	(0.0155)	(0.0157)	(0.0157)
11-20 days		-0.0748	-0.0755	-0.0869	-0.0881
		(0.0130)	(0.0130)	(0.0132)	(0.0132)
Other papers					
1-10 days		-0.0413	-0.0401	-0.0573	-0.0555
		(0.0157)	(0.0157)	(0.0161)	(0.0161)
11-20 days		0.0128	0.0137	0.00891	0.0104
		(0.0151)	(0.0151)	(0.0155)	(0.0155)
Observations	1,795,265	1,795,265	1,795,265	3,220,809	3,220,809
R-squared	0.972	0.972	0.972	0.964	0.964

Notes: Regression of Amazon log daily sales rank on its one-day lag, as well as the log price, log number of ratings, the log of the Amazon star rating, and indicators for the number of days since the title was reviewed by the New York Times or another major US outlet ("other"). The first three columns include only data from Amazon's US site. Columns (4) and (5) include data from Amazon's US, Canada, and UK sites. Columns (4) and (5) also include unreported interactions of the NYT and other review variables with indicators for Canada and Great Britain. All specifications include country-specific title fixed effects as well as polynomial functions of the days until and since the book's publication. Robust standard errors in parentheses.

Table 3—Causal quantity effects

	effect	se
Price elasticity	-0.166	0.022
Amazon stars elasticity (25^{th} pctile)	0.392	0.054
Amazon stars elasticity (50^{th} pctile)	0.621	0.084
Amazon stars elasticity (75^{th} pctile)	0.839	0.114
Amazon stars elasticity (mean)	0.616	0.084
NYT 0-5, not recommended	0.438	0.061
NYT 6-10, not recommended	0.165	0.031
NYT 11-20, not recommended	0.082	0.023
NYT 0-5, recommended	0.584	0.084
NYT 6-10, recommended	0.370	0.057
NYT 11-20, recommended	0.181	0.035
Other 0-10	0.114	0.037
Other 11-20	-0.021	0.034
% effect of review on annual q		
Other only	0.710	0.44
NYT (not rec'd) only	2.189	0.43
NYT (rec'd) only	4.301	0.92
NYT not rec'd and other	3.833	0.94
NYT rec'd and other	6.189	1.44
average	2.641	0.52

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 reported according to percentiles of the number of underlying ratings. The NYT and Other rows show percentage impacts of reviews on sales during the relevant numbers of days after the reviews at the New York Times and other newspapers, respectively. 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.

Table 4—Welfare impacts of professional reviews and Amazon star ratings

	Stars	Reviews	Ratio
ΔRevenue (net)	27.51	19.98	
	(11.79)	(2.10)	
$R > \hat{R}$	92.56		
	(17.00)		
$R < \hat{R}$	-65.05		
	(6.29)		
ΔCS (baseline)	35.83	3.18	11.27
	(6.98)	(0.41)	(2.28)
ΔCS (50 categories)	35.83	3.22	11.13
	(7.13)	(0.42)	(2.23)
ΔCS (reviewed books)	1.68	3.18	0.53
	(0.30)	(0.41)	(0.13)
ΔCS ($\sigma = 0$)	36.58	3.27	11.18
	(7.43)	(0.42)	(2.34)
ΔCS ($\sigma = 0.95$)	34.67	3.03	11.42
	(6.29)	(0.39)	(2.20)
ΔCS (Marshallian: unconstrained)	37.66	4.76	7.92
	(8.01)	(0.60)	(1.80)
ΔCS (Marshallian: $\Delta q = 0$)	40.52	3.89	10.41
	(7.83)	(0.48)	(2.24)

Notes: All dollar figures in millions. Figures for stars are based on scaling sample results up to Amazon's share of 2018 book sales of 695 million units. 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 ΔCS from these reviews requires no scaling. Figures are based on estimates in column (5) of Table 2. Standard errors are based on 500 parametric bootstrap draws.