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

Imke Reimers
Northeastern University

Joel Waldfogel
University of Minnesota, NBER, and ZEW

October 6, 2019

Digitization has led to product proliferation, straining traditional institutions for providing consumers with pre-purchase information; 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 and prices for thousands of the top-selling books of 2018, along with information on the books' Amazon star ratings and 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 3.9 percent overall; and the elasticity of sales with respect to an Amazon star is about 1/6. We use these causal estimates to calibrate structural models of demand for welfare analysis. Under the view that reviews and ratings change preferences rather than just providing information, professional reviews raise consumer surplus by about 3 percent of revenue, or by about $85 million overall, while Amazon star ratings raise consumer surplus by 1 percent of Amazon US book revenue, or by $58 million overall. Under the view that ratings and reviews are purely informative, the increases in surplus are much smaller, although they are roughly a tenth as large as the respective effects of information on revenue. Crowd ratings add significantly to the benefits delivered by professional reviews and do so without diminishing the effect of traditional review sources.

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. The number of new books has always exceeded the capacity of professionals to review them, and 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 facilitate discovery and allow the realization of welfare gains from discovery of new products.

These considerations lead to the question of how pre-purchase product information 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 the sales of books; 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, books are consumed? Third, how do these pre-purchase information sources – professional reviews and crowd ratings – affect the welfare of consumers?

[1] See Deutschman (2004), Pompeo (2017), McG. Thomas (1999), 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).

This paper explores these questions in the market for books. Books at Amazon provide an auspicious context for study for a few reasons. The number of professional reviews, and particularly the number appearing in highly visible outlets, is relatively small and therefore feasible to observe and quantify. Second, and perhaps most important, we have high frequency 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 sells 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 in two ways. First, we make a cross-platform longitudinal comparison for measuring causal impacts of star ratings. Second, we employ 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. We 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, say, the elasticity of quantity sold with respect to the Amazon price or the star rating, or 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. We calculate welfare effects of ratings and reviews under both persuasive and purely informative interpretations of their effects.

Because books are viewed as serious cultural products, professional critics have traditionally viewed themselves as guardians of culture, steering readers toward worthy art and literature. The growing availability of lay opinion as a guide to consumers raises a question of taste leadership, in particular whether lay readers or professional critics will shape consumption decisions. This, in turn, raises a question of whether the impact of professional critics has fallen with the growth of digitally enabled crowd ratings.

The paper proceeds in six sections. Section 1 provides background on the book market, and how its information environment has evolved with digitization. This includes data on the number of books released in the U.S. over time, the numbers of titles reviewed by major traditional review outlets, and information on the visibility of these review outlets. Section 1 also describes 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 from *The New York Times, The Wall Street Journal, The Washington Post, The Chicago Tribune, The Boston Globe,* and *The Los Angeles Times*. Section 4 presents our empirical strategies for measuring causal impacts of professional reviews and star ratings, on sales ranks and quantities sold. Section 4 also presents empirical estimates, on both the causal impacts of ratings, prices, and reviews, as well as the relative impacts of professional vs crowd reviews on the sorts of books promoted and consumed. Section 5 then turns to welfare analysis. We calibrate structural demand models using our causal estimates. We then estimate the respective welfare gains arising from Amazon star ratings and

professional reviews. Section 6 provides estimates of the effect of New York Times reviews on sales over time, 2004-2018.

We have five broad findings. First, professional reviewer outlets, notably the New York Times and to a smaller 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. Over the entire year, a New York Times review raises sales by 3.9 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 is about 1/6. Third, because professional critics focus on a subset of genres while the crowd reviews everything, professional reviews – which systematically raise sales - shift attention toward "serious" genres such as biography and social science. Fourth, both professional critics and crowd ratings affect consumer welfare. Under the persuasive interpretation, professional reviews raise consumer surplus by 2.8 percent of spending on sample books, or by $88 million overall, while crowd-based Amazon star ratings raise consumer surplus by almost 1 percent of book expenditure at Amazon, or by $55 million overall. Under the purely informative view, welfare effects are much smaller: reviews raise CS by $2.84 million, or 0.09$ of revenue, while star ratings raise CS by $1.06 million, or by 0.02% of revenue; but both star ratings and reviews raise CS by about a tenth of their respective effects on revenue. Fifth, while the welfare benefit of the crowd adds substantially, at least proportionally, 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 reviews has not waned. We conclude that digitization has not only delivered a proliferation of new products but has also provided new information mechanisms that, in relative terms at least, add substantially to the information from traditional

review sources. These crowd-based reviews provide pre-purchase information on all books and genres, including those neglected by professional critics, and do so without undermining effects of professional critics on the books and genres they cover.

I. Background

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.[3] 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 carry roughly 200,000 titles, so only a small fraction of new titles have traditionally been marketed directly to consumers.[4] Even before digitization, product discovery was a significant challenge; the challenge has grown substantially since.[5]

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 B2B reviews targeted at libraries and bookstores, from outlets such as Publishers Weekly, Library Journal, and Kirkus. These outlets

[3] 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 U.S.
[4] See Greenfield (2012).
[5] See Waldfogel and Reimers (2015) and Waldfogel (2017) for additional data on the growth in new books since digitization.

review relatively large numbers of titles – although a small share of releases – but have rather limited audiences. The consumer-facing part of the reviewing environment consists mainly of reviews in daily newspapers.

We can 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. Table 1 provides a list of major review outlets, including both the number of titles they reviewed in 2018, as well as two measures of visibility of the outlet: Google searches on their names, and Similarweb data on traffic to their domains in December 2018.

Among B2B book review outlets, Kirkus and Booklist reviewed over 7,500 titles each, Publishers Weekly reviewed 5,603, School Library Journal reviewed 4,919; and Library Journal reviewed 3,692. A number of major U.S. newspapers contain many fewer but still substantial numbers of book reviews. Excepting the New York Times, the major newspapers in Table 1 (in Boston, Los Angeles, Washington, and San Francisco) 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 at some point by The New York Times. Newspapers have far more general traffic and visibility than B2B book review outlets. USA Today and the Washington Post have 190 and 120.5 million monthly visitors, while the New York Times has 302.5. Google search volumes are similar, although the Washington Post had slightly more than

the New York Times during 2018. The New York Times is the most widely circulated outlet among those reviewing books.

While Table 1 shows that the New York Times is not the only mass-market outlet reviewing books, an examination of the co-incidence of reviewing across newspapers reveals that the majority of books reviewed by any of the newspapers in Table 1 are also reviewed by *The New York Times*. Table 2 shows the overlap in hardcover editions reviewed among the major consumer-facing review sources. About 80% of the books prominently reviewed are reviewed by *The New York Times.*

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, Amazon aggregates individuals' ratings into an overall rating, which they report to a tenth of a star. However, the aggregation is not a simple averaging. Rather, "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 are from verified purchasers, and factors that establish reviewer trustworthiness." Second, 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. Finally, the star ratings for a particular book differ across country platforms. Leaving ratings at Amazon is common. In our sample, described in more detail below, the average number of reviews per title is 366.

4. Existing Literature

Our study is related to three existing literatures. First, our study is related to the literature measuring the impact of professional reviews on product sales. Reinstein and Snyder (2005), Sorensen (2007), Berger, Sorensen, and Rasmussen (2010), and Garthwaite (2014) provide three examples of studies employing clever 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 (2009) Duan, Gu, and Whinston (2008), Forman, Ghose, and Wiesenfeld (2008), Helmers, Krishnan, and Patnam (2015), 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 half stars – to measure causal impacts of Yelp ratings on restaurant sales. Below we implement approaches that build on both of these. Finally, our work is related to welfare analyses that make a distinction between ex ante "decision utility" and ex post experienced utility, such as Jin and Sorensen (2006), Alcott (2013), and Train (2015).

2. Theory: Rating and Review Information, Purchase, and Welfare

1. Information and Purchase

Reviews and ratings provide can affects consumers' tendency to purchase products. information on experience goods prior to purchase. 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*, and p_j is the product's price; and x_j contains other observables on product *j*.

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, $\widehat{R_j}$. Utility absent the reviews would then be

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

It's easy to see then that the presence of a positive review – when a product is better than expected so that $R_j > \widehat{R_j}$ – 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 ($\widehat{R_j}$). We address this in section 5.

2. *Review Information and Welfare*

It is of interest to us to measure the welfare benefit to consumers from the availability of review and rating information. The effects of ratings on reviews on welfare depend on whether

we view them as changing preferences – the "persuasive" case - or as simply providing information.

Consider first the persuasive case. Then obtaining information that a product is better than consumers had expected changes preferences so that people attach higher value to the product. In the absence of information, people would consume Q_1 units, obtaining CS = A. In the presence of the review or rating information, people would attach a higher value to the product, would consume Q* units, and would derive CS = A + B + C. Then the welfare benefit of learning that a product is better than expected would be B + C. (There is an analogous case in which consumers would have consumed Q* without reviews and ratings, achieving CS = A + B+ C. When they obtain information, then consume Q1, delivering surplus of A. The arrival of information then reduces CS by B + C).

The purely informative case, which makes a distinction between the ex ante "decision utility" animating purchase and the ex post "decision utility" that consumers experience from products, works differently. Instead of changing their preferences – and how highly they value products – the ratings and reviews merely give them pre-purchase information they would obtained themselves if they had consumed the product.[6]

To see how the purely informative approach works, consider the case in which a product is better than consumers expect it to be. Consumers who purchase the product would experience the full value of the product, but their purchase decisions are based on the expectation that the product is not as good as it actually is. This is depicted in Figure 1. The solid line depicts the

[6] Studies making a distinction between "decision utility" and the ex post "experienced utility" of consumers include Jin and Sorensen (2006), Alcott (2013), and Train (2015).

demand curve for a product that would prevail if consumers were accurately informed prior to purchase. Then they would purchase Q^* units of the product, and they would experience consumer surplus consisting of regions $A + B + C$.

Suppose, instead, that consumers were uninformed prior to purchase and, in particular, that they believed the quality to be lower than its true quality. Then their ex ante demand curve would be given by the dashed curve, and they would choose Q_1 units. Above, we calculated these consumers' experienced consumer surplus as region A. Instead, here, the consumers purchase Q_1 units, which upon consumption they perceive at their true value. Hence, the ex post experienced consumer surplus is 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. Having pre-purchase information therefore allows CS to be higher by region C; and the value of these consumers of getting access to this review information is region C.

There is an analogous case, in which consumers believe the product is better than it actually is 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 loss arises from a "triangle" associated with either consuming too much or too little of the product. 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 we have two measures of the change in welfare from reviews and ratings. If the reviews and ratings change preferences, then:

$$\Delta CS_{persuasive} = CS_{ratings} - CS_{no\ ratings}$$

Alternatively, the purely informative change in CS adjusts the above for the ex post surprise:

$$\Delta CS_{informative} = CS_{ratings} - \left[CS_{no\ ratings} + adjustment\right].$$

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, $CS_{no\ ratings}$ 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.[7]

In what follows, we incorporate the parameter estimates from our causal analysis into a random utility model analogous to the above to develop estimates of the welfare benefits arising from professional reviews, and from crowd ratings, respectively.

3. Data

1. Data Set Construction

The ideal dataset for addressing our questions would be a 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. 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 books are included in the study. Second, we describe how we obtain professional

[7] When consumers choose Q_1 units but should have chosen Q^*, then $CS_{informed} = A+B+C$, $CS_{uninformed}=A$, and adjustment = B.

outlet's review timing. Third, we explain how we obtain lists of ISBN numbers for particular editions, 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 bestseller list we know. 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 of the books reviewed in the New York Times (or any of the other major review outlets listed in Table 2) 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 books (4,920 editions) reviewed by widely followed users of the site Goodreads (all reviewers on the most-popular reviewers list with more than 10,000 followers). 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 in 2018 were published in, say, 2017.

We then obtain the review dates using the newspapers' websites as well as Bowker's Books in Print. We obtain review dates for books reviewed in the New York Times directly from the newspaper's website. For the other newspapers (Boston Globe, Chicago Tribune, Los Angeles Times, Wall Street Journal, and Washington Post), we began with the Bowker lists and then found the reviews for 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, we obtain a list of the books' ISBN numbers from Bowker, which we use to retrieve Amazon data on the respective editions' sales ranks, prices, and Amazon stars. Our data on sales ranks, as well as star ratings and prices, are from keepa.com, which provides Amazon data on physical book editions. A title can have multiple physical editions with separate sales ranks, and we include all of these editions in the sample.

Most previous studies of books make use of Nielsen data, which are available weekly by edition. While Nielsen includes list prices, it does not provide information on the prices actually charged for books. Our daily Amazon data allow us to take a different approach. We obtain the Amazon data for both 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. 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 findings

represent a major part of the market, albeit not its entirety.[8] Second, 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; Brynjolffson, Hu, and Smith, 2003). Ranks are valuable measures of quantity, but many of our analyses below require a way to translate ranks into estimates of sales quantities.

2. Sales data

Amazon does not disclose how it calculates its sales rank, but a few things are clear.[9] 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 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.

Table 3 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 8,631editions, 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.

[8] See https://www.publishersweekly.com/pw/by-topic/industry-news/financial-reporting/article/78929-print-unit-sales-increased-1-3-in-2018.html.
[9] https://www.amazon.com/gp/help/customer/display.html?nodeId=525376

3. Supplementary data

In addition to the main analysis sample consisting of daily sales ranks, we also make use of Nielsen weekly sales data for three ancillary analyses. We have weekly US sales quantities for the top 100-selling physical editions of each week. We employ these data for 2015-2018 for estimating the nested logit substitution parameter, in the appendix. We use just the data for 2018 for estimating the elasticity of sales quantities with respect to ranks.

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 "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 four goals in this section. First, we ascertain whether there is credible causal evidence linking professional reviews and crowd ratings with sales ranks. Second, we quantify these estimates as elasticities of sales ranks with respect to rating and review variables. Third, 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 quantify effects and calibrate structural models for welfare analysis. Fourth, we compare the types of books promoted by professional reviewers versus crowd raters.

1. Documenting the Impact of Professional Reviews

In order to measure the impact of reviews we need a model how log sales ranks would have evolved but for the reviews. To this end define:

$\ln(r_{jt}) =$ log sales rank of title j on date t,

$t_{jp} =$ publication date for title j,

$t_{jr} =$ review date for title ,

$\mu_j =$ a title fixed effect.

We begin with a simple approach to ascertaining whether professional review outlets have an impact on sales. We regress a title's log rank ($\ln(r_{jt})$) on a title fixed effect (μ_j), a lagged sales rank to deal with the serial correlation in the dependent variable, dummies for the number of days before and after publication ($\pi_{t-t_{jp}}$), and dummies for the number of days before and after the appearance of the review ($h_{t-t_{jr}}$).

$$\ln(r_{jt}) = \theta \ln(r_{j,t-1}) + \mu_j + \pi_{t-t_{jp}} + h_{t-t_{jr}} + \varepsilon_{jt}. (1)$$

These regressions include all of the titles in the sample but only the US-domain data. Figure 3 reports the $h_{t-t_{jr}}$ coefficients on the dummies for the days before and after the review's appearance (setting the last pre-review day's coefficient to zero). We estimate this separately for New York Times reviews and for the reviews of the other professional outlets. As the top panel shows, a New York Times review delivers a large an 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 bottom panel shows, professional reviews at other outlets also have detectable effects, but they appear to be much smaller.

Our finding that major sources of traditional reviews have causal impacts on sales has numerous antecedents in existing research. For example, Berger et al (2010) and Garthwaite (2016) both find impacts of traditional book reviews on book sales using Nielsen weekly data.

2. *Effects of the Crowd "Word of Mouth" via Amazon Stars*

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). We work toward our preferred approach by discussing a sequence of possible empirical strategies.

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:

$$\log(r_j) = b_0 + a\ln(p_j) + g\ln(R_j) + \varepsilon_j \qquad (2)$$

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 the problem with (2) 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. Because the sales rank is based on an average of current and past sales, we need to include a lagged dependent variable to account for the serial dependence. The model takes the form:

$$\log(r_{jt}) = \theta \log(r_{j,t-1}) + \mu_j + a\ln(p_{jt}) + g\ln(R_{jt}) + \varepsilon_{jt} \quad (3).$$

Then the effects of the log price (p) and star rating (R) on the log rales rank would be identified from the within-title changes. This approach, while more attractive than (2), 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, where c indexes countries/platforms. With a single point in time, this is:

$$\log(r_{jc}) = \mu_j + a\ln(p_{jc}) + g\ln(R_{jc}) + \varepsilon_{jc} \quad (4)$$

This approach implicitly assumes that the clienteles of the two platforms have the same preferences, so that the quality of the title (μ_j) is the same to the users of both platforms. Implicitly, the on-average identical users of the two platforms are assumed to be subjected to the "experiment" that the different platforms have exogenously different ratings of the same title. Identification here is threatened by the possibility that consumers using the different platforms differ in their attitudes toward particular books.

This is one of the approaches pursued by Chevalier and Mayzlin (2006) in their study of word of mouth in books. Using a cross-sectional sample of roughly 2,500 titles for a point in time, they compare sales ranks and ratings at Amazon and Barnes and Noble (BN) and find that titles with a higher rating on, say, Amazon, have a better sales rank on Amazon relative to BN.

A third possibility is to 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 rankings and reviews across the platforms' environments. That is, one can estimate:

$$\log(r_{jct}) = \theta\log(r_{jc,t-1}) + \mu_{jc} + \mu_{jt} + a\ln(p_{jct}) + g\ln(R_{jct}) + \varepsilon_{jct} \quad (5)$$

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.

Table 4 reports estimates of the longitudinal regression of log sales ranks on star reviews and prices described above. The first column uses only US platform data and includes no edition fixed effects. We offer this column for comparison rather than out because its identification assumptions are justified. The second column includes only US data but adds title/edition fixed effects. Perhaps not surprisingly, the inclusion of title fixed effects changes the coefficient estimates rather substantially. The third column adds fixed effects for the days until, and since, publication. The last two columns use data for all three Amazon platforms and include platform-specific edition fixed effects. In these specifications, the coefficient estimates fall somewhat in absolute value. Column (5) is our preferred specification, as it is the most conservative, accounting for country-specific edition effects as well as effects of time until and since publication.

3. Star Rating Effects Using Discontinuities

The way that Amazon reports its star ratings gives rise to an additional identification strategy for measuring the impact of star ratings on sales. The approach arises from the fact that Amazon reports its star ratings in two ways. 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 (2006). 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 models (3) and (5) above where we also include a series of dummies for each of the possible decimal star ratings along with the continuous measure $ln(R_{jt})$. Ninety percent of these ratings fall between 3.4 and 5, so we focus on this range. Figure 5 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. The coefficients are significantly negative at each of the discontinuities, whereas almost none of the others are significant. We take this as evidence that star ratings have a causal impact.

Because consumers can easily see both the decimal and half-star ratings, it is not reasonable to treat the discontinuity-based evidence as the entire impact of the star ratings. Still, we can derive a discontinuity-based estimate of the impact of stars as follows. We regress the log sales rank on its lag as well as the log price, log reviews, and the log number of ratings (along with various configurations of fixed effects). We also include an indicator for being at a decimal rating that is just above a half-star threshold (2.3, 2.8, etc). The coefficients on these indicators are consistently negative. For comparability we would like a coefficient that is an elasticity. We can accomplish that by interacting our "just-above" dummy with the log of the visible rank less the log of the next lowest visible rank. That is, when R=3.3, this variable is ln(3.5/3); and when R=3.8, this variable is ln(4/3.5). At non-threshold decimal star ratings, this this variable is zero. Using this approach, in Table 5, we obtain discontinuity coefficients between -0.07 and -0.08. The discontinuity-based estimate of the elasticity of the rank with respect to Amazon stars is very similar to the column (5) estimates in Table 4, lending additional credibility to the estimate from our preferred specification.

4. Unified Measurement

We would like to summarize the crowd and professional effect estimates parsimoniously, and we would also like to account for the various different review outlets simultaneously to avoid misattribution of the effect of one outlet to another. To that end we estimate models that include indicators for 0-5 days after the appearance of a review, 6-10 days after its appearance, and 11-20 days after appearance for the New York Times and indicators for 0-10 and 11-20 days for the other professional reviews. 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. Generically, we estimate models linking log sales ranks to their lag as well as three variable groups of interest: the log star rating (R), the log price (p), the number of reviews, and indicators for the professional reviews described above, along with various different configurations of fixed effects.

We begin, in the first two columns of Table 6, with models estimated on only US data, which limits the empirical strategies available for identifying the crowd effect. In particular, we can only make use of within-title variation for identification. The first column includes title fixed effects, and the second column also includes fixed effects for the time until and since publication. In the first five days after the appearance of a New York Times review, sales ranks improve by 22 to 24 percent, on average. In the next five days after a New York Times review, sales ranks are 8 to 13 percent better than prior to the review, and they are 6 to 9 percent than before the review in the next 10 days. Effects of other professional outlets are much smaller: 4-6 percent in the first 10 days, and there is no significant effect thereafter.

Columns (3) and (4) use data for all three platforms and platform-specific edition fixed effects. We allow the professional reviews to have different effects on the different domains, and

the data support this idea. While the effect of the NYT in the first five days is -24 to -26 percent in the US, it is about half as large in Canada, and small and insignificant in Great Britain. Similarly, the other professional outlets have effects in the first ten days but only in the US. Column (4) is analogous to column (5) of Table 4.

So far we have not distinguished among professional reviews according to how positive they are. 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, and review effects for both groups of books are positive, although the effects are larger for the recommended books.

5. Translating Ranks into Quantities

The evidence above indicates that reviews have an impact on sales ranks, but a few steps are required to translate coefficients from our models into elasticities of quantity with respect to, say, Amazon stars. Our Amazon quantity data are rank and not quantity data, and 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. We can summarize these data by a regression of log quantities on log ranks: $q_j = Ar_j^B e^{\varepsilon_j}$. We report this in Table 7.

Second, the rank data reflect not just the current sales quantity but its lag as well. That is, in simplified form:

$$\ln(r_{jt}) = \theta \ln(r_{j,t-1}) + a\ln(p_{jt}) + g\ln(R_{jt}) + \omega_{jt}.$$

This 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 price is $a/(1-\theta)$, while the derivative of the rank with respect to the rating is $g/(1-\theta)$.

Combining the above, the reduced form derivatives of quantity with respect to price and the rating are, respectively:

$$\varepsilon_p = \frac{\partial \ln(q_j)}{\partial \ln(p_j)} = B\, a/(1-\theta) \text{ and}$$

$$\varepsilon_R = \frac{\partial \ln(q_j)}{\partial \ln(R_j)} = B\, g/(1-\theta).$$

If h is a coefficient on a review dummy in a log rank regression, the effect of a review on the log sales quantity is analogously $B\, h/(1-\theta)$.

Table 8 reports estimates of quantity effects from the four specifications in Table 6. The second row reports elasticities of the quantity sold with respect to the Amazon star rating, and they vary between 0.15 and 0.46, or between 0.18 and 0.26 using the discontinuity approach. We obtain standard errors for these estimates by taking 500 parametric bootstrap draws from the estimated joint distributions of the parameters from Table 6, as well as from the independent distributions of B and σ from Table 7. 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 fourth column of the NYT 0-5 row indicates that estimated log sales rise by 0.58 during the 0-5 days after the appearance of a NYT review. Hence, our estimates indicates that sales increase by 78 percent during this period ($e^{0.5788} - 1 = 0.78$).

The first row of Table 8 reports price elasticities of demand, and they vary between -0.37 and -0.49. These are title-level elasticities which, on their face, appear to be rather inelastic. 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."[10] We find similarly inelastic estimates in Reimers 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 and crowd reviews is invariant to it.

Table 8 also report percentage impacts of reviews on annual simulated sales. To calculate effects on sales, we need to translate ranks into sales quantities (or a measure proportional to sales quantities). We do this using the Nielsen data on the sales of the top 100 editions by week. A regression of log sales quantities on log ranks yields a rank yields a coefficient of -0.54 (with a standard error of 0.004). See Table 7. Hence, we estimate daily sales quantities for an edition j as $q_{jt} = \frac{1}{\exp(\ln(rank_{jt})^{B})}$. We estimate the counterfactual quantity of sales absent star rating information by substituting the following for the ln rank: $\ln(rank_{jt}) - B\, g/(1-\theta)\,(R_{jt} - \hat{R}_t)$. We estimate the counterfactual sales absent professional reviews by substituting the following for the rank: $\ln(rank_{jt}) - B\, h_k/(1-\theta) * (review\ indicator\ k)_{jt}$. 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 3.92 percent during 2018. In addition to their intrinsic interest, the estimates in this table are also direct inputs into the calibration of our structural model below.

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

6. What Sorts of Products Are Promoted and Consumed?

In our estimates, professional reviews have only positive impacts on sales. Hence, to the extent that professional outlets' reviews are concentrated in certain genres, those reviews will steer consumption toward those genres. We explore this by comparing the genre distribution of sample books by whether they are professionally reviewed. Figure 6 reports the difference in these genre shares (for books reviewed in professional outlets vs those that are not), for the genres with substantial differences. For example, 12.5 percent of professionally reviewed books are in the political science genre, compared with 2.5 percent of other books, giving rise to a ten percent differential. Other genres more heavily reviewed by professional outlets include biography (8 percent differential), history, and social science (3.5 percent each). Other genres are represented more heavily among the works not professionally reviewed: juvenile fiction makes up 11.8 percent of the books not reviewed by professionals, compared with 6.1 percent of the works reviewed by professionals, for a -5.7 percent differential. Other genres underrepresented in the professional reviews include cooking (-5) and self-help, romance, and religion (all around -2 to -3 percent).

The impacts of Amazon stars are more complicated to quantify without an explicit model. Some books have higher star ratings than users might have expected while others have lower star ratings; and by construction these effects largely cancel. A zero overall effect on sales within a genre would not mean than ratings had not affected consumers. As our theoretical model highlights, the welfare benefit of ratings – the gain in consumer surplus – depends on the size of the difference between predicted and realized ratings. Hence the impact of ratings on consumer welfare will depend on the variance of the prediction errors.

We can therefore get a rough sense of which genres are most affected by comparing the standard errors of ratings across genres. Figure 7 makes this comparison, including only genres with at least 50 sample titles. For the figure, we calculate the standard error of the star ratings using only the final observed rating for each title, based on the largest accumulated number of reviews. Genres with the largest standard deviations are poetry (standard deviation = 0.62), comics & graphic novels (0.60), and women's fiction and literary criticism (both roughly 0.58) and medical (0.56). Genres with the smallest standard deviations are romance (0.34), cooking (0.35), and nature (0.38). The results in Figure 7 suggest that the benefits of stars operate disproportionately through the genres with large star rating variability.

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 benefit of, say, professional reviews versus the Amazon stars. While we document that reviews have only positive effects, Amazon stars can have positive or negative effects on sales. Hence an "effect on quantity sold" metric is inadequate. As our theoretical model suggests, however, the change in consumer surplus is a more natural basis for comparison. A structural demand model allows us to undertake this calculation; below, 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.

1. Preliminaries

In order to build a structural model of demand 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 sales per sample book, a nested logit substitution parameter σ, and measures of book quality, denominated in stars, that consumers would have expected absent the stars' existence. We discuss the derivation of these quantities before introducing the model.

First, for market size, we assume that each member of the US population is making a monthly decision of whether to purchase a books, so M = 12*327 million. Second, we observe that total US physical book sales were 695 million units in 2018.[11] Third, we determine the total 2018 sales of sample books as follows. We know the quantity of sales accounted for by the top 100 editions each week in the Nielsen data. These come to 9.84 percent of total physical sales. We can also create an estimate of how the total sample sales relate to the top 100 sample sales. Our sample includes the weekly top 150 according to USA Today, along with other titles. Thus, assuming that Nielsen and USA Today are similar, we can calculate the share of our sample's sales accounted for by the weekly top 100 in our sample. To this, we create a daily "sales" variable for each title as $q_j = 1/r_{jt}^B$, then aggregate the days (t) to weeks. We can then calculate that our weekly top 100 account for 35.57 percent of "sales" in our sample. This implies that our sample accounts for 192 million of the 695 million physical titles sold in 2018. In the structural model, then, we treat the 192 million sample units as the inside sales (Q). Because we know the total sample book sales, we can measure the 2018 sales of each edition (q_j) as well. We estimate the nested logit parameter σ to be 0.433 (with a standard error of 0.0475) using an empirical model we describe in the appendix.

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

We model consumers' beliefs about book quality in the absence of stars ($\hat{R}_j$) via a regression of log Amazon stars for books on the US platform on publisher fixed effects, genre fixed effects, and dummies for authors' prior experience. We use only one observation per edition, the final observation in the sample when the users have had the most time to leave ratings. The resulting regression explains 24.6 percent of the variation in log stars. We then treat the exponentiated fitted value as a measure of the quality of each book that consumers would have expected absent the star rating system. We explore the sensitivity of our results to the share of variation explain below.

2. A Simple Structural Model

To perform welfare analysis we calibrate a simple structural model of consumer utility to the estimated elasticities. In the nested logit model, we can transform quantities as follows:

$$\delta_j = \ln(s_j) - \sigma \ln(s_{j|g}) - \ln(s_0) = \ln(q_j/M) - \sigma \ln(q_j/Q) - \ln(1 - Q/M),$$

where $s_j = q_j/M$, $s_{j|g} = q_j/Q$, and $s_0 = 1 - Q/M$. Each product's share is then $s_j = \frac{e^{\delta_j/(1-\sigma)}}{1+\sum e^{\delta_j/(1-\sigma)}} \frac{D^{1-\sigma}}{1+D^{1-\sigma}}$, 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 $\delta_j = \delta_j^0 - \alpha p_j - \gamma R_j$, where α and γ are unknown utility function parameters related to estimated parameters *a* and *g*, respectively. We can then 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.

The nested logit gives a simple expression for the price elasticity of demand:

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

Given ε_p from the descriptive analysis above, σ (from the appendix), $s_{j|g}$, s_j, and p_j (data), this formula gives us a parameter estimate of α for each j, which we average for our estimate of utility function parameter α. Similarly, we can infer γ by solving the related elasticity

$$\varepsilon_R = \gamma_j \frac{R_j}{1-\sigma}\left(1 - \sigma s_{j|g} - (1-\sigma)s_j\right).$$

Here, again, we average the parameter estimates to obtain the utility function parameter γ.

We can solve for the utility function parameters associated with reviews in a related way. Reviews are binary rather than continuous, so we cannot use the derivative approach and an elasticity formula. Instead, we can set our empirical measurement of the effect of reviews on sales equal to their logit model analogs. That is, our descriptive analysis tells how each sales quantity q_j would have been different in the absence of reviews, q'_j. More precisely, the descriptive measure $\ln(q_j/q'_j)$ is the change the sales of each book in the presence of the reviews. By construction, $q'_j < q_j$, as removing sales-stimulating reviews simply reduces sales. The descriptive model includes no measure of the impact of reviews on unreviewed books, instead implicitly identifying 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 $\ln(q_j/q'_j)$ is the percentage change in sales for reviewed books, relative to the 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'^r_j as sales of reviewed books in the absence of reviews, and s'^u_j as sales of unreviewed books in the absence of reviews. Then the equation of the descriptive fact and its model analogue is

$\ln(q_j/q'_j) = \ln(s_j^r/s'^r_j) - \ln(s_j^u/s'^u_j)$ for all reviewed j.

A few lines of tedious algebra show that $\ln(q_j/q'_j) = \psi_j(1-\alpha)$. Given α, 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 perform the counterfactuals of interest. That is, 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 $\delta_j = \ln(s_j) - \sigma \ln(s_{j|g}) - \ln(s_0)$, while counterfactual utility absent Amazon stars is given by: $\delta_j^s = \delta_j - \gamma(R_j - \hat{R}_j)$; and counterfactual utility absent professional reviews is given by $\delta_j^p = \delta_j - \psi_j$.

Under the "persuasive" approach the effect of star ratings on CS is given by:

$$\Delta CS_{persuasive} = \frac{M}{\alpha}\left[ln\left(1+\left(\sum \exp\left(\frac{\delta_j}{1-\sigma}\right)\right)^{1-\sigma}\right) - ln\left(1+\left(\sum \exp\left(\frac{\delta_j^s}{1-\sigma}\right)\right)^{1-\sigma}\right)\right]$$

Under the purely informative approach, the change in CS associated with star ratings is given by

$$\Delta CS_{informative} = \frac{M}{\alpha}\left[ln\left(1+\left(\sum \exp\left(\frac{\delta_j}{1-\sigma}\right)\right)^{1-\sigma}\right) - ln\left(1+\left(\sum \exp\left(\frac{\delta_j^s}{1-\sigma}\right)\right)^{1-\sigma}\right) - \sum \gamma\left(R_j - \hat{R}_j\right)q_j'\right]$$, where the term $\sum \frac{\gamma}{\alpha}\left(R_j - \hat{R}_j\right)q_j'$ is the adjustment reflecting the possibility that what's 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.

Analogously, the respective changes in consumer surplus from the presence of professional reviews is given by:

$$\Delta CS_{persuasive} = \frac{M}{\alpha}\left[ln\left(1+\left(\sum \exp\left(\frac{\delta_j}{1-\sigma}\right)\right)^{1-\sigma}\right) - ln\left(1+\left(\sum \exp\left(\frac{\delta_j^p}{1-\sigma}\right)\right)^{1-\sigma}\right)\right]$$

and:

$$\Delta CS_{informative} = \frac{M}{\alpha}\left[ln\left(1+\left(\sum \exp\left(\frac{\delta_j}{1-\sigma}\right)\right)^{1-\sigma}\right) - ln\left(1+\left(\sum \exp\left(\frac{\delta_j^p}{1-\sigma}\right)\right)^{1-\sigma}\right) - \sum \psi_j q_j^p\right],$$

where q_j^p is the quantity of product j chosen in the absence of professional reviews.

Table 10 shows the welfare results. Sample books account for 192 million units sold and 3.06 billion in baseline revenue. Using our preferred specification (in column 4), the addition of professional reviews raises the quantity of books sold by 0.69% and raises revenue by $25.83 million. The addition of Amazon stars has smaller impacts on quantity (0.23%) and revenue $5.97 million.

The size of the welfare effects depend on which approach we use as well as how we normalize. In absolute dollar terms, the change in CS is much higher with the persuasive approach,

in which the effects of reviews and ratings on consumption are assumed only to operate if the reviews and ratings are present. That is, the reviews and ratings change consumers' preferences rather than simply informing consumers of product quality prior to purchase. In that case, reviews raise CS by $88 million, and when scaled to all Amazon books, the presence of Amazon stars raises CS by $58 million. These changes in consumer surplus are quite large in comparison with the changes in revenue, nearly five times higher for stars and over three times for reviews.

The absolute changes in CS under the informative approach are over an order of magnitude smaller. When compared against the changes in revenue, they are nontrivial however. For example, the change in CS with stars is 10 percent as large as the change in revenue brought about by the presence of stars. Similarly, the change in CS effected by reviews is 11 percent as large as the change in revenue from reviews.

A few points are in order. First, by a variety of measures, the changes in welfare stemming from Amazon stars are nearly as big as the changes stemming from professional reviews. Second, if one views ratings and reviews to operate through changing preferences, then their effects are substantial, on the order of 1 and 3 percent of revenue, respectively. Third, even if one views ratings and reviews as merely providing information that consumers would learn upon consumption, the changes in CS brought about by ratings and reviews are substantial in relation to the associated changes in revenue.

The welfare estimates in Table 10 are random variables, and we can create measures of precision for them. To this end we perform 500 parametric bootstrap replications of the regressions in Table 6. The standard errors are small.

3. Robustness

Our welfare analyses depend on estimated parameters. Here we explore the sensitivity of our basic results to different parameter values. 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. The impact of the substitution parameter σ on ΔCS is less obvious. We have experimented with values of σ between 0 and 0.9, and results change only minimally.

The measured welfare benefits of Amazon star ratings also depend on the accuracy of consumers' beliefs about product quality absent star ratings. We have modelled this using a regression of log stars on observables, and the regression explains 24.6 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: $\ln(R_j) + \kappa * \epsilon_j$, 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 regression R^2. Figure 8 shows three measures of CS from the presence of Amazon stars as a function of prediction accuracy, with a vertical line at our baseline persuasive estimate of $55 million. If consumers had no ability to predict quality – corresponding to an R^2 of 0 – then the welfare benefit would be about $58 million per year. 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 $38 million; and the change in CS would be about 5 percent of the change in revenue. If consumers could perfectly predict quality absent star ratings (R^2=100 percent), the star ratings would deliver no welfare benefit.

If one takes the view that reviews change preferences so that in the absence of the reviews, the consumers' ex ante and ex post utility would be the same, then the effects of review are different.

6. Do Professional Review Effects Wane 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. 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 NYT review dates for all books that made the NYT 100 Notable Books of the Year list, for all even years from 2004 to 2018. We manually search for these books' ISBNs in the NPD (formerly Nielsen) Bookscan database to collect weekly unit sales.[12] We then estimate regressions of the form

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

[12] We limit our analysis to the NYT 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.

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

Table 11 shows the results from our regressions. The first column considers all books that were reviewed after their original publication date. This constraint becomes stricter as we move to the right in the table: the second column additionally drops books reviewed at most one week after publication, the next column drops all books with reviews at most two weeks after publication, and so on. We find that NYT reviews had a positive effect on the rate of change in sales in all years, with positive coefficients that are almost always statistically significant. 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 NYT has increased over the last 15 years. This may in fact be due to digitization: it is possible that the NYT was able to utilize digitization to improve the reach of its book reviews. For example, the number of digital-only subscriptions to the NYT 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).[13]

[13] 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 NYT.

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 editions, along with information on review timing in professional review outlets, we document causal impacts on reviews on sales ranks. We then transform these estimates in 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 on book sales – NYT reviews raise sales by over 75 percent in the five days after a review and by 3.9 percent over the year. We also document that the causal elasticy of quantity sold with respect to Amazon stars is about 1/6. Because these two forms of pre-purchase information have causal impacts on buying behavior, they also affect welfare. If we take the view that reviews and ratings change preferences, then reviews raise CS by $88 million, while ratings raise CS by $55 million. If we view reviews and ratings as purely informative, then absolute effects are much smaller; but they each raise CS by about a tenth of their respective effects on revenue. Crowd based ratings made possible by digitization add substantially to the effects of traditional reviews on consumer welfare; at the same time, they do not diminish the impact of traditional reviews.

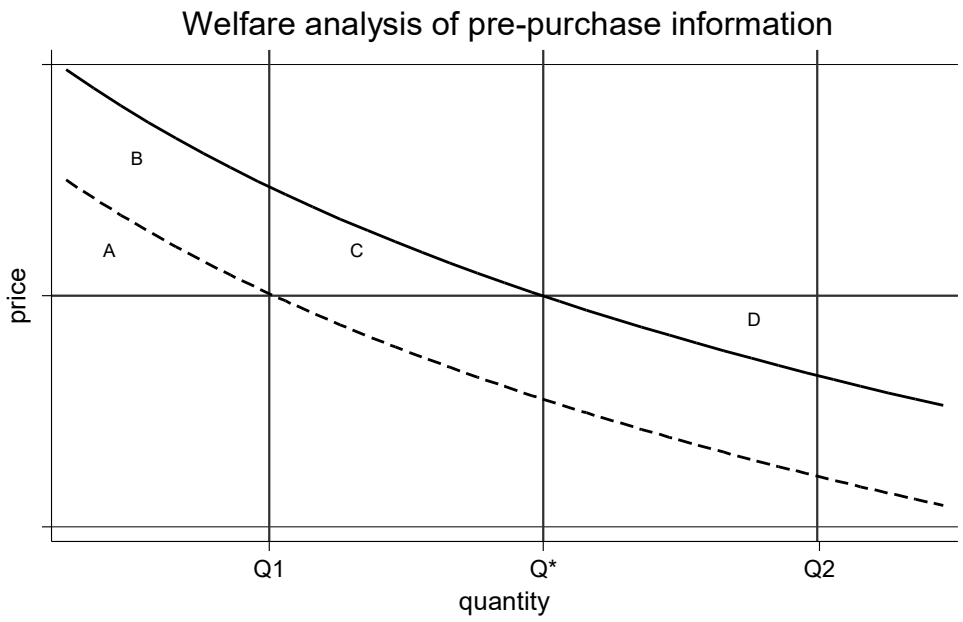


Figure 1: Illustration of the theoretical model

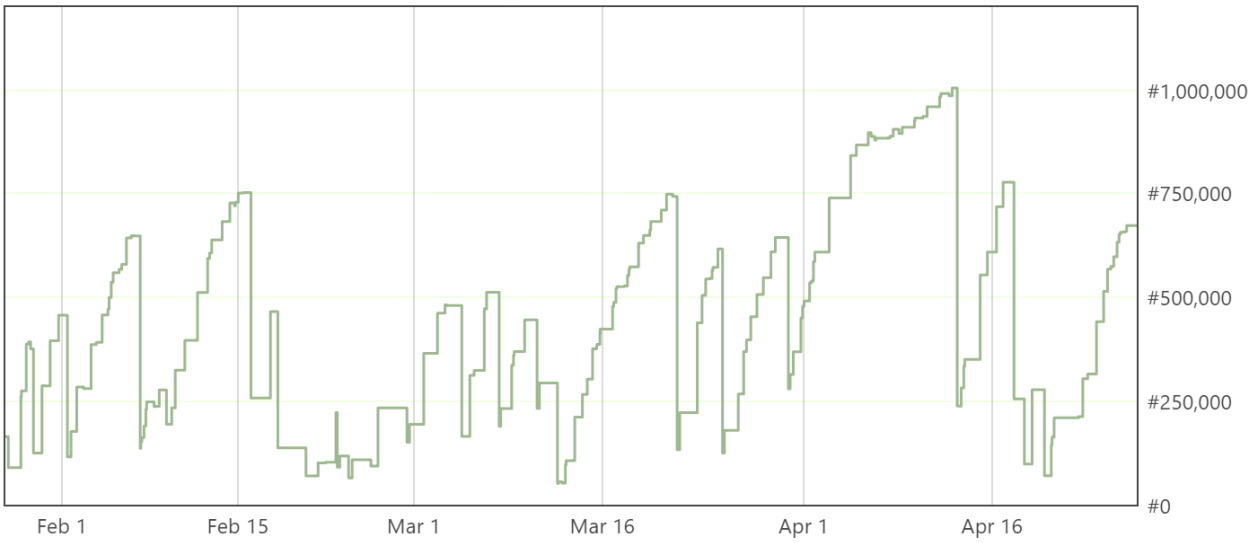


Figure 2: Amazon sales ranks have long memories

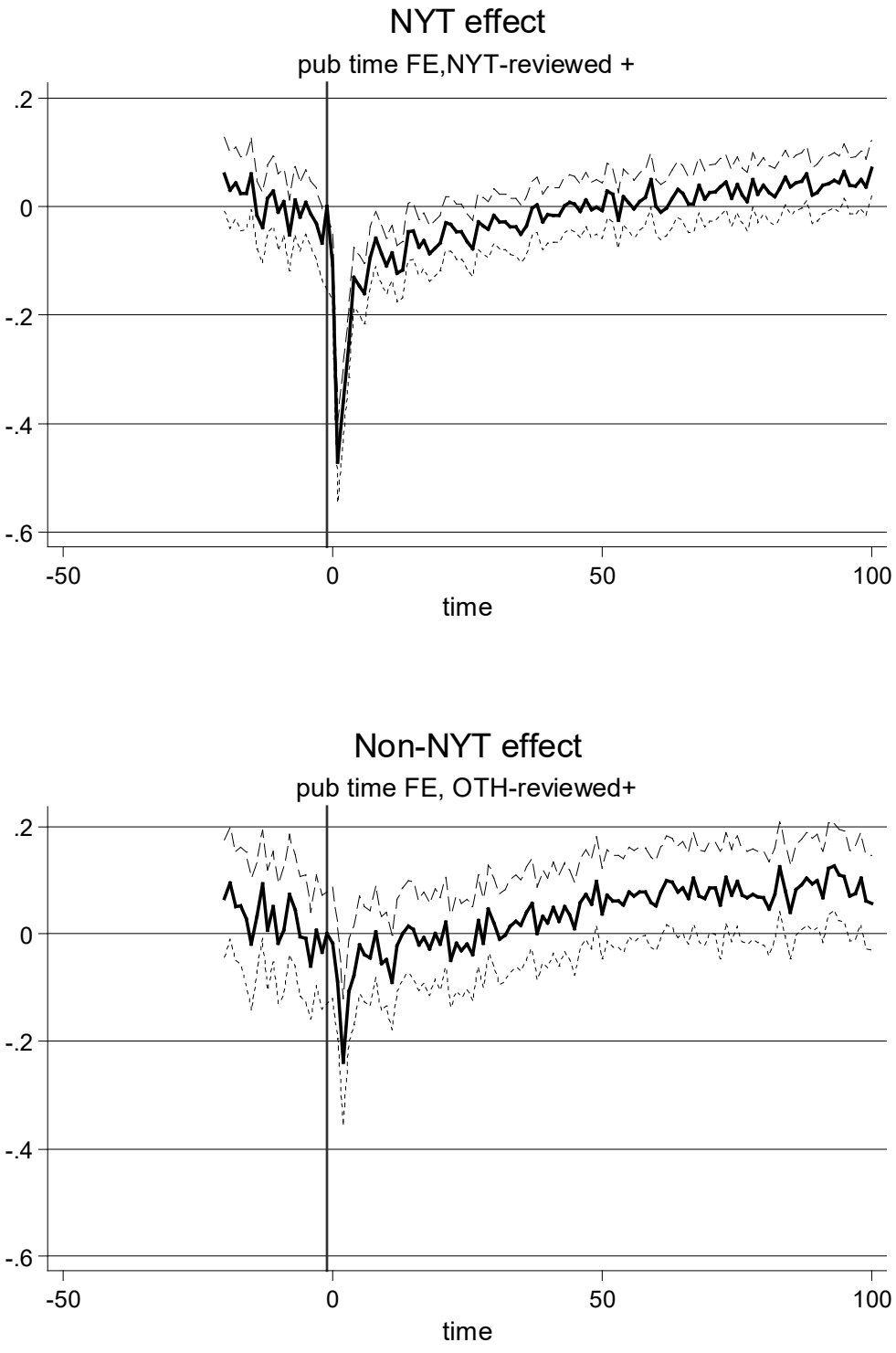


Figure 3: Effects of professional reviews on sales rank

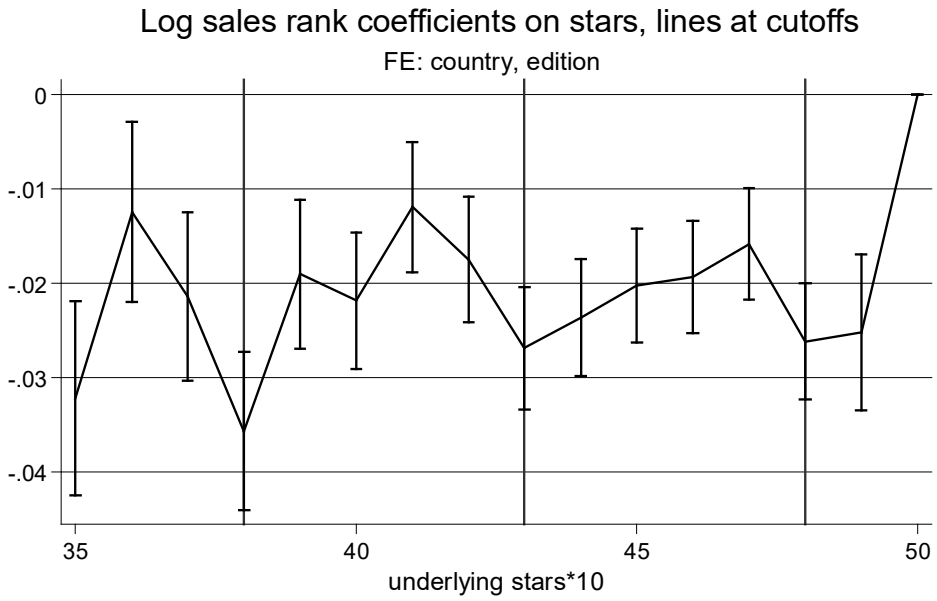


Figure 4: Effects of Amazon star ratings on sales rank

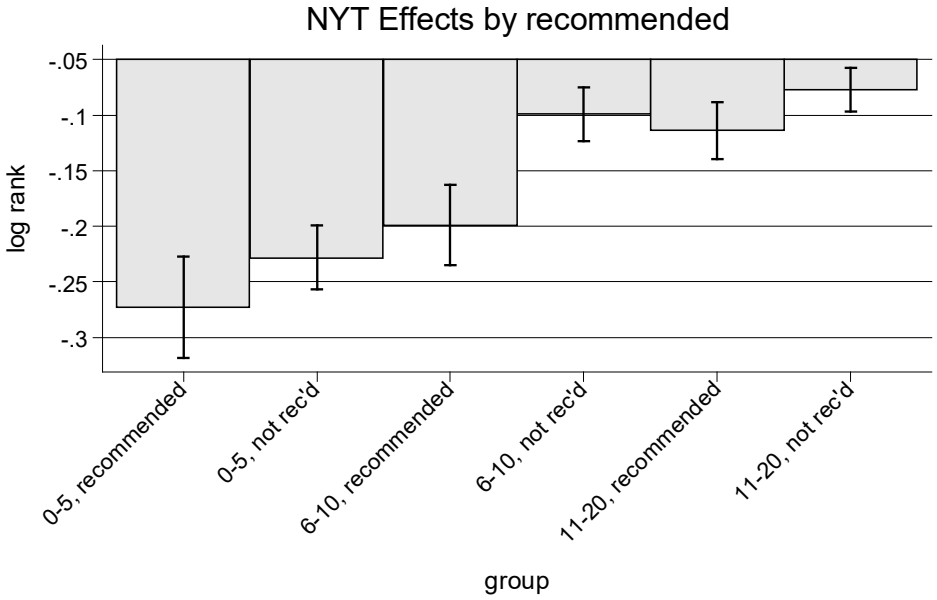
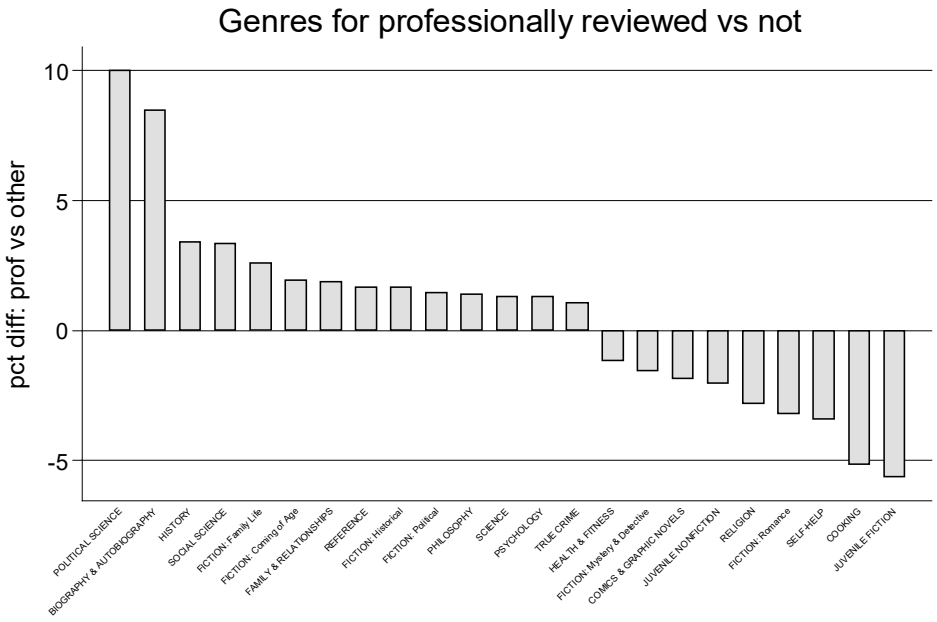


Figure 5: Effect of New York Times reviews by review quality



Note: 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 6: Composition of genres – reviewed vs. not

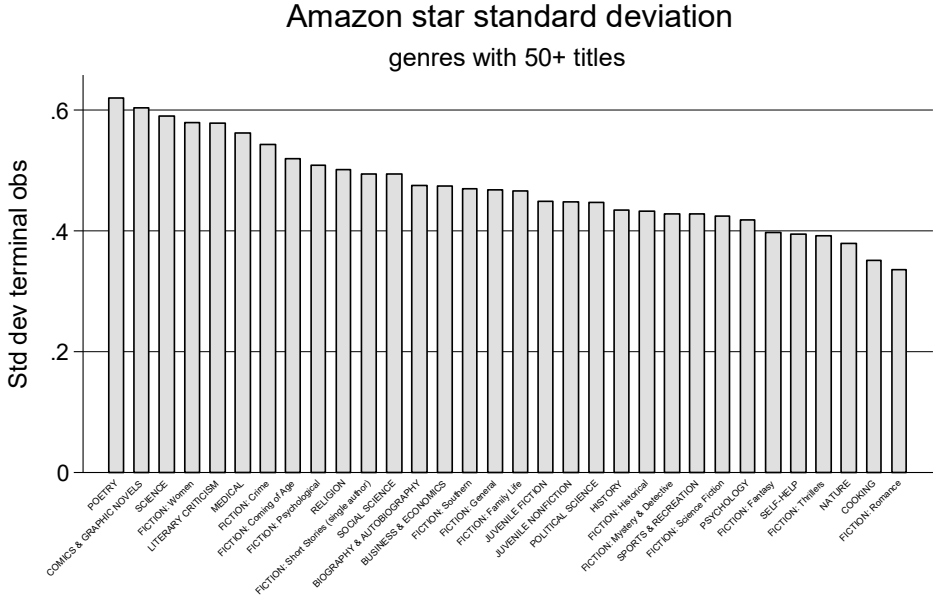
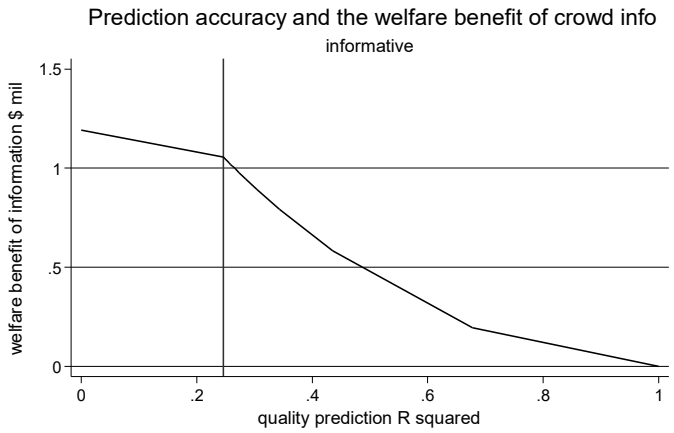
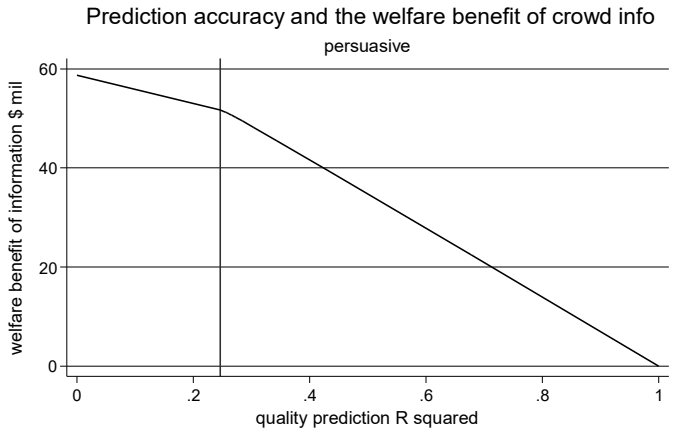


Figure 7: Standard deviation of Amazon stars by genre


Prediction accuracy and the welfare benefit of crowd info
informative
welfare benefit of information $ mil
1.5
1
.5
0
0
.2
.4
.6
.8
1
quality prediction R squared


Prediction accuracy and the welfare benefit of crowd info
persuasive
welfare benefit of information $ mil
60
40
20
0
0
.2
.4
.6
.8
1
quality prediction R squared

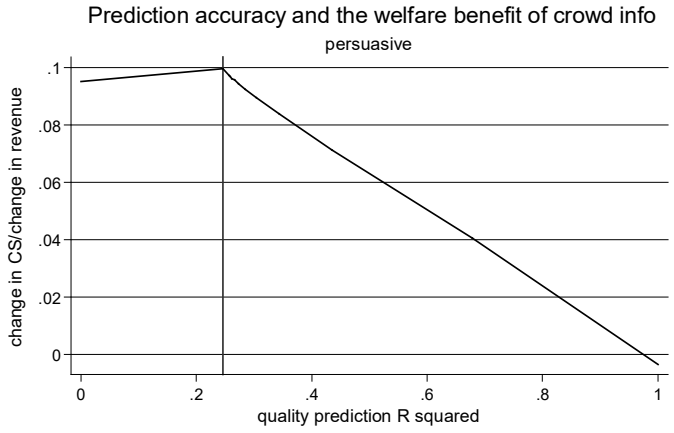

Prediction accuracy and the welfare benefit of crowd info
persuasive
change in CS/change in revenue
.1
.08
.06
.04
.02
0
0
.2
.4
.6
.8
1
quality prediction R squared

Figure 8

Table 1: Major Book Review Sources

Outlet	2018 hardback editions reviewed	Google Trends relative to NYT	Monthly site visits 12/18 (mil)
Kirkus Reviews	7,689	0.01	1.15
Booklist	7,611	0.01	NA
Publishers Weekly	5,603	0.01	2.15
School Library Journal	4,919	0.00	NA
Library Journal	3,692	0.00	0.9
New York Times Full Text Review	1,633	1.00	302.5
New York Times Book Review	183	0.01	NA
Chicago Tribune	248	0.21	19.9
Boston Globe	160	0.15	12.9
Wall Street Journal	159	0.44	67.6
Washington Post	156	1.06	190
USA Today	133	0.58	120.5
Los Angeles Times	106	0.32	35.2
San Francisco Chronicle	93	0.18	5.25

Notes: Number of editions reviewed in 2018 is drawn from Bowker's Books in Print. Bowker lists each edition as a separate entry. To reduce duplication, we restrict attention to hardback editions. The numbers here still include some title repetition but are useful for comparison across review sources. Site visitors are from Similarweb, based on December 2018 (https://www.similarweb.com). Monthly site visits to the New York Times Full Text Review are simply overall New York Times visits.

Table 2: Review overlap across major media outlets, 2018

("Of the titles reviewed by row outlet, how many were also reviewed by the column outlet?")

	Boston Globe	Chicago Tribune	LA Times	NY Times	SF Chronicle	WSJ	Wash Post
Boston Globe	**117**	19	16	71	14	6	17
Chicago Tribune	19	**201**	16	95	18	7	17
LA Times	16	16	**81**	43	11	6	10
New York Times	71	95	43	**1,325**	39	37	52
SF Chronicle	14	18	11	39	**78**	4	12
WSJ	6	7	6	37	4	**115**	6
Washington Post	17	17	10	52	12	6	**122**

Notes: Entries are unique title-author combinations among the hardback editions that Bowker indicates that each outlet reviewed during 2018. Main diagonal shows the unique titles reviewed in the outlet. Off-diagonal outlets show overlap. For example, the Boston Globe reviewed 117 titles; of these 117, the Chicago Tribune reviewed 19. The New York Times reviewed 1,325. Of these, 71 were also reviewed by the Boston Globe. Because the Boston Globe reviewed 117 titles, the NYT reviewed 61 percent (71/117) of those reviewed by the Globe. The outlets collectively reviewed 1,655 distinct title, of which the NYT reviewed 80 percent (1,325/1,655).

Table 3: Sample characteristics

	all	Canada	Great Britain	US	Professionally reviewed	USA Today	neither
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
star rating percentiles							
10^{th}	3.7	3.6	3.6	3.8	3.5	3.8	3.8
25^{th}	4.1	4.1	4	4.2	4	4.2	4.1
50^{th}	4.5	4.5	4.5	4.5	4.4	4.5	4.5
75^{th}	4.7	4.9	4.8	4.7	4.7	4.7	4.8
90^{th}	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

Table 4: Effects of star ratings and prices: various approaches to identification

	(1)	(2)	(3)	(4)	(5)
Log rank lagged	0.979	0.849	0.789	0.819	0.762
	(0.000)**	(0.001)**	(0.001)**	(0.001)**	(0.001)**
Log price	0.028	0.135	0.192	0.122	0.187
	(0.001)**	(0.004)**	(0.004)**	(0.003)**	(0.003)**
log reviews	-0.002	0.095	0.051	0.089	0.040
	(0.000)**	(0.001)**	(0.001)**	(0.001)**	(0.001)**
log star rating	-0.038	-0.138	-0.112	-0.100	-0.071
	(0.004)**	(0.012)**	(0.012)**	(0.009)**	(0.009)**
Fixed effects	none	title	Title, time since publication	Country x title	Country x title, time since publication
countries	US	US	US	all	All
R^2	0.96	0.97	0.97	0.96	0.96
N	1,021,953	1,021,765	1,021,765	1,612,014	1,612,014

* $p<0.05$; ** $p<0.01$

Notes: regression of Amazon log daily sales rank on its one-day lag, as well as the log price, log number of reviews, and the log of the star rating. 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. 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.

Table 5: Discontinuity evidence on Amazon stars

	log sales rank	log sales rank	log sales rank	log sales rank
Log rank lagged	0.849	0.789	0.819	0.762
	(0.001)**	(0.001)**	(0.001)**	(0.001)**
log Amazon price	0.135	0.192	0.122	0.188
	(0.004)**	(0.004)**	(0.003)**	(0.003)**
log reviews	0.095	0.052	0.089	0.040
	(0.001)**	(0.001)**	(0.001)**	(0.001)**
Log star rating	-0.139	-0.113	-0.100	-0.072
	(0.012)**	(0.012)**	(0.009)**	(0.009)**
just above x ln(vR/(vR-0.5))	-0.073	-0.082	-0.083	-0.078
	(0.014)**	(0.013)**	(0.013)**	(0.012)**
Fixed effects	title	Title, time since publication	Country x title	Country x title, time since publication
countries	US	US	all	All
R^2	0.97	0.97	0.96	0.96
N	1,021,765	1,021,765	1,612,014	1,612,014

* $p<0.05$; ** $p<0.01$

Table 6: Effects of crowd and professions reviews on book log sales ranks

	(1)	(2)	(3)	(4)
Sales rank lagged one day	0.846	0.786	0.815	0.760
	(0.001)**	(0.001)**	(0.001)**	(0.001)**
log Amazon price	0.137	0.193	0.125	0.188
	(0.004)**	(0.004)**	(0.003)**	(0.003)**
log reviews	0.091	0.049	0.085	0.038
	(0.001)**	(0.001)**	(0.001)**	(0.001)**
log star rating	-0.131	-0.107	-0.095	-0.068
	(0.012)**	(0.012)**	(0.009)**	(0.009)**
US: NYT, 0-5 days	-0.216	-0.240	-0.237	-0.257
	(0.013)**	(0.013)**	(0.013)**	(0.013)**
US: NYT, 6-10 days	-0.082	-0.129	-0.109	-0.152
	(0.011)**	(0.011)**	(0.011)**	(0.011)**
US: NYT, 11-20 days	-0.055	-0.088	-0.070	-0.101
	(0.009)**	(0.009)**	(0.010)**	(0.009)**
US: other, 1-10 days	-0.031	-0.030	-0.043	-0.037
	(0.018)	(0.017)	(0.018)*	(0.018)*
US: other, 11-20 days	0.004	0.001	-0.001	-0.001
	(0.017)	(0.017)	(0.017)	(0.017)
CA: NYT, 0-5 days			-0.112	-0.112
			(0.043)**	(0.042)**
GB: NYT, 0-5 days			-0.033	-0.030
			(0.026)	(0.026)
CA: NYT, 6-10 days			-0.046	-0.064
			(0.041)	(0.041)
GB: NYT, 6-10 days			-0.008	-0.000
			(0.024)	(0.024)
CA: NYT, 11-20 days			0.021	-0.010
			(0.035)	(0.035)
GB: NYT, 11-20 days			0.014	0.013
			(0.021)	(0.021)
CA: other, 1-10 days			0.016	0.041
			(0.072)	(0.071)
GB: other, 1-10 days			0.054	0.060
			(0.063)	(0.061)
CA: other, 11-20 days			0.064	0.071
			(0.069)	(0.068)

GB: other, 11-20 days			0.059	0.074
			(0.063)	(0.062)
FE	title	Title, time since publication	Country x title	Country x title, time since pub.
R^2	0.97	0.97	0.96	0.96
N	1,021,765	1,021,765	1,612,014	1,612,014

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. The first two columns include only data from Amazon's US site. Columns (3) and (4) include data from Amazon's US, Canadian, and Great Britain sites. * $p<0.05$; ** $p<0.01$

Table 7: Estimates using Nielsen data

	log Nielsen weekly sales	dv	
log Nielsen weekly rank	-0.540		
	(0.004)**		
Log of inside share $\ln(s_{j\|g})$		0.147	0.483
		(0.042)**	(0.046)**
R^2	0.75	0.37	
N	5,200	12,401	12,401

* $p<0.05$; ** $p<0.01$

Table 8: Quantity effects

	1	2	3	4
Price elasticity	-0.4804	-0.4882	-0.3655	-0.4247
std err	0.0172	0.0120	0.0102	0.0069
Amazon stars elasticity	0.4586	0.2697	0.2779	0.1533
std err	0.0519	0.0329	0.0289	0.0198
NYT 0-5	0.7562	0.606	0.6933	0.5788
std err	0.0527	0.0336	0.0418	0.0295
NYT 6-10	0.2858	0.3257	0.3202	0.3425
std err	0.0441	0.0307	0.0351	0.0235
NYT 11-20	0.1933	0.2229	0.2037	0.2278
std err	0.0368	0.0243	0.0297	0.0201
OTH 0-10	0.1072	0.0759	0.1243	0.0834
std err	0.0585	0.0459	0.0510	0.0416
OTH 11-20	-0.0149	-0.0017	0.0019	0.0026
std err	0.0368	0.0243	0.0297	0.0201
Amazon stars elas (discont.)	0.2616	0.2104	0.2485	0.1763
std err				
% effect of being reviewed				
Other only	0.6215	0.4931	0.819	0.564
std err				
NYT only	3.9891	3.9195	4.0071	3.9247
std err				
both	4.795	4.5949	4.9959	4.6633
std err				

Notes: price and Amazon star rows show estimated elasticities of quantity sold with respect to price and Amazon stars, respectively. NYT and OTH rows show percentage impacts on reviews on sales during the relevant numbers of days after the reviews. The bottom three rows show 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 6, as well as from the independent distributions of B and σ from Table 7.

Table 9: Model inputs

variable	value
2018 US unit sales (mil)[14]	695
share of US sales in Nielsen top 100	0.0984
top 100 share of sample sales	0.3557
US unit sales of sample titles (mil)[15]	192.28
Amazon unit sales	
share of physical sales (2017)[16]	0.455
all titles (mil)	316.2
sample titles (mil)	81.6
US pop 2018 (mil)	327.2
market size (12*pop)	3,926.4

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

[15] =(1/.3557)*0.0984*695.

[16] https://www.idealog.com/blog/changing-book-business-seems-flowing-downhill-amazon/

Table 10: Welfare impacts of professional reviews and Amazon star ratings

	informative				persuasive			
	1	2	3	4	1	2	3	4
ΔCS – stars	4.76	1.61	2.28	0.59	73.84	43.88	60.32	29.12
ΔCS - reviews	2.67	2.45	3.62	2.84	79.26	75.99	107.14	88.24
Δ CS / rev - stars	0.16%	0.05%	0.07%	0.02%	2.42%	1.43%	1.97%	0.95%
Δ CS / rev - reviews	0.09%	0.08%	0.12%	0.09%	2.59%	2.49%	3.50%	2.89%
Δ CS – stars (scaled)	8.44	2.86	4.05	1.06	131.01	77.84	107.01	51.66
Δ CS – reviews (scaled)	2.67	2.45	3.62	2.84	79.26	75.99	107.14	88.24
ΔCS/ΔRev - stars	27.98%	15.61%	21.51%	9.96%	434.1%	425.3%	568.2%	487.6%
ΔCS/ΔRev - reviews	10.13%	9.59%	13.40%	11.01%	301.3%	297.1%	396.3%	341.7%
% ΔQ - stars	0.65%	0.39%	0.40%	0.23%				
% ΔQ - reviews	0.70%	0.68%	0.72%	0.69%				
Δ Rev - stars	17.01	10.32	10.62	5.97				
Δ Rev - reviews	26.31	25.57	27.03	25.83				

Notes: all dollar figures in millions. Baseline revenue is $3,057.74 million. "Informative" calculations assume that consumers perceive true quality upon consumption absent review information. "Persuasive" calculations assume that ratings and reviews change consumers' valuations. The absolute 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. Columns (1) – (4) contain estimates corresponding to the estimates in analogous columns of Table 6.

Table 11: Impact of New York Times reviews on log-sales change

Review at least x weeks after pub:	0 weeks	1 weeks	2 weeks	3 weeks	4 weeks
2004 review	0.219	0.209	0.274	0.459	0.436
	(0.061)***	(0.064)***	(0.087)***	(0.089)***	(0.094)***
2006 review	0.121	0.125	0.115	0.155	0.163
	(0.053)**	(0.063)**	(0.075)	(0.095)	(0.117)
2008 review	0.264	0.278	0.295	0.268	0.270
	(0.052)***	(0.062)***	(0.079)***	(0.091)***	(0.120)**
2010 review	0.213	0.202	0.232	0.127	-0.002
	(0.048)***	(0.073)***	(0.092)**	(0.191)	(0.213)
2012 review	0.201	0.251	0.244	0.313	0.288
	(0.044)***	(0.047)***	(0.050)***	(0.062)***	(0.078)***
2014 review	0.088	0.196	0.212	0.297	0.334
	(0.069)	(0.080)**	(0.091)**	(0.106)***	(0.121)***
2016 review	0.168	0.199	0.358	0.295	0.220
	(0.058)***	(0.076)***	(0.085)***	(0.084)***	(0.108)**
2018 review	0.514	0.577	0.612	0.665	0.731
	(0.097)***	(0.109)***	(0.127)***	(0.191)***	(0.204)***
After publication	-2.021	-2.046	-2.026	-2.108	-2.232
	(0.050)***	(0.063)***	(0.078)***	(0.113)***	(0.152)***
Book age (weeks)	0.000	0.000	0.000	0.000	0.000
	(0.000)***	(0.000)**	(0.000)*	(0.000)*	(0.000)
Constant	1.921	1.958	1.945	2.031	2.160
	(0.048)***	(0.060)***	(0.075)***	(0.109)***	(0.148)***
R^2	0.45	0.45	0.45	0.46	0.47
N	8,880	6,925	5,440	3,833	2,908

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

References

Aguiar, L. and Waldfogel, J., 2018. Quality predictability and the welfare benefits from new products: Evidence from the digitization of recorded music. *Journal of Political Economy*, *126*(2), pp.492-524.

Allcott, H., 2011. Consumers' perceptions and misperceptions of energy costs. *American Economic Review*, *101*(3), pp.98-104.

Arellano, M. and Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The review of economic studies*, *58*(2), pp.277-297.

Belloni, A., Chernozhukov, V. and Hansen, C., 2014. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), pp.29-50.

Berger, J., Sorensen, A.T. and Rasmussen, S.J., 2010. Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*, *29*(5), pp.815-827.

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

Brynjolfsson, E., Hu, Y. and Smith, M.D., 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, *49*(11), pp.1580-1596.

Chevalier, J. A. and Goolsbee, A., 2003. Measuring prices and price competition online: Amazon. com and BarnesandNoble. com. *Quantitative marketing and Economics*, *1*(2), pp.203-222.

Chevalier, J.A. and Mayzlin, D., 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, *43*(3), pp.345-354.

Dargis, Manohla. 2014. "As Indies Explode, an Appeal for Sanity: Flooding Theaters Isn't Good for Filmmakers or Filmgoers." *New York Times*, January 9.

Deutschman, Alan. 2004. "The Kingmaker." Wired Magazine, May 1. https://www.wired.com/2004/05/mossberg/?pg=1

Digital Reader Staff. 2016. "Goodreads Reaches New Milestone: Fifty Million Reviews." Digital Reader, April 7. https://the-digital-reader.com/2016/04/07/goodreads-reaches-new-milestone-fifty-million-reviews/

Duan, W., Gu, B. and Whinston, A.B., 2008. Do online reviews matter?—An empirical investigation of panel data. *Decision support systems*, *45*(4), pp.1007-1016.

Eliashberg, J. and Shugan, S.M., 1997. Film critics: Influencers or predictors?. *Journal of marketing*, *61*(2), pp.68-78.

Forman, C., Ghose, A. and Wiesenfeld, B., 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information systems research*, *19*(3), pp.291-313.

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

Greenfield, Jeremy. 2012. "Seven Advantages Barnes & Noble Has in the Bookseller Wars." *Digital Book World*, January 3. http://www.digitalbookworld.com/2012/seven-advantages-barnes-noble-has-in-the-bookseller-wars/.

Helmers, Christian, Pramila Krishnan and Manasa Patnam. 2015 "Attention and Saliency on the Internet: Evidence from an Online Recommendation System, CEPR Discussion Paper No. DP10939, November.

Jin, G.Z. and Sorensen, A.T., 2006. Information and consumer choice: the value of publicized health plan ratings. *Journal of health economics*, *25*(2), pp.f248-275.

Luca, M., 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." Atlantic, Sept 14. https://www.theatlantic.com/national/archive/2011/09/sam-sifton-departure-and-the-end-of-the-career-food-critic/337844/.

McG. Thomas, Robert. 1999. "Gene Siskel, Half of a Famed Movie-Review Team, Dies at 53." New York Times, Feb. 21. https://www.nytimes.com/1999/02/21/nyregion/gene-siskel-half-of-a-famed-movie-review-team-dies-at-53.html

Narula, Svati Kirsten. 2014. "Millions of People Reading Alone, Together: The Rise of Goodreads." Atlantic Monthly. February 12. https://www.theatlantic.com/entertainment/archive/2014/02/millions-of-people-reading-alone-together-the-rise-of-goodreads/283662/.

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, July 27. HTTPS://WWW.VANITYFAIR.COM/NEWS/2017/07/MICHIKO-KAKUTANI-LEAVING-THE-NEW-YORK-TIMES

Reinstein, D.A. and Snyder, C.M., 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), pp.27-51.

Richter, F. (2018). The "Failing" NY Times Passes 2.5 Million Digital Subscriptions. Statista. Statista Inc.. Accessed: August 26, 2019. https://www.statista.com/chart/3755/digital-subscribers-of-the-new-york-times/

Senecal, S. and Nantel, J., 2004. The influence of online product recommendations on consumers' online choices. *Journal of retailing*, *80*(2), pp.159-169.

Sorensen, A.T., 2007. Bestseller lists and product variety. *The journal of industrial economics*, *55*(4), pp.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*, pp.15-22.

Waldfogel, J. and Reimers, I., 2015. Storming the gatekeepers: Digital disintermediation in the market for books. *Information economics and policy*, *31*, pp.47-58.

Waldfogel, J., 2017. How digitization has created a golden age of music, movies, books, and television. *Journal of economic perspectives*, *31*(3), pp.195-214.

Appendix: 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 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:

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

As in Berry (1994), we seek to obtain 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 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, Hansen, and Chernozukov (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 0.433 (with a standard error of 0.0475). We report these estimates in Table 7.