The Welfare Effects of Gender-Inclusive Intellectual Property Creation: Evidence from Books*

Joel Waldfogel[†1]

[1]Carlson School of Management, Department of Economics,
University of Minnesota, NBER, and ZEW.

September 10, 2023

PRELIMINARY DRAFT

Abstract

Women have traditionally participated in intellectual property creation at depressed rates relative to men. Book authorship is now an exception. In 1970, women published a third as many books as men. By 2020, women wrote the majority of new books. Adding new products can have significant welfare benefits, particularly when product quality is unpredictable. Using data on sales of over 8 million individual titles at Amazon, 2018-2021, along with information on 200 million ratings of 1.8 million books by 800,000 Goodreads users, I develop measures of both the supply of new books by male and female authors, as well as their usage by heterogeneous consumers. I show that growth in female-authored books has delivered substantial increases in the female-authored shares of consumption, book awards, and other measures of success, indicating both that the additional female-authored books are useful to consumers and that product quality is unpredictable. I calibrate a simple structural model of demand and product entry with unpredictable quality to quantify the welfare benefit from the additional female-authored books. While revenue gains to female authors come partly at the expense of male authors, the gains from inclusive innovation accrue to a wide range of consumers.

*This research was initiated while I was the Kaminstein Scholar at the US Copyright Office. I thank Office staff for help in getting access to the copyright registration data. I am grateful to Chris Buccafusco, Rem Koning, and Margaret Kyle for discussant comments and to seminar participants at the University of Pennsylvania, the University of Minnesota, the University of Toronto, Washington University, the Paris Digital Conference, the Munich Summer Institute, and the NBER Summer Institute for helpful comments.

[†]E-mail: jwaldfog@umn.edu

1 Introduction

In many creative or innovative areas, women participate at depressed levels relative to men. Relative to white men, women, Blacks, and Hispanics are under-represented among inventors listed on patents; and women account for relatively few movie directors, to cite just a few examples.[1] This is potentially costly, as a growing body of evidence indicates that more inclusive involvement in product and intellectual property (IP) creation could deliver more valuable inventions and greater economic well-being (Bell et al., 2019; Hsieh et al., 2019; Cook, 2011).

Until recently, women had been largely absent from book authorship.[2] As Figure 1 shows, just 10 percent of the 19^{th} century books in the Library of Congress (LOC) had authors with female first names, and the female-authored share reached only 18 percent by 1960. But for books published after 1960, growth in female authorship accelerated sharply, reaching nearly 40 percent by 2010 (in the LOC) and over 50 percent for new US book copyright registrations a few years later.[3] In half a century, women went from producing one book for every three produced by men to output parity: Women have tripled their creative output relative to men, so that recent vintages may be 50 percent larger than they would have been, absent the growth in female authorship.[4]

Broader inclusion in innovative activity has many possible effects, including redistributing income among potential creators. Because most people are consumers rather than producers, however, the impact of inclusive creation on consumers may be more important than its impact on producers. Yet, the vast majority of books attract little use, however, so it is far from obvious that even a large growth in the number of books in the market would have much effect on either buyers or other sellers.

The welfare effect of an influx of new products depends on the quality of the additional

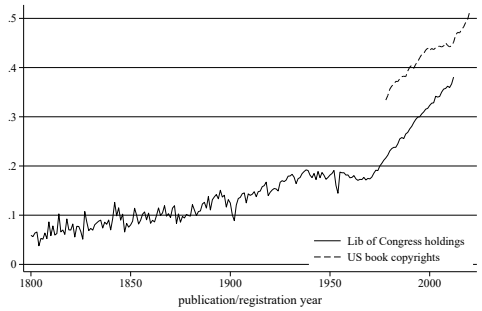
[1] See, for example, (Hunt et al., 2013; Frietsch et al., 2009).

[2] Notable exceptions include Jane Austen, Virginia Woolf, the Bronte sisters, and more.

[3] These data are described below at Section 4.

[4] This calculation presumes that fewer female-authored books would not beget more male-authored entry, an assumption explored empirically below.

Figure 1: Female-authored share of books published



Notes: Share of books in the Library of Congress and among US book copyright registrations whose authors' first names are female.

products which, in turn, can hinge on the predictability of new product quality. Falling entry costs facilitate entry of products with lower expected revenue. If the quality of new products were perfectly predictable before launch, then an expansion in the number of products would bring only products less valuable than the least appealing pre-expansion product. But the value of innovations tends to be highly unpredictable prior to launch, and this may be particularly true in creative contexts (Caves, 2000).[5] As a result, the release of more products is akin to taking more 'draws from the urn'; and it can deliver valuable additional products. The growing participation of women in book authorship has facilitated substantially more draws, which one might reasonably expect to be valuable.[6]

Hence, the book market should provide a useful "experiment" for documenting welfare effects of more inclusive IP creation. The context also has the advantage that, unlike with patents, IP can be readily linked to associated books with observable usage, allowing measurement of effects of more inclusive IP creation on consumption, revenue, and welfare. This leads to five empirical questions. First, how large is the female influx, and does it occur across book genres? Second, are the additional female-authored books valuable to consumers? That is, as as the share of female-authored products has grown, have the shares of consumption and recognition garnered by female-authored books grown similarly? Third, how predictable is

[5]In a famous description of Hollywood movie making decisions, William Goldman declared that "nobody knows anything" (Goldman, 2012).

[6]This approach parallels a tradition of viewing entrepreneurship as experimentation in, for example, Arrow (1969), Weitzman (1979), Bergemann and Hege (2005), Manso (2011), Manso (2016), and Kerr et al. (2014).

product quality at entry? Fourth, does the influx of female-authored books deliver welfare benefits that would not have ensued from male-authored books that the female influx might displace? Finally, how large are welfare effects on consumers and producers from inclusive IP creation, and how do they vary across types of authors and consumers?

To study these questions, one needs not only data on the number of new works created, by author gender over time, but also data on consumption of each work. Moreover, drawing inferences about the possibly heterogeneous welfare benefit of new products requires usage data for different types of consumers. I have two data sources – each with distinct strengths – which allow me to document both supply and demand. First, I have Bookstat data on over 8 million distinct ebook and print editions (those appearing in at least one daily top half million) sold at Amazon between 2018 and 2021, along with each edition's annual Amazon sales over this time period. I also observe author name, book genre, and publication date. Second, I have information on nearly 230 million user interactions with 1.8 million books by over 800,000 individual consumers at Goodreads between 2007 and 2016 (Wan and McAuley, 2018). On the supply side, I use both the Bookstat and Goodreads data to create time series on the numbers of books published annually back to 1960, by author gender and genre. On the demand side, the Bookstat data allow me to directly observe sales of books by author gender, book genre, and calendar year for years of different original publication vintages. The individual-level user-book interaction measures in Goodreads (based on the number of persons rating or "shelving" each book) allow me to create usage measures for different groups of consumers. Although I do not observe consumer gender in the Goodreads data, I can classify consumers according to the sorts of books they use, for example according to that share of the books they use that are written by women, or their use of books in particular genres. This allows me to draw inferences about effects of the female influx not only on overall consumer welfare but also on different kinds of consumers whose preferences differ by product type.

The paper proceeds in seven sections after the introduction. Section 2 provides background on womens' growing participation in the economy during the 20^{th} century as well as a discussion of the literature on female participation in innovative and creative activity in

particular. Section 3 sketches an equilibrium model of consumer demand and entry with quality unpredictability. I use the model, which builds on Aguiar and Waldfogel (2018), to generate descriptive questions I explore in Section 5 and to guide the structural measurement of equilibrium welfare effects of the female influx in Section 6. Section 4 describes the data sources used in the study, including Bookstat, Goodreads, as well as information on the Library of Congress holdings, US copyright registrations, New York Times fiction bestseller lists, and nominations for National Book Awards and Pulitzer Prizes. Section 5 addresses the descriptive questions derived from the model, the female authorship influx and its impact on the usage and success attained by female-authored work, the predictability of product success at entry, and the possible displacement of male entry by female entry. I then turn, in Section 6, to explicit quantification of the welfare effects. I present a calibrated nested logit model of demand, in which revenue and consumer surplus (CS) depend on the distribution of product qualities and the degree of substitutability across books. I implement a model of equilibrium entry with imperfect quality predictability for comparing the status quo choice set with a counterfactual environment without the female authorship influx. I also explore the sensitivity of the welfare estimates to alternative parameter values.

I have five descriptive findings, along with welfare estimates. First, the female influx is both large and widespread, occurring in all genres. Second, the female influx delivered valuable new products: As the share of new books by women rose in all genres, the share of usage – and awards – garnered by female-authored works has grown nearly proportionally. Third, these consumption effects of the female influx are present for a wide variety of consumers, according to their tastes in books. Fourth, the success of new products is largely, but not completely, unpredictable at entry. Available observables explain perhaps a third of ex post product success. Fifth, I find no direct evidence that that the growth in female-authored books displaced male-authored entry. Finally, using the equilibrium model to compare the status quo (including the female influx) to a counterfactual environment without the female influx – and allowing for endogenous male entry response – the influx raises the sales of female authors and depresses the sales of male authors. Perhaps more important, consumer surplus rises not overall but also for both heavy and light users of the various book genres.

2 Background

2.1 The role of women in IP creation

As Goldin (2006) and Costa (2000) document, the 20^{th} century brought a revolution in women's participation in the US economy. While women's labor force participation was under 20 percent in 1900, it reached nearly 80 percent by 2000. Various technological developments, including home appliances and birth control have facilitated female economic activity (see Greenwood et al. 2005 and Bailey 2006). The greater participation of women (and others) has, in turn, contributed substantially to economic growth (Hsieh et al., 2019).

Despite growing female participation in the labor force, female participation in creative and innovative activity is depressed relative to male participation. In science, technology, engineering, and math (STEM)-related areas, the differential is particularly large, as women account for 10-15 percent of the inventors on patents.[7] The role of women in the copyright-protected creative industries has received less attention from researchers, although creative community participants have raised concerns about bias against women, for example in the music and movie industries.[8] Brauneis and Oliar (2018) document that the female-authored share of US copyright authors rose from 30 to 36 percent between 1978 and 2012.[9]

The inclusiveness of innovation is a topic whose urgency has grown with findings that environmental factors affect the tendency for people to engage in innovation, suggesting that more inclusive participation in innovation would deliver additional valuable inventions (Bell et al., 2019; Cook, 2011). Moreover, there is reason for concern that the absence of women in

[7] Hunt et al. (2013) finds that 7.5 percent of patents are granted to women and that much of the gender gap is attributable to lower propensity to patent among "holders of a science or engineering degree." See also Ding et al. (2006) and Frietsch et al. (2009). More recently, (Toole et al., 2021) finds that between 2016 and 2019, the US "women inventor rate grew from 12.1% in 2016 to 12.8%. See also Martínez et al. (2016). Kim and Moser (2020) explore the role of child-bearing in the productivity of women scientists relative to their male peers. Hoisl et al. (2023) explore the role of patental influence on children's tendency to invent. Koffi and Marx (2023) explore the role of differential commercialization by gender.

[8] For example, the Annenberg Inclusion Initiative has highlighted shares of women among people producing music.

[9] Other recent work examines possible gender bias in the promotion of music. Aguiar et al. (2021) measure Spotify's potential bias by label status and whether artists are women, finding that Spotify's New Music lists rank songs in ways that incorporate bias in favor of independent-label artists and, to a lesser extent, women artists.

innovation may affect not only the value but also the direction of innovative activity (Koning et al., 2021). While female participation tends to fall short of male participation in many creative areas, books now stand out for gender-inclusive creation. For most of the past decade, women have authored more than half of new books according to US copyright registrations. More than in most creative or innovative IP contexts, the book market provides a useful test case for measuring the welfare impact of more inclusive IP creation.

3 Theory

This section outlines a theory of entry with unpredictable product quality, which guides the paper's analyses of the female authorship influx empirical exercises in two ways. First, I derive testable implications of the influx for usage and success outcomes, which I investigate in Section 5. Second, the setup here guides the structural model I implement in Section 6 for measuring the equilibrium welfare impact of the influx.

3.1 Model and empirical questions

Authors release books if the books' expected revenues exceed the cost of entry, T. Because consumers can choose among potentially substitute products, revenue for a particular product depends on its own quality as well as the quality of other products in the market. It is natural to view the female influx as a response to a reduction in the female entry threshold. The effect of a female influx on the demand for the new products – and the value that consumers derive from the augmented choice set – depends on both the realized quality of the additional products and whether the additional products displace entry that would otherwise have occurred.

Whether the additional products are valuable, in turn, depends on the predictability of product quality at entry. To see this, suppose quality is completely predictable. Then, if the entry threshold falls from T to T', all of the additional female products would have lower quality than the former threshold; and their values would fall between T and T'.

Even a large influx of female products would have a small effect on welfare and the share of sales garnered by female products. Hence, the change in share of sales attracted by female-authored products (s^f) with the change in the share of books by female authors (n^f) would be modest. That is, we would expect the derivative of s^f with respect to n^f ($\frac{\partial s^f}{\partial n^f}$) to be positive but small.

At another extreme, suppose quality were completely unpredictable at entry. Then, while the new products entering would be as valuable, on average, as existing female products. Each product would face more competition, so all products would have lower revenue than they would have faced in a less competitive environment. But additional female products would raise the female share of sales proportionally. And if male and female products were drawn from the same quality distribution, we would expect a large effect of additional female entry on the share of usage attracted by female-authored books, or $\frac{\partial s^f}{\partial n^f} = 1$.

The size of the equilibrium effect of the female influx on consumer welfare also depends on whether the choice set is changed by the influx. It is possible, in principle, that the female influx simply displaces male-authored entry that would otherwise have occurred. On the other hand, the influx deliver might products that would not have arisen in its absence.

This discussion leads to three important descriptive empirical questions. First, this discussion places attention on $\frac{\partial s^f}{\partial n^f}$, which is an important object of study for us. Much of Section 5 is aimed at measuring the impact of n^f on s^f and whether it is close to zero (as would arise with high predictablity), or to one (as with low predictability). Second, we also study predictability directly, asking how much of the realized success of books can be explained with observable characteristics known at the time of entry. Third, we explore the possible displacement of male-authored entry by female entry.

3.2 Model and welfare analysis

This framework also leads to an approach to equilibrium welfare analysis, which I undertake explicitly in Section 6. I offer a sketch of the approach here. The set of entering products has realized qualities that give rise to consumption levels for each product, as well as consumer

surplus associated with the available choice set. Each potentially entering product has an expected quality, which will in general deviate from realized quality. Products enter if their expected revenue exceeds cost. Equilibrium obtains when all products with expected revenue above the threshold have entered, and no additional products could profitably expect to enter.

For welfare analysis, I compare a status quo choice set – which contains the female influx – to a counterfactual choice set that excludes the female influx. I remove female products from each vintage so that the ratio of female to male products stands at its level for the 1960s. Implementing this requires answers to two questions. First, for the no-female-influx counterfactual, which female products do I remove from the status quo choice set? The choice of which products to remove depends on quality predictability at entry: If quality were completely unpredictable, I could simulate the no-female-influx choice set by removing female products at random. If quality is at least somewhat predictable, I can remove the products of lowest expected quality. Second, with the removal of the female influx, the return to male entry will rise; and how do I endogenously add male products to equilibrate? This process, too, depends on predictability. If quality were completely unpredictable, I could simulate additional male products by drawing from the distribution of existing products. If quality were somewhat predictable, then expected quality would fall in entry order (based on expected quality). I use this idea to predict expected quality for male product entry beyond the products observed in the status quo.

4 Data

This section describes the various data sources used in the study. Section 4.1 describes the name data used to infer author gender. Section 4.2 describes the Bookstat data on books published and their Amazon sales, 2018-2021. Section 4.3 describes the Goodreads book usage data and provides a comparison of usage patterns in these data with patterns in other sources of information on top-selling books. Section 4.4.1 describes the other data sources used in the study, including both additional measures of the supply of new books (from the Library of Congress card catalog and from US copyright registrations), as well as additional

measures of success, including Pulitzer Prize and National Book Award nominations, and the New York Times fiction bestseller list.

4.1 Social Security and WIPO name data for inferring gender

In much of what follows I have lists of works with author first names. I match those first names with the name/sex data to obtain the shares of people with that name who are women. I then calculate the number of women in a group as the sum of the female share across products.[10]

I obtain gender correspondences from two sources. The US Social Security Administration (SSA) and WIPO both maintain data on the distribution of names by sex of child.[11] WIPO maintains a list of 173,723 names which they determine to be either male of female. These data have been used to identify genders of patent inventors (Martínez et al., 2016). The national Social Security names files, covering births from 1880-2021 contain 100,364 distinct first names; and the data indicate the share of persons with each name who are men vs. women.[12] I combine the WIPO and Social Security data. The Social Security data contain information on an additional 9,244 names. Collectively, these data sources give me information on the genders associated with each of 182,967 first names.

Automated matching of first names leaves two challenges to be addressed, authors who use their initials rather than their names, and authors who use pseudonyms not associated with their gender. To address these challenges, I supplement the automated matches in two ways. First, I determine author gender by hand for the top non-matching authors, most of

[10] A word about sex and gender is in order. Sex is based on biological attributes, while gender describes socially constructed roles (see https://cihr-irsc.gc.ca/e/48642.html). To the extent that people use the names assigned them at birth, names would reflect sex understood by parents at birth. If, by contrast, creators employ names they have chosen for themselves, the names might reflect gender as distinct from sex. I have no information about how authors identify, so I cannot distinguish gender from sex. I will therefore – and somewhat inexactly – refer to gender and sex interchangeably. My interest is in the characteristics of populations, such as the authors on copyrights during a particular year, rather than individuals. What matters for these measures is not being correct in each instance but rather in being accurate in the aggregate.

[11] See https://www.ssa.gov/oact/babynames/limits.html for the Social Security data. The WIPO gender data are available at https://www.wipo.int/publications/en/details.jsp?id=4125. I use the file wgnd_langctry.csv.

[12] Of these names, 57,797 are associated only with females, and 31,459 only with males. The remaining 11,108 appear with both sexes. Of these, 8,501 are more than 75 percent associated with a single gender.

whom use initials rather than names. Second, I match author names in the database with a pseudonym database (https://www.trussel.com/books/aka.htm) containing real names associated with 6,892 author pennames.

4.2 Bookstat data

The Bookstat data extract I use includes annual edition-level 2018-2021 Amazon sales, as well as prices, star ratings, and numbers of reviews, for editions appearing in roughly the top 400,000 print editions, and the top 300,000 ebook editions, per day. I include editions published between 1960 and 2021. This is a total of over ten million underlying editions; for each edition, I also observe the author's name, the publisher, the publication date, and the genre (Bookstat includes 41 genres). For the purpose of measuring the supply of new books released each year, I treat the multiple editions of the title as a single book; for usage measurement, I aggregate the sales from all editions.[13]

I create two kinds of measures from the Bookstat data. First, I create supply measures reflecting the numbers of new books published per year, or N_v, where v refers to vintage. I also create this supply measure separately by genre: N_{vg}. I calculate the female-authored share of books from vintage v as $n_v^f = \frac{N_v^f}{N_v}$, where N_v^f is the number of female-authored books published in vintage v. Analogously, $n_{vg}^f = \frac{N_{gv}^f}{N_{vg}}$. Second, I create usage measures by calendar time t as well as vintage v. Define q_{tv} as the sales of books from vintage v during year t, and define q_{tv}^f as the sales of female-authored books from vintage v during year t. Then s_{tv}^f is the female-authored share of vintage v sales during year t , or ($s_{tv}^f = \frac{q_{tv}^f}{q_{tv}}$. Analogously, s_{tvg}^f is the female-authored share of sales for vintage v books in genre g during year t, where $s_{tvg}^f = \frac{q_{tvg}^f}{q_{tvg}}$.

Table 1 summarizes these data. The Bookstat data contain 8.4 million distinct titles published between 1960 and 2021 and 2.6 billion sales of print plus ebooks for the period 2018-2021. I am able to identify the author gender for 79 percent of titles accounting for 87.3 percent of sales. Female authors account for 33.2 percent of overall titles (42.0 percent of

[13] I associate editions together as the same title if they share an author, one edition contains the other's title (or vice versa), and the two edition titles share the same first three letters.

identified titles) and 45.9 (52.6) percent of (gender-identified) sales.

Both female authorship and the shares of sales attracted by female-authored books vary across genres. Table 2 shows female shares of authorship and consumption for the Bookstat data, by genre. Genres differ in their female shares. In the romance genre, women produce 78.3 percent of titles and garner 80.2 percent of sales. In engineering and transport, by contrast, women produce 10.8 percent of titles and attract 10.6 percent of sales.

I use the Bookstat data to create three datasets for analysis. First, I create a dataset with the number of new editions and overall Amazon sales, by book original release vintage and calendar year. For each vintage, I have the total number of titles whose underlying titles were originally released in the vintage; I also have the numbers of titles for books written by women, men, and authors whose genders I cannot determine. For each vintage $\times$ year, I have total sales, as well as the sales by gender of author. Second, I create an analogous dataset where the cells are vintage, calendar year, and genre. Third, for the welfare analysis, I use an edition-level dataset for the last full year of data, 2021, the year that I treat as the status quo (including the female influx) in Bookstat. For each edition, I have Amazon sales during 2021, the book's publication vintage, and author gender.

4.3 Goodreads data

4.3.1 Data description

Like Bookstat, the Goodreads data include a long list of books with metadata (author name, genre, publication data), which I use to create measures of new supply by vintage, genre, and author gender. Goodreads is a site devoted to user ratings and reviews of books. Launched in 2006, it was acquired by Amazon in 2013, when the site had 20 million members.[14] The Goodreads data are from Wan and McAuley (2018) and Wan et al. (2019), and the data were collected in during 2017. Rather than sales measures, the Goodreads data include 230 million interactions with the 2.3 million books made by 800,000 Goodreads users. The

[14] See https://en.wikipedia.org/wiki/Goodreads.

data reflect users' "public shelves," the information anyone can see without logging in.[15] Interactions include rating, reviewing, and "shelving" (indicating an intention to read). Of the user-book interactions in the sample – instances in which a users either rates a book or adds it their "shelf" – 203 million (covering books published between 1960 and 2017) were left by Goodreads users between 2007, the first year with substantial numbers of ratings, and 2016, the last full year of data. I use these data to create "purchase histories" for these users and books. This, in turn, allows me to create measures of the usage of each underlying title (for titles published as early as 1960) during each calendar year from 2007 to 2016. The Goodreads data also include "original publication dates" for 1,268,258 volumes. I replace publication years with these dates' years when these original years are earlier than edition publication years.

As Table 1 shows, the resulting dataset covers 1.9 million titles and 175.5 million instances of "usage," where usage reflects user interactions with books. I am able to match 86 percent of Goodreads titles, accounting for 93.1 percent of usage, with name-gender information. Women authors account for 42.2 percent of the Goodreads titles, or 49.1 percent of gender-identified titles. Female-authored books account for 55.1 percent of usage and for 59.1 percent of identified usage in the Goodreads data.

I do not observe characteristics of the individual consumers, but I can classify the consumers according to the books they use. First, I divide the users according to the share of their books by female-authors. I divide at the median female share (56.73 percent), and this gives 337,044 "female-leaning" users and 539,101 "male-leaning" users. Just over half of the average annual overall usage (13.07, with a median of 1) comes from female-leaning users. Second, I divide users according to above- or below-median usage of each of the ten Goodreads genres.

The Goodreads genres are: children, comics, fantasy, fiction, history, mystery, non-fiction, poetry, romance, young-adult, plus another called missing. Table 3 shows that, as in the Bookstat data, author gender varies across genres. The vast majority (78 percent) of romance books are female-authored. Comics, non-fiction, and history have higher male authorship.

[15] These data are available at https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home.

Second, genres also vary in the shares of sales garnered by women authors.

4.3.2 Goodreads data quality

The Goodreads data are not straightforward measures of sales or consumption, so it is worthwhile to compare those data to other measures of usage to validate their use in this study. A few points are in order for thinking about the comparison. First, the usual measure of book consumption – purchase – is not consumption per se but rather the purchase of a durable good. The Goodreads measure, by contrast, is more nearly reflective of consumption in the sense of reading or intending to read. Second, while there are multiple ostensibly authoritative sources of book sales information, even they are not perfectly correlated.

To assess the Goodreads data, I compare book rankings based on the Goodreads usage information for 2016 with sales ranking information derived from the New York Times and USA Today. These include weekly New York Times fiction bestsellers during 2016 and the USA Today weekly top 150 books during 2016. I transform these weekly rankings into annual rankings as follows. During each week, I calculate "pseudo-sales" (q^p_{tj}) of title j as $\frac{1}{r_j}$, where r_j is the bestseller rank of title j. I sum these pseudo sales across weeks to create annual q^p_j for each title j, and I rank these annual totals. I use three correlation measures: the correlations of the ranks and log ranks, as well as the Spearman rank correlations. For the New York Times and USA Today measures, these correlations are 0.51, 0.56, and 0.53, respectively. The analogous measures for Goodreads and the New York Times are 0.26, 0.50, and 0.71. The measures for Goodreads and USA Today are 0.26, 0.37, and 0.26. I infer that the Goodreads data contain an informative signal.

4.4 Other measures of book production and success

4.4.1 US Copyright registrations

US copyright registrations for books ("nondramatic library works") provide another measure of the number of books created over time. Because they include author names, they can be

used to create measures of the numbers of books created, by author gender, over time. A few caveats are in order, however. First, not all published books have copyright registrations. Second, some authors seek registration for written works that they have not released. Third, the copyright registration data are only available in usable form since 1978. I have data on 6.7 US copyright registrations for "non-dramatic literary works." Of these, I can match author names for 73.7 percent. As Table 1 shows, authors with female first names account for 31.4 percent of registrations between 1978 and 2020 and for 42.6 percent of those with identified genders.[16]

4.4.2 Award nominees

Awards provide an indication of success for cultural products. Two prominent books awards are the National Book Award and the Pulitzer Prize.[17] The institutions granting these awards recognize work in various categories, and the grantors report not only the winner but also the nominees. The National Book Award is given separately for fiction, nonfiction, and poetry; and there are typically five nominees in each category. Pulitzer awards prizes in fiction, history, biography, and general nonfiction and generally lists two nominees along with each winner. I obtained data on 1,067 National Book Award nominations for 1960-2020.[18] Of these, I can match author names for 93.4 percent, and women received 31.8 percent of overall nominations (34.0 percent of gender-identified nominations). I obtained data on 998 Pulitzer nominations for 1960-2020. Of these nominations, I can match author names to gender for 93.4 percent, and women received 35.2 percent of nominations. See Table 1.

4.4.3 Published books in the Library of Congress (LOC)

The LOC made the 2016 version of its card catalog publicly available as data.[19] The card catalog files contain 8.5 million records. I am able to match author gender for 72.5 percent

[16]The US copyright registration are available at https://www.copyright.gov/policy/women-in-copyright-system/.

[17]For details about these awards, see https://www.nationalbook.org/ and https://www.pulitzer.org/prize-winners-by-year.

[18]The data are available at https://www.nationalbook.org/national-book-awards/years/.

[19]See https://www.loc.gov/cds/products/marcDist.php.

of the titles in the card catalog, and 14.3 percent of the titles have female authors (19.7 percent of identified authors). Not all books receiving copyright registrations are included in the LOC collection. Inclusion reflects the Library's judgment that a work is likely to have significance or usefulness. While the LOC is the world's largest library, the Library does not acquire all published, or copyrighted, books. Rather the Library "selects from copyright deposits and other sources" in order "to ensure that the Library acquires important and scholarly works."[20] Hence, the female-authored share of the library's collection, by vintage, provides one measure of the importance of the female contributions to those vintages.

4.4.4 Commercially successful books

I have New York Times fiction bestseller data – covering 15 titles per week – or 44,276 listings overall for 1960-2020. I obtain gender matches for 99.7 of listings, and women account for 35.2 percent of them.[21]

5 Evidence

This section provides direct evidence on four questions. First, I document the female influx into authorship. Second, I explore the impact of the female influx on various the shares of consumption garnered by female-authored books, both overall and for consumers with different preferences. I also explore the impact of the female influx on other measures of female-authored book success. Third, I explore the predictability of realized product success at entry. Fourth, I investigate the possible displacement of male-authored book entry by increased female-authored entry.

[20] According to its policy statements, the "Library should possess in some useful form, the records of other societies, past and present... ...whose experience is of most immediate concern to the people of the United States." See https://www.loc.gov/acq/devpol/cps.html.

[21] The data are available at https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2022/2022-05-10/nyt_full.tsv and are described at https://data.post45.org/wp-content/uploads/2022/01/NYT-Data-Description.pdf.

5.1 Evidence of the female influx and its effects

My various data sources allow me to characterize the female share of supply over time. I do this, in Figure 2, with the female share of books published in each year from 1960 until 2021 for Bookstat and 2016 for Goodreads, along with the female share of authors on US copyright registrations, 1978-2020. The female supply shares rises substantially using all three data sources, from roughly 20 percent in the 1960s to roughly 50 percent by 2010. In the left panel of Figure 2, the female supply share is the ratio of female-authored books to books whose authors are gender-identified. The denominator in the right panel is total books, making the female author share in the right panel a more conservative measure. I use this conservative measure throughout the paper.[22]

Figure 3 summarizes the growth in the female authorship share (n^f), by genre, in the Bookstat and Goodreads data. I regress the log of the female share of authors by publication vintage and genre on vintage. The resulting coefficients, showing percentage annual growth, rise statistically significantly for almost all genres. Female authorship grows subtantially even in genres with traditionally low female shares, such as textbooks, political science, and history. Not only has female authorship risen overall, but this increase occurs for in both genres consumed primarily by women as well as those consumed primarily by men. This raises the possibility that the female influx will raise the value of the choice set for consumers with a wide range of preferences.

5.2 Does the female influx attract usage?

5.2.1 Empirical Strategy

We want to know the causal impact of a vintage's femaleness in authorship on the share of sales that the female books of the vintage garner. To this end we can use the variables s^f_{tv} and n^f_{tv} defined above. For brevity we refer to these variables as "femaleness in demand" and

[22] I have run the statistical exercises using both measures, and none of the substantitive results of the study change.

"femaleness in supply," respectively. We want to measure $\frac{\partial s_{tv}}{\partial n_v}$, the derivative of femaleness in demand with respect to femaleness in supply.

To see what's challenging about measuring this causal relationship – and its possible solution – it is helpful to consider a sequence of measurement approaches. First, we could treat the data as a cross section of observations on original-release vintages v, and run a regression of the form: $s_v = \alpha_0 + \alpha_1 n_v + \epsilon_v$. This approach would ask whether vintages that are more female in supply are also more female in demand. Said another way, this approach would ask whether female-authored books account for a larger share of sales from vintages in which a higher share of books are female-authored. A natural concern about this approach is that tastes may shift over time – and therefore also across vintage – toward the sorts of book that are more predominantly written by women (e.g. romance novels). The shift in demand could elicit an increase in the supply of female-authored books. If so, the coefficient α_1 could reflect the impact of demand on the tendency for women to write books, rather than other way around.

We have a few alternative approaches for addressing this. First, becaues we observe genre, we can calculate s^f_{tvg} and n^f_{vg}, where $s^f_{tvg} = \frac{\text{sales of female-authored, vintage } v\text{, genre } g \text{ books in year } t}{\text{sales of vintage } v\text{, genre } g \text{ books in year } t}$, and $n^f_{vg} = \frac{\text{\# of female-authored books from vintage } v \text{ and genre } g}{\text{\# of books from vintage } v \text{ and genre } g}$. The reverse causality concern above was that, say, growing demand for gender-imbalanced genre could give rise to entry into that genre. This, in turn, could deliver a relationship between s^f and n^f running from s^f to n^f, rather than the other way around. By looking within genre, I avoid the problem of growing demand for genres driving the female-authored share of supply.

Second, if the endogeneity is driven by absolute changes in demand for some genres, I can exclude growing or shrinking genres, measuring the relationship between the female shares of supply and demand in genres that are not growing.

Third, rather than using female-authored shares of total sales s^f as the outcome, I can look at various extreme outcomes, including award-winning and books reaching the right tail of the sales distribution. As the female share of supply changes across vintages, what happens to the female share of nominees for Pulitzer Prizes and National Book Awards?

These awards are given to contemporary works, so that $v = t$ and the regressions take the form $w_v = \lambda_0 + \lambda_1 n_v^f + \varepsilon_v$, where w_v is the female share of nominees for some category of award for books published during v. Related, I examine the relationship between n^f and s^f across deciles of the sales or usage distributions. Does the female influx deliver similar share of female-authored sales throughout the distributionsin Goodreads and Bookstat data? Finally, and related, does the growing female share of authors appear in the right-tail of the sales distribution, as reflected in the New York Times fiction bestseller list?

5.2.2 Effect of female authorship on usage of female-authored books ($\frac{\partial s^f}{\partial n^f}$)

Table 4 reports results for regressions of various measures of the female demand share (s^f) on the female supply share (n^f). The first two columns use data by vintage and year but not by genre. Column (1) uses Bookstat data, while column (2) uses Goodreads data. Both specifications include calendar year fixed effects. The coefficients of interest are 1.02 (standard error = 0.07) and 1.27 (0.04), respectively. These regressions, which suggest more-than-proportionate female sales growth with the growth in the female supply share, are vulnerable to a concern that demand is shifting toward across genres, attracting entry in ways that affect the female shares of supply.

Data disaggregated by genre give us a few ways to better measure the causal impact of the female supply share on the female demand share. First, columns (3) and (4) use data by time, vintage, and genre, along with fixed effects for genre, time, and vintage. The coefficient of interest falls in specifications from both datasets, to 0.91 (0.02) for Bookstat and 1.07 (0.03) for Goodreads. These results show that as the female share of supply rises, the female-authored share of demand rises proportionally, or nearly proportionally.

Second, genre data allow us to estimate the coefficient of interest separately by genre; and Figure 5 reports these for the Bookstat and Goodreads, respectively. There is some variation across genre in the coefficient, although all are statistically larger than 0. While coefficients are particularly high for some of the female-dominated genres, they are also high for traditionally male genres such as history.

Third, if we can concerned that shifts in demand across genres are inducing changes in the female supply share, we can restrict attention to genres that are not growing quickly. Figure 4 shows variation across genres in sales growth, based on genre-specific coefficients from regressions of π_{gv} on v, where the dependent variable is the share of vintage-v sales in genre g, or $\pi_{gv} = \frac{q_{gv}}{Q_v}$ ($Q_v = \Sigma_g q_{gv}$). We run the genre-specific regression $\pi_{gv} = \gamma_0^g + \gamma_1^g v + e_{gv}$. The coefficient γ_1^g shows the growth rate of each genre's share of total vintage sales. As the Figure shows, there is a wide range of growth rates in genres. Romance and young adult, two heavily female-authored genres, are growing more quickly than others. Others are shrinking as shares of vintage sales. The figure lends credence to the concern that demand growth could drive growth in supply; it also suggests a useful robustness check. Columns (5)-(7) report the Bookstat regression in column (3) separately for growing, stable, and shrinking genres. Results are similarly positive for all three groups of genres. Hence, it does not appear that changing demand for particular genres – differentially attracting women to authorship – is responsible for the result.

Results in Table 4 indicate that the influx of female authorship is valuable to consumers. Moreover, the near-proportionality of s_{tv} to n_{tv} indicates that the additional female-authored books are nearly as useful, on average, as the inframarginal female books. The results therefore also suggest that product quality is rather unpredictable.

5.2.3 Heterogeneous consumers

Columns (8) and (9) of Table 4 use the Goodreads measures of usage for users differing in their consumption of female-authored books. The coefficients of interest are large for both male and female-leaning consumers, albeit higher for female-leaning users. This indicates that the female influx's impact is felt similarly for heterogeneous users. Figure 6 takes the male- and female-leaning reader idea a step further with deciles of readers according to the gender shares of the authors whose books they use. Rather than just above or below median use of female-authored work, Figure 6 reports the coefficient on n_{vt}^f from ten separate regressions for readers in different usage deciles for female-authored work. The coefficient is uniformly high until the 9^{th} and 10^{th} deciles, the readers whose usage is most concentrated

in books by male authors. That is, the readers who rely more heavily on male-authored books have smaller coefficients, indicating that they experience a smaller benefit from the female influx. Note that even these readers' coefficients are significantly positive, however.

Rather than dividing readers according to the gender of the authors they use, I can instead divide users according to their usage of books by genre. I divide users into heavy and light consumers of each genre, then I regress separate measures of s^f for the heavy and light users groups on the female share of authors by vintage. Figure 7 does this for all of the Goodreads genres. In most genres, heavy and light users experience similarly large effects (roughly unity). Two exceptions are romance, which is heavily female in both authorship and readership, and non-fiction, which is heavily male. Heavy romance users derive larger benefits from the female influx, as do light users of non-fiction. Even in these genres, both heavy and light users derive substantial benefits from the female influx.

5.2.4 The female influx and additional measures of female author success

Effects of female entry on the usage of the new books provides direct evidence of an effect on welfare. Here we examine other evidence of whether the female influx brings valuable products into the choice set, based on expert/curator judgments.

First, do book vintages with higher female-authored shares have greater female representation in the Library of Congress (LOC) collection. Column (1) of Table 5 reports a regression of the female-authored shares of LOC books, by vintage, on the female-authored share of books published at each vintage, for 1960-2016. The coefficient is 0.66 (0.02). As the female-authored share of supply rises, the female authored share of books chosen for inclusion in the LOC collection rises as well. This provides additional evidence that the female influx adds valuable products to the choice set, albeit at a lower rate than for usage or sales measures.

Books are eligible for awards, and two major book awards at the Pulitzer Prizes, awarded annually for fiction, general non-fiction, history, and biography, and the National Book Award, awarded annually for fiction, nonfiction, and poetry. The next columns of Table 5 turn to book award nominees, including the Pulitzer Prizes, awarded annually for fiction,

general non-fiction, history, and biography, and the National Book Award, awarded annually for fiction, nonfiction, and poetry. Columns (2) reports a regression of the female share of Pulitzer Prize nominees on the female share of authors, by vintage; and the coefficient indistinguishable from 1. Column (3) uses National Book Award nominees and produces another coefficient indistinguishable from 1.

Separate from the judgment of curators are right-tail commercial outcomes. Column (4) of Table 5 examines the impact of the female influx on the female-authored shares of New York Times fiction bestsellers. The coefficient is 1.07 (0.10), indicating that the female share of bestseller authors rises proportionally with the female authorship influx. Finally, I also measure the separate impact of the female supply influx on the female-authored shares of sales across deciles of the sales distribution in the Bookstat data. As Figure 8 shows, the coefficient is similarly high – and precisely estimated – for all deciles of the sales distribution. This reinforces the bestseller results, showing that the female influx is also visible at the top of the distribution.[23]

The judgments of curators and awards committees, along with consumer behavior, indicate that the female influx is valuable to society. Moreover, right-tail female-authored success, like female success overall, grows with female supply shares. Not only is additional female participation valuable; it is roughly as valuable as inframarginal female participation. This finding, consistent with complete unpredictability of product success at entry, also suggests the large welfare benefit from the female influx.

5.3 Predictability of product quality at entry

The framework in Section 3 highlights the importance of product quality predictability for welfare effects. The large effects of n^f on s^f suggest a substantial degree of unpredictability. Still, some of the estimates of $\frac{\partial s^f}{\partial n^f}$ are less than unity, indicating that success is not entirely unpredictable. In this section, I explore predictability directly. This is of interest in itself and also as an input into the entry model in Section 6.

[23] I obtain a similar result with the Goodreads data.

I characterize the realized quality of products with the log of their usage (or sales) measures. In addition to its intuitive appeal, this approach also corresponds closely to the "mean utility" measures used in logit models below.[24]

The 2016 Goodreads data, and the 2021 Bookstat data, include both books published in the current year as well as books published earlier. I am interested in seeing how much of books' success can be explained by factors known before publication. For clarity, I begin by including only books published in the current year in the estimation samples. See Table 6. This ensures that the explanatory variables – and in particular, the author's sales of previously published books – include only information known prior to publication. Using this approach I can can explain 15.2 percent of the variation on success for Goodreads and 29.9 percent for Bookstat.[25]

In columns (2) and (6) I include all of the books observed in the respective Goodreads and Bookstat samples, not just those published in the current year. For these regressions I calculate authors "prior" sales as current-year sales of books published before each book's publication year. This approach has the benefit of allowing me to include all of the books in the samples. The disadvantage of this approach is that, for example, the 2016 sales of a book originally published in 2010 may be affected by an author's post-2010 success. Hence, the author prior sales variable may reflect information not available prior to release. In principle, over-predicting success leads to conservative effects of welfare benefits of new products. In reality, the share of variance explained by using this prior author sales variable (and the entire samples) raises R^2 only modestly.

The remainder of Table 6 (columns (3), (4), (7), and (8)) explores how predictability varies with expected quality. I calculatate $\hat{\sigma}_j = [(\ln \hat{q}_j - \ln_j)^2]^{0.5}$, and I regress this on $\hat{lnq}_j$. In all four cases the error is higher as expected quality is higher.[26] Success is somewhat predictable, so it is not literally true that "nobody knows anything" approach is not literally correct.

[24] In the plain logit, mean utility is $\delta_j = \ln(q_j/M) - \ln(1 - \Sigma q_j/M)$. With constant market size M and a single market observation (e.g. year), quantity q_j and δ_j are perfectly correlated.

[25] The shares of variation explained fall to 13.0 and 15.1 percent, respectively, without the authors' past sales measures. This suggests that the success of authors' first books is less predictable than success of subsequent books.

[26] This will be important in the structural model because expected sales of a product arises from an integral over a quality distribution, and the expectation depends on the variance as well as the mean.

Moreover, the degree of unpredictability is lower for marginal entrants that for inframarginal (higher expected quality) entrants.

5.4 Does the female influx displace male entry?

A growing number of female-authored books – all else equal – necessarily raises the value of the choice set, but it is possible that additional female-authored books displace male-authored entry that would otherwise have occurred. While I will allow for endogenous male entry in the structural model below, I explore displacement directly by regressing the number of new male-authored books in each vintage and genre on the numbers of female-authored and unknown-gender-authored books in the vintage and gender. Table 7 reports a sequence of regressions differing in the included fixed effects. Regardless of specification, the coefficients on female and unknown-gender entry are positive. That is, as the number of new female-authored books in a genre grows, so does the number of new male-authored books. Hence, there is no indication that increased female entry has displaced male entry. This suggests, in turn, that the growth in female entry has augmented the value of the choice set.

Average sales per book, by author gender, provides additional indication that female-authored books add something to the choice set that male-authored books do not. Using the Bookstat data, average sales for female-authored books during 2021 was 188, compared with 117 for male-authored books. Restricting attention to the books with positive sales during 2021, the female average was 309, compared with 198 for male-authored books.

6 Quantifying welfare effects of the female influx

Quantifying the welfare effects of the female influx requires two basic things. First, I need to be able to calculate the revenue and consumer surplus associated with a choice set. Second, I need a way to determine the choice sets for comparison. The status quo choice sets (2016 in Goodreads and 2021 in Bookstat) include the female influx. To measure the impact of

the female influx, I need to compare the status quo to counterfactual environments without the female influx. This, in turn, requires determination of which female products to remove as well as how to model the equilibrium male entry response.

6.1 Nested logit demand model

A nested logit demand model gives rise to a simple calibration approach for calculating the CS for any choice set, as well as the quantity of each product sold. Consumer i derives utility from choice j given by $u_{ij} = x_j\beta - \alpha p_j + \xi_j + \zeta_g + (1-\sigma)\epsilon_j$, along with $u_{i0} = 0$ for the outside good. In this setup, ζ is common to books j, and has a distribution function that depends on σ (with $0 < \sigma < 1$) such that the distribution of ζ is the unique distribution with the property that, if ϵ is an extreme value random variable, then $[\zeta + (1-\sigma)\epsilon]$ is also an extreme value random variable (Berry, 1994). Define product j's "quality" as $\delta_j = x_j\beta - \alpha p_j + \xi_j$. For any a choice set characterized by a set of product qualities $\{\delta_j\}$, I can calculate the usage of each product q_j, as well as the CS for the choice set.

Given a substitution parameter σ, the mean utility of each product j is given by $\delta_j(\sigma) = \ln(s_j) - \ln(s_0) - \sigma\ln(\frac{s_j}{1-s_0})$, where $q_j = M\frac{e^{\delta_j/(1-\sigma)}}{D}\frac{D^{1-\sigma}}{1+D^{1-\sigma}}$, $D = \Sigma e^{\delta_j/(1-\sigma)}$, $CS = \frac{M}{\alpha}\ln(1 + D^{1-\sigma})$. Even without the price parameter α, I can still calculate the percent change in CS with the female influx.

I want to want to compare status quo choice sets containing the female influx to counterfactual choice sets without the female influx (resembling the choice sets for the period 1960-1970). Determing the counterfactual choice sets presents four difficulties. The first is determining which female-authored books to remove from the status quo choice set. It is instructive to first consider a simple, albeit incorrect, approach corresponding, literally, to "nobody knows anything." If product success were entirely predictable, then we could eliminate the female influx by removing products at random. This approach, which the evidence of product quality predictability shows to be incorrect, would remove products as good, on average, as those that remain. Hence, this approach would lead to an overstatement of the benefit of the influx.

The second difficult question is whether – and how – male-authored entry responds to the counterfactual removal of the female influx. A simple, and also likely incorrect, approach would be to assume no additional male entry in response to the no-female-influx counterfactual. Thus, the counterfactual environment would sacrifice high-quality female products while allowing no additional male entry to offset the loss. This approach, too, would overstate the welfare benefit of the female influx. Still, the simple approach is useful to implement as a transparent upper-bound estimate.

Third, I need a substitution parameter σ in order to implement this. I use the estimate that Reimers and Waldfogel (2021) obtain for books (0.373) for my baseline estimates, and I explore how results change with a range of substitution parameters.

Fourth, I need to choose how may of the female-authored books to remove from the status quo choice sets to counterfactually simulate environments without the female influx. I remove female-authored books to bring each post-1970 publication year's ratio of female-authored to male-authored books to its average for the 1960s.

Table 8 shows the upper-bound results based on random removal of female books and no endogenous male author response. Using the baseline substitution parameter, overall revenue rises by 10.8 percent using Bookstat and 21.6 percent using Goodreads. In both cases, male revenue falls, while female author revenue rises. Using the Bookstat (Goodreads) data and the baseline substitution parameter, the female influx raises overall CS by 18.4 (26.7) percent. The CS of female-leaning consumers (using Goodreads) rises by 41.1 percent. Importantly, the CS for male-leaning consumers also rises, albeit less than female-leaning consumers (by 15.4 percent).

6.2 Incorporating partial predictability

The evidence above that product quality is at least somewhat predictable calls for more nuanced approaches to both the removal of female-authored books from the status quo choice set and to the determination of equilibrium male entry in the no-female-influx counterfactual. I need to remove female books of appropriately lower-than-average quality, and I need

to allow for an endogenous male entry response. The elimination of the female influx is accomplished in the model by raising the female entry threshold high enough to bring the female share of books to its 1960-1969 relationship with total new books released.

A diagram is helpful. Panel 1 of Figure 9 shows two schedules reflecting the expected revenue of the marginal male and female entering products as functions of the amount of entry. Panel 2 shows the effect of removing the female influx. Because the removal of the female influx leaves fewer products in the choice set, the expected returns to male and female products both rise, shown by the schedules labelled $M^{'}$ and $F^{'}$, respectively. The expected revenue of the marginal entrant has risen, for both male and female products. Female products, by assumption, cannot respond with additional entry. Male entry can occur, until the expected revenue of the marginal entering product falls to the original male entry threshold. The additional male entry is reflected by the dark line segment in panel 3.

We implement this approach as follows. For removal of female-authored books, I order status quo entry according to predicted quality $\delta_j^{'}$. To predict product quality, I regress realized product quality δ_j on the explanatory variables in columns (2) and (5) of Table 6. I remove the lowest-expected quality books from each publication year to bring female authorship to its relationship with male authorship in the 1960s.

Endogenous male entry beyond what's observed in the status quo is a slightly tougher problem. The generic problem is that we need estimates of the quality of products that don't exist. We can get some evidence from the relationship between expected quality and entry order among products we do observe. In particular, we can fit a line to the relationship between predicted quality and entry order among, the last 5 percent of male entering products: $\delta_j = A + Bn_j + e_j$, where n_j is entry order according to expected quality. We can then simulate the qualities of additional products as follows. If N_0^M is the number of male products in the status quo, then the expected quality of the $(N_0^M + k)^{th}$ product is given by the fitted line $\delta_k^{'} = A + B(N_0^M + k)$, while realized quality is expected quality plus an error term. We also need simulated draws of realized quality; and to produce this we draw an error from $N(0, \sigma_{N+k})$ where the heteroscedasticity has been parameterized as $\hat{\sigma}_k = C + D(N_0^M + k)$.

To determine the extent of additional male entry, we need to find the value of k for which the expected sales of the $(N_0^M + k)^{th}$ product equals the status quo male entry threshold, T^M. This, in turn, requires calculation of the expected sales of the marginal entering product. Using simple logit for illustration, define $q(k)$ as the realized sales of the k^{th} additional entrant. Then the realized value of sales for the $(N_0^M + k)^{th}$ entrant is given by

$$q(k) = \frac{e^{\delta_{N_0^M+k}}}{1 + \Sigma_{j\in F_1} e^{\delta_j} + \Sigma_{j\in M_0} e^{\delta_j} + \Sigma_{j=N_0^M+1}^{N+k} e^{\delta_j}},$$

where M_0 is the status quo male product set, and F_1 is the female product set after removal of the female influx.

I approximate the relationship between realized $q(k)$ and the number of entering male products by regressing $q(k)$ on $(N_0^M + k)$ from a simulated draw of realized qualities. This regression averages across quantities $q(k)$ that contain random realizations. I then choose k such that expected revenue equals the status quo male entry threshold, or $E[q(k)] = T_0^M$. Then, $N_0^M + k$ is the equilibrium number of male products in the no-female influx counterfactual.[27]

Table 9 reports estimates of the welfare effect of the female influx based on the imperfect predictability approach. Using the Bookstat data, the influx raises overall revenue by 1.9 percent in the baseline. Female author revenue rises more, by 11.7 percent, while male author revenue falls by 5.7 percent. Consumer surplus rises by 3.2 percent. Estimates using the Goodreads data are larger in magnitude. Revenue rises by 6.6 percent in the baseline, with female revenue rising 22.9 percent and male revenue falling 8.4 percent. Overall CS rises by 8.1 percent. Both male-leaning and female-leaning CS rise with the female influx. Of course, the proportionate magnitude varies with σ, but the important point is that the well being of disparate kinds of consumers rises with the female influx.

Figure 10 examines impact on the CS of heterogeneous consumers according to whether they are heavy or light users of each genre. The leftmost points indicate that the female influx delivers heavy users of nonfiction a 6.5 percent increase in CS, while light users obtain a

[27] The Bookstat baseline contains 0.813 million books by women and 1.418 million books either by males or authors without gender attribution. When the female influx is removed, non-female authorship equilibrates with with 0.241 million additional books.

9.5 percent increase. At the other extreme, heavy users of romance derive a more than ten percent increase in CS while light users of romance derive just over 6 percent. While there is some heteroegeneity across the different genre breakdowns, it is noteworthy that heavy and light users of each genre derive benefits from the female influx.

Both Tables 8 and 9 show how the results vary according to substitutabilty (σ). Whlie the particular numbers vary across specifications and datasets, the finding of consumer benefit across a wide range of preferences is robust.

7 Conclusion

While women's participation in IP creation continues, generally, to lag men's, the past half century has brought a revolution in gender-inclusive book creation. Women's authorship has grown three times faster than men's, and recent vintages are 50 percent larger than they would have been absent the growing participation of women.

In this paper I document that the growth in female authorship has delivered products that a wide range of consumers finds valuable. As the share of books authored by women has grown from under a quarter to more than a half, so has the share of usage – and other measures of success – garnered by female-authored works. Using a simple structural model, I quantify effects on consumers and producers. I find that the influx of female-authored books raises the welfare of diverse consumers, providing value to consumers the male-authored books would not have delivered in their absence. Effects on revenue are different: Compared to a counterfactual environment with less female authorship, aggregate revenue rises for female authors while falling for male authors.

This paper adds to an emerging body of research (Hsieh et al., 2019; Bell et al., 2019) finding that inclusion is beneficial for innovation and growth. Importantly, the influx of new products from female authors benefits a wide wariety of consumers. Substantial growth in female authorship makes books a useful test case, and these results suggest that the benefits from more inclusive creation and innovation in other contexts might be large.

References

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

Aguiar, L., J. Waldfogel, and S. Waldfogel (2021): "Playlisting Favorites: Measuring Platform Bias in the Music Industry," *International Journal of Industrial Organization*, 102765.

Arrow, K. J. (1969): "Classificatory notes on the production and transmission of technological knowledge," *The American Economic Review*, 59, 29–35.

Bailey, M. J. (2006): "More power to the pill: The impact of contraceptive freedom on women's life cycle labor supply," *The quarterly journal of economics*, 121, 289–320.

Bell, A., R. Chetty, X. Jaravel, N. Petkova, and J. Van Reenen (2019): "Who becomes an inventor in America? The importance of exposure to innovation," *The Quarterly Journal of Economics*, 134, 647–713.

Bergemann, D. and U. Hege (2005): "The financing of innovation: Learning and stopping," *RAND Journal of Economics*, 719–752.

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

Brauneis, R. and D. Oliar (2018): "An Empirical Study of the Race, Ethnicity, Gender, and Age of Copyright Registrants," *Geo. Wash. L. Rev.*, 86, 46.

Caves, R. E. (2000): *Creative industries: Contracts between art and commerce*, 20, Harvard university press.

Cook, L. D. (2011): "Inventing social capital: Evidence from African American inventors, 1843–1930," *Explorations in Economic History*, 48, 507–518.

Costa, D. L. (2000): "From mill town to board room: The rise of women's paid labor," *Journal of Economic Perspectives*, 14, 101–122.

Ding, W. W., F. Murray, and T. E. Stuart (2006): "Gender differences in patenting in the academic life sciences," *science*, 313, 665–667.

Frietsch, R., I. Haller, M. Funken-Vrohlings, and H. Grupp (2009): "Gender-specific patterns in patenting and publishing," *Research policy*, 38, 590–599.

Goldin, C. (2006): "The quiet revolution that transformed women's employment, education, and family," *American economic review*, 96, 1–21.

Goldman, W. (2012): *Adventures in the screen trade*, Hachette UK.

Greenwood, J., A. Seshadri, and M. Yorukoglu (2005): "Engines of liberation," *The Review of Economic Studies*, 72, 109–133.

Hoisl, K., H. C. Kongsted, and M. Mariani (2023): "Lost Marie Curies: Parental impact on the probability of becoming an inventor," *Management Science*, 69, 1714–1738.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019): "The allocation of talent and us economic growth," *Econometrica*, 87, 1439–1474.

Hunt, J., J.-P. Garant, H. Herman, and D. J. Munroe (2013): "Why are women underrepresented amongst patentees?" *Research Policy*, 42, 831–843.

Kerr, W. R., R. Nanda, and M. Rhodes-Kropf (2014): "Entrepreneurship as experimentation," *Journal of Economic Perspectives*, 28, 25–48.

Kim, S. and P. Moser (2020): "Women in science: Lessons from the Baby Boom," .

Koffi, M. and M. Marx (2023): "Cassatts in the Attic," Tech. rep., National Bureau of Economic Research.

Koning, R., S. Samila, and J.-P. Ferguson (2021): "Who do we invent for? Patents by women focus more on women's health, but few women get to invent," *Science*, 372, 1345–1348.

Manso, G. (2011): "Motivating innovation," *The journal of finance*, 66, 1823–1860.

——— (2016): "Experimentation and the Returns to Entrepreneurship," *The Review of Financial Studies*, 29, 2319–2340.

Martínez, G. L., J. Raffo, K. Saito, et al. (2016): *Identifying the gender of PCT inventors*, vol. 33, WIPO.

Reimers, I. and J. Waldfogel (2021): "Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings," *American Economic Review*, 111, 1944–71.

Toole, A. A., M. J. Saksena, C. A. Degrazia, K. P. Black, F. Lissoni, E. Miguelez, G. Tarasconi, et al. (2021): "Progress and Potential: 2020 update on US women inventor-patentees," Tech. rep.

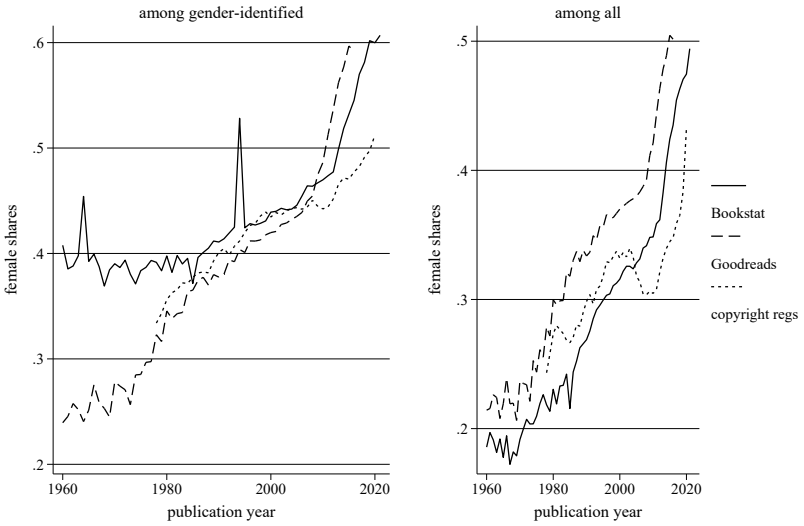
Wan, M. and J. McAuley (2018): "Item recommendation on monotonic behavior chains," in *Proceedings of the 12th ACM conference on recommender systems*, 86–94.

Wan, M., R. Misra, N. Nakashole, and J. McAuley (2019): "Fine-grained spoiler detection from large-scale review corpora," *arXiv preprint arXiv:1905.13416*.

Weitzman, M. L. (1979): "Optimal search for the best alternative," *Econometrica: Journal of the Econometric Society*, 641–654.

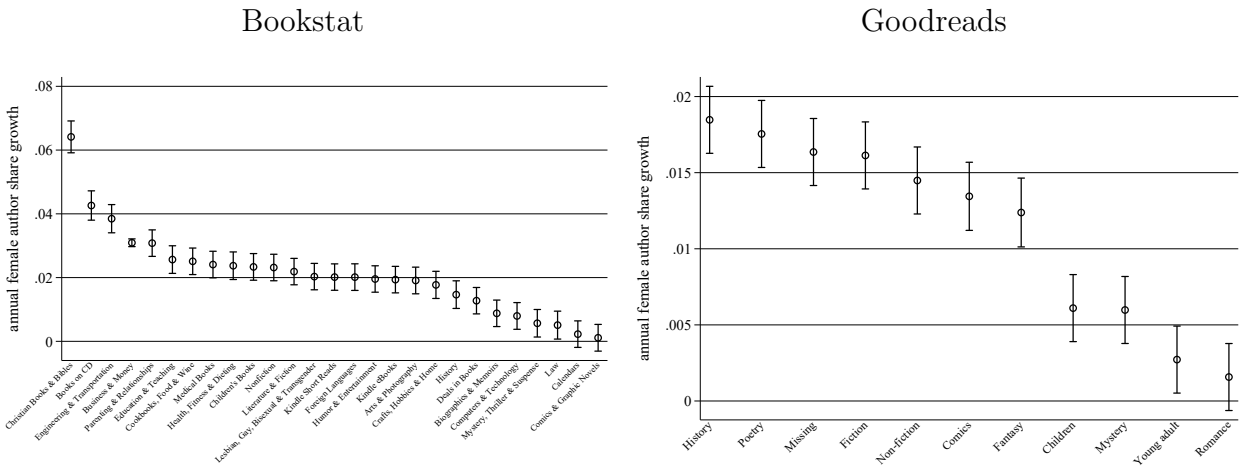
A Figures and Tables

Figure 2: Female-authored share of books by publication year



Notes: Female-authored share of books, by publication year, in Goodreads, Bookstat, and US copyright registration data. The female shares in the left figure are shares of books with identified author genders. In the right figure, books whose authors are not gender-identified are treated as male-authored.

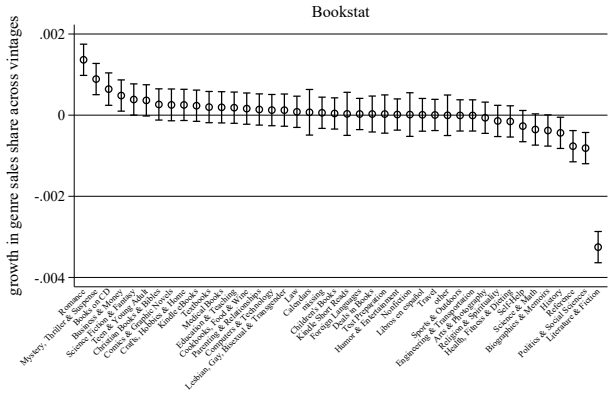
Figure 3: Growth in female-authored share by genre



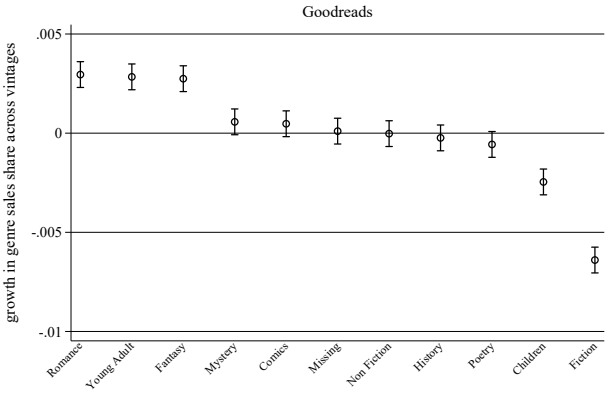
Note: Growth rate of female-authored share of new titles across publication years in Bookstat and Goodreads data.

Figure 4: Genre sales growth

Bookstat



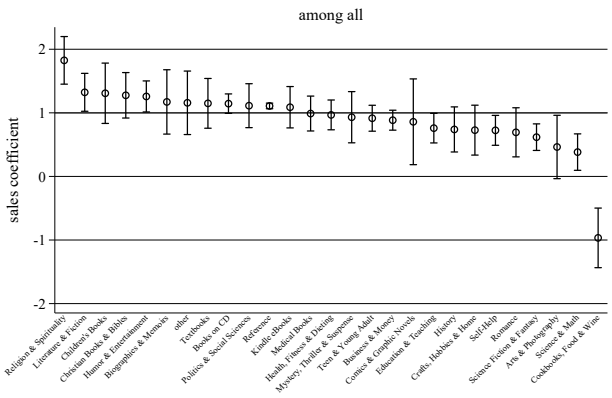
Goodreads



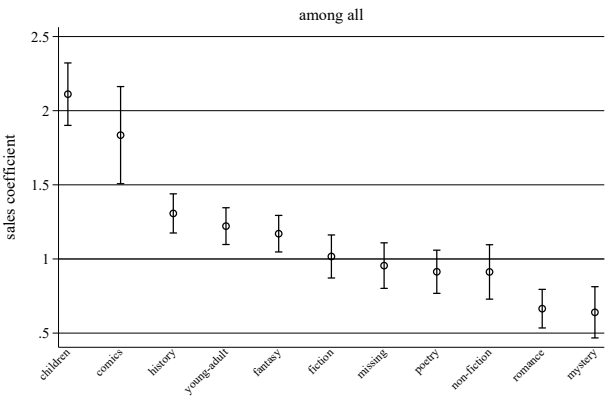
Note: Genre-specific coefficients from regression of genre's share of vintage sales on vintage. The regression includes genre fixed effects.

Figure 5: Coefficient of female share of sales on female share of works

Bookstat

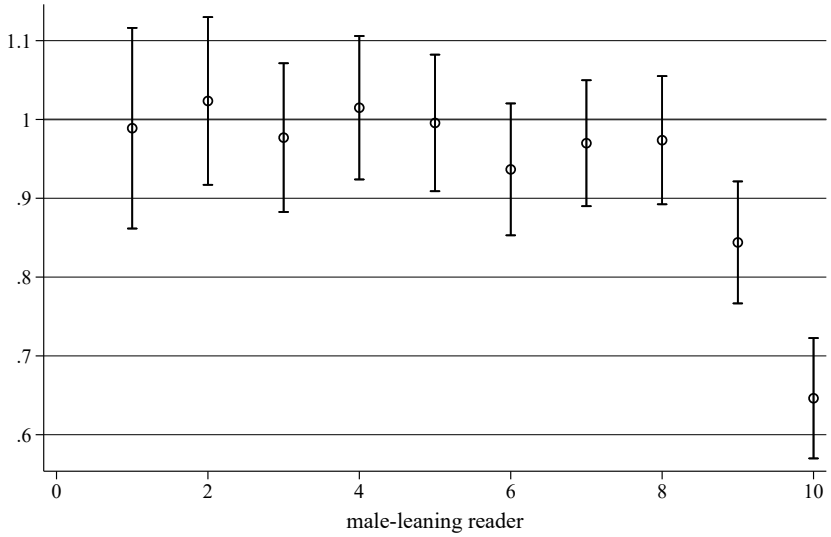


Goodreads



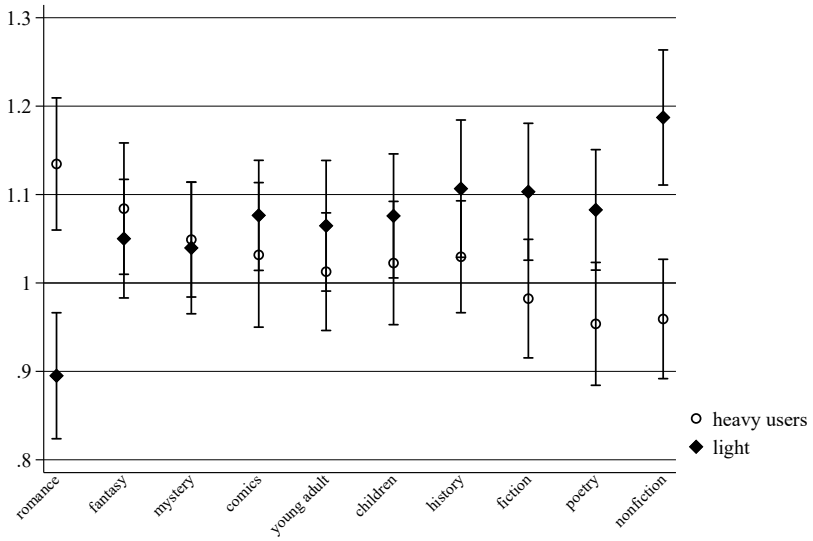
Note: Genre-specific coefficients from regression of female share of sales on female share of works. The regression includes genre and year fixed effects.

Figure 6: Coefficient of % female sales on female works % by "maleness" of consumption (Goodreads)



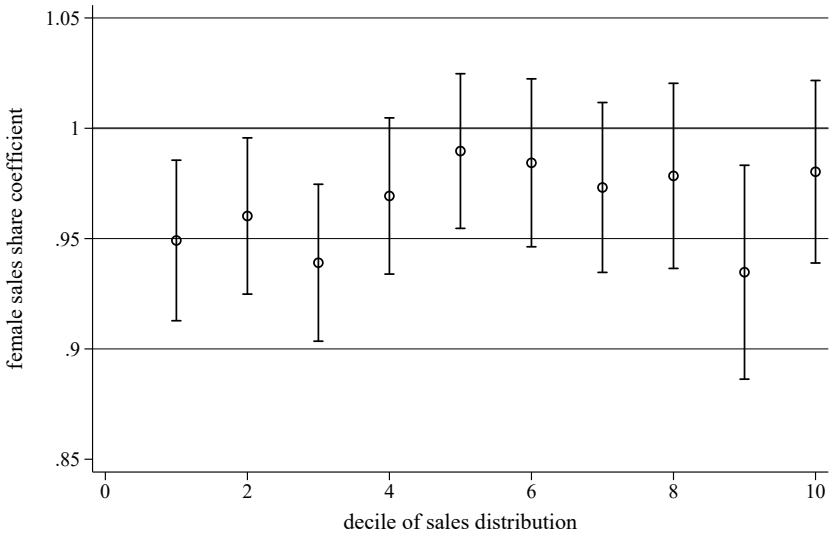
Notes: Reader types are deciles of readers according to the male-authored share of their usage. The regression includes genre and year fixed effects.

Figure 7: Coefficient of % female sales on female works % by reader type (Goodreads)



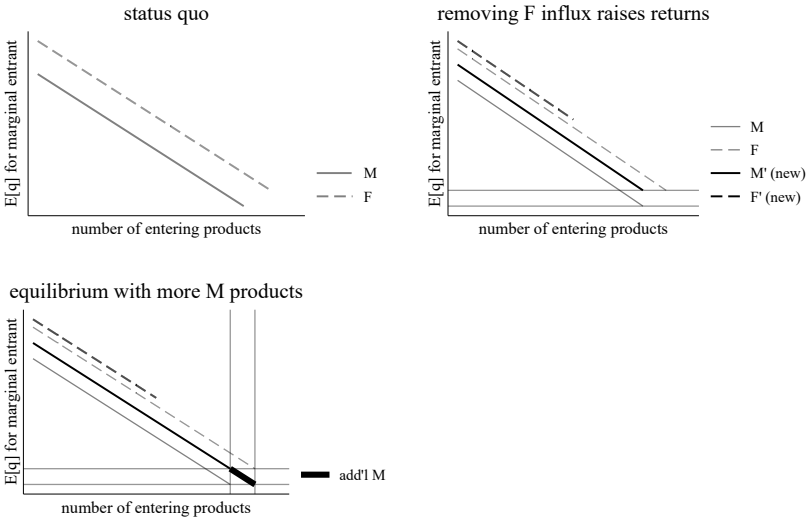
Notes: Coefficients from regressions of female-authored share of usage (s^f) on female-authored share of works (n^f), separately for heavs vs light users of each Goodreads genre. Regressions include genre, year, and publication year FE.

Figure 8: Coefficient of % female sales on female works % by sales decile (Bookstat)



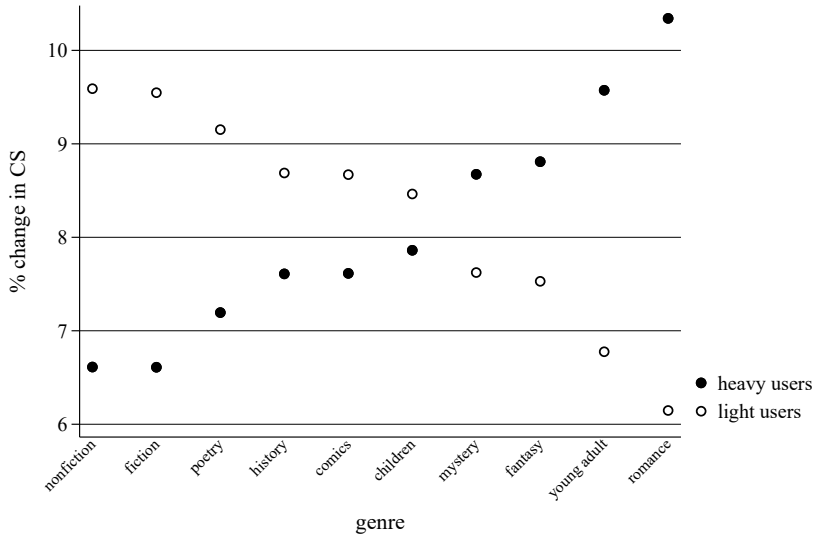
Notes: Sales-decile-specific coefficient relating the female-authored share of sales (in the genre, vintage, year, and sales decile) on the female share of authors by genre and vintage.

Figure 9: Welfare counterfactual illustration



Notes: In the upper left panel the solid and dashed lines represent expected sales of marginal books by male and female authors respectively. The upper right panel depicts the removal of the female influx, which both shortens the female schedule and raises expected returns for both male and female-authored books at each level of entry. The bottom panel shows, in bold, the additional male entry needed to bring the expected sales of the marginal male-authored book to the original male entry threshold.

Figure 10: Effect of female authorship expansion on CS by reader genre type



Notes: The figure shows how heavy vs light genre user-specific welfare estimates vary for different genres, based on the model with imperfect predictability. For example, the leftmost dots shows that light users of nonfiction derive a nearly 10 percent increase in CS from the female influx, while heavy users derive a roughly 6.5 percent increase.

Table 1: Summary statistics

dataset	concept	number	gender-identified titles	female-authored % of titles	quantity (000)	gender-identified quantity	female-authored % of quantity	earliest	latest
Bookstat	books	8,389,304	79.0%	33.2%	2,581,979	87.3%	45.9%	1960	2021
Goodreads	books	1,865,137	86.0%	42.2%	175,465	93.1%	55.1%	1960	2016
Library of Congress	books	8,455,429	72.5%	14.3%				1801	2016
Copyright registrations	listings	6,703,729	73.7%	31.4%				1978	2020
Pulitzer Prize	books	998	93.4%	24.4%				1960	2020
National Book Award	nominations	1,067	93.4%	31.8%				1960	2021
NYT fiction	nominations	44,276	99.7%	35.2%				1960	2020

Notes: Bookstat data are estimates of Amazon sales and include both ebooks and print sales. The quantity figure covers 2018-2021 sales. The Goodreads usage measure reflects interactions between users and books, including the registration of an intent to read a title. The usage occurs between 2007 and 2016. The Library of Congress data refer to its publicly available card catalog. The copyright registration data reflect US copyrights for "non-dramatic literary works." Pulitzer Prize and National Book Award entries refer to nominations for these honors. The NYT row reflects the 15 weekly New York Times fiction bestsellers.

Table 2: Female authorship and sales: combined editions in Bookstat

genre	% fem authors	female aut % of sales
Romance	78.3	80.2
Cookbooks, Food & Wine	51.4	56.1
Parenting & Relationships	49.4	55.7
Lesbian, Gay, Bisexual & Transgender	49.3	54.9
Teen & Young Adult	47.1	62.7
Children's Books	46.0	43.8
Kindle eBooks	43.6	58.5
Literature & Fiction	43.6	58.1
Mystery, Thriller & Suspense	41.7	47.5
Deals in Books	41.0	65.6
Self-Help	40.1	42.2
Kindle Short Reads	39.0	67.6
Health, Fitness & Dieting	38.6	40.8
Nonfiction	38.4	39.6
Crafts, Hobbies & Home	37.3	45.4
Books on CD	32.5	53.4
Christian Books & Bibles	30.4	39.1
missing	30.4	26.8
Education & Teaching	30.1	32.3
Foreign Languages	29.3	33.0
Biographies & Memoirs	29.2	33.3
Religion & Spirituality	28.1	35.4
Science Fiction & Fantasy	27.4	29.9
Reference	26.9	30.6
Medical Books	26.4	29.0
Arts & Photography	26.4	31.8
Travel	26.0	22.0
Textbooks	24.1	27.4
Comics & Graphic Novels	23.7	15.7
Politics & Social Sciences	23.7	28.4
other	21.3	19.8
Calendars	21.2	29.6
Law	21.1	24.5
Business & Money	21.1	18.7
Humor & Entertainment	19.4	23.0
Test Preparation	17.9	12.2
History	17.6	17.3
Science & Math	17.0	20.1
Sports & Outdoors	15.6	15.6
Computers & Technology	14.8	14.4
Engineering & Transportation	10.8	10.6

Notes: The first column shows the share of authors who, according to their first names, are apparently female. The denominator includes books whose author genders cannot be inferred. The second column shows the share of sales accruing to books whose authors are apparently female.

Table 3: Female authorship and sales: Goodreads

genre	N	female-authored % of books	female-authored % of sales
romance	195194	0.783	0.829
young-adult	64869	0.643	0.759
fantasy	172555	0.511	0.613
children	77557	0.479	0.378
mystery	132605	0.422	0.529
fiction	369491	0.370	0.430
non-fiction	285309	0.336	0.372
missing	295069	0.333	0.320
history	130460	0.331	0.431
poetry	36366	0.310	0.287
comics	61906	0.165	0.208
total	1821381	0.423	0.552

Notes: The first column shows the share of authors who, according to their first names, are apparently female. The denominator includes books whose author genders cannot be inferred. The second column shows the share of sales accruing to books whose authors are apparently female.

Table 4: Female authorship and success regressions

	(1) BS (tv)	(2) GR (tv)	(3) BS (tvg)	(4) GR (tvg)	(5) declining	(6) stable	(7) growing	(8) GR fem (tvg)	(9) GR men (tvg)
female-authored share of new products	1.199***	1.272***	0.911***	1.066***	1.176***	0.914***	0.905***	1.040***	0.960***
	(0.0721)	(0.0430)	(0.0249)	(0.0332)	(0.0922)	(0.0343)	(0.0428)	(0.0414)	(0.0341)
Observations	248	570	9664	6199	2228	4775	2661	6183	6198
$\overline{R^2}$	0.525	0.611	0.456	0.717	0.346	0.401	0.611	0.606	0.624

Notes: Regressions of the female-authored share of consumption for vintage v books in year t on the female share of books published at vintage v. All specifications include year fixed effects. All specifications except columns (2) and (4)-(6) use Bookstat data. Columns (1) and (2) use time × vintage data; the remaining columns use time × vintage × genre data. All specifications include time fixed effects; specification beginning with column (3) use time, vintage, and genre fixed effects. In columns (4) and (5) the dependent variables are the female-authored usage shares among female- and male-leaning users. Column (7) includes only the bottom quartile of genres according to sales growth, column (8) uses the middle 50 percent, and column (9) includes only growing genres. The female shares of supply and demand are calculated as the share of authors with female-identified first names relative to all. The unidentified authors, some of whom are female, are in the denominators.

Table 5: Female authorship and recognition regressions

	(1) LOC %fem	(2) Pul %fem	(3) NBA %fem	(4) NYT %fem
% fem-aut'd	0.659***	1.079***	1.212***	1.069***
	(0.0232)	(0.191)	(0.182)	(0.103)
Observations	1018	408	185	61
$\overline{R^2}$	0.631	0.136	0.210	0.640

Notes: The first three columns report regressions of the female-authored shares of Library of Congress holdings, Pulizter Prizes, and National Book Awards in each year on the female shares of books released in those years (Bookstat). Regressions in columns (1)-(3) include prize category (fiction, etc.) fixed effects. The last two columns report regressions of the female shares of bestselling authors (New York Times fiction authors and Publishers Weekly) on the female shares of books published in that year.

Table 6: Predictability regressions

	(1) ln q, 2016 GR	(2) ln q, GR	(3) σ, sparse	(4) σ, saturated	(5) ln q, 2021 BS	(6) ln q, BS	(7) σ, sparse	(8) σ, saturated
log author prior sales	0.125***	0.218***			0.298***			
	(0.00254)	(0.000506)			(0.00121)			
missing author sales	0.339***	0.620***			0.427***			
	(0.0159)	(0.00879)			(0.00884)			
female author	0.240***	0.0592***			0.229***	0.214***		
	(0.0103)	(0.00249)			(0.00527)	(0.00270)		
log author prior sales						0.359***		
						(0.000429)		
missing author sales						0.163***		
						(0.00440)		
ln qhat			0.519***	0.517***			0.393***	0.389***
			(0.000955)	(0.000942)			(0.000516)	(0.000516)
Constant	1.494***	0.418***	0.210***	0.209***	0.880***	0.627***	0.544***	0.523***
	(0.0292)	(0.0351)	(0.00162)	(0.00160)	(0.0138)	(0.112)	(0.00138)	(0.00138)
Observations	110756	1245754	1245754	1245754	558853	2232294	2232294	2232294
$\overline{R^2}$	0.152	0.217	0.192	0.195	0.299	0.336	0.206	0.203

Notes: Columns (1) and (2) show regression of log quantity on factors potentially predictive of success. Column (1) uses Goodeads data for 2016 and includes only books published in 2016. Column (2) uses Goodreads data for 2016 and includes all publication years. Columns (3) and (4) report regressions of the standard errors of the residuals from (1) and (2) on predicted quantity, using the Goodreads data. Columns (5)-(8) repeat the exercise using Bookstat data for 2021.

Table 7: Male author entry displacement

	(1) OLS	(2) genre FE	(3) vintage FE	(4) all FE
female-authored books	0.294***	0.292***	0.276***	0.275***
	(0.00560)	(0.00561)	(0.00503)	(0.00497)
Unknown gender-authored books	1.163***	1.171***	1.005***	0.998***
	(0.00994)	(0.01000)	(0.00957)	(0.00953)
Observations	17113	17113	16720	16720
$\overline{R^2}$	0.891	0.893	0.915	0.920

Notes: Regressions of the male-authored books entering by vintage on female entry and unknown gender entry, along with vintage and genre fixed effects, as indicated. All columns use Bookstat data.

Table 8: Effects of female influx on CS and revenue: no quality predictability and no endogenous male entry

% change in CS				% change in revenue		
			Bookstat			
σ	overall			overall	female	male
0.25	22.6			13.25	160.54	-25.35
0.373	18.44			10.79	154.89	-26.97
0.5	14.34			8.38	149.35	-28.56
0.75	6.83			4	139.26	-31.45
			Goodreads			
sigma	overall	female-leaning	male-leaning	overall	female	male
0.25	32.91	51.3	18.72	26.62	135.98	-19.44
0.373	26.73	41.11	15.39	21.62	126.65	-22.62
0.5	20.7	31.4	12.06	16.73	117.54	-25.73
0.75	9.78	14.45	5.83	7.89	101.07	-31.35

Notes: Model simulations based on both the basline substitution parameter ($\sigma = 0.373$) as well as a range from 0.25 to 0.75 with no product quality predictability and no endogenous entry. Female products are removed from status quo choice sets at random, and no endogenous male entry is allowed in response. These are upper-bound estimates.

Table 9: Effects of female influx on CS and revenue with endogenous counterfactual male entry

	% change in CS			% change in Revenue		
σ	overall	female	male	overall	female	male
			Bookstat, 2021			
0.25	3.87			2.25	12.14	-5.39
0.373	3.22			1.87	11.73	-5.74
0.5	2.56			1.49	11.3	-6.09
0.75	1.27			0.74	10.49	-6.78
			Goodreads, 2016			
0.25	9.82	12.82	6.85	7.93	24.41	-7.23
0.373	8.14	10.61	5.69	6.58	22.88	-8.41
0.5	6.43	8.35	4.51	5.19	21.26	-9.59
0.75	3.15	4.06	2.22	2.54	18.18	-11.85

Notes: Model simulations based on both the basline substitution parameter ($\sigma = 0.373$) as well as a range from 0.25 to 0.75 with endogenous male entry and removal of the female influx. To simulate the environment without the female influx, female-authored books are removed in accordance with entry order from the prediction model. Moreover, the counterfactual environment includes additional male-authored books up to the level of male entry at which the marginal entering male book has the status quo usage. Consumers are classified as "male"- or "female-leaning" according to the female-authored share of the books they use.