NBER WORKING PAPER SERIES

THE WELFARE EFFECT OF GENDER-INCLUSIVE INTELLECTUAL PROPERTY CREATION: EVIDENCE FROM BOOKS

Joel Waldfogel

Working Paper 30987
http://www.nber.org/papers/w30987

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2023

Some of the work on this project was undertaken while I was the Kaminstein Scholar at the US Copyright Office. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

The Welfare Effect of Gender-Inclusive Intellectual Property Creation: Evidence from Books
Joel Waldfogel
NBER Working Paper No. 30987
February 2023
JEL No. J16,L82,O3

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 produced the majority of books. Adding new products can have significant welfare benefits, particularly when product quality is unpredictable. Using data on sales of 8.9 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 a roughly equal proportionate increase 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 with unpredictable product 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, gains to consumers from inclusive innovation are experienced by a wide range of consumers.

Joel Waldfogel
Frederick R. Kappel Chair in Applied Economics
3-177 Carlson School of Management
University of Minnesota
321 19th Avenue South
Minneapolis, MN 55455
and NBER
jwaldfog@umn.edu

1 Introduction

In many areas of creative or innovative endeavor, 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 a more inclusive involvement in 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 are 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 distributing income more evenly among potential creators. Because most people are consumers rather than producers of any given product, the impact of inclusive creation on consumers may be more important than its impact on producers. But the vast majority of books attract little use, 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 heavily on the predictability of

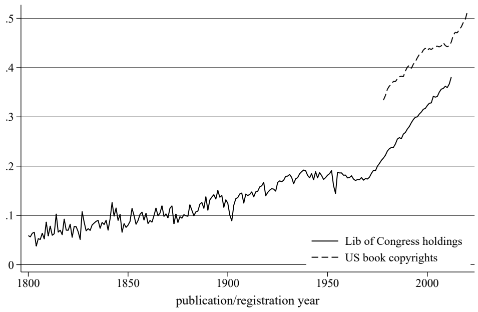
[1] See, for example, (Hunt et al., 2013; Frietsch et al., 2009).

[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.

new product quality (Aguiar and Waldfogel, 2018). If the quality of new products were predictable before launch, then an expansion in the number of products would bring only products less valuable than the least-useful 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, taking more 'draws from the urn' can deliver valuable additional products; and the growing participation of women has facilitated substantially more draws.[6]

Hence, the book market should provide a useful "experiment" in inclusive IP creation. The context also has the advantage that, unlike with patents, IP can be readily linked to associated books which have observable usage, allowing documentation of effects of more inclusive IP creation on consumption, revenue, and welfare. This leads to four empirical questions. First, 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? Second, 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? Third, how large are welfare effects on consumers and producers from inclusive IP creation? Finally, do heterogeneous consumers gain

[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).

from the female influx, and what happens to the revenue of male and female authors?

To study these questions, one needs not only data on the number of works created over time, by gender of author, one also needs data on consumption of each work. Moreover, to draw inferences about the possibly heterogeneous welfare benefit of new products, I require usage data for different types of consumers. I have two data sources, which allow me a fairly comprehensive look at both supply and demand. First, I have Bookstat data on over 10 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 their annual Amazon sales over this time period. I also observe author name, genre, and publication date. Second, I have information on nearly 230 million book ratings of 1.8 million books by over 800,000 individual consumers at Goodreads between 2007 and 2016 (Wan and McAuley, 2018). I use both the Bookstat and Goodreads data to create time series on the numbers of books published by year back to 1960, by author gender and genre. The Bookstat data also allow me to directly observe sales of books by author gender, time, and vintage. The individual-level usage 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 on not only 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 existing literatures on womens' growing participation in the economy during the 20^{th} century as well as work on female participation in innovative and creative activity in particular. Section 3 presents a simple theoretical framework drawing on Aguiar and Waldfogel (2018) illustrating the role of quality predictability in the effects of entry on both the usage and welfare benefits from newly entering products. 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 book bestseller lists, and

National Book Awards, and Pulitzer Prizes. Section 5 turns to documenting the female authorship influx and evidence of its impact on welfare, in particular through its effect on the female-authored share of consumption. Not only does the female-authored share of books rise quickly after 1970, it rises in all genres. I use a variety of approaches to measure the impact of the new female-authored supply on female-authored consumption shares. I address possible concerns that growing demand for "female" books drives the female author influx, rather than the other way around, in various ways. First, I document a within-genre relationship between the female-authored shares of supply and demand, showing that the relationship is not driven by growing demand for heavily female-authored genres such as romance. Second, I estimate the relationship between female-authored supply and demand separately for growing and stable genres, showing that the relationship arises even in genres that are not growing so that demand growth is not likely driving the shifting gender composition of supply. I also show that the female influx has raised the shares of female-authored books amongs award nominees and best sellers. I conclude that the influx of female-authored books is consequential for consumers: As the share of female-authored books has risen, the share of consumption garnered by female-authored books has risen in roughly equal proportion. This has two distinct implications. First, the influx of female-authored works is valuable to consumers. Second, the quality of new books is unpredictable.

Next, I explore whether the female influx delivers benefits not available from the male-authored works it might displace. Using data on entry by author gender, vintage, and genre, I find no evidence that entry of female-authored books displaces male-authored entry. Moreover, the female-authored books have higher average usage than the average male-authored book. These facts suggest that the female influx changes the composition and raises the size and equilibrium appeal of the choice set.

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 measure the welfare effects of the female influx by comparing the status quo choice set to a choice set that would have arisen absent the female author influx. For the baseline estimates, I remove female-authored books at random to bring each vintage's female-authored share to its average for 1960-1970. Using an estimate of the nested logit substitution parameter from Reimers and Waldfogel (2021), I find that the female influx would raise revenue by between a tenth and a fifth overall and CS by about a quarter overall. Revenue to female authors more than doubles, while revenue to male authors falls by about a quarter. While gains in female author revenue come partly at the expense of male authors, both male-leaning and female-leaning consumers see increases in consumer suplus, by 15 and 41 percent, respectively. Moreover, CS gains are experienced by both heavy and light users of each of ten book genres in the Goodreads data. Similar results – including welfare benefits for heterogeneous consumers – arise from a wide range of substitution parameters and from models incorporating some predictability of product quality, as Section 7 shows. I conclude that inclusive innovation in books is "win-win": not only is the female authorship influx welfare-improving for women (as authors and readers) but also for a wide range of consumers, including those reliant on traditionally male 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 had risen to 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 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). 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: quality predictability and the welfare benefit of new products

In usual ways of thinking about product entry, creators launch products if their expected revenue exceeds the cost of bringing the product to market. Hence, one can view the influx of female-authored books as arising from some combination of rising expected revenue and falling costs. Whatever the cause, when entry conditions change and new products become available, the effect of an influx of new products of consumption and welfare depends on the

[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.

[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.

predictability of product quality at entry (Aguiar and Waldfogel, 2018).

To see this, suppose that product quality (and therefore each product's realized revenue) were completely predictable. Then if the cost of entry were K, all products with quality such that their expected revenue $\geq K$ would enter. If something happened that reduced the entry threshold from K to K', then additional entry would occur. But the additional products would all have realized revenue between K and K'. While revenue and CS would rise, the increases would be modest. In particular, all additional products would be "worse" than the lowest-appeal product previously on offer.

Effects of entry on consumption and welfare can be very different when product quality is unpredictable. In the extreme case of complete unpredictability, additional products would attract as much average use, per product, as pre-existing products. Then an increase in the number of products would raise welfare much more than with complete predictability. Rather than all products having realized sales between K and K', the additional products would have realized revenue lying throughout the distribution. With complete unpredictability, a new set of products accounting for, say $x\%$ of total products would collectively garner $x\%$ of total sales.

This simple description provides a natural way to understand the female influx into authorship. The large increase in the tendency for women to write books, relative to men, might plausibly be rationalized by falling costs for women, relative to expected revenue and in comparison to the the costs for men. These costs could include both direct costs and opportunity costs. The factors attracting women in the labor force generally, and into book authorship in particular, during the 20^{th} century could stem from declining opportunity costs for women. As K for female authors has fallen, more women have project ideas for which they expect revenue to exceed costs.

Effects of a female influx on the various components of welfare depend not only on predictability but also on the substitutability of products for one another. Suppose that quality is completely unpredictable, so that the additional female-authored books attract their proportionate shares of usage. Then an influx will cause female authors to attract a higher share of revenue. While the fact that new products attract usage is sufficient to demonstrate effects on the distribution of revenue, effects on consumer surplus are different. If the new products simply displace equivalent products that would have entered absent the female influx, then the influx will have no effect on consumer surplus.

This framework leads to a series of empirical questions that I address below. First, how has the female influx affected the number of books available? Second, has the female influx attracted users to consumption, thereby affecting revenue and CS? Third, does the female influx change the value of the choice set, or does it simply displace otherwise equivalent male-authored works? Fourth, how large are the welfare effects of the female influx?

4 Data

The main data for this study, information on the supply of new books by year as well as consumers' usage of these books over time, are drawn from two sources, Bookstat and Goodreads. I also use data on book success derived from bestseller lists, the Library of Congress catalog, US copyright registrations, and nominations for major books prizes. Finally, I use Social Security and WIPO data mapping first names to gender to infer author genders.

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]

[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

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. I also match the authors with more than 10,000 instances of usage but non-matching names by hand.

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 distinct 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 have 8.86 million distinct titles in the data, and I am able to match first names of 87.7 percent of sales in the Bookstat data to the name-gender database.

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^f, where v refers to vintage.

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.

[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.

I also create this supply measure separately by genre: N^f_{vg}. I calculate the female-authored share of books from vintage v as $n^f_v = \frac{N^f_v}{N_v}$, where N^f_v is the number of female-authored books published in vintage v. 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^f_{tv} as the sales of female-authored books from vintage v during year t. Then s^f_{tv}, (where $(s^f_{tv} = \frac{q^f_{tv}}{q_{tv}})$ is the female-authored share of vintage v sales during year t. Analogously, s^f_{tvg} is the female-authored share of sales for vintage v books in genre g during year t, where $s^f_{tvg} = \frac{q^f_{tvg}}{q_{tvg}}$.

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 editions whose underlying titles were originally released in the vintage; I also have the numbers of editions for books written by women, men, and authors whose genders I cannot determine. For each vintage × 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 2021. For each edition, I have Amazon sales during 2021, the book's publication vintage, and author gender. Table 1 summarizes these data. Across vintages, an average of 22.3 percent of books are authored by women, and 31.2 percent of sales accrue to female-authored books. (Shares are higher when the denominator includes only gender-identified books).

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.

4.3 Goodreads data

Like Bookstat, the Goodreads data include a long list of (2.3 million) books with metadata (author name, genre, publication data), which I use to create measures of new supply by vintage, genre, and author gender. I am able to match 93.0 percent of the usage in Goodreads with name-gender information.

Rather than sales measures, the Goodreads data include 230 million interactions with the 2.3 million books made by 800,000 Goodreads users' 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. The data include 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 includes "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.

Table 3 summarizes these data. Across vintages, an average of 33.7 percent of books are authored by women, and 39.1 percent of sales accrue to female-authored books. (Shares are higher when the denominator includes only gender-identified books).

The Goodreads data include 10 genres: children, comics, fantasy, fiction, history, mystery, non-fiction, poetry, romance, young-adult, plus another called missing.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

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

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

according to above- or below-median usage of each of the ten Goodreads genres.

Restricting attention to observations with an author-gender match as well as valid data on publication dates, the data consist of 175 millions individual-book interactions (shelving or rating a book) on 1,821,381 distinct books (by 561,952 distinct authors). As with the Bookstat data, I use the Goodreads data to create measures of the female-authored shares of books from each vintage (n_v^f) as well as the share of usage garnered by the female-authored books from each vintage during each year (s_{vt}^f). I also create these measures by genre. Moreover, I create usage measures separately for different user groups (e.g. those whose book usage is high on female-authored books or high on uses of particular genres).

Table 4 shows two things. First, the table shows how 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. Second, genres also vary in the shares of sales garnered by women authors. The female authored shares of demand and supply are highly correlated across genres.

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 6,703,729 US book copyright registrations. Of these, I can match author names for 73.7 percent. Authors with female first names account for 42.6 percent of registrations between 1978 and 2020.[16]

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

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 (for books).[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 1,389 National Book Award nominations for 1960-2020.[18] Of these, I could match author names for 92.2 percent. I obtained data on 998 Pulitzer nominations for 1960-2020. Of these nominations, I can match author names to gender for 93.3 percent.

4.4.3 Published books in the Library of Congress

Not all books receiving copyright registrations are included in the Library of Congress (LOC) collection. Inclusion presumably 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."[19] Hence, we can use the female-authored share of the library's collection, by vintage, as a gauge of the importance of the female contributions to those vintages.

The LOC made the 2016 version of its card catalog publicly available as data.[20] The card catalog files contain 8.5 million records. Restricting attention to books published between 1960 and 2016, there are 7.4 million records. Of these, 71.1 percent have first names; and 75.7 percent of these match with the name database.

[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] 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.

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

4.4.4 Successful books

I have data on 750 annual New York Times fiction bestsellers for 1931-2020. I use the 44,276 listings for 1960-2020. These data include roughly 15 titles per week, along with author names, and I match 93.9 percent of listings with gender data.[21]

5 Evidence of the female influx and its effects

5.1 The female influx

My data 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. 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 female authorship levels and growth across all of the narrow genre shares in the Bookstat data based on regressions of 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.

It is clear not only that female participation in the creation of books has increased sub-

[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.

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

stantially but also that this increase occurs for all sorts of books. The increase in female production occurs not only for genres consumed primarily by women but also those consumed primarily by men. This raises the possibility that the female influx will raise the value of the choice set for a wide range of consumers.

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 "femaleness in supply," respectively. We want to know $\frac{\partial s_{tv}}{\partial n_v}$, 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 original-release vintages v, and run a regression of the form: $s_v = \alpha_0 + \alpha_1 n_v + \epsilon_v$. This approach asks whether vintages that are more female in supply are also more female in demand. Said another way, this approach asks 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 product entry rather than the object of interest, the impact of supply on demand.

We have a few alternative approaches for addressing this. First, we observe genre. Hence, 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, we avoid the problem of growing demand for genres driving 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 total sales s^f as the outcome, I can look at various kinds of positive 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 Pulitzer Prizes and National Book Awards? These awards are given to contemporary works, so that $v = t$ and the regressions take the form $w_v = \lambda_0 + \lambda_1 f_v + \varepsilon_v$, where w_v is the female share of nominees for some category of award for books published during v. Related, I look at the relationship between n^f and s^f across deciles of the sales distribution. Does the female influx deliver similar share of sales of female-authored sales throughout the distribution? Finally, and related, does the growing female share of authors appear in the right-tail of the sales distribution reflected in bestseller lists?

5.2.2 Results on $\frac{\partial s^f}{\partial n^f}$

Table 5 reports regression 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 1.48 (standard error = 0.07) and 1.24 (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.

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 across-vintage 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.

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, we can simply use data by genre; and columns (3) and (4) of Table 5 use data by time, vintage, and genre. These specifications include genre, time, and vintage fixed effects. The coefficient of interest falls in specifications from both datasets, to 1.01 (0.03) for Bookstat and 1.02 (0.03) for Goodreads. These results show that as the female share of supply rises, the female-authored share of demand rises proportionally.

Second, genre data allow us to estimate the coefficient of interest separately by genre; and Figures 5 and 6 reports these for the Bookstat and Goodreads data, 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. 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 5 indicate that the influx of female authorship is valuable to consumers. Moreover, the proportionality of s_{tv} to n_{tv} indicates that the additional female-authored books are as useful, on average, as the inframarginal female books. The results therefore suggests that the female influx is valuable and, in addition, that product quality is unpredictable.

5.2.3 Heterogeneous consumers

Columns (8) and (9) of Table 5 use the Goodreads measures of usage by user type (male- vs female-leaning). The coefficients of interest are similarly large for both male and female-leaning consumers. This indicates that the female influx's impact is felt similarly for heterogeneous users. Figure 7 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 7 reports the coefficient on n_{vt}^{f} from 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 genre usage. I divide users into heavy and light consumers of each genre, then I regress separate measures of s^{f} for the user groups on the female share of authors by vintage. Figure 8 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 groups derive substantial benefits from the female influx.

5.3 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 6 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 0.65 (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. In addition to the winners, the Pulitzer committee accounces a few nominees in each category; and the National Book Award committee announces a winner and four nominees for fiction, nonfiction, and poetry.[23] The female shares of award nominees rise substantially over this time period, from roughly 20 to 50 percent for the National Book Awards and from 20-30 percent for the Pulitzer Prizes. Columns (2)-(3) report regressions of female shares of award nominees on the female share of authors, by vintage. Coefficients for the Pulitzer Prices are indistinguishable from 1, as the coefficients for the National Book Awards.

Column (4) of Table 6 examines the impact of the female influx on the female-authored shares of bestsellers. Column (4) uses roughly 750 NYT fiction bestsellers per year for 1960-2020, and coefficient is 0.934 (0.09).

Finally, I also measure the separate impact of the female supply influx on the female-authored

[23] See https://www.pulitzer.org/prize-winners-by-year/ and https://www.nationalbook.org/national-book-awards/years/.

shares of sales across deciles of the sales distribution. As Figure 9 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.

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, is proportional to female supply shares. Not only is additional female participation valuable; it is 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.

5.4 Could additional male entry substitute for the female influx?

A growing number of female-authored books is a necessary, but not a sufficient, condition for the female influx to affect user welfare. It is possible, for example, that additional female-authored books displace male-authored entry that would otherwise have occurred. I explore this 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. 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 were 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. Not only does the female influx appear not to displace male-authored entry, the female-authored books are on average more appealing to consumers.

6 Quantifying welfare effects

The evidence above clearly indicates that the influx of books by female authors is valuable to consumers and generates revenue for female authors. Moreover, the rough proportionality of the female-authored sales share with the female-authored share of new products is consistent with new product quality being highly unpredictable. Finally, the additional female books appear to be net additions to the choice set. These observations motivate a simple approach for measuring the welfare benefit of the female influx: Compare a recent post-female-influx choice set with a counterfactual choice set removing female-authored books at random to make the female-authored share resemble its level in pre-1970 vintages.

What's required is a way to calculate the revenue, and consumer surplus, associated with any choice set. To this end, I employ an easily-calibrated nested logit model of demand. During some some time period consumers choose among J books in the choice set, or they choose the outside good. (For the sake of notational simplicity, I introduce the functional form of demand now). 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 $\delta_j = x_j\beta - \alpha p_j + \xi_j$. Given a choice set characterized by a set of product qualities $\{\delta_j\}$, I calculate the usage of each product q_j, as well as the CS for the choice set.

Given an estimate of σ, I can construct an estimate of mean utility: $\delta_j = \ln(s_j) - \ln(s_0) - \sigma ln(s_{j|in})$. This, along with a specification of which products j are included in the choice set, allows calculation of CS and revenue for any choice set. That is, $CS = \frac{M}{\alpha}\ln(1 + (\Sigma e^{\delta_j/(1-\sigma)})^{1-\sigma}$, and $REV = \Sigma p_j \frac{e^{\delta_j/(1-\sigma)}}{D}\frac{D^{1-\sigma}}{1+D^{1-\sigma}}$, where $D = \Sigma e^{\delta_j/(1-\sigma)}$, where $M =$ market size.

Calculation of CS requires an estimate of the price parameter α, but the proportional change in CS does not depend on the price parameter. That is, $\frac{CS}{CS_0}$ (where CS refers to the status quo, and CS_0 describes the counterfactual without the female influx) is invariant with respect to α. Instead, the change in CS depends on the parameter σ along with the full distribution of product qualities in the baseline and counterfactual scenarios. While this aproach seems very parsimonious, it is important to note that CS and revenue in the status quo and counterfactual simulations depend on the distributions of product qualities included. Given an estimate of σ, both Bookstat and Goodreads datasets allow calculation of $\%\Delta CS$. The Goodreads data allow separate calculation of $\%\Delta CS$ for different types of consumers.

In principle, my data allow estimation of σ. Credible inference on σ requires plausibly exogenous variation in the number of products in the choice set across time, to see the impact of more or fewer products on the tendency for consumers to purchase books overall. It is not clear whether the data at hand cover a context with credibly exogenous variation in the number of products. The existing literature provides some evidence about the substitution parameter relevant to books. Reimers and Waldfogel (2021) presents an estimate of 0.373, which I use as a baseline; I also show how all measures of $\%\Delta CS$ vary across a wide range of possible estimates of σ.

With the Bookstat data, I compare the 2021 choice set of ebook and print editions with a counterfactual choice set in which books are removed to simulate a choice set in which the female influx did not occur. In particular, I calculate the ratio of female to male-authored books published during the 1960s. For each vintage after 1970, I keep just enough female-authored books to maintain the ratio for that publication year. I do the same thing for Goodreads data using the 2016 choice set For these calculations, I take the conservative approach of treating the books for which author gender attribution is not possible as male-authored.

Table 8 shows how CS and revenue change with the female influx present, relative to counterfactual environments in which it did not happen. At the baseline substitution parameter, CS rises by 19.57 percent using Bookstat and by 26.77 percent using Goodreads. Also at the baseline, revenue rises by 11.88 percent using Bookstat and 21.65 percent using Goodreads. Both datasets also allow calculation of the distinct changes in the revenue accruing to male and female-authored books. At the baseline, female author revenue rises by 183.46 percent in Bookstat and 126.74 percent in Goodreads. Also at the baseline, male author revenue declines by 26.96 percent using Bookstat and 22.65 percent using Goodreads. Finally, the Goodreads data allow us to calculate separate effects on different groups of consumers differing in the extent to which they read female-authored books. As Table 8 shows, CS rises for both male-leaning and female-leaning consumers. 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 20 percent increase in CS, while light users obtain a 35 percent increase. At the other extreme, heavy users of romance derive a 37 percent increase in CS while light users of romance derive less than 20 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.

7 Robustness

7.1 Varying substitution parameters σ

Estimates above are reported for a particular value of the substitution parameter σ. It is of interest to know how the study's results vary for different degrees of substitutability among books. Figures 11 - 13 show how the welfare estimates vary with different estimates of the substitution parameter σ. For a wide range of σ values, the effects of these substantial increases in the numbers of products in the choice set are large.

7.2 Imperfect predictability

The foregoing analyses assume no quality predictability at entry. This means that an influx of new entry brings additional products as good, on average, as the products already entering.

The products removed from the status quo choice set are randomly chosen. This is reasonable given the results showing proportionality of s^f and n_f. Still, we can explore the consequences of quality predictability for the welfare estimates. We assume that products enter when their expected revenue exceeds costs. With greater predictability, growth in entry delivers additional products less valuable than those otherwise entering. In the extreme, with perfect predictability, all additional products would be worse than the marginal previously entering product. Below I develop a simple model showing how welfare effects change as predictability rises.

Define the expected quality of product j as $\delta_j^{'} = \delta_j + \kappa\varepsilon_j$, where δ_j is the true (realized) quality of product j, ε_j is a standard normal error, and κ is a scaling parameter. When $\kappa = 0$, quality is perfectly predictable. As κ grows less predictable; when κ is very large, quality is completely unpredictable.

The parameter κ thus affects which products are removed in the no-female-influx counterfactual. In general, the counterfactual removes lower-expected-quality products. When κ is large, the lower-expected-value products have high realized value, with the consequence that the female influx gives rise to large increases in CS and overall revenue. As κ declines toward zero, the lower-expected-value products have lower realized value. Then the influx gives rise to smaller changes in CS and overall revenue.

Figure 14 shows how $\%\Delta CS$ and $\%\Delta rev$ evolve with predictability, separately for Bookstat and Goodreads samples. The horizontal axes show the correlation ρ of expected and realized quality (corr(δ_j,$\delta_j^{`}$)). When κ is zero (perfect predictability), the correlation is zero; as κ increases, the correlation falls. The baseline case – no predictability – has $\rho = 0$; and $\%\Delta CS$ is 18-25 percent, while $\%\Delta rev$ is 12-20 percent. As ρ increases, these both fall. For example, when $\rho = 0.5$, $\%\Delta CS$ is 6-9 percent and $\%\Delta rev$ is 3-5 percent. It is worth noting that the evidence in Section 5, that increases in the female shares of sales are proportional to increases in the female shares of works, is consistent with little or no quality predictability. Hence, it is likely that the welfare effects of the female influx are substantial.

8 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 roughly a quarter to 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 findings (Hsieh et al., 2019; Bell et al., 2019) indicating that inclusion is beneficial for innovation and growth. Importantly, the influx of new products from female authors benefits not only women but also men in their capacity as readers. Books provide an unusually good example. 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.

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,” .

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.

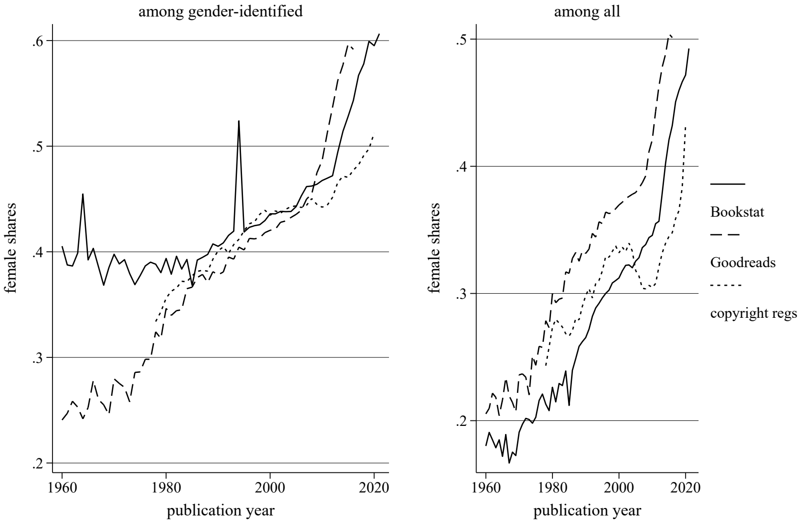
Wan, M. and J. McAuley (2018): “Item recommendation on monotonic behavior chains,” in *Proceedings of the 12th ACM conference on recommender systems*, 86–94.

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.

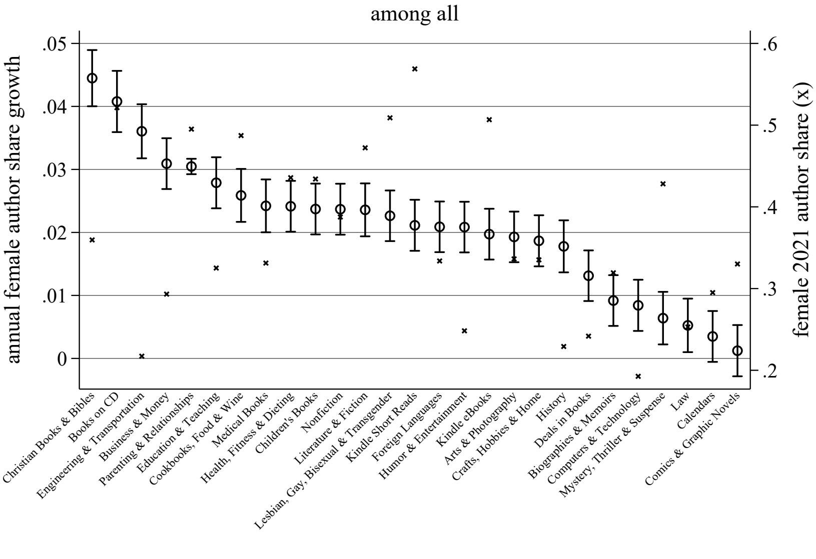
A Figures and Tables

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



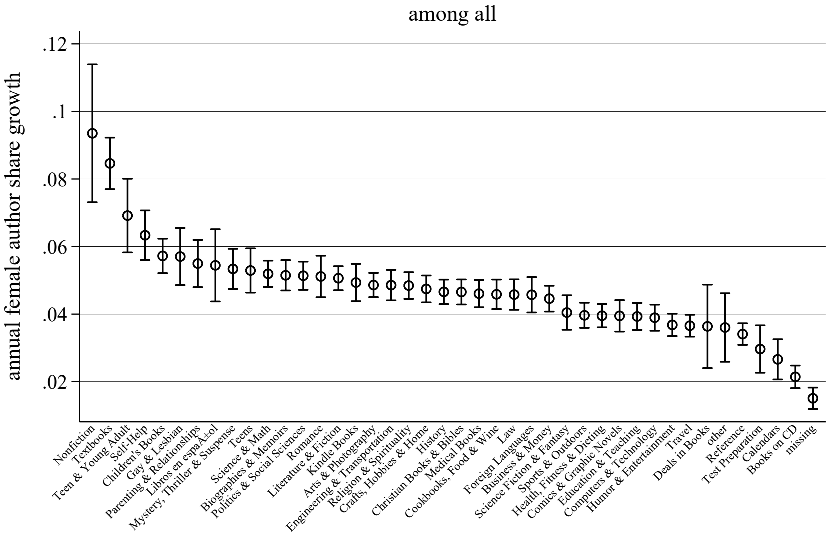
Notes: Female-authored share of books, by publication year, in Goodreads, Bookstat, and US copyright registration data. The female shares in the left figure are shares of books with identified author genders. In the right figure, books whose authors are not gender-identified are treated as male-authored.

Figure 3: Growth in female-authored share by genre (Bookstat)



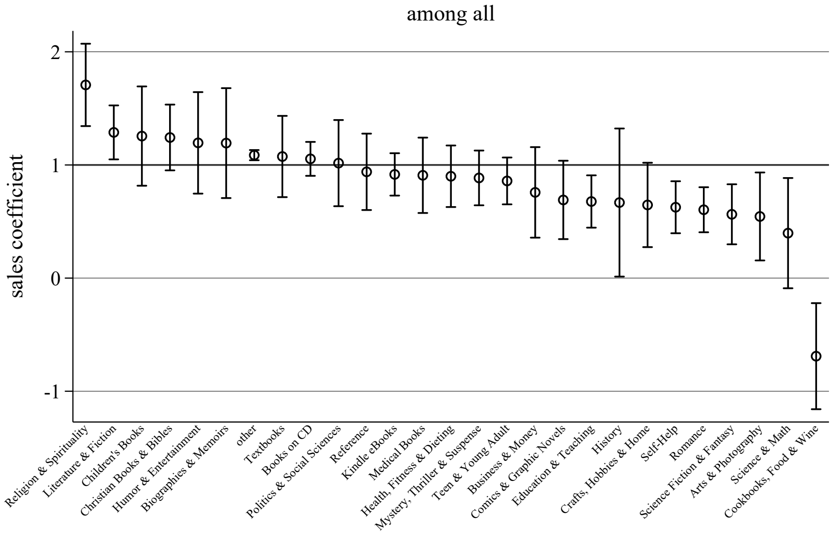
Notes: Growth is coefficient of a regression of log female share in a vintage on vintage. The level (x) is the 2021 vintage level.

Figure 4: Genre sales growth (Bookstat)



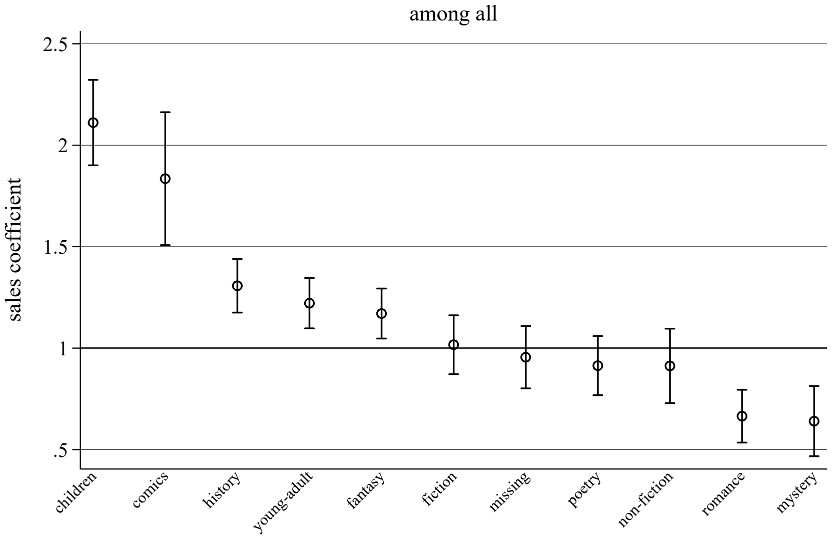
Notes: Genre-specific coefficients from regression of genre's share of vintage sales on vintage. The regression includes genre fixed effects.

Figure 5: Coefficient of female share of sales on female share of works (Bookstat)



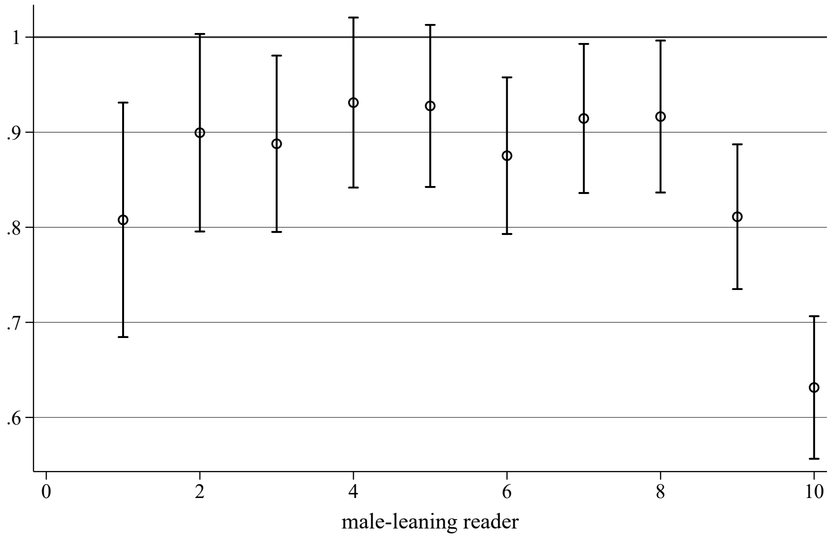
Notes: Genre-specific coefficients from regression of female share of sales on female share of works. The regression includes genre and year fixed effects.

Figure 6: Coefficient of female share of sales on female share of works (Goodreads)



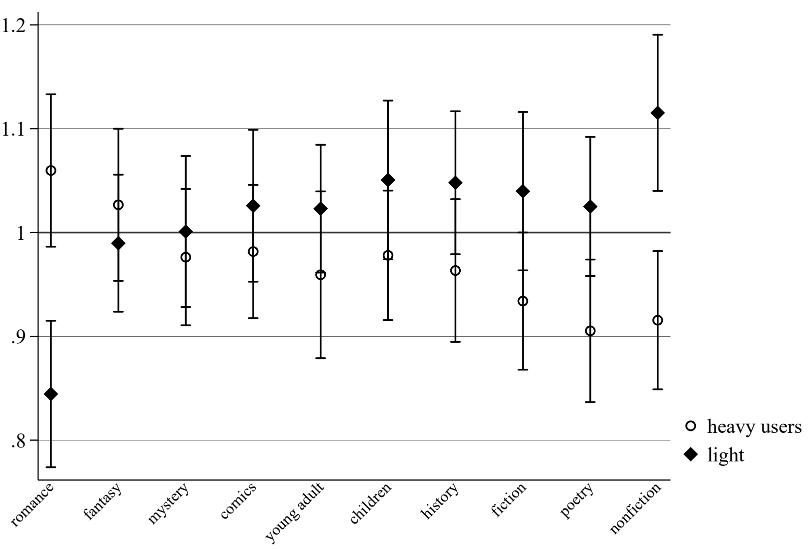
Notes: Genre-specific coefficients from regression of female share of sales on female share of works. The regression includes genre and year fixed effects.

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



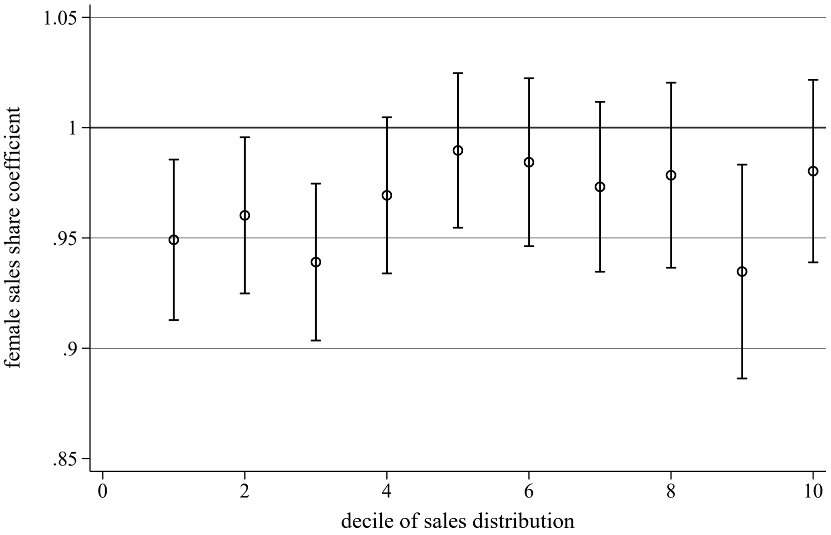
Notes: Reader types are deciles of readers according to the male-authored share of their usage. The regression includes genre and year fixed effects.

Figure 8: Coefficient of % female sales on female works % by reader types (Goodreads)



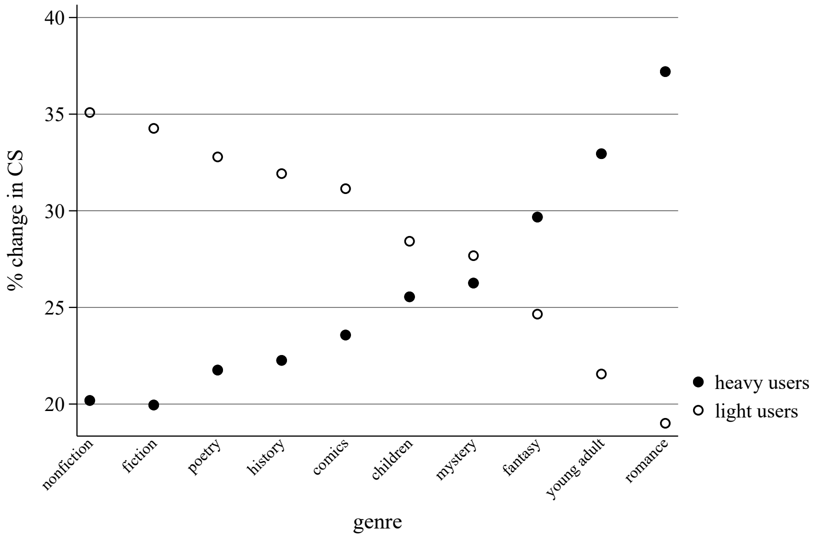
Notes: Heavy users are above-median users of the genre.

Figure 9: Coefficient of % female sales on female works % by sales decile (Bookstat)



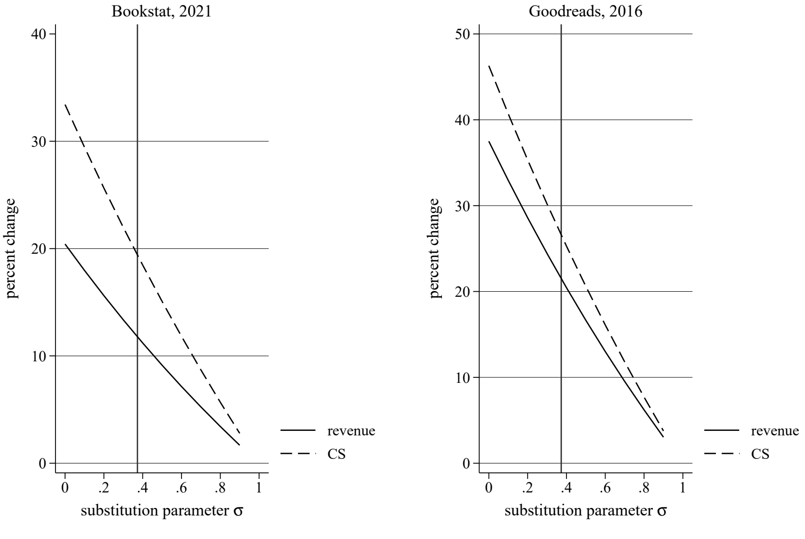
Notes: .

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



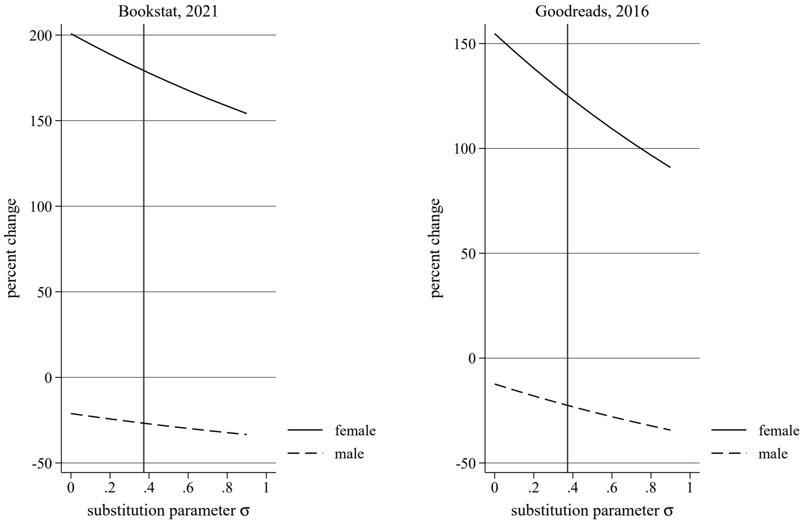
Notes: The figure shows how heavy vs light genre user-specific welfare estimates vary for different genres. For example, the leftmost dots shows that light users of nonfiction derive a 35 percent increase in CS from the female influx, while heavy users derive a 20 percent increase.

Figure 11: Substitution and the effect of female authorship expansion on revenue and CS



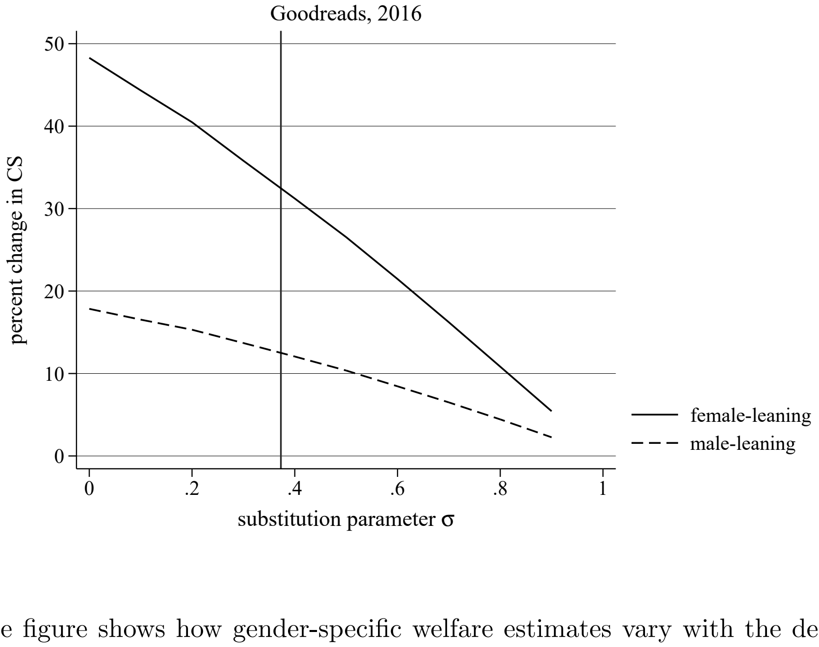
Notes: The figure shows how welfare estimates vary with the degree of substitutability among books (σ). Overall welfare effects are smaller, the more substitutable are books.

Figure 12: Substitution and the effect of female authorship male and female author revenue



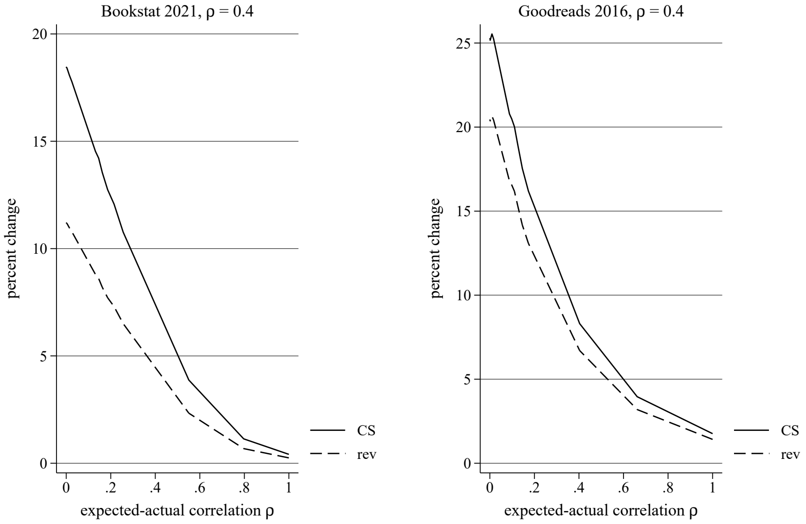
Notes: The figures show how male and female author revenue vary with the female influx, for varying degrees of substitutability among books.

Figure 13: Effect of female authorship expansion on CS by reader gender



Notes: The figure shows how gender-specific welfare estimates vary with the degree of substitutability among books (σ). "Male-leaning" refers to users who read predominantly male-authored books. Male authors are those with male first names.

Figure 14: Welfare effects and quality predictability



Notes: The figure shows how welfare estimates vary with product quality predictability. The horizontal axis depicts the correlation between expected and realized product quality. The figure is drawn for $\sigma = 0.4$.

Table 1: Summary statistics (Bookstat)

	vintage × year		vintage × year × genre	
	mean	sd	mean	sd
fem-aut'd % of id'd eds	28.1	8.7	30.7	18.9
fem-aut'd % of id'd sales	34.0	13.4	31.6	27.8
fem-aut'd % of eds	22.3	7.2	24.9	16.7
fem-aut'd % of sales	31.2	11.8	27.5	26.1
pub year (vintage)	1,990.50	17.9	1,992.60	17.5
observations	248		9555	

Notes: The first two columns describe the vintage × year data based on sales 2018-2021 of books published 1960-2021. The latter two columns repeat the exercise for vintage × year × genre cells.

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:

Table 3: Summary statistics (Goodreads)

	vintage × year		vintage × year × genre	
	mean	sd	mean	sd
fem-aut'd % of id'd titles	37.9	9.5	41.6	20.7
fem-aut'd % of id'd sales	41.7	14.8	42.9	24.5
fem-aut'd % of total titles	33.7	8.7	37.1	19.9
fem-aut'd % of total sales	39.1	14.0	38.7	24.1
pub year	1,988	16.5	1,988	16.4
observations	570		6260	

Notes: The first two columns describe the vintage × year data based on usage 2007-2016 of books published 1960-2021. The latter two columns repeat the exercise for vintage × year × genre cells.

Table 4: Consumption by genre and readers' gender leanings: 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: Female-leaning readers are those whose shelves include more than the median share of female-authored books.

Table 5: 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	1.482***	1.244***	1.008***	1.017***	0.918***	0.996***	1.188***	0.955***	0.923***
	(0.0727)	(0.0422)	(0.0250)	(0.0327)	(0.0622)	(0.0306)	(0.0674)	(0.0406)	(0.0337)
Observations	248	570	9320	6198	2657	4317	2346	6182	6197
$\overline{R}^2$	0.625	0.609	0.463	0.715	0.286	0.481	0.519	0.604	0.620

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 6: Female authorship and recognition regressions

	(1) LOC %fem	(2) Pul %fem	(3) NBA %fem	(4) NYT %fem
% fem-aut'd	0.648***	0.990***	1.201***	0.934***
	(0.0229)	(0.189)	(0.179)	(0.0917)
Observations	1018	408	185	61
$\overline{R^2}$	0.630	0.120	0.211	0.631

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 7: Male author entry displacement

	(1) OLS	(2) genre FE	(3) vintage FE	(4) all FE
female-authored books	0.324***	0.290***	0.296***	0.262***
	(0.00593)	(0.00611)	(0.00526)	(0.00538)
Unknown gender-authored books	1.098***	1.226***	0.935***	1.046***
	(0.0103)	(0.0111)	(0.00967)	(0.0106)
Observations	16976	16323	16563	15975
$\overline{R^2}$	0.881	0.880	0.910	0.910

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

	% change in CS	
	Bookstat	Goodreads
all	19.57	26.77
male-leaning consumers		15.34
female-leaning consumers		41.3
	% change in revenue	
all	11.88	21.65
male authors	-26.96	-22.65
female authors	183.46	126.74

Notes: Model simulations based on $\sigma = 0.373$ and random removal ($\rho = 0$). The Bookstat column compares 2021 with a hypothetical alternative in which female-authored books are removed to mimic the extent of female-authored entry during the 1960s. The Goodreads column does an analogous exercise for the 2016 Goodreads usage data. Consumers are classified as "male"- or "female-leaning" according to the female-authored share of the books they use.