Canadian Labour Economics Forum

WORKING PAPER SERIES

Learning in Creative Tasks: Evidence from a Digital Platform

Jiatong Zhong (University of Alberta)

CLEF WP #83

Learning in Creative Tasks: Evidence from a Digital Platform

Jiatong Zhong*

This version: December 3, 2024

Abstract

This paper explores learning-by-doing in the context of creative tasks using detailed data on fiction writers from a digital publishing platform. I construct measures to quantify authors' performance over time and document significant variation in both starting levels and rates of improvement. Learning manifests as an improvement in quality instead of the speed of production. Quality improvement can last for several years, much longer than typically observed in manufacturing settings, and authors do not improve fastest at the beginning. These findings show that human capital accumulation has distinctive features in the context of creative and complex tasks, which institutions should consider in the training and evaluation of new workers in creative occupations.

Keywords: learning by doing, learning curves, creative tasks
JEL Classification: D83, J24, J46, L82

*Corresponding author. Department of Economics, University of Alberta. jzhong5@ualberta.ca

1 Introduction

Learning by Doing (LBD) is an economic phenomenon that associates productivity improvement with the accumulation of experience for individuals or organizations. At the country level, learning by doing is considered a source of total factor productivity improvement, which leads to long-run sustainable growth (Arrow 1962; Romer 1986). At the organizational level, learning by doing reduces worker separation rates and the cost of production. At the individual level, learning by doing is a form of human capital accumulation that improves wages and the outcomes of tasks.[1] Empirical studies have documented learning by doing in a variety of settings in the manufacturing sector, but there is relatively sparse evidence of individual learning by doing on creative and non-routine tasks.

With the rise of digital platforms that allow individuals to reach a large audience, more individuals have replaced firms as the units of production, especially in the provision of services and creative products.[2] It is, therefore, crucial to understand how individuals improve with experience in creative tasks, as learning drives productivity growth. Recent advances in generative Artificial Intelligence (AI) also prompt us to ask how fast and how well we can learn a complex and creative skill, as the differential between human and AI improvement determines whether AI will complement or substitute labour in the task (Acemoglu 2024). The effects of LBD in routine and relatively simple tasks cannot be readily generalized to non-routine and complex tasks because theoretical arguments have suggested that complex tasks take longer to learn (Jovanovic and Nyarko 1995). Empirically, we still know little about how individuals learn in creative tasks because the value of creative output is hard to quantify, and few datasets track these outcomes over time.

This paper provides new evidence on learning by doing in creative tasks using detailed records of authors writing for profit on a digital publishing platform from 2005 to 2022. The data set includes information on authors and each book they wrote.[3] This dataset offers an excellent context for studying individual learning by doing. Existing empirical literature on learning-by-doing seeks to disentangle the impact of individual experience on cost reduction or quality improvement from confounding factors like the increase in capital investment, changes in input prices, and organizational learning. The digital publishing industry has sev-

[1] There are many examples in various contexts, and the literature is too vast to be fully cited here. A few recent examples include Levitt et al. (2013); Haggag et al. (2017); Ost (2014); Rocha et al. (2015), and the survey by Thompson (2010) provides many more.

[2] For example, an individual can upload promotional content to their social media platform, effectively producing a service that was previously provided by advertising agencies.

[3] Here, information on authors refers to the information about their books available on the publishing platform rather than their personal information.

eral attractive features that mitigate these concerns. First, creative writing is highly labour-intensive, and the capital requirement for each agent is relatively small, so the capital stock does not affect productivity and the quality of output. Second, price is fixed and fully observable, so strategic pricing is irrelevant in this industry and will not affect productivity in any period. The lack of price fluctuation also means that productivity or quality does not fluctuate with quantity supplied responding to prices. Third, in a traditional setting, an organization can improve organizational-level productivity without individual learning by reallocating workers or displacing the least efficient workers. In our context, the digital platform neither reallocates nor displaces authors, and the authors make independent entry-exit decisions.[4]

This paper performs analysis on a sample containing 57,206 books written by 6,561 authors who have more than 1,000 followers by November 9, 2022. This criterion screened out authors who have only written briefly and cannot be used to identify individual learning.[5] An author's experience is measured by the number of characters they have written before a new book starts. I use reader retention rates as the main learning outcomes, as data suggest that this measure can reflect book quality regardless of an author's accumulated reputation. I supplement the confidential records with two data sets scraped from the website, where one includes the universe of all authors who signed a contract with the platform, and the other tracks the initial reader growth for a set of new books uploaded between January and February 2021.

I find that book quality improves as authors accumulate more writing experience, and learning continues up to when the author has written at least five million characters, which can take three to five years even for the most prolific writers. As predicted by Jovanovic and Nyarko (1995), this duration is much longer than what has been found in manufacturing settings or nonroutine manual tasks, where learning typically plateaus in less than six months (Nagypál 2007; Levitt et al. 2013; Haggag et al. 2017). The magnitude of improvement, however, is small—the reader retention rate increases by 7.6 percent after five million characters. I also find patterns of learning that contrast with previous findings in other contexts: new authors do not learn fastest initially, and the improvement is exclusively in the quality of output, not the quantity produced per unit of time. Learning is primarily driven by general writing experience instead of genre-specific experience.

Author heterogeneity affects both the speed and magnitude of improvement. I find that learning primarily occurred among authors whose first book is in the bottom two quality

[4] If an author violates the platform's guidelines, the platform can terminate their contracts. Such instances are rare and typically involve serious offences like plagiarism.

[5] The data set for the main analysis is confidential and provided by the digital platform. The platform only provided book-level information for authors with at least 1000 followers.

quartiles. Writing habits also affect authors' speed of learning: authors who wait longer between books learn faster. Surprisingly, authors who write more per day do not necessarily learn faster after controlling for experience.

Finally, I examine alternative mechanisms that can explain the improvement of book performance as authors accumulate experience: survival bias and the learning of consumer preference. My results rule out both explanations as the main reason for the learning patterns we observe.

This paper contributes to a growing empirical literature on individual learning by doing.[6] Levitt et al. (2013) and Hendel and Spiegel (2014) estimate the effect of learning by individual agents performing routine tasks in a manufacturing setting.[7] This paper provides evidence of learning in a creative task and shows that the learning patterns are quite different in manufacturing and our context. The literature has also studied learning in non-routine tasks, such as entrepreneurship, teaching, and taxi driving. Studies on serial entrepreneurs found that learning contributes to a higher survival rate of new businesses (Rocha et al. 2015; Lafontaine and Shaw 2016). This paper moves beyond the existence of learning to describe the dynamic of the process. The literature on teachers' learning has found strong evidence that students benefit from having experienced teachers (Jackson and Bruegmann 2009; Ost 2014). Yet even recent studies using matched teacher-student panel data offer limited insights into what individual behaviours enhance learning. This paper offers new evidence on the relationship between learning and individuals' work habits, taking advantage of detailed records of outcomes and production history. Haggag et al. (2017) documented how New York taxi drivers improve their income by accumulating neighbourhood-specific knowledge. This paper found that, among nonroutine tasks, the learning patterns of manual and cognitive tasks have important distinctions.[8] To the best of my knowledge, there is only one paper that exploits the writing history of individuals to document how performance evolves with experience. Yu et al. (2023) studies the life cycle of academic researchers in the biomedical field by collecting their publications and citations. Their paper focuses on the life-cycle productivity, while this paper focuses on the trajectory of improvement.

In addition to the learning-by-doing literature, this paper provides new evidence to a broader literature on human capital accumulation and the return of experience. Adda and Dustmann (2023) examined the wage dynamics over the life cycle of workers using task-

[6] There are more papers on learning by doing in general than what I can discuss here. Thompson (2010) provides a good review of this literature.

[7] Levitt et al. (2013) studies the productivity and quality improvement of workers in automobile assembly lines. Hendel and Spiegel (2014) document productivity growth in a steel mini mill.

[8] I adopt the definition of routine, nonroutine, manual, and cognitive tasks from Autor et al. (2003). Driving is categorized as a nonroutine manual task in Autor et al. (2003) while writing as a nonroutine cognitive task.

specific measures of experience. Gathmann and Schönberg (2010) addresses the costs of moving across occupations for workers. This paper complements previous works by dissecting the dynamic of human capital accumulation within a job and shedding light on the time cost of building human capital within a skilled and non-routine occupation.

This paper is organized as follows. In section 2, I introduce the market of online publishing. Section 3 describes the datasets and data patterns. Section 4 introduces the empirical strategy, and section 5 presents the main results. Section 6 discusses alternative mechanisms through which the main results can emerge, and section 7 concludes.

2 The market of online publishing

With increasing access to digital technology, online literature—novels and other forms of writing available primarily in digital formats—has grown from a somewhat niche category to a significant sector engaging nearly half of China's internet users. By 2023, the sector had a market value of 5.62 billion U.S. dollars, with 537 million active users in China(CASS 2023). In our context, online publishing is narrowly defined as publishing on digital platforms for profit.

This paper used data from one of the largest Chinese online literature publishing platforms. Unlike e-commerce platforms such as Amazon, which typically sells entire books, the online publishing platform used in this study sells books chapter by chapter. Authors upload their work to the platform anonymously, usually in the form of novel serials.[9] Authors don't set their own prices as they would in traditional publishing, and there is almost no barrier to entry: anyone with internet access can register and publish under a pen name. For an author to get paid for their work, however, they need a contract with the company running the platform.

To sign a contract, a new author can submit a writing sample, typically the beginning of a story, to an editor hired by the platform. The editor then reviews the sample and decides whether to offer a contract.[10] Although the signing process is selective, there are no requirements on the author's educational attainment, hours of work, prior experience, or even age.[11]

Authors on this platform earn income from three main sources: subscriptions, patronage, and royalties. Subscription provides the most reliable income stream for most authors. Readers are initially provided free access to the first 20-30 chapters of a novel and, if in-

[9] Authors provide their legal names when signing contracts with the platform, but their personal information remains confidential and is not used in this paper.

[10] Authors who were declined can re-apply as many times as they want. Contracts are standardized, and new authors typically cannot negotiate terms.

[11] An author under the age of 18 can sign a contract with the website under the supervision of their guardians.

terested, can subscribe to access additional pay-to-read chapters (also called "VIP chapters" henceforth) at a flat rate of 0.03 yuan per 1000 characters. Patronage offers an additional income stream, with readers making voluntary donations starting at 1 yuan. The donation is common, but the value is usually small: 98.3% of books received some patronage, and the median patronage value is 109 yuan (16.2 U.S. dollars).[12] The most successful authors also earn royalties from physical book sales, audiobooks, or even television and movie adaptations. Selling adaptation is rare: only 8.6% of books from authors with over 1000 followers have sold physical copies or some form of adaptation. The publishing website retains 50% of all the above-mentioned streams of income. In return, they offer editorial services, legal services, and a platform to reach a large audience.

3 Data Description

The data used in this paper is derived from three main sources. The first of these sources is confidential firm records that cover all contracted authors with over 1000 followers. These records provide a cross-sectional view of the chapter-by-chapter performance of every book by authors in the sample, including confidential variables that allow me to measure their market performance. Since the firm records only cover a subset of authors, I supplement them with two datasets constructed by scraping the online publishing website. The first supplementary dataset covers all authors who have a contract with the website, but it does not include confidential variables used to construct quality measures. The second supplementary dataset tracks the performance of a set of new books. The empirical analysis in Section 4 uses the confidential data set exclusively, while Section 3.4 uses the scraped data sets to provide a more comprehensive overview of author and book characteristics.

3.1 Confidential records

The confidential data set was collected and compiled by the online publishing platform. It covers information for all contracted authors with over 1000 followers as of November 10, 2022. The sample contains 6,561 authors and 57,206 books written between February 21, 2005 and November 9, 2022. This dataset provides detailed information on authors' activities and their books' performance. The publicly available variables describe the genre, the number of chapters, word counts by chapter, the status of completion, clicks on free chapters, number of reviews and scores, amount of donations contributed by its readers, the last time each chapter is edited, and the number of followers of each author and every book. In addition

[12] I use the Chinese Yuan to U.S. dollar exchange rate in 2022, reported by the Federal Reserve Bank of St. Louis, which is 6.729 yuan = 1 dollar.

Table 1: Summary Statistics from Confidential Data

	>=5000 followers			1000–4999 followers			All >=1000 followers		
	N=1972			N=4509			N=6481		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Panel I: Author level summary statistics									
Num. of characters (millions)	5.25	4.16	4	2.53	2.05	2	3.36	2.51	3.03
Experience (years)	7.48	6.58	3.47	6.42	5.6	3.61	6.74	5.88	3.6
Avg. characters/year (millions)	0.75	0.67	0.48	0.5	0.41	0.39	0.58	0.48	0.43
Followers	27004.34	10934.5	65591.15	2224.84	1914	1046.81	9764.6	2724	37938.78
Number of books	12.45	11	8.16	7.22	6	4.89	8.81	7	6.53
Number of complete books	11.83	10	8.08	6.76	6	4.79	8.3	7	6.43
Days between books	204.63	150.53	207.12	220.02	146.88	287.69	215.25	148.5	265.39
Avg. characters/book (millions)	0.44	0.39	0.24	0.37	0.32	0.19	0.39	0.34	0.21

Panel II: Book level summary statistics for authors with >1000 followers

	>=5000 followers			<5000 followers			All >=1000		
	# books=24546			# books=32646			# books=57192		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Book following (thousands)	27.17	10.11	62.38	7.33	3.88	10.24	15.85	5.56	42.73
Number of reviews (thousands)	11.87	3.69	35.34	2.24	1.09	3.96	6.37	1.73	23.83
Number of subscribers (thousands)	7.7255	3.333	16.33584	2.27069	1.248	3.05839	4.61	1.88	11.28
Patronage values (CNY)	3456.61	261	20852.82	889.24	61	4946.86	1991.12	109.00	14219.92
Number of patrons	174.32	47	800.15	31.28	15	52.2	92.67	23.00	530.42
First chapter clicks (thousands)	142.49	76.33	229.11	48.91	31.84	52.97	89.07	44.90	162.10
Complete? (Y=1)	0.95	1	0.22	0.93	1	0.25	0.94	1.00	0.24
Click-follow ratio	12.09	6.91	171.15	10.11	7.56	8.53	10.96	7.26	112.31
VIP chp10/1st chp click	0.14	0.12	0.09	0.12	0.11	0.09	0.13	0.11	0.09
Good comment ratio	0.94	0.94	0.04	0.93	0.94	0.05	0.93	0.94	0.05
Subscribers/1st chp click	0.05	0.05	0.04	0.05	0.04	0.05	0.05	0.04	0.05

Note: The sample consists of authors who have finished at least one book. Titles with no content were dropped when calculating the average length per title. The book sample contains novels, novellas, and short stories written by authors with more than 1000 followers. "Days between books" reports the average number of days between the first upload dates of two consecutive books by the same author.

to public variables, this dataset includes confidential market performance measures: the total number of subscribers and the clicks of the first ten pay-to-read chapters. The subscriber count measures the number of readers who have purchased at least one chapter of this book, and the clicks measure the traffic to each of the pay-to-read chapters. Although these variables do not directly reveal the author's income, they provide meaningful indicators for how commercially successful each book is.[13]

Panel II of Table 1 presents differences in books written by authors with more than 1,000

[13]Since the website charges each reader a fixed price per 1000 characters and the word count per VIP chapter is public information, knowing the number of subscribers per chapter will allow one to infer the author's subscription income from this book. To protect the authors' privacy, the platform only provided part of the information to prevent researchers from uncovering the authors' actual income streams.

followers. I further divide these books into titles by authors with over 5,000 followers ("top authors") and those with 1,000 to 4,999 followers. Top authors have a more engaged reader base, as indicated by an average reviews-to-following ratio of 0.44 compared to 0.31. Their readers are also more generous: among top authors, it takes an average of 41.23 readers to generate one patron, compared to 55.02 among the rest. Overall, top authors outperform the rest by a substantial margin across nearly all data metrics.

The gap narrows, however, when we look at the confidential variables that involve reader purchases. The ratio of clicks of the tenth pay-to-read chapter to the first chapter, which captures the fraction of readers who are willing to buy at least ten chapters, averages 0.14 among top authors and 0.13 in the full sample. While top authors still outperform in general, the gap between groups is much smaller despite substantial variation of the ratio within each group (standard deviations = 0.09). This pattern suggests that a ratio that measures paying reader retention may be a better metric for book quality, as high reader following does not necessarily lead to higher retention rates.

3.2 Public information of all signed authors

The first supplementary dataset was constructed by scraping the publicly available information on all authors who have signed a contract with the website as of January 7, 2021, and the books they uploaded from August 2003 to January 2021. Starting with a full sample of 415,629 pieces of writing uploaded by 48,110 contracted authors, I excluded all essays and book reviews and kept only novels, novellas, and short stories. I then limited the sample to authors with at least one complete novel, reducing the dataset to 33,540 authors and 323,264 novels. For each author, I tracked the time they uploaded each novel, the last time they edited each novel, their follower count, and how many books and how many characters they have written.

In addition to author-level information, I tracked book-level information for every book written by authors with over 1,000 followers by January 7, 2021. For each book, I tracked the list of public variables described in Section 3.1. There are a total of 92,681 titles, among which only 57,462 have complete book-level performance statistics. Among the 35,219 titles that are missing following, reviews, and click information, 46.8% are removed by the authors, and the rest are preview titles.[14] I can observe some book-level variables for removed titles, such as length, genre, and completion status, but not the chapter-level variables like chapter clicks. Among the books that the authors removed, 70% of the titles are complete. I collected

[14]Most of the titles are removed per the platform's content guidelines. Titles that contain content deemed inappropriate for younger audiences (e.g., violence, suicide, or drug use) are censored and their authors can revise or remove them. Some books are removed due to copyright concerns.

book-level and author-level variables based on the titles that I can obtain information on, and the removed titles are only kept in the dataset so I can correctly account for an author's experience.

Author experience is calculated for each book and can be measured in time or quantity produced. The time-based measure is the difference between the day this author uploaded for the first time on this platform and the upload date of the first chapter of this book. The quantity-based measure is the number of Chinese characters this author has written on this platform before they uploaded the first chapter of this book. When measuring experience, I include incomplete books, essays, book reviews, and removed titles. It is worth noting that the contracts offered by this digital platform mandate that every novel and novella they write must be published on this website exclusively. This requirement implies that the count of characters published on this platform captures the complete novel-writing experience of the authors from the time they signed the contract to the time of observation.[15]

Table 1 summarizes key features of the confidential data set and Table 2 summarizes the supplementary public data set. The confidential data includes more authors with over 1000 followers than the public data because it covers almost two more years of data, and the author's following and the number of contracted authors grow over time. However, the confidential data contains fewer books if we compare Panel II of Table 2 with Panel II of Table 1. The discrepancy is mainly due to preview titles in the public dataset. A preview title is a "placeholder book" where authors describe an idea or premise without developing substantial content. The authors typically put the titles there to accumulate followers and assess market interest in several potential new titles. Some also use the time stamps on record to prevent their idea from being "scooped" by other authors. The confidential data set does not contain preview titles and they are excluded from all data analysis in this paper.[16] If we focus on complete titles in both data sets, 94 percent of titles in the confidential data set are marked as completed (53,788), while only 62 percent of titles in the public dataset are (57,343). Overall, the preview titles explain most of the gap between the number of books in the data set I compiled and the data provided by the online publishing platform.

[15]The platform allows authors to publish creative writing pieces shorter than 30,000 characters on any media form of their choice. Pieces of that length are usually short stories, poetry, prose, essays, or reviews.

[16]The preview titles have a special tag on the web page. New books that are still short will not be mistaken for a preview title as the author will need to remove the "preview" tag to give readers access to their content. Authors can start writing and uploading content to a preview-only title without revealing the content to readers.

Table 2: Summary Statistics from Public Data

	>=1000 followers			<1000 followers			All authors		
	N=4814			*N*=28726			*N*=33540		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Panel I: Author level summary statistics									
Experience (years)	6.1	5.36	3.48	3.69	2.82	3	4.03	3.15	3.19
Num. of characters (millions)	3.57	2.76	2.86	0.89	0.65	0.83	1.27	0.77	1.63
Avg. characters/year (millions)	0.68	0.58	0.46	0.39	0.26	0.44	0.44	0.3	0.46
Followers	8034	2501	28649.17	154.15	64	207.7	1285.27	93	11200.75
Number of books	19.34	16	13.75	9	7	7.28	10.49	8	9.25
Number of complete books	11.96	9	10.03	4.27	3	4.56	5.37	3	6.28
Days between books	171.34	136.83	127.74	194.93	131.5	226.14	191.43	132.83	214.57
Avg. characters/book (millions)	0.27	0.24	0.17	0.15	0.12	0.11	0.16	0.14	0.13
Panel II: Book level summary statistics for authors with >1000 followers									
	>=5000 followers			<5000 followers			All >=1000		
	N=30666			*N*=62015			*N*= 92681		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Book following (thousands)	17.80	5.83	44.25	4.28	1.86	7.36	8.96	2.68	27.47
Number of reviews (thousands)	9.58	2.36	31.09	1.56	0.58	3.17	4.34	0.89	18.86
Patronage values (CNY)	1751.88	34.00	15448.76	302.09	4.00	2625.75	781.79	8.00	9167.63
Number of patrons	102.79	12.00	612.86	14.95	2.00	39.22	44.02	4.00	356.38
First chapter clicks (thousands)	117.19	58.90	205.18	36.28	19.93	48.93	64.30	28.20	132.77
Completion status (Y=1)	0.68	1.00	0.47	0.59	1.00	0.49	0.62	1.00	0.49

Note: The sample consists of authors who have finished at least one book. Titles with no content were dropped when calculating the average length per title. The book sample contains novels, novellas, and short stories written by authors with more than 1000 followers. "Days between books" reports the average number of days between the first upload dates of two consecutive books by the same author.

3.3 Daily progression of new books

The final dataset tracks the day-to-day performance of newly published books, revealing how books differ at the beginning of their life cycle and how they build a following over time. Unlike the previous two data sets, which capture the performance of books months or years after publication, this data set offers a dynamic view of new books' initial accumulation of readership.

I chose the set of authors who have over 1000 followers as of January 7, 2021 and tracked all new books they started uploading between January 7 and February 25, 2021. This sample contains 941 titles. I tracked the book-length, book following, average chapter clicks, the number of reviews, the first three free chapter clicks, and word counts daily from February 26 to September 17, 2021. The books from the first sample were uploaded in a stretch of six weeks, which still introduced some variation in initial upload time. To mitigate the variation in book performance introduced by upload time instead of book characteristics, I supplement

the first set of books by adding a second sample of 179 books uploaded within a single week (May 1-May 8, 2021) by authors with over 1000 followers. For the second group, I tracked the same variables from May 9 to October 20, 2021.

3.4 Empirical Patterns

This section documents the writer's characteristics revealed by the three data sets. These patterns inform the empirical strategy used in the next section.

Pattern 1: While successful authors have more experience on average, they can also emerge from less experienced cohorts.

Panel 1 of Table 2 compares the summary statistics of authors that have more than 1000 followers ("successful authors") with those that do not. Successful authors tend to have written more characters and books than the rest. An average author with more than 1000 followers has written 3.57 million characters by January 2021, while an average author with fewer followers has written less than a million.[17] Successful authors also write longer books and publish more frequently. For both groups of authors, the mean of all performance metrics is higher than the median, indicating that both groups have high-performing outliers. This feature prompts me to use logarithmic normalization in the subsequent analysis.

FigureA1 plots the number of authors joining the website over time and the fraction of authors with over 1000 followers. For authors joining between 2007 and 2019, the fraction of successful authors in each cohort hovers around 10%.[18] The pattern is robust to a stricter definition of "successful". FigureA2 plots the entry time of authors with at least 1000 followers and the fraction of authors with at least 5000 followers among them. Similar to the patterns from figure A1, for authors who have been on the platform for over two years, joining earlier does not increase their chance of being among the most popular authors. This is somewhat surprising because an author who joined earlier has more time to hone her skills, build a reputation, and accumulate readers. The pattern suggests that time with the platform is a poor predictor of an author's success and author improvement may not be linear.

[17] To give non-Chinese readers a sense of what one million characters read like, the Chinese translation of *Le Comte de Monte-Cristo*, a novel published initially in newspapers as a serial, is around 856,000 characters. Character counts of *Le Comte de Monte-Cristo* are based on the translation by Kelu Zheng.

[18] Fewer authors joining after 2019 met this criterion, as the public data used for Figure A1 was compiled in January 2021. It appears to be hard to accumulate over 1000 followers within a couple of years. Pre-2007 cohorts are much smaller, so the ratio is also noisier in those cohorts.

Pattern 2: For most authors, the exit rate declines over time, but not for relatively successful authors.

Figure A3 plots the survival function and hazard function of all contracted authors.[19] An author is defined as having "exited" if they have not posted anything since January 1, 2019, approximately two years before the public data is collected. [20]

For the majority of authors, the exit rate falls rapidly after the first year. One-third of all authors have only worked on one novel, which explains the high initial exit rate. Among authors with over 1000 followers, however, the hazard rate increases slowly from an initial near-zero up until nine years after entry, remaining below 5 percent at any level of experience. The overall low exit rate among somewhat successful authors suggests that cultivating even a small readership is enough to motivate authors to keep writing. It also indicates that the quality improvement we observe is unlikely a result of selective attrition, as the exit rate is low for the authors studied in my sample.

Pattern 3: Cumulative output is a better measure for experience than time.

Thompson (2010) remarked that while cumulative output is the most popular measure of experience, alternatives like elapsed time or cumulative investments are also used in the literature. While capital investment hardly matters in our context, it is important to consider if improvement is better predicted by time, as research has suggested that exceptional individual performance requires prolonged periods of effort (Ericsson et al. 1993).[21] Recall that the author's experience is book-specific, as it accounts for time or characters written up to the beginning of each book. I call the time measure the author's "tenure" and denote it by T_{it}, where i is the book subscript and t is the month this book was uploaded. Authors' cumulative output is the number of characters they have written before uploading the current book's

[19] The survival function is defined as $S(t) = Pr(T > t)$. The hazard function is defined as $h(t) = \lim_{\Delta t \to 0} \frac{Pr(t<T\leq t+\Delta t|T>t)}{\Delta t}$

[20] There is no formal definition of whether an author is "active," and I chose January 2019 for two reasons. First, while it is common for authors to take breaks between books, authors in this industry publish more frequently than in traditional publishing. The median author takes 133 days between publishing the first chapters of two consecutive books and, on average, only 22 days off after completing a book before they start a new book. Only 6% of authors who took breaks of two or more years returned to write anything afterwards. Second, January 2019 predates the outbreak of COVID-19 at the end of 2019, which forced many people to stay at home and can change the supply of writers. I also run a robustness check using January 1, 2018, as the cutoff date for "active", and it yields a similar pattern.

[21] Ericsson et al. (1993) argued that expert-level individual performance in many domains is a result of deliberate and intense practice for a prolonged period of time–usually 10 or more years. Their research was later popularized by the book *Outliers* written by Malcolm Gladwell, who summarized part of their findings–albeit with a generalization–as the "10,000-hours rule."

first chapter. Formally, the past characters count is $E_{it} \equiv \sum_{\tau=0}^{t-1} w_{i,\tau}$, where $w_{i,\tau}$ are the word counts of previous books published by this author.

Following the discussion on Thompson (2010) p.445, I estimate the model:

$$\ln y_{it}^{j} = \alpha + \beta \ln E_{it} + \gamma \ln T_{it} + \eta_t + \xi_j + \epsilon_i \tag{1}$$

where y_{it}^{j} is the outcome of a book i uploaded at month t written by the author j, η_t is the upload year-month fixed effect that captures the time trend common to all books, and ξ_j is the author fixed effect that captures the individual time-invariant characteristics.[22] The outcome variables are the number of book followers to the first chapter clicks ratio, VIP chapter 10 to the first chapter clicks ratio, the number of subscribers to the first chapter clicks ratio, and projected income per 10,000 characters.[23]

Table 3 shows that past character count is a much better predictor of book outcomes. The results are consistent with the patterns in figures A1 and A2 and likely reflect the variation in writing speed among authors. Pane I of Table 1 shows that the average number of characters an author can write per year has a standard deviation close to the size of the median. Given this significant variation, author tenure is only a loose measure of writing experience, whereas past character count serves as a more direct measure. Therefore, in Section 4, the main analysis uses past character counts as the preferred measure of experience.

Pattern 4: Books show differences early on.

Table 1 shows considerable variation in the performance of books. Since the confidential firm records offer a cross-sectional view of book performance, this variation can reflect both inherent differences in book quality and the accumulation of readers over time. Figure A8 provides direct evidence that books vary in popularity even without the impact of time by tracking new titles published within the same time window.

Figure A8 plots the distribution of the average daily growth rate of a book's followers. The left panel measures the growth rate as a percentage and the right panel measures the average number of new followers per day on a log scale. There is substantial heterogeneity

[22] The first book of every author is dropped from the observations because the cumulative output and elapsed time before the first book are both zero. The outcome of the first books, which is absorbed by the author's fixed effect, measures the authors' initial productivity instead of their learning experience.

[23] Projected income is calculated as $\text{Income}_b = \text{Subscribers}_i \times \sum_{n=1}^{N_i} \hat{\beta}_i^{\,n-1}$, where $\hat{\beta}_i$ is the geometric average of the click attrition rates from the first 10 VIP chapters; N_i is the number of VIP chapters in book i; and Subscribers_i is the number of subscribers of book i. The number of subscribers is the total number of readers who purchased at least one pay-to-read chapter. Since the subscribers likely have purchased the first pay-to-read chapter, it is a proxy for the number of readers of the first pay-to-read chapter. Projected income per 10,000 characters is calculated as the projected income of book i divided by the length of book i in 10,000 characters.

Table 3: Overall improvement with experience

	ln(fol/chp1 click)	ln(Vchp10/chp1 click)	ln(sub/chp1 click)	ln(proj.inc/10k)
	(1)	(2)	(3)	(4)
ln(characters)	0.110***	0.0452***	0.0439***	0.0397***
	(0.00507)	(0.00718)	(0.00797)	(0.0150)
ln(tenure)	0.00266	0.00869*	-0.0160***	0.00329
	(0.00371)	(0.00526)	(0.00583)	(0.0109)
Author FE	Yes	Yes	Yes	Yes
Upload year-month FE	Yes	Yes	Yes	Yes
Observations	49214	48632	49212	37039

Note: Standard errors in parentheses. ln(characters) is the natural logarithm of millions of characters written before the current book is published. ln(tenure) is the log number of days this author has been on the publishing platform before the current book is published. The outcome variables are the log of followers to first chapter click ratio, the log of VIP chapter 10 click to first chapter click ratio, the log of subscribers to first chapter click ratio, and the log of projected subscription income per 10,000 characters. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

in the growth of popularity among new titles. The most successful title accumulates over 400 new followers a day on average, while the mode accumulates around 20 per day.

Interestingly, the least successful new titles can have average growth rates near zero or even negative, losing followers they accumulated during the preview period as their plots unfold. This implies that even an author with an established fan base can produce books that fail, and readers on this platform are quick to abandon books that don't meet their expectations, regardless of the author's reputation.

4 Empirical Strategy

In the learning-by-doing literature, the pattern of learning is often described with a power function. This functional form is well-supported by empirical evidence in manufacturing industries (Thompson 2012). In this section, we use the patterns discussed earlier to motivate our empirical strategy for estimating authors' learning curves.

The first and foremost question is, "is there evidence of learning?" Learning by doing can explain why successful authors are more experienced on average, but Pattern 1 also reveals that relatively inexperienced authors can still be very successful, and Pattern 3 again confirms that time is a poor indicator of performance. Heterogeneity in authors (e.g., their innate talent or writing habits) is, therefore, a necessary part of the analysis, and author fixed effect are included in all my preferred specifications.

Pattern 2 sheds some light on whether the observed improvement is from learning or

survival bias. This question is relevant because learning by doing is not the only mechanism that can lead to improvement in outcomes over time. If the worst writers drop out over time, the average book quality can increase even if the surviving authors do not improve. Pattern 3 shows that the exit rate is low for authors in the sample we study. This pattern mitigates the concern for survival bias and supports further investigation into author learning. Section 6 presents more robustness checks on survival bias.

Pattern 3 suggests that, in the empirical analysis following, the experience measure should be the cumulative number of characters written by the author before a book is published. Lastly, Pattern 4 shows that variation in book performance is not solely due to differences in publication timing. The substantial heterogeneity among new books confirms that we have sufficient book-level variation for the empirical strategy.

4.1 Traffic-neutral measures of book performance

A challenge in measuring book performance is that indicators for market reception, such as the number of followers, often increase over time as authors build their reputations. While follower count is a good predictor of an author's income, it is not an ideal outcome variable to measure individual learning. An author who experienced no improvement in writing quality can still accumulate more followers simply by letting more readers stumble upon their books over time. To circumvent this issue, I choose measures that are "neutral" to authors' number of followers and the traffic to their books (measured by clicks to the first chapter) to capture book quality and reception. In the empirical analysis, I use the ratio of the VIP chapter 10 clicks to the first (free) chapter clicks to measure the quality of books. It captures the reader retention rate of a book, regardless of its following and popularity.

A subtle assumption necessary to claim the measure as "follower-neutral" is that if a book is not captivating, readers will lose interest even if it is written by a reputable author. This argument might not hold if readers had more patience for a reputable author and would purchase additional pay-to-read chapters even if they did not enjoy the book. Panel II of Table 1 suggests that readers do not show more tolerance to reputable authors. If we compare the first chapter clicks of very popular authors (those with over 5000 followers) to the rest, the former has three times more clicks on their first chapter than the latter. When we compare the VIP chapter 10 to the first chapter click ratio, however, the difference between these two groups is small and statistically insignificant. The number of subscribers to the first chapter click ratio displays the same pattern. It appears that the author's popularity alone is not enough to retain readers.

To further test this assumption, I examine the reader retention rate of new books from

authors with different levels of experience using data on new books' daily progression. The outcome variable is the third-to-first chapter click ratio 90 days after the book is published. I regressed the new book retention rate on the authors' characteristics, including their experiences measured in the number of characters written before this book, the number of followers they had before writing this book, and the average chapter 10 to chapter 1 click ratio from all previous books written by this author. Table B3 reports the results of this regression and confirms that neither prior experience nor the number of followers guarantees a high reader retention rate for new books.

In addition to the VIP chapter 10 to chapter 1 click ratio, Section 5 and Section 6 include the subscription-to-click ratio, follow-to-click ratio, and projected income per 10,000 characters as outcome variables. The subscription-to-click ratio is another measure of reader retention that captures the fraction of readers who purchase at least one pay-to-read chapter out of all who visited the first chapter. The follow-to-click ratio is the ratio of the book follower count to the first chapter clicks. While it is a quality measure like other outcome variables, I consider it less informative than the subscription-to-click ratio and the VIP chapter 10 to chapter 1 click ratio because following a book does not incur a monetary cost. The last outcome variable, the projected income per 10,000 characters, approximates the author's income from book subscriptions. While this measure is the most relevant to the financial return of authors, it is not a traffic-neutral measure, because, for two books of the same quality, the one written by the more popular author may have more subscribers and yield higher projected income. I use this measure in robustness checks but avoid it when analyzing the speed and magnitude of learning. The danger in using income as an outcome variable is that while learning will increase income, income can increase without learning when authors accumulate readers over time.

4.2 Empirical Implementations

Combining all patterns from the previous discussion, equation 2 presents a baseline model that controls for book characteristics and various fixed effect. Each observation is a book written by author i uploaded in month t, and for simplicity, I omit the book subscript.

$$\ln y_{it} = \alpha + g(E_{it}; \beta) + X_{it} + \eta_t + \xi_i + \epsilon_{it} \tag{2}$$

where X_{it} includes the length and genre fixed effect of a book published by author i in time t. There are 234 distinct genres in my data set.[24] η_t is the upload year-month fixed effect

[24] A genre is defined by four tags that describe the originality, orientation, time period, and fiction sub-genre of the book. An example of a genre in my data set reads like "original-romance-alternate history-science fiction".

that capture the time-specific supply and demand factors that affect all books, such as the number of new books released in that month or the number of readers active at the time. ξ_i is the author fixed effect that captures time-invariant author characteristics like initial talent or writing style, and ϵ_{it} is the error term. After including the author fixed effect, the learning this model estimates is within-author: it captures their improvement as they accumulate more experiences on the platform. I only include books that are uploaded from 2006 to 2021. Books uploaded after 2022 may not be complete by the time the dataset is compiled, and very few books were written before 2006. I also exclude the outliers where the value of the outcome variables exceeds the 99th percentile or is below the first percentile.

The term $g(E_{it};\beta)$ represents a function that maps the author's experience into the outcome of interest. In the parametric specification, I adopt a log form and define $g(E_{it};\boldsymbol{\beta}) = \beta \ln(e_{it})$. In addition, I follow the approach of Haggag et al. (2017) and implemented a more flexible nonparametric specification to capture the duration of learning:

$$g(E_{it};\boldsymbol{\beta}) = \beta_1 \mathbb{1}\{0 \leq e_{it} \leq 1\} + \beta_2 \mathbb{1}\{1 < e_{it} \leq 2\} + ... + \beta_{10} \mathbb{1}\{9 < e_{it} \leq 10\} \tag{3}$$

In equation 3, $\mathbb{1}\{.\}$ is an indicator function and e_{it} is the number of characters (in millions) the author had written before the current book was published. In the nonparametric specification, authors who have written over 10 million characters are in the exclusion category.

5 Results

5.1 Quality improvement among authors

Table 4 presents the estimates of the parametric model, including the interaction of author experience with an indicator for whether an author is new. Here, a new author is defined as one who has written less than 1 million characters. Column 1 examines whether new authors improve when they write their first one million characters. Considering that the average length of a book in my sample is 340,000 characters, we are examining whether an author's second or third book is better than her first book. Column 1 of Table 4 shows that there is some evidence of improvement when authors write their first million characters, although when compared to the results in column 2, new authors improve relatively slowly. Columns 2 and 3 present results from the regressions that include all authors, and column 3 also includes the interaction with the new author indicator. For the full sample, reader retention improves faster, and column 3 confirms that authors do not learn faster in their first million characters. An author's second or third book may be better than the first, but the magnitude of improvement seems larger in subsequent books. Columns 4-6 present results from regressions that include book-length and genre fixed effect. Column 5 shows that a writer who

has written five million characters improves her VIP chapter 10 to chapter 1 click ratio by 7.6 percent compared to her first book. Column 6 shows that writers can improve their reader retention by 10 percent when they move from the first million to five million characters.[25] For the first one million characters, however, improvement is slower. The predicted retention rate increases by 1.6 percent when moving from half of a million characters to a million characters, whereas it increases by 2.4 percent when moving from 1 million to 1.5 million characters. This pattern is robust to an alternative definition of new authors. Table B1 presents the results of the same specifications, but defines new authors as those who have participated for less than two years. Both the signs and magnitudes of coefficients in Table B1 are similar to those in Table 4.

The results in Table 4 are somewhat surprising. If learning follows the power law exactly, the largest gain should occur within the first million characters. Though there is evidence for learning, new authors do not learn faster. In contrast, Haggag et al. (2017) found that new taxi drivers improve their income fastest in their first one-hundred shifts. To better understand the slow learning rate at the beginning of a writer's career, I plot the distributions of the first and subsequent books for authors who have written at least five books and have non-missing click ratios for all four books. The final sample contains 4608 authors. Figure A4 plots the kernel density of two reader retention measures, VIP chapter 10 to chapter 1 click ratio and subscription to first chapter click ratio. Compared to the two left panels, the fourth book has higher means and shows more obvious improvement from the first book. Figure A5 plots the cumulative distribution functions of the VIP chapter 10 to chapter 1 click ratio for the first four books written by each author. There is no clear improvement if we compare the second book to the first book, and for the best-performing quartile, the second book is actually slightly worse than the first book. The fourth book does better almost everywhere in the distribution, but the distinction is less clear for the third book. Slower learning for beginners is not a pattern found in previous studies on individual LBD, but it is consistent with anecdotal evidence in creative writing, what some call "the curse of the second book."(Fürst 2022) The "curse" refers to a pattern found in multiple forms of creative careers, from filmmaking to fiction writing, that many creators experience difficulties in creating a second piece. In my analysis, I found that even if writers manage to write again, their second book is not necessarily better than the first.

[25]The difference in reader retention rate is computed from $\exp[\hat{\beta} \times \ln(5)]$-1, where $\hat{\beta}$ is the coefficient estimate.

Table 4: Improvement with experience

Specification	(1)	(2)	(3)	(4)	(5)	(6)
New			-0.0176*			-0.0152
			(0.0105)			(0.0105)
New×ln(characters)			-0.0284**			-0.0341**
			(0.0142)			(0.0144)
ln(characters)	0.0333***	0.0515***	0.0602***	0.0236**	0.0455***	0.0592***
	(0.0115)	(0.00627)	(0.0121)	(0.0116)	(0.00639)	(0.0125)
New author only	Yes	No	No	Yes	No	No
Length	No	No	No	Yes	Yes	Yes
Genre FE	No	No	No	Yes	Yes	Yes
Observations	14768	48545	48545	14530	46748	46748
R^2	0.684	0.571	0.571	0.696	0.588	0.588

Note: The outcome variable is the natural logarithm of VIP chapter 10 click to first chapter click ratio. Robust standard errors in parentheses, clustered at the author level. Author and upload year-month FE included in all specifications. New authors are defined as authors who have written less than 1 million characters.
*** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

5.2 Duration of Learning

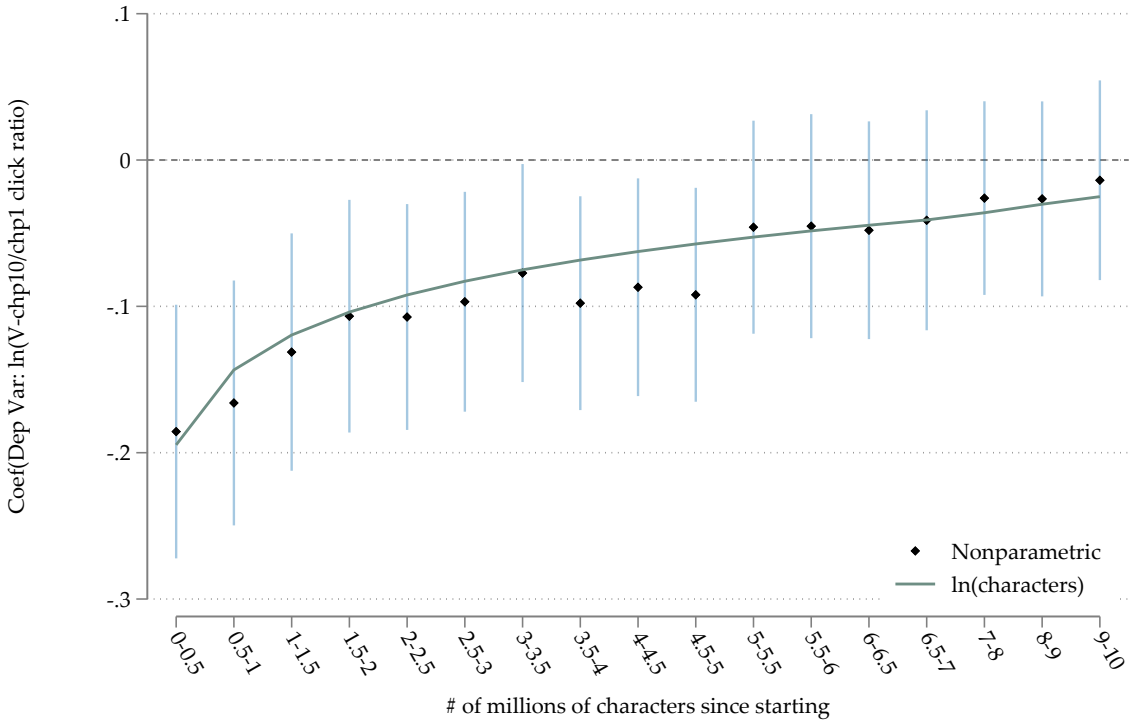
Empirical studies of more routine jobs found that the effect of learning by doing fades a few months after starting the task. Levitt et al. (2013) observed that the defect rate of workers on automobile assembly lines dropped by two-thirds within eight weeks. Similarly, Haggag et al. (2017) found that income gains for new taxi drivers slowed down after 40 shifts.[26] For a broader set of tasks, Nagypál (2007) used French matched employer-employee data and found that learning-by-doing is only present in the first six months of an employment relationship.

The duration of learning varies by industry and task complexity. Jovanovic and Nyarko (1995) suggests that the more decisions a task involves, the longer it will take to learn. I found that, although the quality improvement eventually slows down for writers, the duration of learning is much longer than what was documented in studies on more routine tasks. This finding is consistent with the theoretical prediction of Jovanovic and Nyarko (1995).

Figure 1 presents a scatter plot of the estimates (with 95% confidence intervals) from the nonparametric model alongside a line plot from the parametric model. The outcome variable

[26] In Haggag et al. (2017), the median duration of a shift is 9 hours. If a taxi driver works five shifts a week, 40 shifts take eight weeks to complete.

Figure 1: Parametric versus Nonparametric Approaches



is the natural logarithm of the VIP chapter 10 to chapter 1 clicks ratio. Two patterns immediately stand out. First, reader retention improves significantly up to when the author has written around 5 million characters, a volume that would take three to five years to complete even for the top 10% most prolific authors in my sample.[27] Second, the estimates from the nonparametric model closely imitate the shape of the parametric plot, supporting the use of a log specification in the subsequent analysis.

5.3 Learning and initial performance

I now focus on learning among authors with different characteristics. I divide authors into four quartiles based on the performance of their first book, then estimate equation 1 for each subsample and report the results in Table 5. Author, genre, length, and upload year-month fixed effect are included in all specifications.

Panel 1 and 2 of Table 5 report the results when the outcome variables are the VIP chapter 10-to-chapter 1 click ratio and the number of subscribers-to-first chapter click ratio respec-

[27] The 90th percentile author, in terms of quantity produced, can write 1.14 million characters per year. The 99th percentile author can write 1.96 million characters per year.

Table 5: Learning rates by performance of the first book

	Full Sample	Lowest	Low	High	Highest
	Panel One: VIP Chapter 10 to First Chapter Clicks Ratio				
ln(characters)	0.0428***	0.0928***	0.0444***	0.0222*	-0.00315
	(0.00589)	(0.0123)	(0.0112)	(0.0119)	(0.0121)
Observations	46415	14107	12704	10821	8572
	Panel Two: Subscription to First Chapter Clicks Ratio				
ln(characters)	0.0384***	0.0858***	0.0518***	0.0214*	-0.0118
	(0.00658)	(0.0140)	(0.0119)	(0.0129)	(0.0140)
Observations	47008	13889	11548	10535	11036

Note: This table shows the estimates from equation 2 when authors are divided into four quartiles based on the performance of their first book. Robust standard errors in parentheses, clustered at the author level. Length, genre, author and upload year-month fixed effects included in all specifications. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

tively. Since authors are divided into quartiles based on their first book's ranking in each outcome, the sets of authors in each quartile differ slightly between Panels 1 and 2. I only keep authors for which the corresponding outcome variables are available for their first book.[28]

Table 5 shows that learning is more notable for authors whose first book's performance is at the bottom quartile. The authors who started in the bottom two quartiles can improve the VIP chapter 10-to-chapter 1 click ratio by 16.1 percent after they write five million characters. For the authors in the second quartile, both the coefficient and the level of improvement are approximately halved, as authors in this quartile only improve by 7.4 percent. The magnitude halves again for the third quartile, to 3.6 percent, and the coefficient becomes a precisely estimated zero for the top quartile.

Panel 2 reveals a similar pattern. Writing five million characters improves the level of subscriber count-to-first chapter click ratio of the bottom three quartiles by 14.8 percent, 8.7 percent, and 3.5 percent, respectively. There is again no evidence of improvement for authors who started in the top quartile. The results in Table 5 imply that the authors who wrote a

[28]The outcome variables can be missing for a number of reasons. If the author's first book is free, it will report a missing value for the number of subscribers and the VIP chapter 10 clicks. If their first book ends before it has 10 pay-to-read chapters, it will also report a missing value for the VIP chapter 10 clicks. In either case, that author will be excluded from the sample for this analysis. The difference in the author sample explains the difference between the estimates from column 1 of Table 5 Panel 1 and the estimates reported in column 3 of Table 4.

strong first book either due to talent or previous writing experience may not maintain as big of an advantage as before when their peers accumulate experience.

5.4 Learning and writing habits

In this section, I investigate how writing habits affect the improvement of reader retention. Many online content creators believe that frequent and consistent uploads are crucial to retaining a paying audience (Arriagada and Ibáñez 2020). I constructed two measures of writing habits related to an author's productivity: the average number of characters the author uploads per day and the average time gap between books. The average number of characters per day is only calculated when the author is actively uploading: that is, I divide the length of the book over the number of days between when the book began being published and when it was marked completed. The median author uploads 2,609 characters per day on average, while the most prolific author can upload 20,801 characters a day on average. The time gap between books measures the days between when a book is marked as completed and when the author's next book starts being published. The median author has an average time gap of 22 days.

The average number of characters uploaded per day closely aligns with the traditional measure of productivity as quantity produced in a unit of time. While productivity is not as important as quality in many nonroutine and creative tasks, I examined whether productivity improves with the author's experience. Figure A6A plots the parametric and nonparametric estimates of equation 2, where the outcome variable is the log of the average number of characters uploaded per day. Surprisingly, experienced authors do not write faster than inexperienced authors: they actually write slower. The reduction in speed is not due to new authors strategically uploading an already completed book quickly to attract readers, which would result in an overestimation of new authors' writing speed. Figure A6B plots the estimates with the outcome variable being the average number of characters written between the starting dates of two books, which circumvents the measurement errors mentioned previously. With either measure, it is clear that new authors write faster than experienced authors.

This finding raises the question of whether the quality improvement we observe in Table 4 and Figure 1 can be explained by experienced authors writing more slowly and carefully. Table B2 presents a robustness exercise that includes writing speed as a control. Quality improvement with authors' experience remains similar in magnitude to the results in Table 4 after controlling for both measures of author writing speed, indicating that slower writing by experienced authors is not the main contributor to better book quality.

Next, I examine whether different writing habits affect an author's learning speed by es-

timating the following equation,

$$\ln y_{it}^{j} = \alpha + \beta_1 \ln(e_{it}+) + \beta_2 \ln(e_{it}) \times Z_{it} + \beta_3 Z_{it} + X_{it} + \eta_t + \xi_j + \epsilon_{it} \tag{4}$$

where Z_{it} is the average number of characters uploaded per day and the time gap between books, respectively. I use the parametric form of experience in equation 4 because the parametric and nonparametric estimates are remarkably similar in Figure 1 and the parametric form is computationally more efficient. The results are reported in Table 6, and book-length, genre, and author fixed effect are included in every specification. Columns 1 and 2 report the impact of writing habits on VIP chapter 10-to-chapter 1 click ratio. Consistent with the beliefs of practitioners, readers do indeed appreciate more new content per day: if an author can write 1000 more characters a day, she can improve her VIP chapter 10-to-chapter 1 click ratio by 3.5 percent. The authors who write faster, however, do not seem to learn faster after controlling for experience. Although the coefficient estimate for the interaction between experience and writing speed is not statistically significant, the sign of the coefficient suggests that if anything, faster writers may learn slightly slower than their peers. There is no evidence that more characters per day helps accumulate followers conditional on traffic: the coefficient estimates for both characters per day and the interaction term are very small and statistically insignificant in column 3.

Column 2 reports the interaction of the time gap between books and the accumulation of experience. Contrary to popular belief, there is no evidence for readers penalizing authors who take longer breaks between books. Whether measured by reader retention of paying readers (VIP chapter 10 to chapter 1 click ratio) or by popularity conditional on traffic (followers to first chapter click ratio), the coefficient for the gaps between books is either zero or positive. This result suggests that readers do not mind waiting between books. Authors who take some time between books also seem to learn faster than their peers as shown in columns 2 and 4, although the interaction term coefficient is again small and not statistically significant when the outcome is VIP chapter 10 to chapter 1 click ratio.

5.5 Improvement with general and specific experience

In the previous sections, I broadly defined author experience as cumulative general novel writing experience. However, in addition to the general experience, authors accumulate genre-specific experience. Specializing in a single genre can attract loyal readers, potentially securing traffic for new publications, while writing across multiple genres can bring fresh writing experience and allow authors to experiment and find the genres that best match their skills.

Table 6: Learning and Writing Habits

Dependent variable	ln(Vchp10/chp1 click) (1)	ln(Vchp10/chp1 click) (2)	ln(follow/chp1 click) (3)	ln(follow/chp1 click) (4)
ln(characters)	0.0597***	0.0468***	0.0951***	0.112***
	(0.0149)	(0.00666)	(0.00594)	(0.00474)
characters/day (000')	0.0346***		0.00189	
	(0.00607)		(0.00214)	
ln(charac.)×charac./day	-0.00530		0.00207	
	(0.00458)		(0.00134)	
book gap (months)		0.000544		0.00259***
		(0.000377)		(0.000294)
ln(charac.)×book gap		0.000367		0.00211***
		(0.000365)		(0.000267)
Observations	42324	43415	42635	43871
R^2	0.599	0.585	0.739	0.737

Note: Robust standard errors in parentheses, clustered at the author level. Length of the book, author FE, upload year-month FE, and genre FE included in all specifications. ln(characters) measures the log of millions of characters written before the current book. The outcome variable in columns 1 and 2 is the natural logarithm of VIP chapter 10 click to first chapter click ratio. The outcome variable in columns 3 and 4 is the natural logarithm of followers to first chapter click ratio. Book gap measures the months between the completion date of a published book and the first published date of the next book. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

In this section, I explore how general or specific experiences contribute to authors' improvements. In this dataset, a genre is defined by the tags chosen by the author of each book. A broad definition of genre uses the first two tags–originality (e.g., original or inspired by existing works) and category (e.g., romance, children's story, etc.). A narrower definition includes four tags, providing a more detailed categorization (e.g., thriller, horror, historical). Half of all authors have written in more than one broad genre. The variation in genres within authors allows me to explore the role of genre-specific writing experience alongside general experience.

To measure genre-specific experience, I calculate the number of characters an author has previously written within each genre. I estimate a model that controls for both general experience E_{it} and genre-specific experiences E^s_{it}:

$$\ln y_{it} = \alpha + g(E_{it};\beta) + h(E^s_{it};\gamma) + X_{it} + \eta_t + \xi_i + \epsilon_{it} \tag{5}$$

where both g(.) and h(.) are represented by a series of dummy variables, similar to those specified in equation 3. For both general and genre-specific experiences, the category where authors have written over 8 million characters is excluded as the baseline.

Table 7 reports the results from estimating equation 5. Consistent with results from Table 4 and Figure 1, reader retention rate, as measured by the VIP chapter 10-to-chapter 1 click ratio, improves with general writing experience. Column 2 shows that reader retention rate improves with genre-specific experience, but the effect is not statistically significant. General writing experience accounts for most of the initial improvement in reader retention.

Column 3 reports the results from the same estimation when genre is measured with more detailed categories. With narrowly defined genres, genre-specific experience can explain an even smaller fraction of improvement, and genre-specific experience remains statistically insignificant. Specific experience appears to have no impact in my context, which contrasts with previous studies on other complex and non-routine tasks. Ost (2014), for example, found that teachers with more grade-specific experiences can improve their students' math scores more than their peers with similar levels of general experience but less specific experience.

Columns 4 and 5 report the results of equation 5 when the outcome variable is the log subscriber count-to-first chapter click ratio. Column 4 shows that the subscription rate also improves with the general writing experience, although the effects are smaller and not statistically significant. Genre-specific experiences, however, explain almost all the improvement in this scenario. The contrast between results in columns 2-3 and column 5 show that specific experience can matter, but on different aspects of a book's performance. Specializing in a genre indeed attracts more paying readers, but may not be enough to retain them.

6 Alternative explanations and robustness checks

In this section, I explore mechanisms other than learning that can potentially explain the improvement of book performance with experience. I address two alternative explanations: survival bias and learning of consumer preference.

Learning-by-doing or survival bias

Even without individual learning, overall book quality can improve as the worst authors voluntarily exit the platform. Less capable authors, after dropping out, cease to accumulate experience and their relatively weak performance may be incorrectly attributed to their lack of experience. While including author fixed effect and exploring within-author variation mitigates this concern, it does not completely eliminate the threat of this survival bias.

To test this alternative mechanism, I construct a measure for whether an author has dropped out in my sample. I define an author as currently "active" if they have had some activity on the platform after 2021. The activity need not be uploading new content. If an author has

been revising an existing book and updating the revised chapters online, her activity will be recorded and she is considered "active". The data set collects author information up to November 2022, so if an author has not updated anything since January 2021, almost two years before the date of observations, she is classified as inactive.[29] Among all authors, only 17 percent are inactive, and they wrote 11 percent of the books in the sample. Table B4 compares the characteristics of the active and inactive authors. Inactive authors tend to be less popular and have a lower reader retention rate. Since the full sample included books uploaded in or before 2021, the baseline analysis includes all authors who have exited and caused potential bias in my analysis. In this robustness check, I re-estimate equation 2 with a subsample of authors who were active at the time the confidential data was compiled. If survival bias is the main reason for quality improvement instead of learning, then excluding the inactive authors should eliminate most, if not all, learning effects in my estimates.

Columns 1 and 3 in Table 8 report the results from the full sample and columns 2 and 4 restrict the sample to active authors only. Panel 1 shows that, after excluding inactive authors, the estimate corresponding to the rate of learning increases instead of decreases. Panel 2 and 3 use alternative outcome variables and there is not a statistically significant difference between samples with and without the inactive authors. After excluding inactive authors, the coefficient in column 4 in panel two preserves 76.4 percent of the size of the estimate in column 3. In panel three, the estimate of active authors is 89 percent of the size of the full sample estimate in column 3. Table 8 reveals that survival bias is unlikely the main cause of quality improvement.

Additionally, I estimate equation 2 by active status and the year they joined the platform. I include authors who joined the platform from 2008 to 2019 because the 2005 to 2007 cohorts are too small to estimate equation 2 (see Figure A1). I also exclude authors who joined in 2020 or after, as they are likely to still be active by 2021, so there is no distinction between the active sample and the full sample.

Less capable authors from earlier cohorts are more likely to have exited as they had the time to learn about their own ability by the time of observation. If survival bias is the main mechanism, we will observe that (1) the coefficient estimates for all authors decrease over time and (2) the coefficient estimates of active authors in earlier cohorts should be close to zero.

[29] Exit is defined differently in Section 3.4. Authors who had no activities after January 2019 are considered inactive. The analysis in Section 3.4 is done with public data, which was compiled in January 2021. January 2019 is approximately two years before the observation. The analysis in this section is done with the confidential data, which was compiled in November 2022. Both definitions chose approximately two years before the observation date as the cut-off for the author's exit.

Figure A9 plots the estimates by cohort and their 95% confidence intervals. With the exception of a higher learning rate in 2008, the rest of the cohorts do not see decreasing full sample estimates over time. In addition, the difference between estimates from active authors and all authors is small and statistically insignificant. For the 2008 cohort, where we observe the largest full sample estimate and a notable discrepancy between active authors and the full sample, active authors seem to learn faster than the full set of authors. If survival bias is the main mechanism, the opposite should be true, and active authors should have a smaller, even near-zero estimate. This pattern remains true in Panel B and C where I used other outcome variables to measure book performance. The cohort-specific regressions also suggest that survival bias is unlikely to be the source of quality improvement.

Improvement of quality versus learning about consumer preferences

Another possible explanation for the improvement in book performance is that, rather than honing their writing skills, authors learned about consumer preferences and started writing books that cater to popular interest. If that is the case, better market performance of the book may reflect a better match with audience tastes instead of any intrinsic improvement in the authors' skills. To investigate whether authors actively cater to popular interests, I test if similar books were written after one extremely successful title was published. The idea is that a tremendously successful title can reveal or even create audience interest in a theme, and other authors are then incentivized to produce similar books to attract readers who are interested in the same theme. If we find evidence for such behaviour and if the market does reward books of similar themes, then some of the observed learning may be attributable to improvements in preference matching instead of an improvement in quality.

I started by identifying the "superstar" titles that are successful enough to inspire followers. I define a "superstar" as a book that is in the top 1 percent in terms of book following within a broad genre in a quarter. Book following, unlike the number of subscribers or VIP chapter clicks, is a public performance metric observable to all authors, so it is more likely to influence author behaviours. There are 12 unique broad genres, but I limit the search for superstars to genre-quarter pairs with over 100 books published, which left me four genres with at least one superstar. I focused on superstars published before 2021 to allow enough time to observe the response of new books published in the year following the appearance of a superstar. The final sample contains 377 superstar titles and 56,710 other titles.

I use author-selected keywords to identify an author's tendency to cater to a popular theme. For each book, the author can choose up to five keywords to describe the content and style of the book. The keywords are specific enough to inform readers of its content (e.g.,

"detective mystery thriller, horror, adventure, inspired by Sherlock Holmes"). I measured the relative monthly frequency of each keyword that was used by the superstar titles. For the one year leading up to a superstar being published and for one year after, I calculate the relative frequency of a keyword among all books newly published in that month in the same broad genre. I focus on the same broad genre because authors competing for the same audience are more likely to study and follow a superstar from that genre. The equation I estimate is the following:

$$\mathbb{1}\{\text{keyword}\}_{it}^{n} = \sum_{\tau=-12}^{\tau=12} \beta_\tau \mathbb{1}\{t - t^* = \tau\} + \eta_t + \epsilon_{it} \tag{6}$$

where t^* is the publishing time of a superstar, η_t is the upload year-month fixed effect, and $\mathbb{1}\{\text{keyword}\}_{it}^{n}$ is an indicator that takes value 1 if book i includes the nth keyword of a superstar in its description and it is uploaded in month t.

Figure A7 plots the coefficient estimates for β_τ and their 95% confidence intervals. If authors systematically choose themes that the readers may be interested in, then we should see an increase in the relative frequency of a keyword after a superstar with that keyword is published. Figure A7 shows neither an increase nor a decrease in keyword relative frequency after the publication of a superstar title. I found no evidence of authors catering to a popular theme: the keyword relative frequency remains remarkably flat for all four keywords.[30] Learning and adapting to consumer preferences is, therefore, unlikely to be a main driver of improvement in book performance.

7 Conclusion

How do individuals improve when working on a non-routine and creative task, and what factors help them learn? This study finds significant, though gradual, improvement in writing quality as writers gain experience. Unlike the rapid learning observed in more routine tasks, learning in non-routine and creative tasks is extended, with slower initial improvement but notable gains in quality occurring over several years. The analysis reveals substantial heterogeneity in learning outcomes. Authors who start with performance in the bottom quartile experience the most improvements, indicating that those with more room to grow benefit the most from accumulating experience. I also investigate the relationship between writing habits and learning. Writing speed does not correlate strongly with faster learning, and authors who take longer breaks between books tend to experience quicker improvement.

[30] Authors can choose up to five keywords, but most of them stop at four. The number of observations with a fifth keyword is too small to estimate equation 6.

Compared to previous studies in fields like teacher improvement, specific experience does not explain much of the author's improvement in writing quality. I also provide evidence that the observed improvements are driven by the accumulation of experience rather than survival bias or preference matching.

High performance in creative tasks has long been attributed to the innate quality of the creators. This view is still common among practitioners: in my conversations with some authors and editors on the online platform, exceptional authors are often referred to as geniuses. This paper shows that learning has a substantial role in author performance, and the authors who started below average actually learn faster, closing the performance gap as they accumulate experience. Since I found that learning can last for several years and new writers do not improve fastest initially, whether creators can achieve their full potential depends on whether they can stay on the creative task for a prolonged period of time. This result highlights the benefits of supporting early career writers or, more generally, workers in occupations where writing is an important output. In a competitive and minimally intervened environment, a large share of new creators may be forced out before substantial improvement can occur. My results have policy implications for institutions that evaluate early-career researchers and creators. If inexperienced creators can receive sufficient support that allows them to stay on their creative path for several years, there can be human capital gains from individual learning.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used Grammarly and ChatGPT in order to improve the readability and language of the manuscript. After using this service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

Acemoglu, D. (2024). The simple macroeconomics of ai*. *Economic Policy*, page eiae042.

Adda, J. and Dustmann, C. (2023). Sources of wage growth. *Journal of Political Economy*, 131(2):456–503.

Arriagada, A. and Ibáñez, F. (2020). "you need at least one picture daily, if not, you're dead": content creators and platform evolution in the social media ecology. *Social Media+ Society*, 6(3):2056305120944624.

Arrow, K. J. (1962). The economic implications of learning by doing. *The review of economic studies*, 29(3):155–173.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, 118(4):1279–1333.

CASS, C. A. o. S. S. (2023). 2023 report on the development of chinese online literature (in chinese). Technical report, Chinese Academy of Social Sciences.

Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363.

Fürst, H. (2022). The curse of the difficult second book: Continuation and discontinuation in early literary careers. *Poetics*, 92:101642.

Gathmann, C. and Schönberg, U. (2010). How general is human capital? a task-based approach. *Journal of Labor Economics*, 28(1):1–49.

Haggag, K., McManus, B., and Paci, G. (2017). Learning by Driving: Productivity Improvements by New York City Taxi Drivers. *American Economic Journal: Applied Economics*.

Hendel, I. and Spiegel, Y. (2014). Small steps for workers, a giant leap for productivity. *American Economic Journal: Applied Economics*, 6(1):73–90.

Jackson, C. K. and Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, 1(4):85–108.

Jovanovic, B. and Nyarko, Y. (1995). A Bayesian Learning Model Fitted to a Variety of Empirical Learning Curves. *Brookings Papers on Economic Activity. Microeconomics*, 1995:247.

Lafontaine, F. and Shaw, K. (2016). Serial Entrepreneurship: Learning by Doing? Technical Report 2.

Levitt, S. D., List, J. A., and Syverson, C. (2013). Toward an Understanding of Learning by Doing: Evidence from an Automobile Assembly Plant. Technical Report 4.

Nagypál, (2007). Learning by Doing vs. Learning About Match Quality: Can We Tell Them Apart? *Review of Economic Studies*, (74):537–566.

Ost, B. (2014). How do teachers improve? the relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, 6(2):127–151.

Rocha, V., Carneiro, A., and Amorim Varum, C. (2015). Serial entrepreneurship, learning by doing and self-selection. *International Journal of Industrial Organization*, 40:91–106.

Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of political economy*, 94(5):1002–1037.

Thompson, P. (2010). LEARNING BY DOING. *Handbooks in Economics*.

Thompson, P. (2012). The Relationship between Unit Cost and Cumulative Quantity and the Evidence for Organizational Learning-by-Doing. *Journal of Economic Perspectives*, 26:203–224.

Yu, H., Marschke, G., Ross, M. B., Staudt, J., and Weinberg, B. A. (2023). Publish or perish: Selective attrition as a unifying explanation for patterns in innovation over the career. *Journal of Human Resources*, 58(4):1307–1346.

Table 7: Improvement and General or General-specific Experience

General or genre-specific	ln(Vchp10/chp1 click)			ln(sub/chp1 click)	
experience (millions of charac.)	(1)	(2)	(3)	(4)	(5)
experience $\in (0,1)$	-0.146***	-0.134***	-0.145***	-0.0434	0.0202
	(0.0328)	(0.0398)	(0.0340)	(0.0377)	(0.0456)
experience $\in [1,2)$	-0.0938***	-0.0749**	-0.0911***	-0.0123	0.0583
	(0.0304)	(0.0371)	(0.0316)	(0.0350)	(0.0423)
experience $\in [2,3)$	-0.0809***	-0.0428	-0.0766***	-0.0116	0.0621
	(0.0284)	(0.0346)	(0.0295)	(0.0324)	(0.0392)
experience $\in [3,4)$	-0.0636**	-0.0226	-0.0603**	-0.0173	0.0498
	(0.0269)	(0.0338)	(0.0284)	(0.0305)	(0.0369)
experience $\in [4,5)$	-0.0684***	-0.0506	-0.0662**	-0.0131	0.0307
	(0.0265)	(0.0320)	(0.0278)	(0.0292)	(0.0335)
experience $\in [5,6)$	-0.0301	-0.0219	-0.0295	0.00652	0.0484
	(0.0264)	(0.0312)	(0.0276)	(0.0280)	(0.0317)
experience $\in [6,7)$	-0.0300	-0.0404	-0.0299	0.0123	0.0238
	(0.0261)	(0.0313)	(0.0273)	(0.0276)	(0.0303)
experience $\in [7,8)$	-0.0108	-0.0361	-0.0108	0.00600	-0.00902
	(0.0253)	(0.0294)	(0.0259)	(0.0300)	(0.0336)
genre exp$\in [0,1)$		-0.00452	-0.00266		-0.0758
		(0.0397)	(0.0486)		(0.0462)
genre exp$\in [0,1)2$		-0.0146	-0.00735		-0.0851*
		(0.0391)	(0.0485)		(0.0456)
genre exp$\in [0,1)$		-0.0423	-0.0106		-0.0881*
		(0.0386)	(0.0489)		(0.0452)
genre exp$\in [0,1)$		-0.0416	-0.0000859		-0.0751*
		(0.0393)	(0.0494)		(0.0444)
genre exp$\in [0,1)$		0.00686	0.00263		-0.0312
		(0.0381)	(0.0492)		(0.0420)
genre exp$\in [0,1)$		0.0188	0.0183		-0.0390
		(0.0379)	(0.0480)		(0.0402)
genre exp$\in [0,1)$		0.0495	-0.00432		0.0214
		(0.0381)	(0.0498)		(0.0397)
genre exp$\in [0,1)$		0.0718**	0.00945		0.0639
		(0.0351)	(0.0526)		(0.0404)
Observations	53,210	53,210	53,210	53,774	53,774

Note: Robust standard errors in parentheses, clustered at the author level. Length of the book, author FE, upload year-month FE, and genre FE included in all specifications. ln(characters) measures the log of millions of characters written before the current book. The outcome variables in columns 1, 2, and 3 are the natural logarithm of VIP chapter 10 click to first chapter click ratio. Genre is defined with two key-words in column (2) and with four key-words in column (3). The outcome variable in columns 3 and 4 is the natural logarithm of number of paid readers (subscribers) to first chapter click ratio. * $p < 0.10, ** p < 0.05, *** p < 0.01$

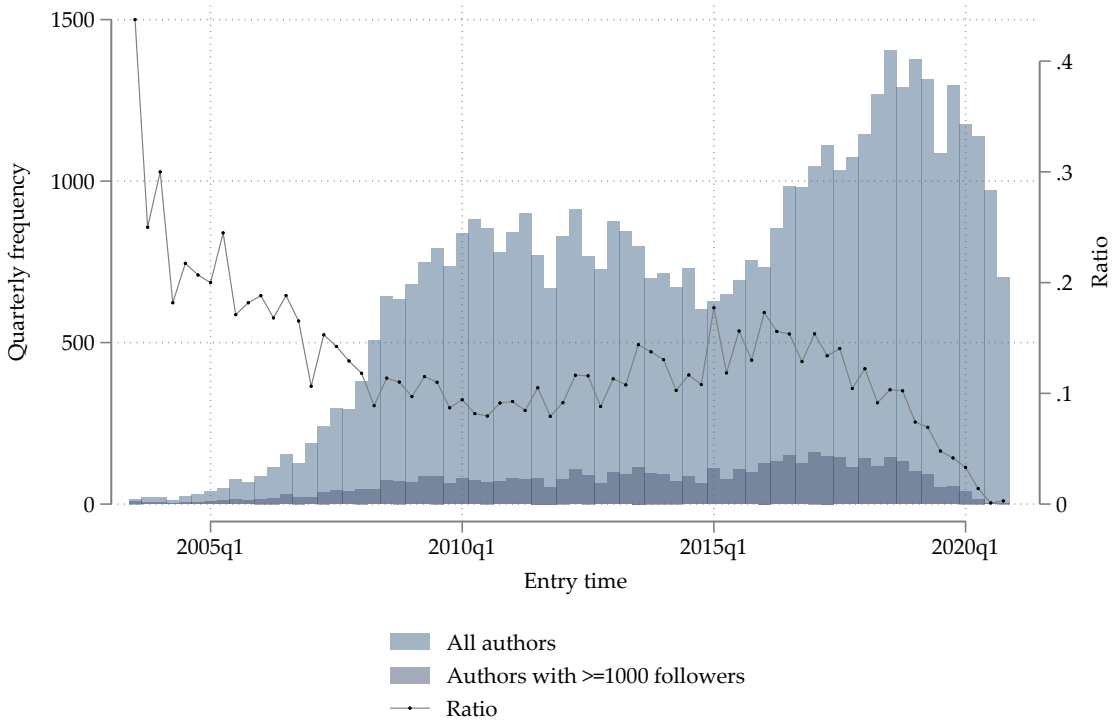
Table 8: Learning of all authors versus currently active authors

Specification	(1)	(2)	(3)	(4)
Panel One: VIP Chapter 10 to First Chapter Clicks Ratio				
ln(characters)	0.0515***	0.0537***	0.0455***	0.0475***
	(0.00627)	(0.00653)	(0.00639)	(0.00670)
Observations	48545	43802	46748	42094
R^2	0.571	0.548	0.588	0.567
Panel Two: Subscription to First Chapter Clicks Ratio				
ln(characters)	0.0275***	0.0235***	0.0174**	0.0133*
	(0.00753)	(0.00777)	(0.00767)	(0.00796)
Observations	49069	44290	47274	42585
R^2	0.462	0.443	0.476	0.458
Panel Three: Projected Income				
ln(characters)	0.0444***	0.0431***	0.0393***	0.0350***
	(0.0127)	(0.0132)	(0.0129)	(0.0134)
Observations	36551	33158	35361	32040
R^2	0.430	0.423	0.448	0.442
Active authors only	No	Yes	No	Yes
Length	No	No	Yes	Yes
Genre FE	No	No	Yes	Yes

Note: Robust standard errors in parentheses, clustered at the author level. Author and upload year-month fixed effects included in all specifications.
*** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

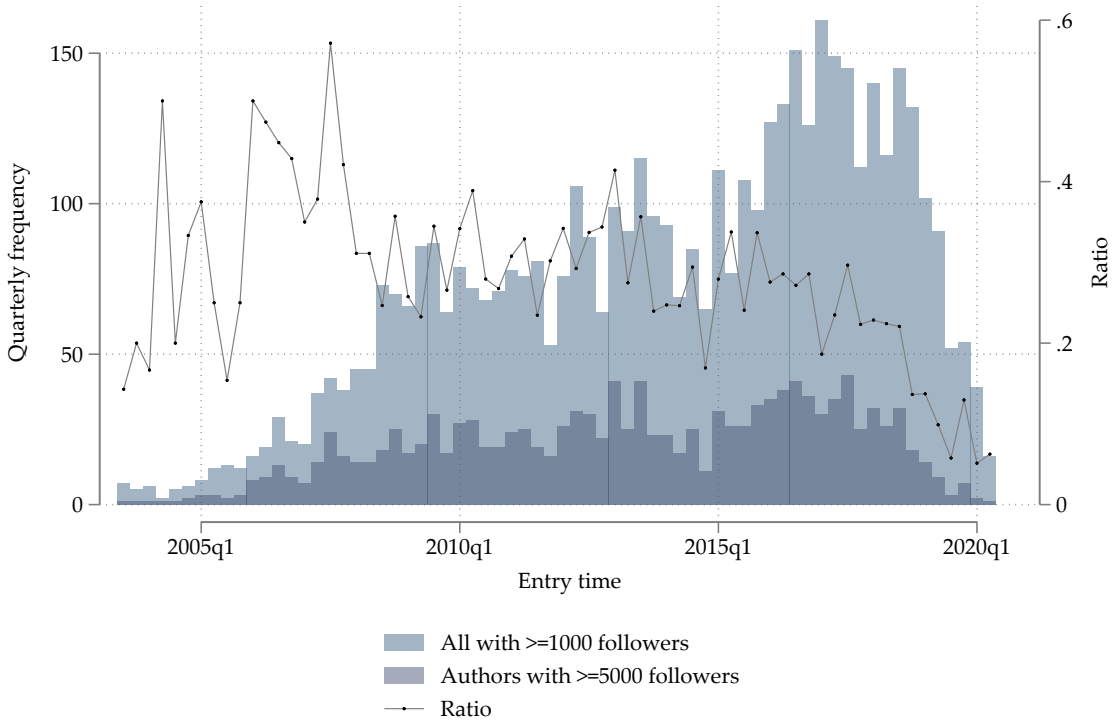
Appendix A Additional figures

Figure A1: Number of authors joining the platform over time



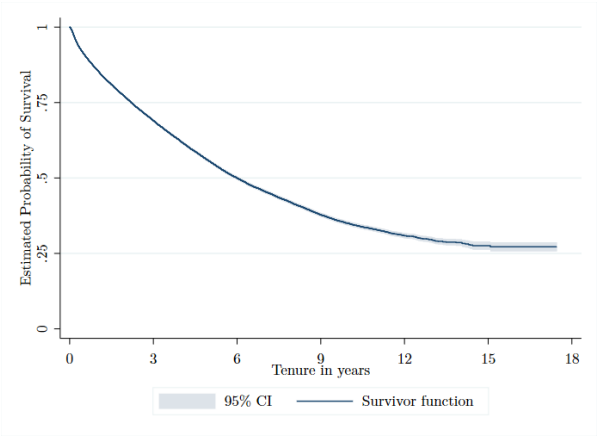
Notes: This figure counts the authors who joined the platform each year. The time of entry is measured as the time this author uploaded the first chapter of his/her first book.

Figure A2: When did top authors join the platform?

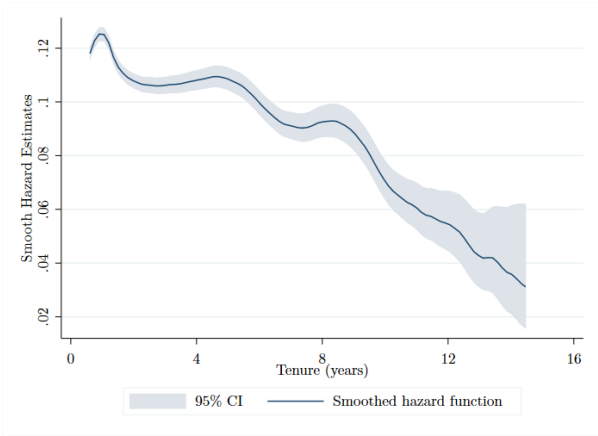


Notes: This figure counts the authors entered over time. The time of entry is measured as the time this author uploaded the first chapter of his/her first book.

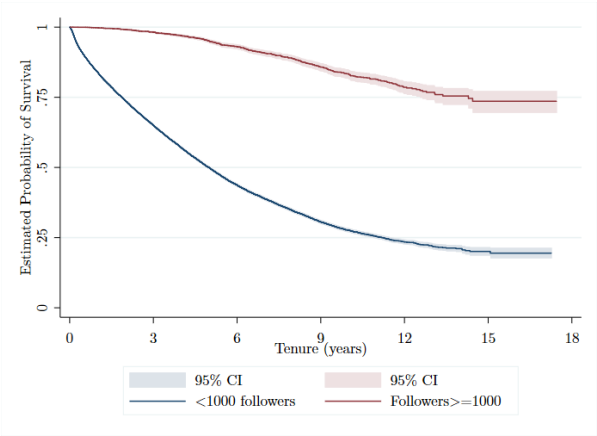
Figure A3: Estimated probability of staying active



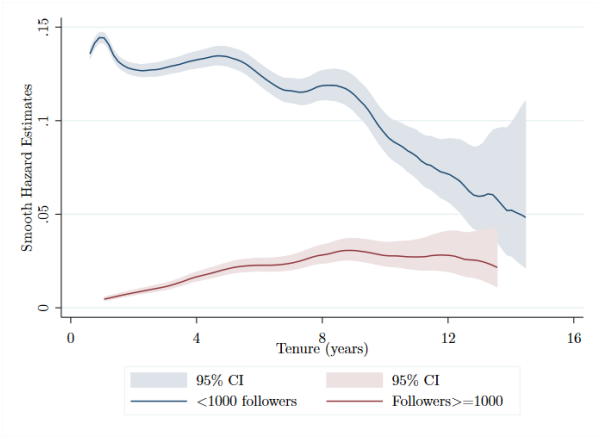
(a) Kaplan-Meier survival function



(b) Hazard function



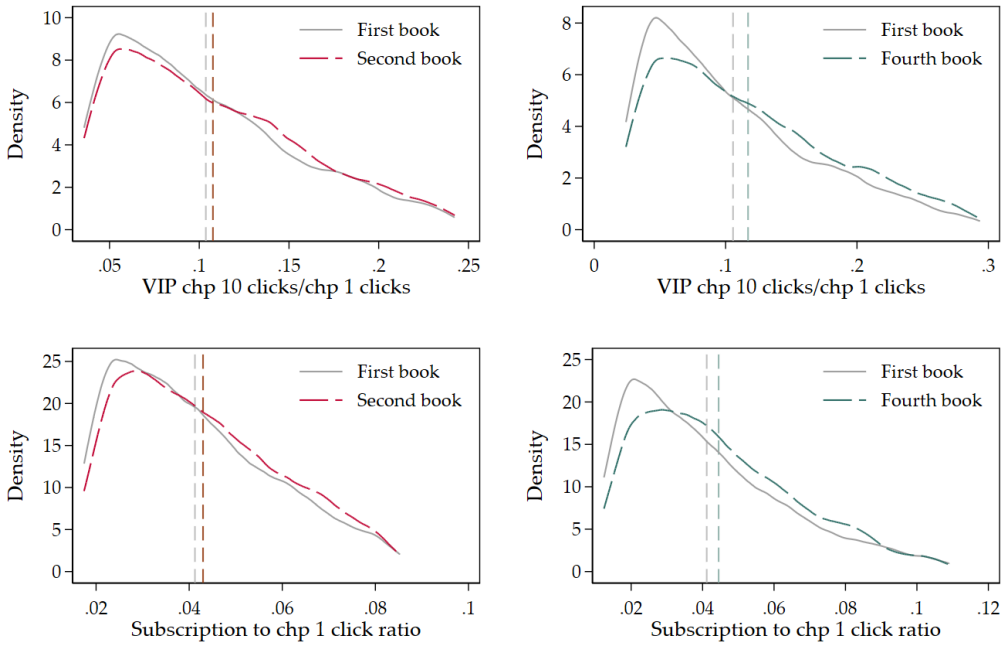
(c) Kaplan-Meier survival function by followers



(d) Hazard function by followers

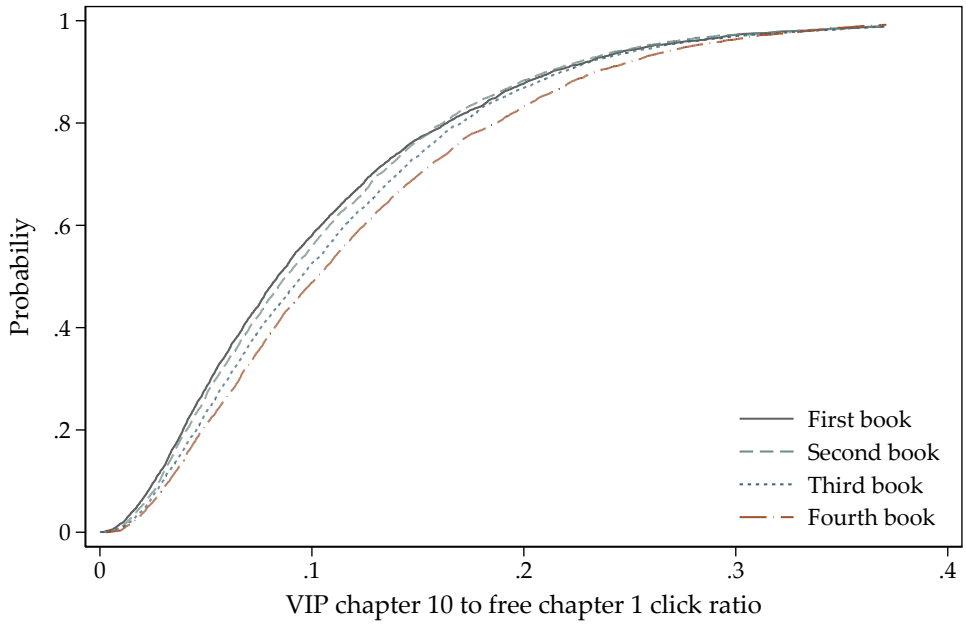
Notes: Panel (a) and (c) plot the Kaplan-Meier survival function over tenure of the authors. An author is considered inactive, or has exited, if he or she has not uploaded anything since January 2020 (inactive for more than 12 months). Panel (b) and (d) plot the smoothed hazard functions.

Figure A4: Distribution Shift Between First and Fourth Book



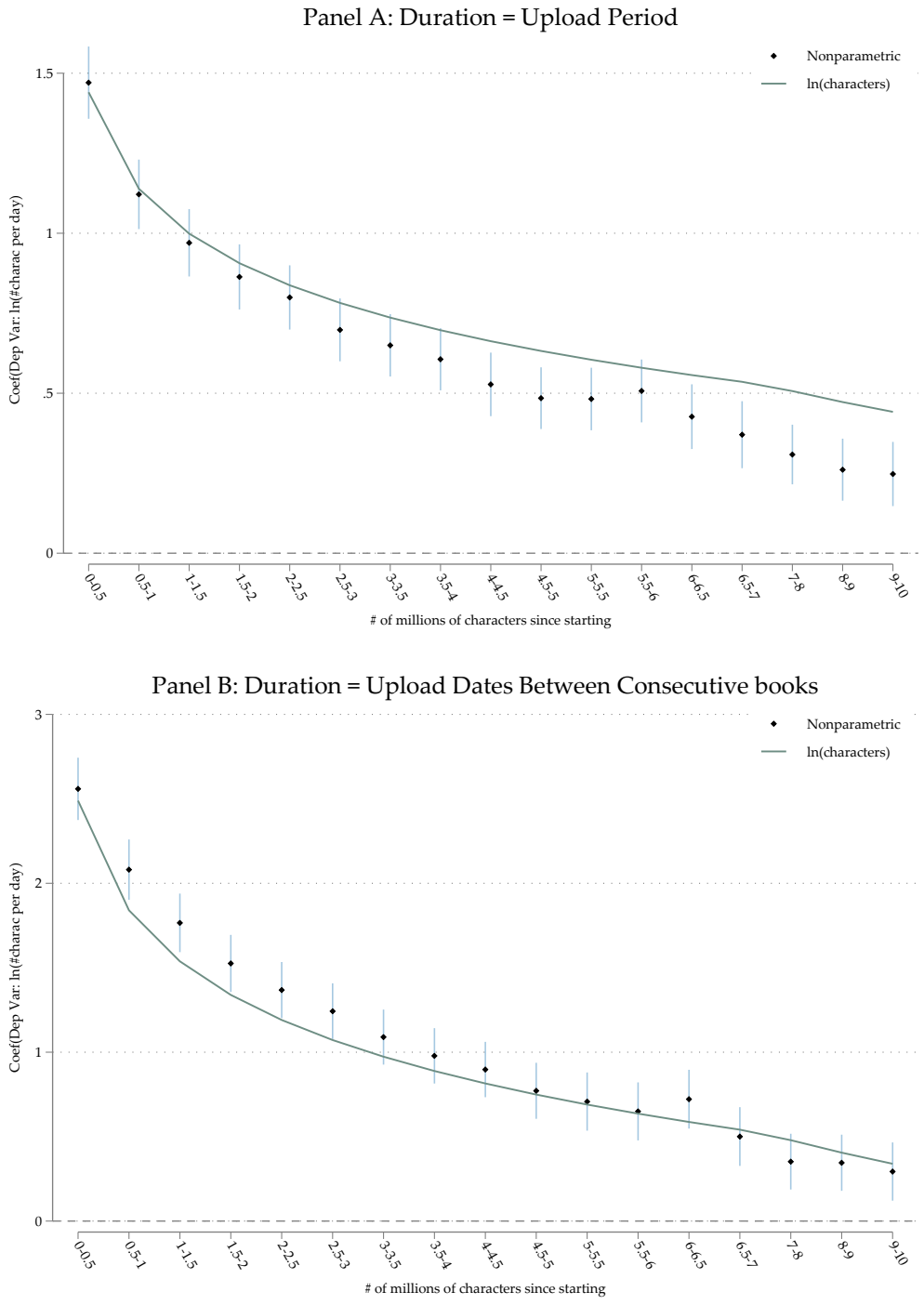
Notes: This figure plots the distribution of the first and the fourth books. I keep authors who have written at least five books, and for clear presentation, I drop books over the 95th and below the 5th percentile. Vertical lines represent the mean in each group.

Figure A5: CDF of the VIP chp 10 to chp 1 click ratio



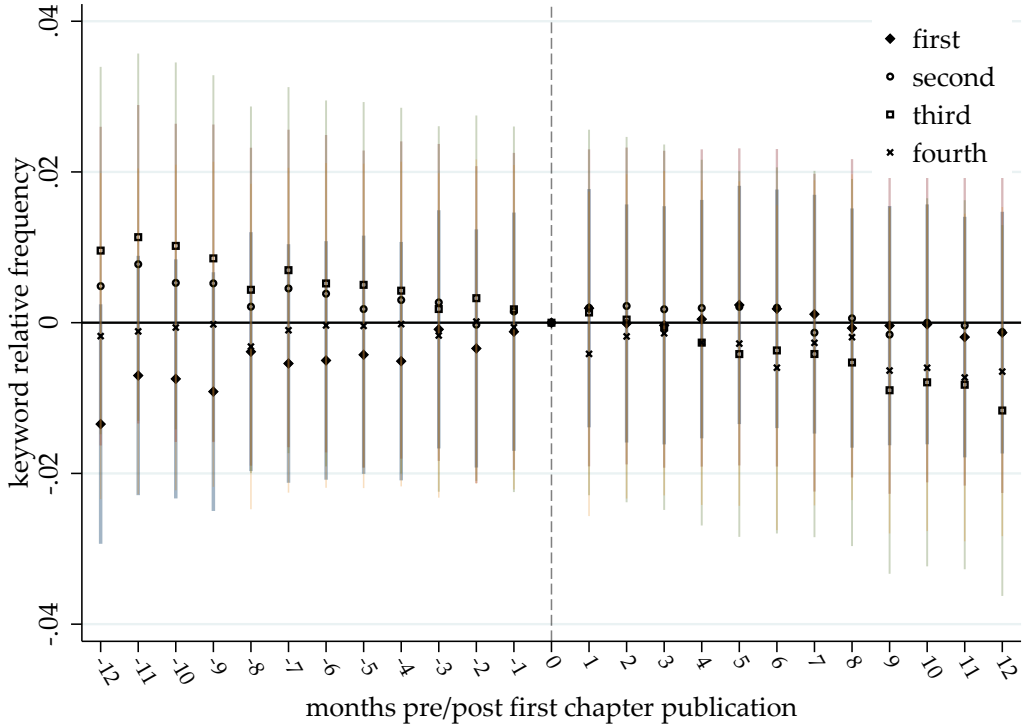
Notes: This figure plots the cumulative distribution function of the VIP chapter 10 clicks to free chapter 1 click ratio for the first four books of the authors. Only authors who (1) have written at least five book; and (2) have a valid VIP chapter 10 to chapter 1 click ratio for all four books are included. The author sample size is 4608.

Figure A6: Writing speed progression with experience



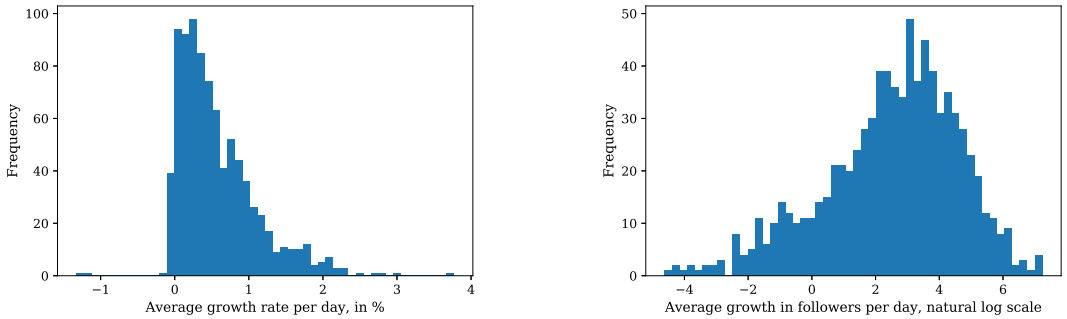
Notes: This figure plots the coefficient estimates of parametric and nonparametric specification shown in equation 2. The outcome variable for Panel A is the average number of characters published per day by the authors during the period when a book is being uploaded. The outcome variable for Panel B is the average number of characters published per day by the authors during the first upload dates of two consecutively published books.

Figure A7: Keyword relative frequency among new titles



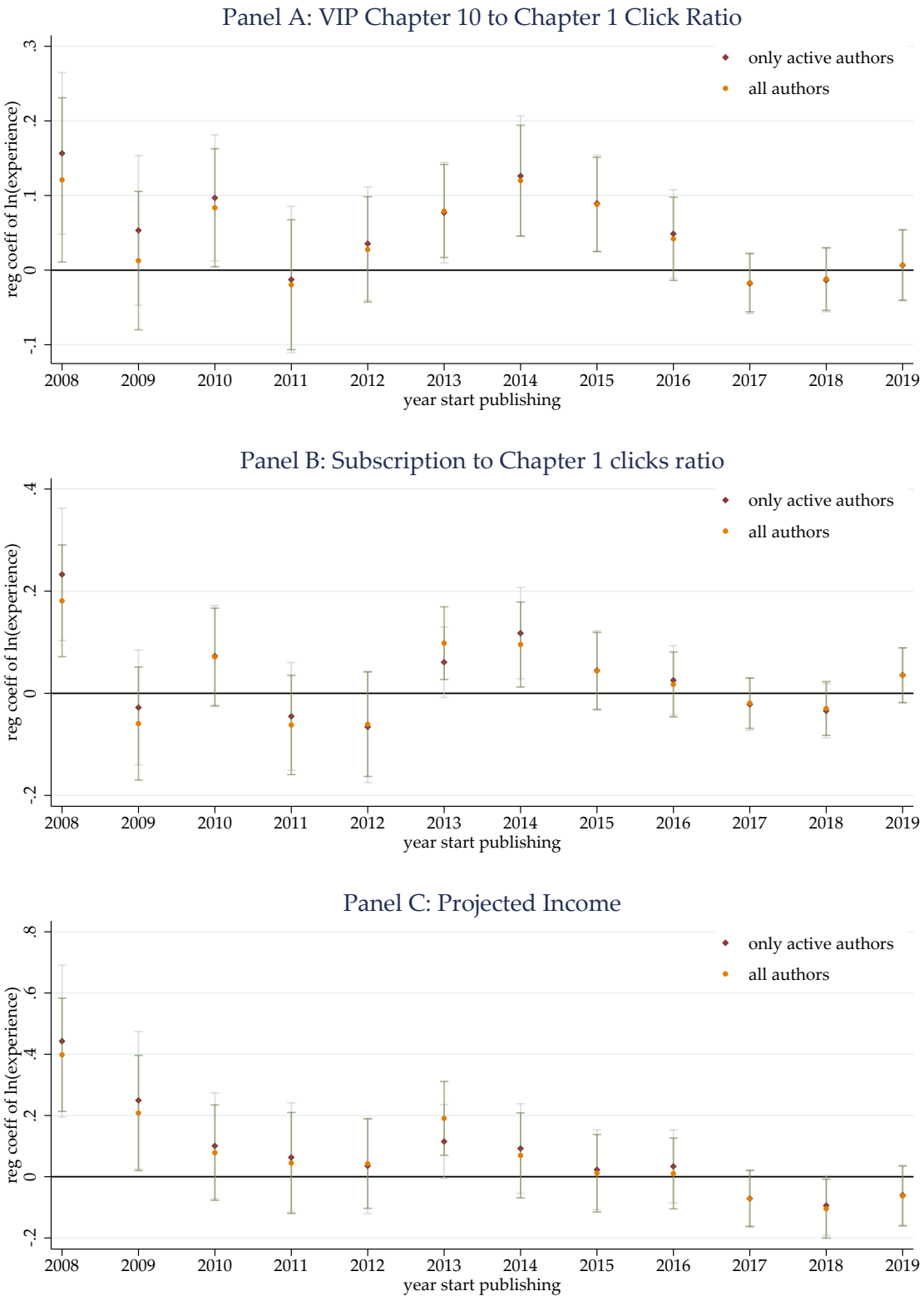
Notes: This figure plots the relative frequency of four keywords of a superstar title among new titles published 12 months before and 12 months after the superstar is published. A superstar is defined as the top 1% title within a broad genre in a quarter if that genre has over 100 titles published in that quarter.

Figure A8: Average daily growth rate of followers



Notes: This figure plots the average daily percentage increase in followers from February 26, 2021 to September 17, 2021. Books sample from new books written by authors with over 1000 followers and the first chapter is uploaded between January 7 and February 25, 2021.

Figure A9: Cohort-specific regression coefficients



Appendix B Additional tables

Table B1: Improvement with experience with alternative definition of new authors

Specification	(1)	(2)	(3)	(4)	(5)	(6)
New			-0.0233*			-0.0125
			(0.0123)			(0.0123)
New×ln(characters)			-0.000275			-0.00611
			(0.0104)			(0.0105)
ln(characters)	0.00708	0.0515***	0.0480***	0.0114	0.0455***	0.0484***
	(0.0104)	(0.00627)	(0.0116)	(0.0105)	(0.00639)	(0.0117)
New author only	Yes	No	No	Yes	No	No
Length	No	No	No	Yes	Yes	Yes
Genre FE	No	No	No	Yes	Yes	Yes
Observations	25712	48545	48545	25017	46748	46748
R^2	0.618	0.571	0.571	0.633	0.588	0.588

Note: The outcome variable is the natural logarithm of VIP chapter 10 click to first chapter click ratio. New authors are defined as authors who have joined the platform for less than two years. Robust standard errors in parentheses, clustered at the author level. Author and upload year-month fixed effects included in all specifications.

*** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table B2: Robustness: author improvement with experience, controlling for writing speed

	ln(Vchp10/chp1 click)	ln(Vchp10/chp1 click)	ln(Vchp10/chp1 click)
	(1)	(2)	(3)
ln(characters)	0.0466***	0.0395***	0.0588***
	(0.00652)	(0.00659)	(0.00676)
ln(charac/day)		-0.0278***	
		(0.00482)	
ln(charac/day)			0.0207***
(alternative)			(0.00225)
Observations	45494	43133	45494
R^2	0.592	0.608	0.593

Notes: Robust standard errors in parentheses, clustered at the author level. Author, length, genre, and upload year-month FE included in all specifications. Characters per day is defined as the average number of characters the author publishes when the book is being uploaded. The alternative measure of characters per day is the average number of characters the author publishes between the publication date of the first chapter of two consecutive books. The outcome variable is the natural logarithm of VIP chapter 10 click to first chapter click ratio.
** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table B3: Reader retention rate of new book and authors' characteristics

	(1)	(2)	(3)	(4)
	chp3/chp1 click	chp3/chp1 click	chp3/chp1 click	chp3/chp1 click
ln(characters)	-0.00389			0.00676
	(0.00641)			(0.00755)
ln(followers)		-0.0111**		-0.0137***
		(0.00431)		(0.00510)
avg chp10/chp1			-0.00573	-0.00623
click ratio			(0.00529)	(0.00528)
Observations	888	888	887	887
R^2	0.202	0.208	0.193	0.201

Notes: Standard errors in parentheses. Length and genre fixed-effects included in all specifications. The outcome variable is the new book's chapter 3 to chapter 1 click ratio measured 90 days after it is published. ln(followers) is the natural logarithm of the author's number of followers. avg chp10/chp1 click ratio is the average chapter 10 to chapter 1 click ratio of all books previously written by that author.
*** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table B4: Summary Statistics: Active versus Inactive Authors

	Active N=5425			Inactive N=1136		
	Mean	Median	SD	Mean	Median	SD
author following	10547.85	2955	37776.82	5557.86	1935.5	37231.95
avg book following (000)	18.32	10.12	30.65	9.23	4.1	40.96
experience (years)	6.13	5.39	3.37	9.43	9.37	3.51
num of characters (millions)	3.66	2.79	3.16	1.76	1.41	1.5
number of books	9.43	8	6.75	5.32	4	4.01
average clicks per chp	17.98	11.5	24.72	15.59	9.85	36.66
VIP chp10/1st chp click	0.15	0.14	0.06	0.09	0.08	0.06
Subscribers/1st chp click	0.06	0.05	0.03	0.04	0.04	0.03

Notes: Sample contains authors who have finished at least one book. Titles with no content were dropped when calculating average length per title. Book sample contains novels, novella, and short stories written by authors with more than 1000 followers. Authors are defined as inactive if they have not updated or revised anything since January 2021.