Title
Essays on Platform Policies, Ratings and Innovation

Author
He, Xiuyi

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Essays on Platform Policies, Ratings and Innovation

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Management

by

Xiuyi He

2023

ABSTRACT OF THE DISSERTATION

Essays on Platform Policies, Ratings and Innovation

by

Xiuyi He

Doctor of Philosophy in Management

University of California, Los Angeles, 2023

Professor Brett William Hollenbeck, Chair

Reputation and feedback systems are commonly integrated as a part of online marketplaces. However, the majority of the literature focuses on the static impact of online reviews (Reimers and Waldfogel, 2019; Tadelis, 2016). There is a growing body of research showing that firms respond to online reviews by taking certain actions, including adjusting advertising strategy accordingly (Hollenbeck et al., 2019), manipulating seller reputation with fake reviews (Mayzlin et al., 2014; Luca and Zervas, 2016), adopting costly short-run action to improve ratings (Hunter, 2020) and replying to reviews and improving product quality based on reviews (Proserpio and Zervas, 2016; Ananthakrishnan et al., 2019). Given the economic significance of two-sided platforms, each platform policy change can have large impacts on consumers, sellers and the platforms themselves. Across two essays, I aim to show two types of firm responses to their ratings and shed light on their corresponding platform rating policy implications.

In Chapter 1, we study the market of fake product reviews on Amazon.com. Reviews are

purchased in large private groups on Facebook and other sites. We hand collected data on these markets and then collected a panel of data on these products' ratings and reviews on Amazon, as well as their sales rank, advertising, and pricing policies. Using detailed data on product outcomes before and after they buy fake reviews, we can directly determine if these are low-quality products using fake reviews to deceive and harm consumers or if they are high-quality products that solicit reviews to establish reputation. We find that a wide array of products purchase fake reviews, including products with many reviews and high average ratings. Buying fake reviews on Facebook is associated with a significant but short-term increase in average rating and number of reviews. We exploit a sharp but temporary policy shift by Amazon to show that rating manipulation has a large causal effect on sales. The theoretical literature on review fraud shows conditions when they are a deceptive form of fraud and conditions where they function as simply another form of advertising. Finally, we examine whether rating manipulation harms consumers or whether it is mainly used by high-quality product producers as an alternative to advertising or by new products trying to solve the cold-start problem. We find that after firms stop buying fake reviews, their average ratings fall and the share of one-star reviews increases significantly, particularly for younger products, indicating rating manipulation is mostly used by low-quality product producers. Finally, we observe that Amazon deletes large numbers of reviews, and we document their deletion policy.

In Chapter 2, we study how rating system design affects innovation incentives. In settings where product quality cannot be observed prior to purchase, online ratings serve as a signal of product quality for consumers and affect demand. Owing to their impact on sales, ratings also motivate firms to innovate. If firms use displayed ratings to guide their investments in improving product quality, then platform rating aggregation policies can play a key role in increasing or decreasing firms' innovation incentives. We study in depth the impact of online rating systems on innovation incentives and, more importantly, the corresponding implications of the design of the rating aggregation policy. After collecting a unique firm-

level dataset from a mobile game app platform, we combined reduced-form analysis and the structural model to show how rating systems can be optimized for innovation. We show that innovation has a positive impact on all key rating system metrics. Building on empirical evidence, we developed a dynamic structural model to represent firms' forward-looking behavior and estimate innovation cost. We then evaluate the impact of alternative rating aggregation policies on innovation incentives. The counterfactual analysis shows that placing greater weight on recent ratings can increase the innovation rate substantially.

Across two chapters, this dissertation contributes substantively and theoretically to our comprehension of how firms respond to their online ratings and how two-sided platforms can design better policies to combat fake reviews and encourage firm innovation. As rating systems are increasingly adopted by platforms and consumers rely on ratings to make decisions, it is important to design better platform rating policies to help consumers, honest firms, and the platforms themselves.

The dissertation of Xiuyi He is approved.

Anand V. Bodapati

Randolph E. Bucklin

Sylvia Hristakeva

Brett William Hollenbeck, Committee Chair

University of California, Los Angeles

2023

To my parents Xi and Jiaxun, and my husband Albert,

for their continuing love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am deeply grateful to Professor Hollenbeck, my advisor, whose guidance has been instrumental in my success throughout the PhD program. His insightful questions and encouragement have pushed me to delve deeper and work more efficiently. Most importantly, he has been a role model, exemplifying what it means to be an exceptional researcher. Working with him over the past few years has been an incredibly fortunate and exhilarating experience, marked by numerous paper publications and media mentions.

I would also like to extend my gratitude to Professor Hristakeva, whose impact on various aspects of my PhD journey has been overwhelmingly positive. Her unwavering support, candid feedback, and influence on my thinking have shaped my academic growth. From our very first meeting on admit day, I knew that I would enjoy a fulfilling PhD life at UCLA due to her presence.

I am indebted to the remaining members of my committee. Professor Bucklin, thank you for your availability, insightful questions, and valuable feedback. Professor Bodapati, your constant encouragement and support in attending conferences and workshops have been invaluable. Through your mentorship, I have learned to proactively pursue opportunities and make things happen.

I extend my gratitude to all the faculty members in Marketing at UCLA for fostering a welcoming and supportive environment. Professor Rossi, thank you for imparting valuable knowledge on teaching and research presentation, which has been crucial for my job market preparations. To Professor Hershfield, your guidance and assistance throughout the PhD program have been immensely helpful. Professor Honka, thank you for your emotional and practical support during my PhD journey and the job market.

I consider myself fortunate to have had a close-knit cohort. Special thanks go to Julia

Levine, Ipek Demirdag, Pedro Makhoul, Joey Reiff, David Zimmerman, and Jon Bogard. Our countless game nights and mutual support have played a pivotal role in my achievements. I would also like to express my gratitude to David Dolifka, Daniel Mirny, Mahsa Paridar, Neha Nair, Eunsun Kim, Tayler Bergstorm, and Kalyan Rallabandi for enriching my life. A special mention goes to Dan Yavorsky, whose mentoring presence has been akin to that of a caring older brother.

I would like to thank Professor Proserpio for the incredible experience of working together on the fake review project. I have gained invaluable research skills and learned about research quality under his guidance. I also want to acknowledge that the first chapter of my dissertation is a joint work with Professor Hollenbeck, and Professor Proserpio. The corresponding DOI link of the final version is https://doi.org/10.1287/mksc.2022.1353.

My friend and coauthor, Jingcun Cao, deserves my heartfelt thanks. Without his support and engaging research discussions, I wouldn't have found the inspiration for the second chapter of my dissertation.

I am immensely grateful to Professor Steenkamp and Professor Feinberg, who introduced me to the field of quantitative marketing and ignited my passion for academic research. Their support during critical moments of my job market journey was invaluable. This PhD experience has been the most remarkable adventure of my life thus far, and I owe it all to them.

I am eternally grateful to my incredible husband, Albert, who has been a constant source of support throughout my PhD journey and the academic job market. He has prepared meals with the perfect sugar ratio, picked me up from the airport late at night on multiple occasions, and provided comfort during moments of anxiety. He unwaveringly believes in me, even during moments of self-doubt. I eagerly anticipate spending a lifetime together.

The utmost gratitude goes to my mother, Xi, who instilled in me the ability to envision a

better future and to embrace resilience in the face of challenges. Her support and belief in me have been unwavering, allowing me to be true to myself. Additionally, I express deep appreciation to my father, Jiaxun, for exemplifying a strong work ethic that has inspired me throughout my journey. Their financial support and encouragement to make my own decisions have been invaluable. They are my greatest supporters, and I am motivated every day to make them even prouder.

VITA

Xiuyi (Sherry) He

EDUCATION

B.A. Economics & Statistics 2017

University of Michigan - Ann Arbor

PUBLICATION

He, Sherry, Brett Hollenbeck, and Davide Proserpio. "The market for fake reviews." *Marketing Science* (2022).

He, Sherry, Brett Hollenbeck, Gijs Overgoor, Davide Proserpio, and Ali Tosyali. "Detecting fake-review buyers using network structure: Direct evidence from Amazon." *Proceedings of the National Academy of Sciences* 119, no. 47 (2022): e2211932119.

Davide Proserpio, Brett Hollenbeck, and Sherry He. "How fake customer reviews do–and don't–work." *Harvard Busines Review* (2020).

He, Sherry, Brett Hollenbeck, and Davide Proserpio. "Exploiting social media for fake reviews: evidence from Amazon and Facebook." *ACM SIGecom Exchanges* 19, no. 2 (2021): 68-74.

CONFERENCE PRESENTATION

Wharton Innovation Doctoral Symposium, 2022

Association for Consumer Research Conference, 2021

19th Conference on the Economics of Information and Communication Technologies, 2021

14th Digital Economics Conference, 2021

Conference on Digital Experimentation, 2020

Conference on Artificial Intelligence, Machine Learning and Business Analytics, 2020

HONORS & AWARDS

Dissertation Year Fellowship, UCLA 2022

AMA-Sheth Foundation Doctoral Consortium Fellow, 2021

Anderson School of Management PhD Fellowship, UCLA 2017-2022

Morrison Center for Marketing Analytics Research Grant ($5,000 in total), UCLA 2019-2020

The Ferrando Honors Prize, University of Michigan, 2017

Chapter 1

The Market for Fake Reviews

1.1 Introduction

Online markets have from their first days struggled to deal with malicious actors. These include consumer scams, piracy, counterfeit products, malware, viruses, and spam.[1] And yet online platforms have become some of the world's largest companies in part by effectively limiting these practices and earning consumer trust. The economics of platforms suggest a difficult trade-off between opening the platform to outside actors such as third-party sellers and retaining strict control over actions taken on the platform. Preventing fraudulent or manipulative actions is key to this trade-off.

One such practice is manipulating reputation systems with fake product reviews. Conventional wisdom holds that fake reviews are particularly harmful because they inject noise and deception into systems designed to alleviate asymmetric information, cause consumers to

[1] Recent work has documented many examples including firms increasing their visibility in search rankings via fake downloads (Li et al., 2016), increasing revenue via bot-driven advertising impressions (Wilbur and Zhu, 2009; Gordon et al., 2021), manipulating social network influence with fake followers, manipulating auction outcomes, or defrauding consumers with false advertising claims (Rao and Wang, 2017; Chiou and Tucker, 2018; Rao, 2018).

purchase products that may be of low quality, and erode the long-term trust in the review platforms that is crucial for online markets to flourish (Cabral and Hortacsu, 2010; Einav et al., 2016; Tadelis, 2016).

We study the economics of rating manipulation and its effect on seller outcomes, consumer welfare, and platform value. Despite being illegal, we document the existence of large and active online markets for fake reviews.[2] Sellers post in private online groups to promote their products and pay customers to purchase them and leave positive reviews. These groups exist for many online retailers, including Walmart and Wayfair, but we focus on Amazon because it is the largest and most developed market. We collect data from this market by sending research assistants into these groups to document which products are buying fake reviews and when.[3] We then track these products' outcomes on Amazon.com, including their reviews, ratings, prices, and sales rank. This is the first data of this kind, providing direct evidence on the fake reviews themselves and on the outcomes from buying fake reviews.

The mere existence of such a large and public market for fake reviews on the largest e-commerce platform presents a puzzle. Given the potential reputation costs, why does Amazon allow this? In the short run, platforms may benefit from allowing fake positive reviews if these reviews increase revenue by generating sales or allowing for higher prices. It is also possible that fraudulent reviews are not misleading on average if high-quality firms are more likely to purchase them than low-quality firms. They could be an efficient way for sellers to solve the "cold-start" problem and establish a good reputation. Indeed, Dellarocas (2006) shows that this is a potential equilibrium outcome. In an extension of the signal-jamming

[2]The FTC has brought cases against firms alleged to have posted fake reviews, including a case against a weight-loss supplement firm buying fake reviews on Amazon in February 2019. See: `https://www.ftc.gov/news-events/press-releases/2019/02/ftc-brings-first-case-challenging-fake-paid-reviews-independent`.
On May 22, 2020, toward the end of our data collection window, the UK Competition and Markets Authority (CMA) announced it was opening an investigation into these practices. See: `https://www.gov.uk/government/news/cma-investigates-misleading-online-reviews`.

[3]While technically the seller buys the fake reviews, not the product, because our analysis is done at the product level and sellers often have many products, for clarity we refer to products buying fake reviews.

literature on how firms can manipulate strategic variables to distort beliefs, he shows that fake reviews are mainly purchased by high-quality sellers and, therefore, increase market information under the condition that demand increases convexly with respect to user rating. Given how ratings influence search results, it is plausible that this condition holds. Other attempts to model fake reviews have also concluded that they may benefit consumers and markets.[4] The mechanism is different, but intuitively this outcome is similar to signaling models of advertising for experience goods. Nelson (1970) and later Milgrom and Roberts (1986) show that separating equilibria exist where higher quality firms are more likely to advertise because the returns from doing so are higher for them. This is because they expect repeat business or positive word-of-mouth once consumers have discovered their true quality. If fake reviews generate sales which, in turn, generate future organic ratings, a similar dynamic could play out. In this case, fake reviews may be seen as harmless substitutes for advertising rather than as malicious. Therefore, we are left with an empirical question as to whether or not to view rating manipulation as representing a significant threat to consumer welfare and platform reputations.

Our research objective is to answer a set of currently unsettled questions about online rating manipulation. How does this market work, in particular, what are the costs and benefits to sellers from buying fake reviews? What types of products buy fake reviews? How effective are they at increasing sales? Does rating manipulation ultimately harm consumers or are they mainly used by high quality products? That is, should they be seen more like advertising or outright fraud? Do fake reviews lead to a self-sustaining increase in sales and organic ratings? These questions can be directly answered using the unique nature of our data.

We construct a sample of approximately $1,500$ products observed soliciting fake reviews over a nine-month period. We find a wide assortment of product types in many categories. Many

[4]These attempts include (Glazer et al., 2020) and Yasui (2020). In addition, both Wu and Geylani (2020) and Rhodes and Wilson (2018) study models of deceptive advertising and conclude that this practice can benefit consumers under the right conditions.

products have a large number of reviews and few are new to Amazon. These products do not have especially low ratings, with an average rating slightly higher than comparable products we do not observe soliciting fake reviews. Almost none of the sellers purchasing reviews in these markets are well-known brands, consistent with research showing that online reviews are more effective and more important for small independent firms than for brand name firms (Hollenbeck, 2018).

We then track the outcomes of these products before and after the buying of fake reviews. In the weeks after they start to purchase fake reviews, the number of reviews posted per week roughly doubles. The average rating and share of five-star reviews also increase substantially, as do search position and sales rank. The increase in average ratings is short-lived, with ratings falling back to the previous level within two to four weeks, but the increase in the weekly number of reviews, sales rank, and position in search listings remains substantially higher more than four weeks later. We also track outcomes after the last observed post soliciting fake reviews and find that the increase in sales is not self-sustaining. Sales begin to fall significantly right after the fake review campaign ends. New products with few reviews, which might be using fake reviews efficiently to solve the cold-start problem, see a larger increase in sales initially but a similar drop-off afterward.

We also document how the platform regulates fake reviews. We see that Amazon ultimately deletes a very large share of reviews. For the products in our data observed buying fake reviews, roughly half of their reviews are eventually deleted, but the deletions occur with an average lag of over 100 days, thus allowing sellers to benefit from the short-term boost in ratings, reviews, and sales.

Next, to understand how effective and profitable this practice is, we leverage review deletions to measure the causal effect of fake reviews on sales. Our previous results are descriptive, and the increase in sales we document could be attributed in part to factors other than fake reviews, include unobserved demand shocks, advertising, or price cuts. To isolate the effect

of rating manipulation on sales, we take advantage of a short period in which Amazon mass deletes a large number of reviews. Products that purchased fake reviews just before this period do not receive the boost in positive reviews that other products buying fake reviews do, but they behave similarly otherwise, allowing us to use these products as a control group. Comparing outcomes across products, we find that rating manipulation causes a significant increase in sales.

Lastly, we turn to the question of whether rating manipulation is efficient or it harms consumers. To do so, we study reviews and ratings posted after the fake review purchases end. If the products continue to receive high ratings from consumers, it suggests that fake reviews are more like advertising and are mainly bought by high-quality products, potentially solving the cold-start problem. If, by contrast, ratings fall and they receive many one-star ratings, it suggests that consumers felt they were deceived into buying products whose true quality was lower than they expected at the time of purchase and, therefore, they overpaid or missed out on a higher quality alternative. While there is an inherent limitation in using ratings to infer welfare, we nevertheless find that the evidence primarily supports the consumer harm view. The share of reviews that are one-star increases by 70% after fake review purchases, relative to before. This pattern is especially true for new products and those with few reviews. Text analysis shows that these one-star reviews are distinctive and place a greater focus on product quality, further confirming that consumers were deceived.

Prior studies of fake reviews include Mayzlin et al. (2014), who argue that in the hotel industry, independent hotels with single-unit owners have a higher net gain from manipulating reviews. They then compare the distribution of reviews for these hotels on Expedia and TripAdvisor and find evidence consistent with review manipulation. Luca and Zervas (2016) use Yelp's review filtering algorithm as a proxy for fake reviews and find that these reviews are more common on pages for firms with low ratings, independent restaurants, and restaurants with more close competitors. Using lab experiments, Ananthakrishnan et al. (2020)

show that a policy of flagging fake reviews but leaving them posted can increase consumer trust in a platform.

We contribute to this literature by documenting the actual market where fake reviews are purchased and the sellers participating in this market. This data gives us a direct look at rating manipulation, rather than merely inferring its existence. Our data on firm outcomes before and after rating manipulation allow us to understand the short- and long-term effectiveness of rating manipulation and assess whether and when consumers are harmed by them.

This research also contributes to the broader academic study of online reviews and reputation. By now, it is well understood that online reviews affect firm outcomes and improve the functioning of online markets (see Tadelis (2016) for a review). There is also a growing body of research showing that firms take actions to respond to online reviews, including by leaving responses directly on review sites (Proserpio and Zervas, 2016) and changing their advertising strategy (Hollenbeck et al., 2019). A difficult tension has always existed in the literature on online reviews, coming from the fact that the reviews and ratings being studied may be manipulated by sellers. By documenting the types of sellers purchasing fake reviews and the size and timing of their effects on ratings and reviews, we provide guidance to future researchers on how to determine whether review manipulation is likely in their setting.

1.2 Data and Settings

In this section, we document the existence and nature of online markets for fake reviews and discuss in detail the data collection process and the data we obtained to study rating manipulation and its effect on seller outcomes, consumer welfare, and platform value. We collected data mainly from two different sources, Facebook and Amazon. From Facebook,

we obtained data about sellers and products buying fake reviews, while from Amazon we collected product information such as reviews, ratings, and sales rank data.

1.2.1 Facebook Groups and Data

Facebook is one of the major platforms that Amazon sellers use to recruit fake reviewers. To do so, sellers create private Facebook groups where they promote their products by soliciting users to purchase their products and leave a five-star review in exchange for a full refund (and in some cases an additional payment). Discovering these groups is straightforward by searching for "Amazon Review." We begin by documenting the nature of these groups and then describe how we collect product information from them.

Discovering groups We collected detailed data on the extent of Facebook group activity from March 28, 2020 to Oct 11, 2020. Each day, we collected the Facebook group statistics for the top 30 groups by search rank. During this period, on average, we identify about 23 fake review related groups every day. These groups are large and quite active, with each having about 16,000 members on average and 568 fake review requests posted per day per group. We observe that Facebook periodically deletes these groups but that they quickly reemerge. Figure 1.1 shows the weekly average number of active groups, number of members, and number of posts between April and October of 2020.[5]

Within these Facebook groups, sellers can obtain a five-star review that looks organic. Figure 1.2 shows examples of Facebook posts aimed at recruiting reviewers. Usually, these posts contain words such as "need reviews," "refund after pp [PayPal]" with product pictures. The reviewer and seller then communicate via Facebook private messages. To avoid being detected by Amazon's algorithm, sellers do not directly give reviewers the product link;

[5]The total number of members and posts likely overstates the true amount of activity due to double-counting the same sellers and reviewers across groups.

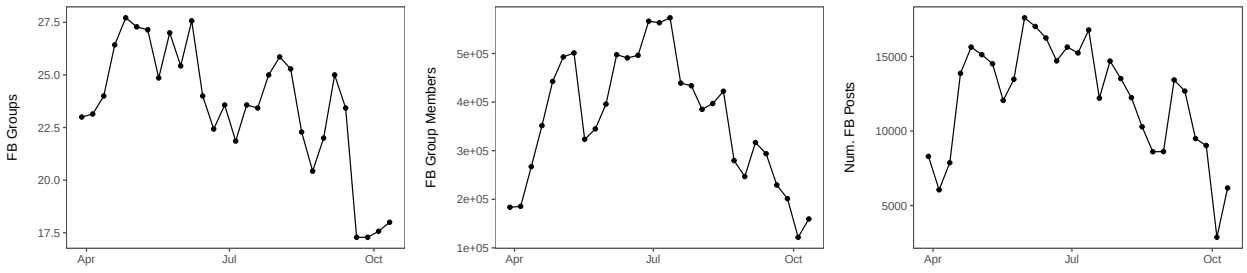


Figure 1.1: Weekly average number of FB groups, members, and seller posts

instead, sellers ask reviewers to search for specific keywords associated with the product and then find it using the title of the product, the product photo, or a combination of the two.

The vast majority of sellers buying fake reviews compensate the reviewer by refunding the cost of the product via a PayPal transaction after the five-star review has been posted (most sellers advertise that they also cover the cost of the PayPal fee and sales tax). Moreover, we observe that roughly 15% of products also offer a commission on top of refunding the cost of the product. The average commission value is $6.24, with the highest observed commission for a review being $15. Therefore, the vast majority of the cost of buying fake reviews is the cost of the product itself.

Reviewers are compensated for creating realistic seeming five-star reviews, unlike reviews posted by bots or cheap foreign workers with limited English skills, which are more likely to be filtered by Amazon's fraud detection algorithms. The fact that the reviewer buys the product means that the Amazon review is listed as a "Verified Purchase" review and reviewers are encouraged to leave lengthy, detailed reviews that include photos and videos to mimic authentic and organic reviews.[6] Finally, sellers recruit only reviewers located in the United States, with an Amazon.com account, and with a history of past reviews.

This process differs from "incentivized reviews," where sellers offer free or discounted products or discounts on future products in exchange for reviews. Several features distinguish fake

[6]The fact that these fake reviews are from verified purchases indicates that an identification strategy like the one used in Mayzlin et al. (2014) will not work in settings like these.

reviews from incentivized reviews. The payment for incentivized reviews is not conditional on the review being positive, whereas reimbursement for fake reviews requires a five-star rating. Incentivized reviews, in principle, contain informative content for consumers, whereas in many cases the reviewer posting a fake review has not used or even opened the product. Finally, incentivized reviews typically involve disclosure in the form of a disclaimer contained in the review itself that the product was received for free or at a discount in exchange for the review. [7]

Discovering products We use a group of research assistants to discover products that are promoted. Facebook displays the posts in a group in an order determined by some algorithm that factors in when the post was made as well as engagement with the post via likes and comments. Likes and comments for these posts are relatively rare and so the order is primarily chronological. We directed our research assistants to randomize which products were selected by scrolling through the groups and selecting products in a quasi-random way while explicitly ignoring the product type/category, amount of engagement with the post, or the text accompanying the product photo.

Given a Facebook post, the goal of the research assistants is to retrieve the Amazon URL of the product. To do so, they use the keywords provided by the seller. For example, in Figure 1.2, the search words would be "shower self," "toilet paper holder," and "cordless vacuum." After a research assistant successfully identifies the product, we ask them to document the search keywords, product ID, product subcategory (from the Amazon product page), date of the Facebook post, the earliest post date from the same seller for the same product (if older posts promoting the same product exist), and the Facebook group name.

We use the earliest Facebook post date as a proxy for when the seller began to recruit

[7] Amazon has at times allowed incentivized reviews and even has formally sponsored them through its Vine program and its "Early Reviewer Program," but the company considers fake reviews a violation of its terms of service by both sellers and reviewers, leaving them subject to being banned from the platform if caught.

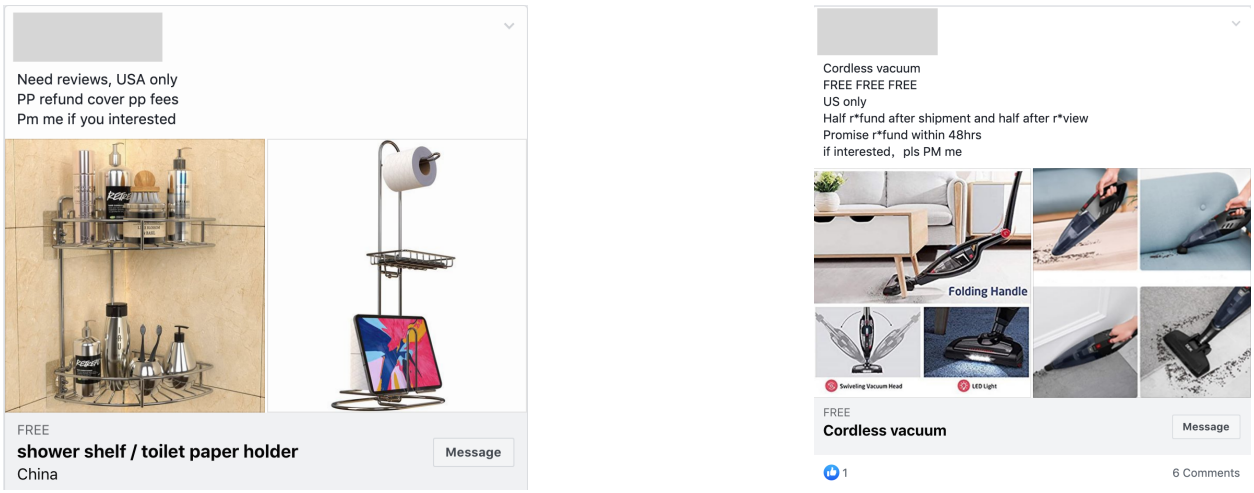


Figure 1.2: Examples of Fake Review Recruiting Posts

fake reviewers. To identify when a seller stops recruiting fake reviews for a product, we continuously monitor each group and record any new posts regarding the same product by searching for the seller's Facebook name and the product keywords. We then use the date of the last observed post as a proxy for when the seller stopped recruiting fake reviews.

We collect data from these random Facebook fake review groups using this procedure on a weekly basis from October 2019 to June 2020, and the result is a sample of roughly 1,500 unique products. This provides us with the rough start and end dates of when fake reviews are solicited, in addition to the product information.

1.2.2 Amazon Data

After identifying products whose ratings are manipulated, we collect data for these products on Amazon.com.

Search Results Data For each product buying fake reviews, we repeatedly collect all information from the keyword search page results, i.e., the list of products returned as a result of a keyword search query. This set of products is useful to form a competitor set

for each focal product. We collect this information daily, including price, coupon, displayed rating, number of reviews, search page number, whether the product buys sponsored listings, and the product position in each page.

Review Data We collect the reviews and ratings for each of the products on a daily basis. For each review, we observe rating, product ID, review text, presence of photos, and helpful votes.

Additionally, twice per month we collect the full set of reviews for each product. The reason for this is that it allows us to measure to what extent Amazon responds by deleting reviews that it deems as potentially fake.

In addition to collecting this data for the focal products, we collect daily and twice-monthly review data for a set of 2,714 competitor products to serve as a comparison set. To do so, for each focal product we select the two competitor products who show up most frequently on the same search page as the focal product in the seven days before and seven days after their first FB post. The rationale is that we want to create a comparison set of products that are in the same subcategory as the focal products and have a similar search rank. We collect these products' reviews data from Aug 14th, 2020 to Jan 22rd, 2021.

Sales Rank Data We rely on Keepa.com and its API to collect sales rank data twice a week for all products. Amazon reports a measure called Best Seller Rank, whose exact formula is a trade secret, but which translates actual sales within a specific period of time into an ordinal ranking of products.

1.2.3 Descriptive Statistics

Here, we provide descriptive statistics on the set of roughly 1,500 products collected between October 2019 to June 2020. We use this sample of products to characterize the types of products that sellers promote with fake reviews. On the one hand, we might expect these products to be primarily products that are new to Amazon.com with few or no reviews whose sellers are trying to jump-start sales by establishing a good online reputation. On the other hand, these might be products with many reviews and low average ratings, whose sellers resort to fake reviews to improve the product reputation and therefore increase sales.

Table 1.1 shows a breakdown of the top 15 categories and subcategories in our sample. Fake reviews are widespread across products and product categories. The top categories are "Beauty & Personal Care," "Health & Household," and "Home & Kitchen," but the full sample of products comes from a wide array of categories, and the most represented product in our sample, Humidifiers, only accounts for roughly 1% of products. Nearly all products are sold by third-party sellers.

Table 1.1: Focal Product Top Categories and Subcategories

Category	N	Subcategory	N
Beauty & Personal Care	193	Humidifiers	17
Health & Household	159	Teeth Whitening Products	15
Home & Kitchen	148	Power Dental Flossers	14
Tools & Home Improvement	120	Sleep Sound Machines	12
Kitchen & Dining	112	Men's Rotary Shavers	11
Cell Phones & Accessories	81	Vacuum Sealers	11
Sports & Outdoors	77	Bug Zappers	10
Pet Supplies	62	Electric Back Massagers	10
Toys & Games	61	Cell Phone Replacement Batteries	9
Patio, Lawn & Garden	59	Light Hair Removal Devices	9
Electronics	57	Outdoor String Lights	9
Baby	42	Cell Phone Charging Stations	8
Office Products	30	Electric Foot Massagers	8

We observe substantial variation in the length of the recruiting period, with some products being promoted for a single day and others for over a month. The average length of the

Facebook promotion period is 23 days and the median is six days.

In Table 1.2, we compare the characteristics of our focal products to a set of competitor products. We define competitor products as those products that appear on the same page of search results for the same product keywords as our focal products. We observe that the focal products are significantly younger than competitor products, with a median age of roughly five months compared with 15 months for products not observed buying fake reviews. But with a mean age of 229 days, the products collecting fake reviews are not generally new to Amazon and without any reputation. Indeed, out of the 1,500 products we observe, only 94 solicit fake reviews in their first month.

Focal products charge slightly lower average prices than their competitors, having a mean price of $33 (compared with $45 for the comparison products). However, this result is mainly driven by the right tail of the price distribution. Fake review products actually charge a higher median price than their competitors, but there are far fewer high-priced products among the fake review products than among competitors.

Turning to ratings, we observe that products purchasing fake reviews have, at the time of their first Facebook post, relatively high product ratings. The mean rating is 4.4 stars and the median is 4.5 stars, which are both higher than the average ratings of competitor products. Only 14% of focal products have ratings below four stars, compared with 19.5% for competitor products. Thus, it appears that products purchasing fake reviews do not seem to do so because they have a bad reputation. Although, we note that ratings may of course be influenced by previous unobserved Facebook campaigns.

We also examine the number of reviews. The mean number of reviews for focal products is 183, which is driven by a long right tail of products with more than 1,000 reviews. The median number of reviews is 45, and roughly 8% of products have zero reviews at the time they are first seen soliciting fake reviews. These numbers are relatively low when compared

Table 1.2: Characteristics of Focal Products and Comparison Products

	Count	Mean	SD	25%	50%	75%
Displayed Rating						
Fake Review Products	1,315.0	4.4	0.5	4.1	4.5	4.8
All Products	203,480.0	4.2	0.6	4.0	4.3	4.6
Number of Reviews						
Fake Review Products	1,425.0	183.1	493.5	10.0	45.0	167.0
All Products	203,485.0	451.4	2,619.0	13.0	59.0	250.0
Price						
Fake Review Products	1,425.0	33.4	45.0	16.0	24.0	35.0
All Products	236,542.0	44.7	154.8	13.0	21.0	40.0
Sponsored						
Fake Review Products	1,425.0	0.1	0.3	0.0	0.0	0.0
All Products	236,542.0	0.1	0.3	0.0	0.0	0.0
Keyword Position						
Fake Review Products	1,425.0	21.4	16.1	8.0	16.0	33.0
All Products	236,542.0	28.2	17.3	13.0	23.0	43.0
Age (days)						
Fake Review Products	1,305.0	229.8	251.1	77.0	156.0	291.0
All Products	153,625.0	757.8	797.1	257.0	466.0	994.0
Sales Rank						
Fake Review Products	1,300.0	73,292.3	151,236.4	7,893.3	26,200.5	74,801.5
All Products	5,647.0	89,926.1	323,028.9	5,495.0	21,610.0	72,563.5

with the set of competitor products, which has a median of 59 reviews and a mean of 451 reviews. Despite these differences, it seems that only a small share of the focal products have very few or no reviews. We also observe that the focal products have slightly lower sales than competitor products as measured by their sales rank, but the difference is relatively minor.

Turning to brand names, we find that almost none of the sellers in these markets are well-known brands. Brand name sellers may still be buying fake reviews via other (more private) channels, or they may avoid buying fake reviews altogether to avoid damages to their reputation. This result is also consistent with research showing that online reviews have larger effects for small independent firms relative to firms with well-known brands (Hollenbeck, 2018).

Table 1.3: Seller Characteristics

	Count	Mean	SD	25%	50%	75%
Focal Sellers						
Number of Products	660.0	23.9	83.9	3.4	7.8	15.2
Number of Reviews	642.0	176.9	297.0	34.0	81.2	201.1
Price	655.0	37.2	71.1	16.4	23.5	37.2
Seller Country						
Mainland China	798.0	0.8				
United States	112.0	0.1				
Hong Kong	13.0	0.0				
Japan	7.0	0.0				
Canada	6.0	0.0				

Note: This table shows information on seller characteristics, where the number of products, number of reviews and price variables are calculated as averages taken over all seller products. Variable counts differ based on the structure of Amazon seller pages making data collection impossible for some sellers. The number of observations for seller country is calculated at the product level.

Finally, to better understand which type of sellers are buying fake reviews, we collect one additional piece of information. We take the sellers' names from Amazon and check the U.S. Trademark Office for records on each seller. We find a match for roughly 70% of products. Of these products, the vast majority, 84%, are located in China, more precisely in Shenzhen or Guangzhou in the Guangdong province, an area associated with manufacturing and exporting. The distribution of sellers by country-of-origin and other seller characteristics are shown in Table 1.3. This table shows that most sellers sell fewer than 15 products, with a median 7.8 products. Their products tend to have fewer than 200 reviews, similar to the focal products. The sellers' other products are also priced similarly to the focal products.

To summarize, we observe purchases of fake reviews from a wide array of products across many categories. These products are slightly younger than their competitors, but only a small share of them are truly new products. They also have relatively high ratings, a large number of reviews, and similar prices to their competitors.

1.3 The Simple Economics of Fake Reviews

We build on the results from the previous section on how the fake review marketplace works, and briefly show the costs and benefits of buying fake reviews. We start by focusing on the costs the sellers incur when buying a fake review.

First, to buy one fake review, a seller must pay to the reviewer:

$$P(1 + \tau + F_{PP}) + Commission \tag{1.1}$$

Where P is the product's list price, τ is the sales tax rate, F_{PP} is the PayPal fee, and Commission refers to the additional cash offered by the seller, which is often zero but is sometimes in the $5-10 range. After the reviewer buys the product, the seller receives a payment from Amazon of:

$$P(1 - c)$$

Where c is Amazon's commission on each sale. So the difference in payments or net financial cost of one review is:

$$P(1 + \tau + F_{PP}) + Commission - P(1 - c) = P(\tau + F_{PP} + c) + Commission$$

This is the share of the list price that is lost to PayPal, Amazon, and taxes, along with the potential cash payment. Along with this financial cost the seller bears the production cost of the product (MC), making the full cost of one fake review:

$$Cost = MC + P(\tau + F_{PP} + c) + Commission \tag{1.2}$$

If we define the gross margins rate as λ such that $\lambda = \frac{P-MC}{P}$, we can show that equation 1.2

becomes

$$Cost = P(1 - \lambda + \tau + F_{PP} + c) + Commission \tag{1.3}$$

This defines the marginal cost of a fake review to the seller. The benefit of receiving one fake review is a function of how many organic sales it creates Q_o and the profit on those sales, which is:

$$Benefit = Q_o P(\lambda - c) \tag{1.4}$$

where again c refers to Amazon's commission from the sale. Setting equations 1.3 and 1.4 equal allows us to calculate the break-even number of organic sales Q_o^{BE}. This is the number of extra incremental sales necessary to exactly justify buying one fake review. If the seller does not offer an additional cash commission, and the vast majority of sellers do not, this can be written as:

$$Q_o^{BE} = \frac{1 - \lambda + \tau + F_{PP} + c}{\lambda - c} \tag{1.5}$$

Where the direct effect of price drops out and this is just a function of the product markup and observable features of the market. We take these market features as known:

- $\tau = .0656$[8]
- $F_{PP} = 2.9\%$
- Amazon commission c varies by category but is either 8% or 15% in almost all cases.[9]

[8] https://taxfoundation.org/2020-sales-taxes/. We aggregate by taking an average of state and local sales taxes.

[9] https://sellercentral.amazon.com/gp/help/external/200336920.

The result for products in the 8% commission categories is:

$$Q_o^{BE} = \frac{1.175 - \lambda}{\lambda - .08} \tag{1.6}$$

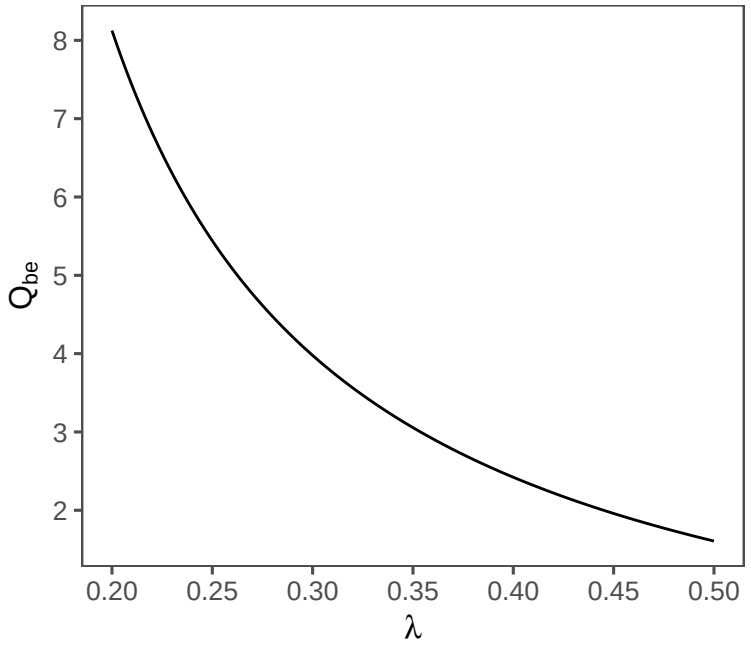
Thus the break-even level of incremental sales needed to justify buying one fake review is a simple expression of a product's price-cost margin. It is clear that products with larger markups require fewer incremental organic sales to justify a fake review purchase. This is for two reasons that this analysis makes clear. First, because the cost of a fake review is lower since, conditional on price, the marginal cost is lower, and second, because the benefit of an organic sale is larger for products with larger markups.

Figure 1.3 plots equation 1.6 where the X-axis is λ and the Y-axis is Q_o^{BE}. It shows that, for products with relatively low markups, the break-even number of organic sales approaches 10, but for products with relatively high markups, this number is below 1.

Note that this is not a theoretical model of the full costs and benefits of fake reviews, many of which are not accounted for, including the risk of punishment and the extent to which Q_o varies as a result of product quality. This is merely a simple description of the direct financial costs and benefits sellers face and how they determine the profitability cutoff for Q_o. Nevertheless, several direct implications follow from this analysis. First, the economics of fake reviews can be quite favorable for sellers since a fairly small number of organic sales are needed to justify their cost. In practice, cheap Chinese imported products often have very large markups such that these sellers only need to generate roughly one additional organic sale to profit from a fake review purchase.

Second, this is especially the case for lower quality products with larger markups. For a concrete example, imagine two products that both list a price of $25. Product A costs $15 to produce and product B costs $20 to produce because A is of lower quality than B. For

Figure 1.3: Organic sales needed to justify one fake review



product A $Q_o^{BE} = 2.4$ and for product B $Q_o^{BE} = 8.1$. The lower cost product needs far fewer organic sales to justify the expense of one fake review.

Third, this analysis makes clear why we are unlikely to observe fake negative reviews applied to competitor products, as in Luca and Zervas (2016) and Mayzlin et al. (2014). The cost of a fake review for a competitor product is significantly higher because it requires the firm buying the review to incur the full price of the competitor's product, and the benefit is likely to be lower because the negative effect on competitor sales is indirect and dispersed across potentially many other products.

1.4 Descriptive Results on Product Outcomes After Buying Fake Reviews

In this section, we quantify the extent to which buying fake reviews is associated with changes in average ratings, number of reviews, and sales rank, as well as other marketing activities such as advertising and promotions. To do so we take advantage of a unique feature of our data in that it contains a detailed panel on firm outcomes observed both before and after sellers buy fake reviews. We stress that, in this section, the results are descriptive in nature. We do not observe the counterfactual outcomes in which these sellers do not buy fake reviews, and so the outcomes we measure are not to be interpreted strictly as causal effects. We present results on the causal effects of fake reviews on sales outcomes in Section 1.5.

We first present results in the short term, i.e., immediately after sellers begin buying fake reviews for their listings. We then show results for the persistence of these effects after the recruitment period has ended. Finally, we show descriptive results on the extent to which Amazon responds to this practice by deleting reviews.

1.4.1 Short-term Outcomes After Buying Fake Reviews

We begin by quantifying the extent to which buying fake reviews is associated with changes in average ratings, reviews, and sales rank in the short term. To evaluate these outcomes, we partition the time around the earliest Facebook recruiting post date (day 0) in 7-day intervals.[10] We then plot outcomes for eight 7-day intervals before and four 7-day intervals after the first fake review recruitment post.

[10]For example, the interval 0 includes the days in the range [0,7) and the interval -1 includes the days in the range [-7,0).

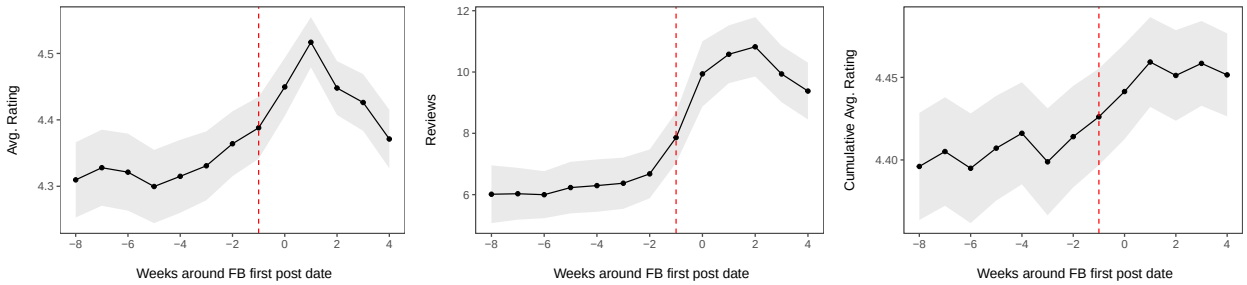


Figure 1.4: 7-day average ratings (left), number of reviews (center), and cumulative average ratings (right) before and after fake reviews recruiting begins. The red dashed line indicates the last week of data before we observe Facebook fake review recruiting.

Ratings and reviews We first examine ratings and reviews. In the left panel of Figure 1.4 we plot the weekly average rating after rating manipulation begins. We see that, first, the average ratings increase by about 5%, from 4.3 stars to 4.5 stars at its peak. Second, this increase in rating is short-lived, and it starts dissipating just two weeks after the beginning of the fake review recruiting; despite this, even after four weeks after the beginning of the promotion, average ratings are still slightly higher than ratings in the pre-promotion period. Third, the average star-rating increases slightly roughly two weeks before the first Facebook post we observe, suggesting that we may not be able to capture with high precision the exact date at which sellers started promoting their products on Facebook. Despite this limitation, our data seems to capture the beginning date of the fake review recruitment fairly well because the largest change in outcome is visible after or on interval zero

Next, we turn to the number of reviews. In the middle panel of Figure 1.4, we plot the weekly average number of posted reviews. We observe that the number of reviews increases substantially around interval zero, nearly doubling, providing suggestive evidence that recruiting fake reviewers is effective at generating new product reviews at a fast pace. Moreover, and differently from the average rating plot, the increase in the weekly number of reviews persists for more than a month. This increase in the number of reviews likely reflects both the fake reviews themselves and additional organic reviews that follow naturally from the increase in sales we document below. Finally, Figure 1.4 confirms that we are not able to capture the

exact date at which the Facebook promotion started.

Does the increase in reviews lead to higher displayed ratings? To answer this question, in the right panel of Figure 1.4, we plot the cumulative average rating before and after the Facebook promotion starts. We observe that ratings increase and then stabilize for about two weeks, after which the increase starts to dissipate.

Sales rank In the left panel of Figure 1.5, we plot the average log of sales rank. The figure shows that the sales rank of these products increases between the intervals -8 and -3, meaning that rating manipulation typically follows a period when sales are falling. When the recruiting period begins, we observe a large increase in weekly sales (i.e. sales rank falls.) This increase is likely reflecting both the initial product purchases by the reviewers paid to leave fake reviews as well as the subsequent increase in organic sales that follow. The increase in sales lasts for at least several weeks.

The center panel of Figure 1.5 plots sales in units sold. Amazon does not display this metric but it is possible to measure sales in units for a subset of products and then estimate the relationship between rank and units. Appendix 1.A describes how we collected this data and modeled the relationship, and more details are available in He and Hollenbeck (2020). We plot the observed sales and point estimates of estimated sales around the time of the first Facebook post and see a sharp increase in average units sold, from around 16 units per week to roughly 20.

Keyword search position So far we have shown that recruiting fake reviews is associated with improvements in ratings, reviews, and sales. One reason for observing higher sales may be that higher ratings signal higher quality to consumers, who then are more likely to buy the product. A second reason is that products recruiting fake reviews will be ranked higher in the Amazon search results due to them having higher ratings and more reviews. To

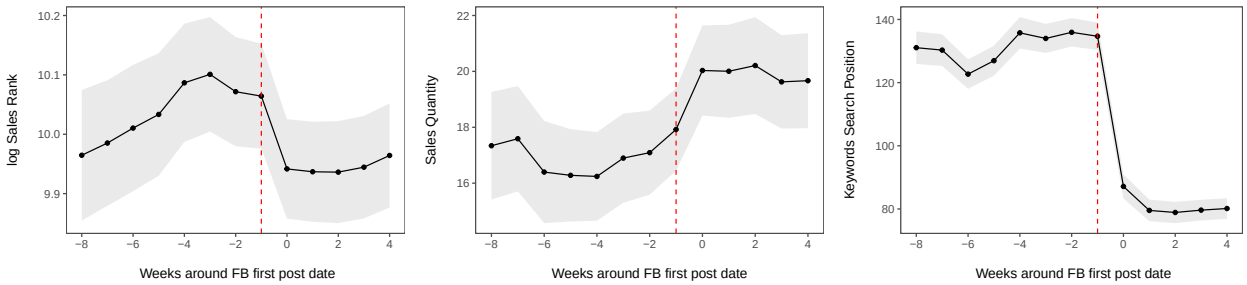


Figure 1.5: 7-day average sales rank (left), sales in units (center), and keyword search position (right) before and after fake reviews recruiting begins. The red dashed line indicates the last week of data before we observe Facebook fake review recruiting.

investigate whether this is the case, in the right panel of Figure 1.5 we plot the search position rank of products recruiting fake reviews. We observe a large drop in search position rank corresponding with the beginning of the Facebook promotions, indicating that products recruiting fake reviews improve their search position substantially. Moreover, this change seems to be long-lasting as the position remains virtually constant for several weeks.

Verified purchases and photos An important aspect of the market for fake reviews is that reviewers actually buy the product and can therefore be listed as a verified reviewers. In addition, they are are compensated for creating realistic reviews, i.e, they are encouraged to post long and detailed reviews including photos and videos. In the left panel of Figure 1.6, we show changes in the average share of verified purchase reviews. Despite being quite noisy in the pre-promotion period, the figure suggests that verified purchases increase with the beginning of the promotion. In the right panel, we observe a sharp increase in the share of reviews containing photos.

Marketing activities Finally, we investigate to what extent rating manipulation is associated with changes in other marketing activities such as promotions (rebates, sponsored listings, and coupons). We plot these quantities in Figure 1.7. We observe a substantial drop in prices (left panel) that persists for several weeks and an increase in the use of sponsored

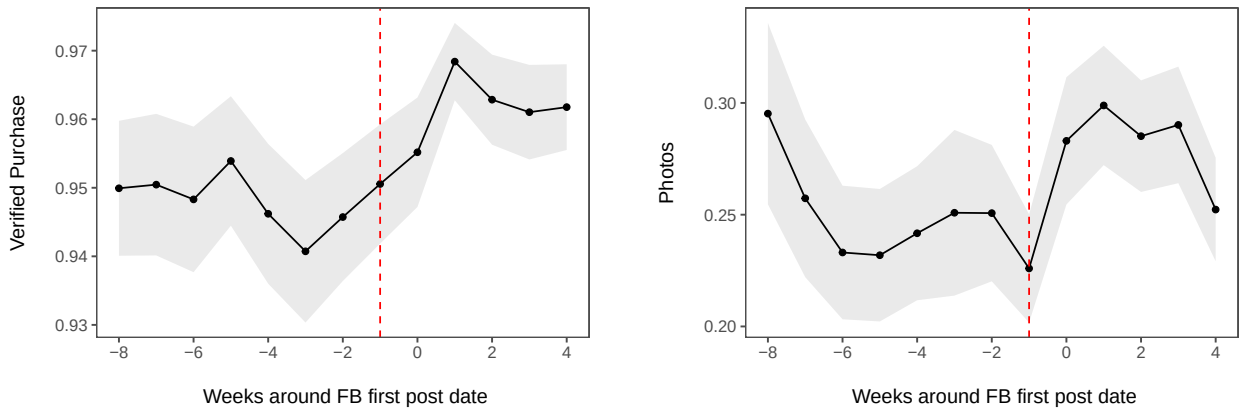


Figure 1.6: 7-day average verified purchase (left) and number of photos (right) before and after fake reviews recruiting begins. The red dashed line indicates the last week of data before we observe Facebook fake review recruiting.

listings, suggesting that Amazon sellers complement the Facebook promotion with advertising activities. This result is in contrast with Hollenbeck et al. (2019) who find that online ratings and advertising are substitutes and not complements in the hotel industry, an offline setting with capacity constraints. Finally, we observe a small negative (albeit noisy) change in the use of coupons.

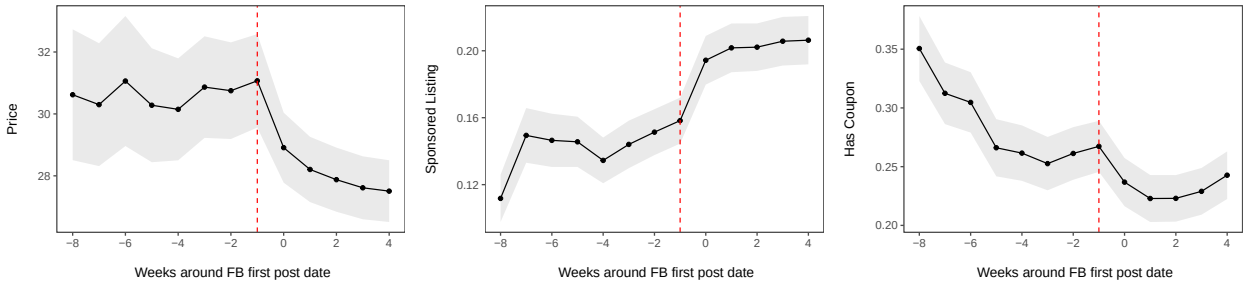


Figure 1.7: 7-day average prices (left), sponsored listings (center) and has coupon (right) before and after fake reviews recruiting begins. The red dashed line indicates the last week of data before we observe Facebook fake review recruiting.

1.4.2 Long-term Outcomes After Buying Fake Reviews

In this subsection, we describe what happens after sellers stop buying fake reviews. We are particularly interested in using the long-term outcomes to assess whether rating manipulation

generates a self-sustaining increase in sales or organic reviews. If we observe that these products continue to receive high organic ratings and have high sales after they stop recruiting fake reviews, we might conclude that fake reviews are a potentially helpful way to solve the cold-start problem of selling online with limited reputation.

We therefore track the long-term trends for ratings, reviews, and sales rank. Similar to Section 1.4.1, we partition the time around the last Facebook recruiting post date in 7-day intervals, and plot the outcomes for four weeks before fake reviews recruiting stop (thus covering most of the period where products recruited fake reviews) and eight weeks after fake reviews recruiting starts. Doing so, we compare the Facebook promotion period (negative intervals) with the post-promotion period (positive intervals).

Ratings and Reviews Long-term trends in ratings and reviews reviews are shown in Figure 1.8. We observe that the increase that occurs when sellers buy fake reviews is fairly short. After one to two weeks from the end of the Facebook promotion, both the weekly average rating and the number of reviews (left and middle panel, respectively) start to decrease substantially. The cumulative average rating (right panel) drops as well. Interestingly, these products end up having average ratings that are significantly worse than when they began recruiting fake reviews (approximately interval -4).

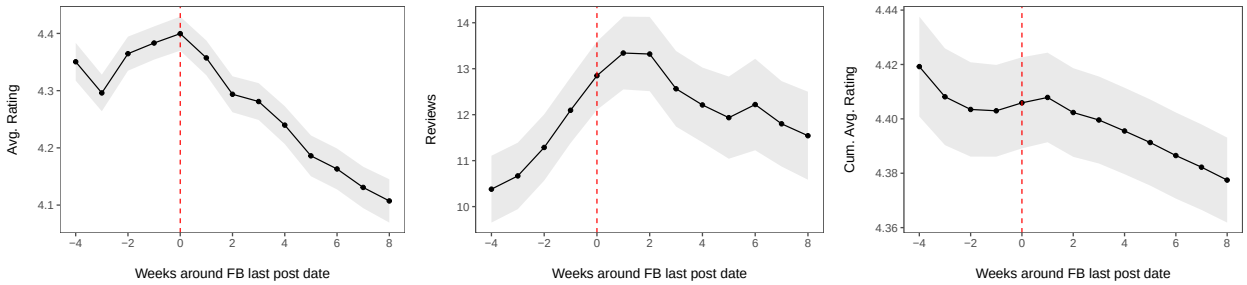


Figure 1.8: 7-day average number of average ratings (left), number of reviews (center), and cumulative average ratings (left) before and after fake reviews recruiting stops. The red dashed line indicates the last week of data in which we observe Facebook fake review recruiting.

Sales Rank The left panel of Figure 1.9 shows the long-term trend in the average log sales rank. It shows that sales decline substantially after the last observed Facebook post. This suggests that the increase associated with recruiting fake reviews is not long lasting as it does not lead to a self-sustaining set of sales and positive reviews.

The middle panel of Figure 1.9 shows sales in units, estimated using the procedure described in Appendix 1.A. The result is consistent with sales rank, showing that sales peak during the week of the last Facebook post and subsequently decline.

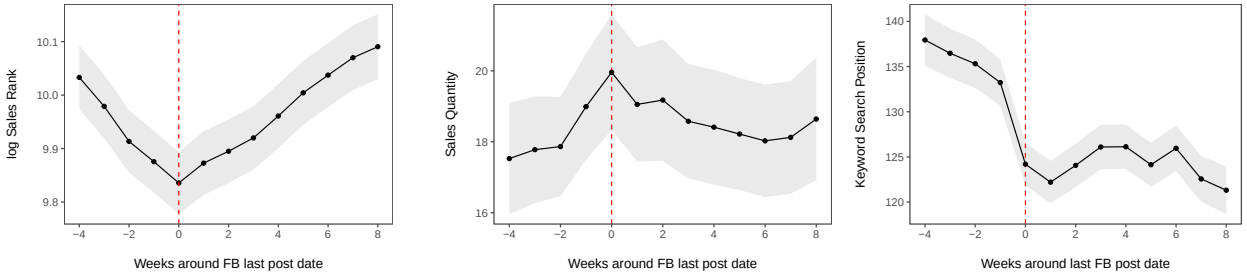


Figure 1.9: 7-day average sales rank (left), sales in units (center), and keyword rank (right) before and after fake review recruiting stops. The red dashed line indicates the last week of data in which we observe Facebook fake review recruiting.

Keyword search position The right panel of Figure 1.9 shows the long-term trend in average keyword search position. We observe that after the Facebook campaign stops, the downward trend in search position stops but does not substantially reverse even after two months. Therefore, products enjoy a better ranking in keyword searches for a relatively long period after fake review recruiting stops.

The relatively stable and persistent increase in search position suggests that this measure may have a high degree of inertia. After an increase in sales and ratings causes a product's keyword rank to improve, it does not decline quickly, even when sales are decreasing. This also suggests that the decrease in sales shown in Figure 1.9 does not come from a reduced product visibility but from the lower ratings and increase in one-star reviews. Finally, while we demonstrate in the next section that Amazon deletes a large share of reviews from

products that recruit fake reviews, the inertia in keyword rank suggests that Amazon does not punish these sellers using the algorithm that determines organic keyword rank. This could therefore serve as an additional policy lever for the platform to regulate fake reviews.

1.4.3 Regression and Heterogeneity Analysis

We have so far shown the outcomes associated with recruiting fake reviews visually. Appendix 1.B shows the same results in a regression context to test whether the changes in outcomes we observe are statistically meaningful when a full set of fixed effects is included as well as to quantify the size of these changes for all products and specific subgroups of products.

Consistent with our visual analysis, we see significant increases in average rating, number of reviews, sales, and search position (keyword rank) after fake review recruiting begins. We also see significantly higher use of sponsored listings in this period and a significant increase in the share of reviews that are from verified purchases. The regression results also confirm that the changes in the number of reviews and search position are especially persistent. Regression results also confirm the visual analysis that shows that average ratings, number of reviews, sales and keyword position all fall after fake review recruiting ends.

Using the regression framework we are also able to test if outcomes differ upon relevant dimensions of product heterogeneity. We are particularly interested in understanding whether there are larger changes in ratings, reviews, and sales for new products with few reviews, as these may buy fake reviews to alleviate the cold-start reputation problem. Regression results shown in Table 1.10 of Appendix 1.B show that, in the short-term period after the first Facebook post for fake reviews, these new products do see their sales increase by a much larger margin than for regular products. They also get a larger increase in number of reviews but do not see an increase in weekly average rating.

After they stop buying fake reviews, we find that these products' ratings fall even further than for regular products, but that their increase in number of weekly reviews is more persistent. The persistence of their increase in weekly reviews corresponds to a larger and more persistent increase in sales. These results combine to suggest that rating manipulation is associated with especially positive outcomes for this type of product.

1.4.4 Amazon's Response

In this subsection, we provide evidence on the extent to which Amazon is aware of the fake review problem and what steps it is taking to remove these reviews.

While we cannot observe reviews that are filtered by Amazon's fraud detection practices and never made public, by collecting review data on a daily and twice-monthly basis, we can observe if reviews are posted and then later deleted. We calculate the share of reviews that are deleted by comparing the full set of observed reviews from our daily scraper with the set of reviews that remain posted at the end of our data collection window. We find that for the set of products observed recruiting fake reviews, the average share of posted reviews that are ultimately deleted is about 43%, compared to 23% for products not observed recruiting fake reviews. This suggests that, to some extent, Amazon can identify fake reviews.

To further characterize Amazon's current policy, we next analyze the characteristics of deleted reviews and the timing of review deletion.

Characteristics of Deleted Reviews In Table 1.4, we report the mean and standard deviation for several review characteristics for deleted and non-deleted reviews, respectively. Following the literature on fake reviews, we focus on characteristics that are often found to be associated with fake reviews. Specifically, we focus on whether the reviewer purchased the product through Amazon (verified purchase), review rating, number of photos associated

with the review, whether the reviewer is part of Amazon's "Early Reviewer Program", i.e., is one of the first users to write a review for a product the length of the review title, and the length of the review.[11]

We find that deleted reviews have higher average ratings than non-deleted reviews. This is driven by the fact that the vast majority of deleted reviews are five-star reviews (see Figure 1.10). Deleted reviews are also associated with more photos, shorter review titles, and longer review text. In general, we might expect longer reviews, those that include photos, and those from verified purchases to be less suspicious. The fact that these reviews are more likely to be deleted suggests that Amazon is fairly sophisticated in targeting potentially fake reviews.[12] Finally, we find no difference for whether the review is associated with a verified purchase or tagged as "Amazon Earlier Reviews."[13]

Table 1.4: Comparing deleted and non-deleted reviews characteristics

	Deleted Reviews	Non-deleted Reviews
Verified purchase	0.98	0.96
	(0.16)	(0.20)
Review rating	4.65	4.24
	(0.98)	(1.37)
Number of photos	0.35	0.19
	(0.93)	(0.72)
Early reviewer	0.00	0.01
	(0.00)	(0.11)
Title length	9.81	21.08
	(13.94)	(13.80)
Review length	236.73	198.75
	(222.88)	(231.68)

Note: Standard deviations in parentheses.

[11]For more details about the "Early Reviewer Program," we refer the reader to `https://smile.amazon.com/gp/help/customer/display.html?nodeId=202094910`.

[12]This result contrasts with Luca and Zervas (2016), who find that longer reviews are less likely to be filtered as fake by Yelp.

[13]We find that Amazon does not delete any reviews tagged as "Amazon Earlier Reviews" potentially because Amazon's process to identify and select early reviewers drastically reduces the possibility of these reviews being fake.

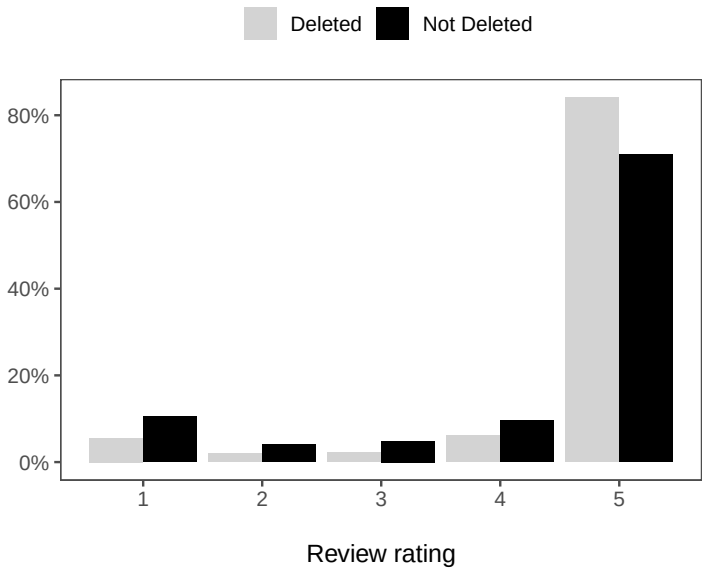


Figure 1.10: Rating distribution for deleted and non deleted reviews

When Are Reviews Deleted? Finally, we analyze when Amazon deletes fake reviews for focal products. We do so by plotting the number of products for which reviews are deleted over time relative to the first Facebook post, i.e., the beginning of the buying of fake reviews. To do so, we partition the time in days around the first Facebook post and then plot the number of products for which reviews are deleted. Because products recruit fake reviews for different time periods, we perform this analysis by segmenting products based on the quartiles of campaign duration. Figure 1.11 shows the results of this analysis.

What emerges from this figure is that Amazon starts deleting reviews for more products after the Facebook campaign begins (red-dashed line) and often it does so only after the campaign terminated (blue-dashed line). Indeed, it seems that most of the review deletion happens during the period covering the two months after the first Facebook post date, but most campaigns are shorter than a month. A simple calculation suggests that reviews are deleted only after a quite large lag. The mean time between when a review is posted and when it is deleted is over 100 days, with a median time of 53 days.

This analysis suggests the deleted reviews may be well-targeted at fake reviews, but that

there is a significant lag between when the reviews are posted and when they are deleted; and this lag allows sellers buying fake reviews to enjoy the short-term benefits of this strategy discussed in Section 1.4.1. In the next section, we show that there is one time period in our data during which Amazon's deletion policy changes significantly; we use this period to identify the causal effects of fake reviews on sales.

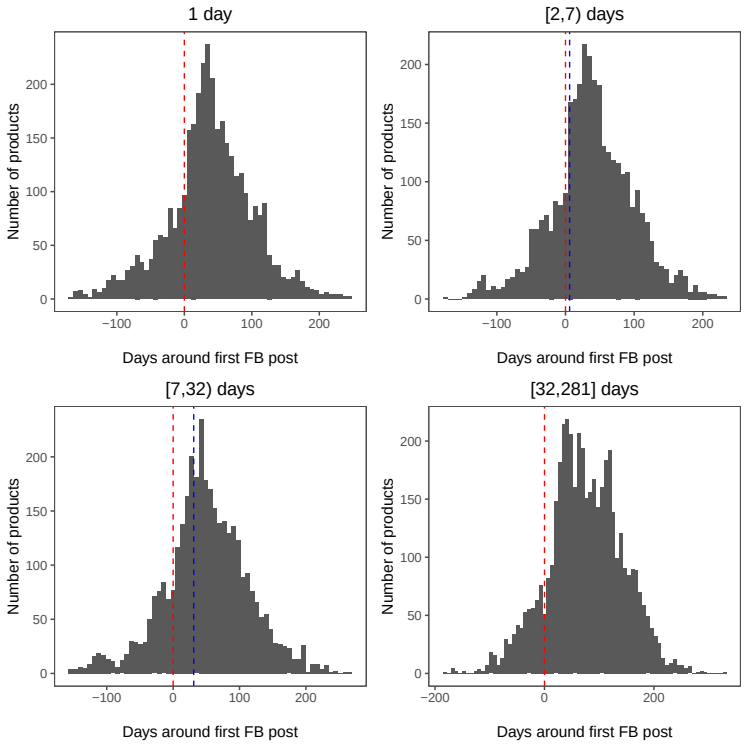


Figure 1.11: Number of products for which reviews are being deleted over time relative to the first Facebook post date. The red dashed line indicates the first time we observe Facebook fake review recruiting, and the blue dashed line indicates the last time we observe Facebook fake review recruiting.

1.5 The Causal Effect of Fake Reviews on Sales

In this section we measure the size of the effect of fake reviews on sales. The results in the previous section are descriptive and may not provide a valid estimate of the effect size. There are two concerns. The first is that sellers buying fake reviews may time these purchases

around unobserved shocks to demand, either positive or negative. While product fixed effects capture time-invariant unobserved heterogeneity, they would not capture these shocks. The second concern is that many sellers change prices and advertising at the same time they recruit fake reviews, making it difficult to isolate the effect of fake reviews on sales. To overcome this, we exploit a temporary change in Amazon policy that allows us to isolate and measure the causal effect of fake review recruiting on sales. This measurement is useful to understand the effects that fake reviews have on sales and to establish that this is a profitable strategy for sellers.

To accomplish this, we take advantage of an event that occurred during our sample period. As we discussed in Section 1.4.4, Amazon deletes a large number of reviews, albeit after a lag. Figure 1.12 shows the amount of review deletion over time for the products seen buying fake reviews. There is one occasion during mid-March 2020 when Amazon undertakes a large-scale purge of reviews with much higher rates of deletion than normal and without a lag.[14] Assuming sellers had no foresight that this review purge was about to be undertaken, a subset of the sellers who recruited fake reviews had the misfortune of doing so during or just before the review purge occurred. Therefore, the products of these unlucky sellers should have no (or a much smaller) increase in positive reviews after they recruited fake reviews compared to the other products. We thus refer to these as control products and all other products that recruited fake reviews at different times as treated products. We can therefore employ a difference-in-differences (DD) strategy that compares the change in sales of treated and control products to estimate the size of the effect of rating manipulation.

In our case, the DD identification strategy requires four assumptions to hold to identify a causal effect. First, Amazon should not have strategically selected the products for which reviews were deleted, i.e., control products should be similar to treated products in both observable and unobservable characteristics. Second, the review purge should be effective

[14]There is another spike in review deletion in May of 2020, but it affects substantially fewer reviews and is not as long-lasting.

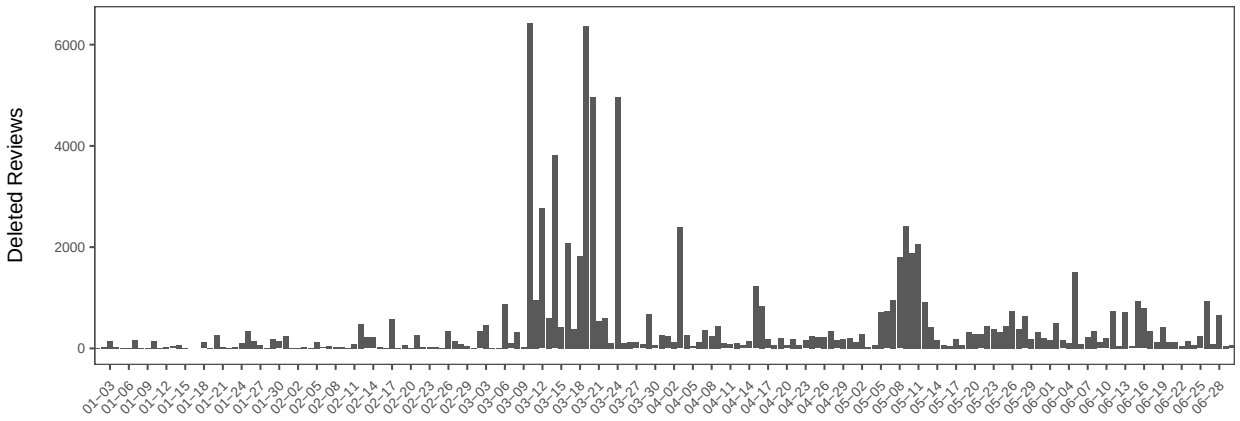


Figure 1.12: Amazon deleted reviews by date

at preventing the control products from acquiring fake reviews. Third, treated and control products should not differ in their use of marketing activities that can affect sales. Fourth, the parallel trends assumption should hold, i.e., pre-treatment sales trends for treated and controls products should be similar. We start by presenting the empirical strategy setup, we then test each of the assumptions discussed above, and then provide estimates and robustness checks.

1.5.1 Empirical strategy setup

We start by taking the midpoint date of the review purge, which is March 15, and defining our set of control products as all products whose first observed Facebook post is in the interval [-2,1] weeks around this date. This results in 78 control products. The 1,412 products whose sellers started recruiting fake reviews outside of this window is the set of treated products.

We then estimate a standard DD regression which takes the following form:

$$y_{it} = \beta_1 \text{ Treated}_i + \beta_2 \text{ After}_{it} + \beta_3 \text{ Treated}_i \times \text{After}_{it} + \alpha_i + \tau_t + X_{it}^{'}\gamma + \epsilon_{it}, \tag{1.7}$$

where y_{it} is the outcome of interest for product i at year-week t, $Treated_i$ is an indicator for

whether product i is treated and $After_{it}$ is an indicator for the period after the first observed Facebook post for product i. α are product fixed effects to account for time-invariant product characteristics, and τ are year-week fixed effects to account for time-varying shocks to the outcome that affect all products (e.g., holidays). The coefficient β_2 measures the effect of fake review recruiting for control products, and the coefficient of interest, β_3, is the classical DD estimate which measures the difference in sales for treated products. We estimate the regression in Equation 1.7 using OLS and clustering standard errors at the product level.

1.5.2 Identification checks

Treated and control products are similar To test this assumption, we show that (1) treated and control products are similar in most of their observable characteristics, and (2) Amazon does not seem to select specific products with the review purge. In Table 1.5 we compare treated and control products over a large set of variables by taking the average over the period [-8,-2) weeks before the products begin to recruit fake reviews.[15] We find that they are largely similiar but that control products are older, with lower average weekly ratings, and more cumulative reviews.

To reduce concerns about differences between treated and control products that could affect the DD estimates, we employ Propensity Score Matching (PSM) (Rosenbaum and Rubin, 1983) to match treated and control products on the observable variables that are different across treatment conditions, i.e., age, weekly average ratings, and cumulative reviews. To do so, for every product, we average these variables over the period [-8,-2) weeks and then implement PSM using the Gaussian kernel matching procedure with a bandwidth of 0.005, and imposing a common support, i.e., we drop treatment observations whose propensity score is higher than the maximum or less than the minimum propensity score of the controls.[16]

[15] We exclude weeks [-2,-1] because the analysis in Section 1.4.1 suggests that for some products, outcomes start to change up to two weeks before the first Facebook post.

[16] We choose a bandwidth that allowed for a good matching, meaning that there are no longer any statis-

Table 1.5: Comparison of Treated and Control Products

	Control	Treated	t-stat
Age	9.84	7.15	2.36*
Weekly Avg. Ratings	4.10	4.32	−2.07*
Cum. Avg. Ratings	4.32	4.43	−1.36
Weekly Reviews	5.21	5.78	−0.33
Cumulative Reviews	234.80	109.90	3.11**
Price	27.10	33.60	−1.38
Coupon	0.23	0.26	−0.37
Verified	0.92	0.93	−0.60
Number of Photos	0.25	0.26	−0.15
Category	41.90	40.50	0.41

Note: t-test for equality of means for treated and control units. Means are computed at the interval level for the period [-8,-2) weeks.

Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

We start with 1,412 treated and 78 control products and, after matching, we are left with 987 treated and 48 control products. We verify that PSM eliminates the imbalance between treated and control units by computing a weighted (using the PSM weights) t-test for equality of means of treated and control products. We report the results of this test in Table 1.6.[17]

Turning to Amazon's criteria of selecting which products' reviews are deleted, in Appendix 1.C, we show that review deletion during the purge period is highly concentrated on individual reviewers and is not targeted at specific products.

Manipulation Check Here we present evidence showing that the review purge creates a valid set of control products. To do so, the purge must prevent these products, who were observed attempting to buy fake reviews, from receiving the treatment of an increase in reviews. We do so by estimating Equation 1.7 with the outcome set to be the log of cumulative reviews. We report these results in column 1 of Table 1.7. As expected, $After$ is

tically significant differences between treated and control units for the variables used for matching.

[17]In Appendix 1.D, we show that our results are not sensitive to the type of matching algorithm used.

Table 1.6: Comparison of Treated and Control Products after matching

	Control	Treated	t-stat
Age	7.78	7.63	0.10
Weekly Avg. Ratings	4.07	4.15	−0.48
Cum. Avg. Ratings	4.34	4.33	0.07
Weekly Reviews	5.11	7.24	−0.72
Cumulative Reviews	109.48	124.69	−0.39
Price	25.74	32.63	−1.20
Coupon	0.24	0.27	−0.30
Verified	0.94	0.95	−0.31
Number of Photos	0.22	0.24	−0.19
Category	43.75	39.61	0.75

Note: Weighted t-test for equalty of means for treated and control units. Means are computed at the interval level for the period [-8,-2) weeks.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

small and close to zero, suggesting that there is no increase in reviews for control products. However, the interaction coefficient $After \times Treated$, is positive and significant and suggests that the number of cumulative reviews for treated products increased by approximately 10% more than control products.

Table 1.7: Diff-in-Diff Estimates

	(1) log Cum. Reviews	(2) Sponsored	(3) Coupon	(4) log Price	(5) log Sales Rank
After	0.047	0.014	0.011	−0.003	0.198*
	(0.036)	(0.026)	(0.047)	(0.009)	(0.097)
After × Treated	0.099*	0.027	−0.031	0.006	−0.375**
	(0.048)	(0.032)	(0.046)	(0.013)	(0.116)
PSM Sample	Yes	Yes	Yes	Yes	Yes
N	12620	7477	7477	7417	11553
R^2	0.96	0.65	0.65	0.99	0.87

Note: All specifications include product and year-week FE. Cluster-robust standard errors (at the product level) in parentheses.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Marketing activities are similar To investigate whether treated and control products' marketing activities are similar, we estimate Equation 1.7 for three different outcomes: (1) whether product i buys sponsored listings: (2) whether product i offers discounts through coupons; and (3) product i price. We report these estimates in columns 2-4 of Table 1.7. We do not observe any statistically significant change in sponsored listings, coupons, and price after the first Facebook post for both treated and control products. Therefore, the assumption about marketing activities being similar across treatment and control products is satisfied.

Parallel trends Finally, we test the parallel assumption. To do so we estimate the following Equation:

$$y_{it} = \beta_1 \text{ Treated}_i + \beta_2 \text{ After}_{it} + \lambda_k \text{ Treated}_i \times \text{Week}_{kit} + \alpha_i + \tau_t + X_{it}^{'}\gamma + \epsilon_{it}, \qquad (1.8)$$

where everything is as in Equation 1.7, and $Week_{kit}$ represents a set of k dummies identifying 7-days intervals around the first Facebook post of each product. The λ_k coefficients can be interpreted as weekly treatment effects estimated before and after the treatment with respect to the baseline week -3.[18] We plot these estimates along with their 95% confidence intervals in Figure 1.13. Two findings emerge from this figure. First, while there is a decreasing trend in the pre-treatment period, the estimates before week -2 are indistinguishable from zero, suggesting that the parallel trends assumption is satisfied during this period. Second, there is a statistically significant increase in sales in weeks -1 and -2 relative to this baseline. This is consistent with the results in section 4.1 showing that for treated products sales begin to increase slightly early, suggesting that our DD analysis contains the same measurement error issue as the descriptive analysis. In our estimates of the size of the causal effects we measure the change in sales occurring after the first observed Facebook post (in week -1) and

[18]We choose to set the baseline week to be -3 because, as we discussed in Section 1.4.1 we observe that for some products outcomes start to change at week -2.

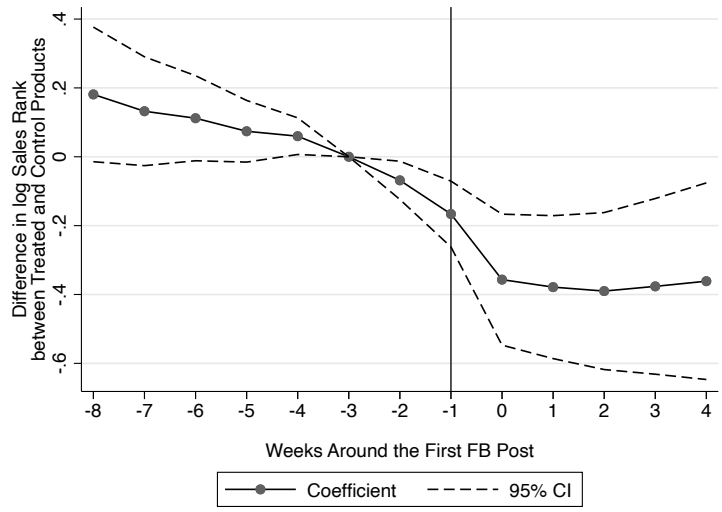


Figure 1.13: The evolution of the treatment effect, i.e., the difference in log Sales Rank between treated and control products.

so to the extent that some of the increase in sales occurs before this, we may underestimate the size of the effect. Nevertheless, we do observe a large decrease in sales rank for treated products after week 0.

1.5.3 The effect of fake reviews on sales

To measure the causal effect of fake reviews on sales, we estimate Equation 1.7 using as the outcome the log of sales rank. We report these estimates in column 5 of Table 1.7. First, we find that the sales rank of control products increases about 22%. This is in line with the evidence we provided in Section 1.4.1 where we showed that products start recruiting fake reviews when sales are falling. In the absence of fake reviews, sales are therefore likely to continue to fall and thus sales rank should increase. Second, and in line with what we observed in Figure 1.13, we estimate that compared to control products, treated products see a reduction in sales rank of 31%. The overall effect of fake reviews on sales rank for treated products ($\beta_1 + \beta_2$) is about 16%.

In Appendix 1.D, we present several robustness checks that reinforce the causal interpretation of our results. First, we show that the sales estimates are not sensitive to the choice of the window around the mid-purge date used to select the set of control products. Second, to reduce concerns about our results being driven by the way in which we select control products, we consider a specification in which we define the treatment as a continuous variable rather than as a binary variable based on a time cutoff around the purge event. Third, we perform placebo tests where we re-estimate our results for fake purge dates, and find no difference in outcomes of treated and control products.

1.6 Evidence of Consumer Harm from Fake Reviews

We conclude the paper by evaluating whether consumers are harmed by fake reviews. To do so, we analyze the products' ratings after they stop buying fake reviews. If they continue receiving high ratings after rating manipulation ends it would be evidence that fake reviews are used by high-quality products in a manner akin to advertising. This would be consistent with the predictions of theoretical results in Dellarocas (2006) and others. If, by contrast, we see declining ratings and observe a large number of one-star reviews, it would suggest fake reviews are bought to mask low product quality and deceive consumers.

There is an inherent limitation in using ratings to infer welfare because consumers leave ratings for many reasons and generally ratings are not a literal expression of utility. But we argue that when products receive low ratings and a large number of one-star reviews, it indicates that the actual quality of these products is lower than what most customers expected at the time of their purchase. The low ratings are either a direct expression of product quality or an attempt to realign the average rating back toward the true level and away from the manipulated level. In this latter case, we still infer consumer harm, either because it indicates consumers paid a higher price than what they would have if the product

was not overrated due to rating manipulation, or because the fake reviews caused them to buy a lower quality product than the closest alternative. This analysis is also important from the platform's perspective. An increase in one-star reviews would indicate that fake reviews are a significant problem since they reflect negative consumer experiences that erodes trust in the platform's reputation system.[19]

1.6.1 One-Star Ratings and Reviews

We previously showed in Figure 1.8 in Section 1.4.2 that average ratings fall after fake review recruiting ends. Figure 1.14 shows why. The share of one-star reviews increases by roughly 70% after fake review recruiting stops. The increase in the share of one-star ratings and the increase in the total number of ratings mean that the absolute number of one-star reviews increases by even more.

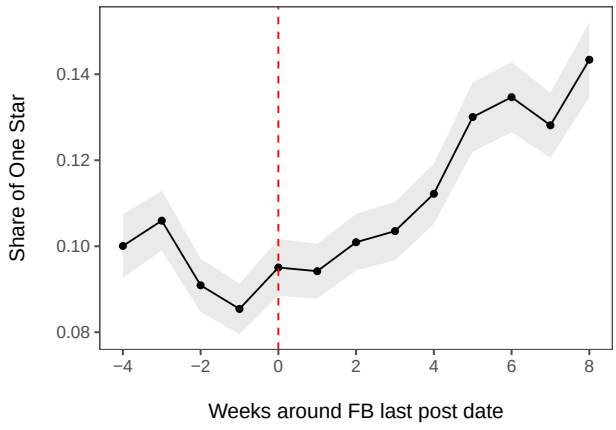


Figure 1.14: 7-day average share of one-star reviews before and after fake reviews recruiting stops. The red dashed line indicates the last time we observe Facebook fake review recruiting.

Next, we explore how this pattern varies for different types of products. It may be the case that ratings stay high for certain products. For example, new products (i.e., products with few reviews or that have been listed on Amazon for a brief period of time) might use fake

[19]Nosko and Tadelis (2015b) show that when a buyer has a bad product experience with a third-party seller on a platform, they are significantly less likely to shop at that platform again.

reviews to bootstrap their reputation, which they can sustain if these products are high quality.

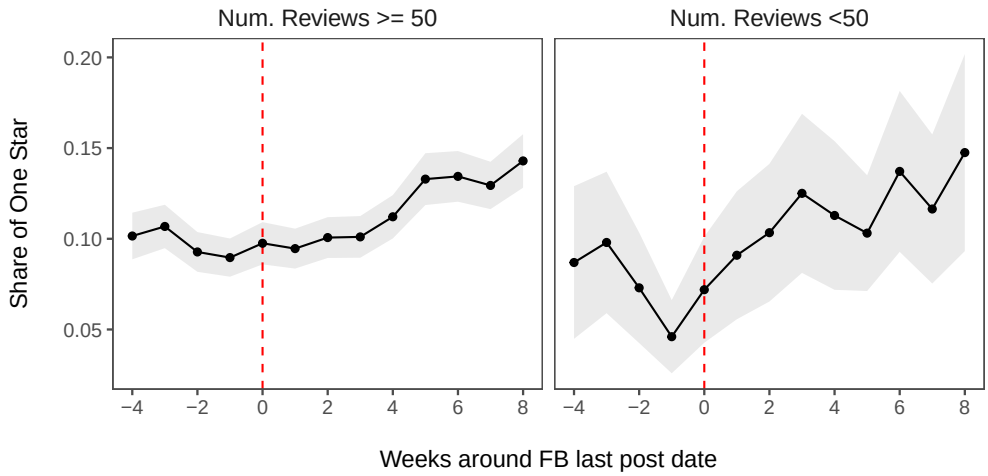


Figure 1.15: 7-day average share of one-star reviews before and after fake reviews recruiting stops by number of reviews accumulated prior to the fake review recruiting. The red dashed line indicates the last time we observe Facebook fake review recruiting.

To test this, we segment products by number of reviews and age. Figure 1.15 shows how the share of one-star reviews changes for products with fewer than 50 reviews. The increase in one-star ratings is sharper for products with few reviews. Figure 1.16 makes the same comparison for products that have been listed on Amazon for fewer than 60 days. The young products experience a much larger increase in one-star reviews than the other products, with more than 20% of their ratings being one-star two months after they stop recruiting fake reviews. Overall, these results refute the idea that "cold-start" products use fake review efficiently. Instead, these products seem to be of especially low quality.

1.6.2 Text Analysis

So far, we have shown increases one-star reviews to provide evidence that consumers are harmed by rating manipulation. Here, we provide additional evidence by using state-of-the-art machine learning algorithms to analyze the text of these negative reviews.

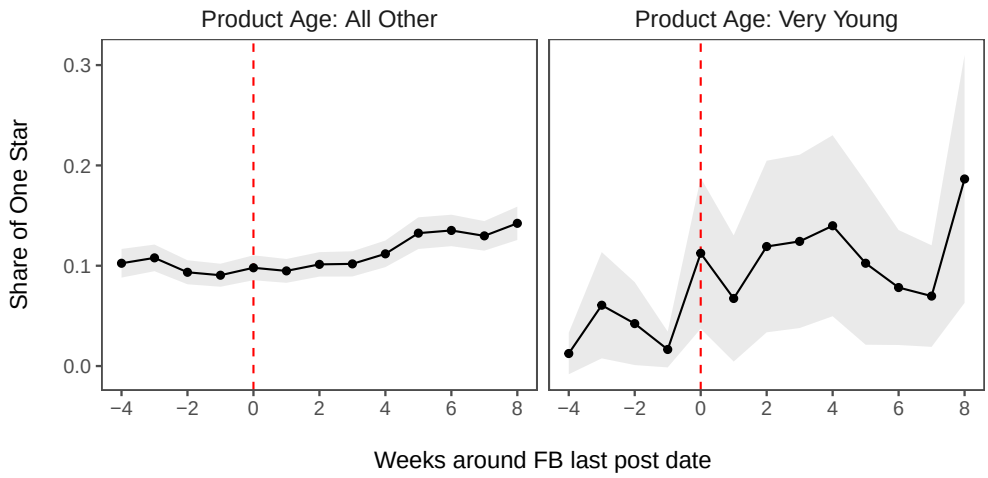


Figure 1.16: 7-day average share of one-star reviews before and after fake reviews recruiting stops by product age (very young products are those listed for fewer than 60 days). The red dashed line indicates the last time we observe Facebook fake review recruiting.

The goal of this analysis is twofold. First, we want to test if the negative reviews posted after a product buys fake reviews are different from other negative reviews. It could be the case that one-star reviews increase after any sales spike and this is not a phenomenon specific to fake reviews. If so, test analysis should not be able to distinguish between them. Second, if they are indeed distinctive, we want to identify what text features differentiate them. Our simple model discussed in Section 1.3 shows that the returns to rating manipulation are higher for products with lower production costs, all else equal. It therefore predicts that negative reviews from these products are likely to focus on quality issues and value relative to price.

We perform two types of comparisons. First, we compare the post-campaign one-star reviews for fake review products to the one-star reviews for these same products prior to their first Facebook post. Second, we compare the post-campaign one-star reviews to one-star reviews for a different set of products that were not observed buying fake reviews.

We start by sampling 5,000 one-star reviews of each type: from products recruiting fake reviews prior to the first Facebook post, from those same products after the last Facebook

post, and from a set of competitor products.[20] Then, we train a text-based classifier to predict whether each review is from either before or after fake review recruiting, or in the second test, from either a product recruiting fake reviews or not. Following standard practice, we split the review dataset into an 80% training sample and a 20% test sample. We present the results using a Naive Bayes Classifier based on tf-idf. Depending on the configuration of the classifier (we can change the number of text features used by the classifier by removing very rare and very popular words), we achieve an accuracy rate that ranges between 61% and 75% and a ROC-AUC score that varies between 66% and 83% for both types of comparisons. These results suggest that in both cases the classifier can distinguish between the different kinds of one-star reviews based on their text. In other words, even holding the products themselves and their star-rating constant, the reviews written for products after fake review recruiting contain significantly different text features compared with those written beforehand. Similarly, these reviews contain a significantly different set of words compared with reviews written for products that did not recruit fake reviews.

We next look at what are the most predictive text features for distinguishing the different product types. In Table 1.8, we compare the text features of negative reviews posted before and after rating manipulation by reporting the top 30 features. What emerges from this table is that one-star reviews written after rating manipulation occurs are predicted by text features mostly related to product quality ("work", "broke", "stop work") or value ("money", "waste money") or else explicitly suggest the consumer felt deceived or harmed ("return", "disappoint"). By contrast, the reviews for the same products prior to rating manipulation are associated with idiosyncratic product features, such as "earplug", "milk frother", or "duvet". Table 1.9 reports the top features for the model trained using fake review products and competitor products. Again, reviews for fake review products are associated with text features mostly related to product quality ("qualiti", "stop work", "work", etc.), value/price

[20] As we discussed in Section 1.2, competitor products are defined as those products appearing on the same results page for a keyword search as the focal products.

("waste money", "money", "disappoint", etc.); instead, competitors' one-star reviews are predicted by text features mostly related to idiosyncratic product characteristic ("second attach", "fade", "reseal", etc.)

Overall, these results are consistent with one another and add further evidence that consumers who bought products that recruited fake reviews felt deceived in thinking that the products were of higher quality than they really were.

Table 1.8: Most Predictive Text Features: Before v After Fake Reviews

Period	Top 30 Text Features
Before recruiting fake reviews	muzzl, around neck, duvet, laundri, earplug, needless, milk frother, foam earplug, rectal, topper, espresso, lightn, like go, keep lick, nois reduct, degre differ, like tri, frizzi, espresso machin, wildli, breath, work never, expect much, concert, time open, stori, octob, inflat collar, unsaf, vinegar
After recruiting fake reviews	work, product, money, return, use, month, wast, time, would, wast money, stop, charg, like, even, disappoint, broke, stop work, week, first, tri, light, back, good, bought, batteri, qualiti, item, recommend, purchas, turn

Note: Model accuracy and ROC-AUC are 61% and 66%, respectively

Table 1.9: Most Predictive Text Features: Focal vs Non-Focal Products

Products	Top 30 Text Features
Recruiting fake reviews	work, product, money, return, use, time, stop, wast, month, would, like, wast money, charg, even, broke, stop work, week, disappoint, good, back, light, first, tri, bought, qualiti, review, turn, batteri, recommend, great
Not recruiting fake reviews	reseal, command, bang, fixtur, apart piec, septemb, product dont, fade, ignit, use never, use standard, terrier, compani make, desktop, love idea, wifi connect, bead, solar panel, inexpens, within year, return sent, compani product, second attach, pure, cycl, thought great, solar charg, blame, bought march, price paid

Note: Model accuracy and ROC-AUC are 63% and 69%, respectively

1.7 Discussion and Conclusions

It has become commonplace for online sellers to manipulate their reputations on online platforms. In this paper, we study the market for fake Amazon product reviews, which takes place in private Facebook groups featuring millions of products. We find that soliciting reviews on Facebook is highly effective at improving several sellers' outcomes, such as number of reviews, ratings, search position rank, and sales rank. However, these effects are often short-lived as many of these outcomes return to pre-promotion levels a few weeks after the fake reviews recruiting stops. In the long run, this boost in sales does not lead to a positive self-sustaining relationship between organic ratings and sales, and both sales and average ratings fall significantly once fake review recruiting ends. Rating manipulation is not used efficiently by sellers to solve a cold-start problem, in other words.

We also find evidence that this practice is likely harmful to consumers, as fake review recruiters ultimate see a large decrease in ratings and increase in their share of one-star reviews. An important implication is that rating manipulation is also likely to harm honest sellers and the platform's reputation itself. If large numbers of low-quality sellers are using fake reviews, the signal value of high ratings could decrease, making consumers more skeptical of new, highly rated products. This, in turn, would make it harder for high-quality sellers to enter the market and would likely reduce innovation.

Firms are continuously improving and perfecting their manipulation strategies so that findings that were true only a few years ago, or strategies that could have worked in the past to eliminate fake reviews, might be outdated today. This is why studying and understanding how firms manipulate their ratings continue to be an extremely important topic of research for both academics and practitioners. As a testament to this, Amazon claims to have spent over $500 million in 2019 alone and employed over 8,000 people to reduce fraud and abuse

on its platform.[21]

We also document that Amazon does delete large numbers of reviews and that these deletions are well-targeted, but there is a large lag before these reviews are deleted. The result is that this deletion policy does not eliminate the short-term profits from these reviews or the consumer harm they cause.

Of course, Amazon has other potential policy levers at its disposal to regulate fake reviews. But we do not observe Amazon deleting products or banning sellers as a result of them manipulating their ratings. Nor do we observe punishment in the products' organic ranking in keyword searches. This keyword ranking stays elevated several months after fake review recruiting has ended, even when Amazon finds and deletes many of the fake reviews posted on the platform. Reducing product visibility in keyword rankings at the time fake reviews are deleted could potentially turn fake reviews from a profitable endeavor into a highly unprofitable one.

It is not obvious whether Amazon is simply under-regulating rating manipulation in a way that allows this market to continue to exist at such a large scale, or if it is assessing the short-term profits that come from the boost in ratings and sales and weighing these against the long-term harm to the platform's reputation. Quantifying these two forces is, therefore, an important area of future research.

1.A Sales Data

In this appendix, we first describe how we collect data on sales in units, and then how we convert sales rank to sales in units for instances in which this is unobserved. Amazon does not display metrics on sales quantities, only on an ordinal Best Seller Rank, a number that

[21] See: `https://themarkup.org/ask-the-markup/2020/07/21/how-to-spot-fake-amazon-product-reviews`

ranks products based on their rate of sales relative to other products in the same category.

To acquire sales quantity data, we exploit a feature of the Amazon website that allows us to infer the number of units of a product that are currently in stock. To observe a product's inventory, one must simply add to an Amazon cart increasing numbers of units of the product until the seller runs out of stock. At this point, Amazon will display an alert telling the buyer the total number of units available. The highest number of units that can be added to an Amazon cart is 999 and so for products with inventories below 1000 this method allows us to observe the number of units currently in stock. We employ research assistants to collect data using this method for a panel of products every 2 days. By observing inventories repeatedly over time, we can infer the rate of sales.

After collecting inventory data, we first remove observations in which the inventory is 0 or at the upper limit of 999 or if the seller has placed a limit on the number of units that can be purchased. We then calculate the difference in inventories between each two day period. We remove any observations where the inventory increases over this period. We use the remaining data to calculate sales per day. A more detailed description of this procedure and the resulting data can be found in He and Hollenbeck (2020). We observe data on sales in units for 683 of the focal products.

These data do not cover every period and, most importantly, we cannot observe sales data prior to the first Facebook post of these products. Therefore we estimate the relationship between sales rank and sales in units using the sales data to approximate the level of sales for these missing periods. To do so, we generalize the approach taken by Chevalier and Goolsbee (2003) and estimate a log-log regression with product fixed effects. This provides a good fit, with an adjusted-R^2 of .89. More details on the estimation and alternative models for estimated sales quantities are available in He and Hollenbeck (2020).

Lastly, we then use the regression estimates to infer the missing data on sales units at

different dates for the same set of products based on their observed rank on those dates. We plot these outcomes in the short run and long run in Figures 1.5 and 1.9.

1.B Descriptive regression analysis

1.B.1 Short-term Analysis

We use data from the interval [-8,4] weeks around the first Facebook post and estimate the following equation on each outcome variable:

$$y_{it} = \beta_1 \text{ After}_{it}^{\leq 2} + \beta_2 \text{ After}_{it}^{>2} + \alpha_i + \tau_t + \epsilon_{it}, \tag{1.9}$$

where $\text{After}_{it}^{\leq 2}$ is a dummy for the time period from zero to two weeks after the beginning of the Facebook promotion and $\text{After}_{it}^{>2}$ is a dummy for the time period after that. This divides up our sample into three periods: a before period, a period in which short-term changes should be present, and a period in which more persistent changes should be present. In each case we include year-week, τ_t, and product fixed effects, α_i. We include data on the 2,714 competitor products for which we have collected daily review data. These products are never observed buying fake reviews, so their After_{it} dummies are all set at zero.

The results for each variable for all products are shown in Table 1.10.[22] Consistent with our visual analysis, we see significant short-term increases in average rating, number of reviews, sales, and search position (keyword rank). The increase in weekly average rating is roughly .11 stars. We also see significantly higher use of sponsored listings in this period and a significant increase in the share of reviews that are from verified purchases. There are

[22]The high R^2 are likely due to the inclusion of product and year-week fixed effects fixed effect.

also positive coefficients for the longer-term dummy for the number of reviews and search position, confirming that the changes in these variables are more persistent.

Table 1.10: Short-term Outcomes After Recruiting Fake Reviews

	(1) Avg. Rating	(2) log Reviews	(3) log Sales Rank	(4) log Keyword Rank	(5) Sponsored	(6) Coupon	(7) log Photos	(8) Verified	(9) log Price
$\leq$ 2 wks	0.107***	0.445***	−0.260***	−0.412***	0.044***	0.002	0.022***	0.022***	−0.013**
	(0.019)	(0.017)	(0.022)	(0.028)	(0.009)	(0.013)	(0.006)	(0.003)	(0.004)
> 2 wks	0.034	0.320***	−0.246***	−0.434***	0.061***	−0.007	0.003	0.018***	−0.016**
	(0.021)	(0.020)	(0.028)	(0.030)	(0.010)	(0.014)	(0.007)	(0.004)	(0.005)
N	186389	247218	193381	91733	94122	94122	186389	186389	92361
R^2	0.22	0.67	0.81	0.64	0.55	0.52	0.15	0.15	0.98
$\leq$ 2 wks	0.117***	0.439***	−0.238***	−0.409***	0.049***	0.002	0.024***	0.016***	−0.013**
	(0.019)	(0.018)	(0.023)	(0.030)	(0.010)	(0.013)	(0.006)	(0.003)	(0.004)
$\leq$ 2 wks × Coldstart	−0.198**	0.091	−0.275**	−0.030	−0.078*	−0.004	−0.033	0.078***	−0.005
	(0.067)	(0.057)	(0.085)	(0.074)	(0.031)	(0.041)	(0.027)	(0.022)	(0.015)
> 2 wks	0.059**	0.309***	−0.217***	−0.440***	0.069***	−0.005	0.006	0.015***	−0.016**
	(0.022)	(0.021)	(0.029)	(0.031)	(0.011)	(0.014)	(0.007)	(0.004)	(0.005)
> 2 wks × Coldstart	−0.360***	0.142*	−0.349***	0.067	−0.114***	−0.021	−0.054	0.050*	0.002
	(0.080)	(0.070)	(0.100)	(0.086)	(0.034)	(0.051)	(0.030)	(0.023)	(0.017)
N	186389	247218	193381	91733	94122	94122	186389	186389	92361
R^2	0.22	0.67	0.81	0.64	0.55	0.52	0.15	0.15	0.98

Note: All specifications include product and year-week FE. Cluster-robust standard errors (at the product level) in parentheses.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Next, we add interactions with an indicator for whether or not the product is new to Amazon with few reviews. New products without established reputations may have different incentives to buy fake reviews or different outcomes afterwards. We define these as "cold-start" products if they have been listed on Amazon for 4 or fewer months and have 8 or fewer reviews. This is roughly 10% of the observed products. We see in Table 1.10 that these products do have different outcomes, specifically that these products' sales increase by a much larger margin than for regular products. They also get a larger increase in number of reviews but do not see an increase in weekly average rating.[23]

[23]This last result is due to the fact that cold-start products frequently start out with a perfect five-star rating. When measured 2 weeks prior to their first Facebook post, we find that 83% of cold-start products have an average rating of 5.0 stars, leading to an overall average rating of 4.65 stars across products. This compares with an average rating of 4.35 stars for non-cold-start products. The high initial rating these products enjoy inevitably decreases as more reviews are added.

1.B.2 Long-term Regressions

Similar to how we presented results for the short-term outcomes, we now show the long-term results in a regression context.

To do so, we take the interval [-4,8] weeks around the last Facebook post and regress each outcome variable on a dummy for the time period from one to three weeks afterward, as well as an additional dummy for the time period after that. In each case, we include year-week and product fixed effects. Results are shown in Table 1.11.

Table 1.11: Long-term Outcomes After Recruiting Fake Reviews

	(1) Avg. Rating	(2) log Reviews	(3) log Sales Rank	(4) log Keyword Rank	(5) Sponsored	(6) Coupon	(7) log Photos	(8) Verified	(9) log Price
$\leq$ 2 wks	−0.033	0.060***	−0.052**	−0.169***	0.021***	−0.003	−0.008	0.009**	−0.007*
	(0.018)	(0.018)	(0.019)	(0.017)	(0.005)	(0.009)	(0.006)	(0.003)	(0.003)
$>$ 2 wks	−0.156***	−0.239***	0.082**	−0.138***	0.036***	−0.005	−0.043***	0.003	−0.016***
	(0.020)	(0.020)	(0.027)	(0.021)	(0.007)	(0.010)	(0.006)	(0.003)	(0.004)
N	187640	249444	194840	97022	99409	99409	187640	187640	97543
R^2	0.22	0.67	0.81	0.65	0.56	0.52	0.15	0.15	0.98
$\leq$ 2 wks	−0.026	0.042*	−0.041*	−0.171***	0.022***	−0.003	−0.007	0.007*	−0.008**
	(0.019)	(0.018)	(0.020)	(0.018)	(0.006)	(0.009)	(0.006)	(0.003)	(0.003)
$\leq$ 2 wks × Coldstart	−0.121	0.259***	−0.151	0.044	−0.035	0.003	−0.012	0.039	0.017
	(0.075)	(0.064)	(0.083)	(0.073)	(0.023)	(0.038)	(0.028)	(0.021)	(0.013)
$>$ 2 wks	−0.146***	−0.251***	0.090**	−0.144***	0.040***	−0.003	−0.040***	0.001	−0.017***
	(0.020)	(0.021)	(0.027)	(0.021)	(0.007)	(0.011)	(0.006)	(0.003)	(0.004)
$>$ 2 wks=1 × Coldstart	−0.180*	0.186**	−0.114	0.111	−0.074*	−0.023	−0.047	0.038*	0.024
	(0.079)	(0.071)	(0.108)	(0.088)	(0.029)	(0.043)	(0.032)	(0.019)	(0.014)
N	187640	249444	194840	97022	99409	99409	187640	187640	97543
R^2	0.22	0.67	0.81	0.65	0.56	0.52	0.15	0.15	0.98

Note: All specifications include product and year-week FE. Cluster-robust standard errors (at the product level) in parentheses.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

The overall results, shown in the first two rows, confirm the visual analysis that shows that average ratings, number of reviews, sales and keyword position all fall after fake review recruiting ends. However, some of the increases in these variables are still present in the first week or two after the last Facebook post.

We also test interactions for "cold-start" products. We find that these products' ratings fall

even further than for regular products, but that their increase in number of weekly reviews is more persistent. This is consistent with the fact that the decrease in sales rank is larger and more persistent for cold-start products. We don't find differences in terms of keyword rank, and find that the use of sponsored listings decreases for cold-start products while it increases for the rest of the products.

1.C Analysis of the mid-march Amazon purge

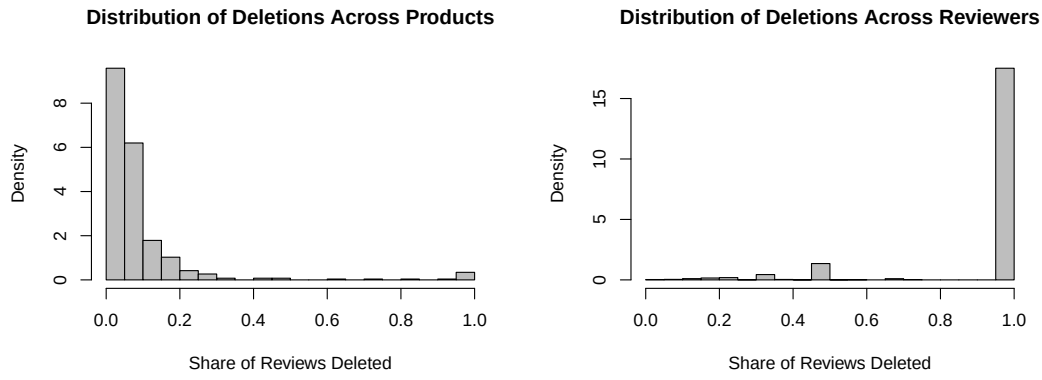
We have done an investigation of the patterns in review deletion across products, time, and reviewers in order to better understand the review "purge" and what selection criterion Amazon is using for these deletions. We focus first on the distribution of deletions across products and across reviewers to determine whether deletions are targeted at specific products buying fake reviews or at reviewers writing them. To give an example of this logic: if 10% of reviews were deleted during the review purge event, it could be that 10% of products were targeted and they all had 100% of their reviews deleted or it could be that specific products were not targeted and all products had about 10% of their reviews deleted. Similar analysis could find if individual reviewers were targeted or if the deletions are uniform across reviewers (of course these are extreme examples - reality must lie somewhere in between.)

We focus our investigation on the focal products (products observed buying fake reviews on Facebook) and find that during the 2-week period we call the review "purge", 3.2% of all 230,000 reviews are deleted. This is a small share of the total stock of reviews but in terms of the flow of deletions is many times higher than during normal periods. These deletions effect 40.6% of products (i.e. they have at least one review deleted) and 3.2% of reviewers.

This suggests deletions are targeted at a small group of specific reviewers and are not targeted at a narrow set of products. We next show the distribution of the share of reviews deleted

at both the product and reviewer levels (conditional on having at least one review deleted.) We plot histograms of each in Figure 1.17.

Figure 1.17: Distribution of Deletions During Purge Event



This figure shows that the vast majority (93%) of products affected by the review deletion event have fewer than 20% of their reviews deleted and nearly half have fewer than 5% of their reviews deleted. Among reviewers, the opposite pattern holds. The vast majority (87.5%) of reviewers have all of their reviews deleted. This evidence is unfortunately biased, however, by the nature of our data collection. We initially only collect reviews at the daily basis for our focal products and so the set of reviewers we analyze here are those who have posted on these products in this time period. We did not scrape these reviewers' other reviews (for non-focal products) at the time, as would be required to track the full share of their reviews deleted at a given point in time. Therefore the vast majority (83%) of these reviewers have only 1 review observed to begin with.

Yet, among reviewers with more than 1 review who have reviews deleted, the same pattern does hold. In this group, reviewers with multiple reviews, at least one of which is deleted in the review purge, 77% have 100% of their reviews deleted. When we condition on reviewers having at least 5 reviews the share with all reviews deleted is 78%.

This analysis strongly suggests that individual products are not targeted when Amazon deleted large numbers of reviews in mid-March 2020 but rather that individual reviewers

were targeted.

1.D DD Robustness checks

Sensitivity to the purge window Here we show that the sales estimates are not too sensitive to the choice of the window around the review purge used to select the set of control products. We do so by reporting in Table 1.12 the estimates for sales rank using three alternative windows around the mid-purge date: [-2,2] weeks, [-1,2] weeks, and [-1,1] weeks.

Table 1.12: Diff-in-Diff using different purge windows

	(1)	(2)	(3)
Purge Window	[-2,2]	[-1,2]	[-1,1]
After	0.166*	0.178*	0.198
	(0.070)	(0.077)	(0.115)
After × Treated	−0.325***	−0.338***	−0.377**
	(0.086)	(0.092)	(0.131)
PSM Sample	Yes	Yes	Yes
N	12512	12512	11553
R^2	0.85	0.85	0.87

Note: All specifications include product and year-week FE. Cluster-robust standard errors (at the product level) in parentheses.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Continuous treatment To further reduce concerns about our results being driven by the way in which we select control products, here we show that our estimates are robust to a continuous definition of the treatment. To do so, for each product, we define a treatment variable, *log Purge Distance*$_i$, which is equal to the log of the absolute value of the difference in days between the mid-purge date (March 15, 2020) and the date of the first Facebook post of each product. We then estimate Equation 1.7, but replacing the binary treatment variable with this new continuous treatment. We report these results in Table 1.13. We observe that

for small values of the treatment variable, i.e., for products whose first Facebook post is very close to the mid-purge date, there is a small and non statistically significant effect on reviews, and a positive effect on sales. However, the opposite is true for products whose first Facebook post is far from the mid-purge date.[24]

Table 1.13: Estimates using a continuous treatment variable

	(1) log Cum. Reviews	(2) Sponsored	(3) Coupon	(4) log Price	(5) log Sales Rank
After	0.040	−0.042	−0.034	−0.025	0.362*
	(0.070)	(0.047)	(0.067)	(0.019)	(0.146)
After × log Purge Distance	0.041*	0.019	0.009	0.004	−0.135***
	(0.019)	(0.013)	(0.018)	(0.005)	(0.037)
N	15789	9543	9543	9463	15077
R^2	0.93	0.64	0.67	0.99	0.87

Note: All specifications include product and year-week FE. Cluster-robust standard errors (at the product level) in parentheses.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Placebo review purge To further reinforce the validity of our estimates, we perform a placebo test in which we create a placebo review purge by moving the mid-purge date either four weeks back or four weeks forward. We estimate Equation 1.7 using these thresholds and report these results in Table 1.14. As expected, we observe that recruiting fake reviews has a negative effect on sales rank for control products and that this effect is not different for treated products.[25]

Alternative Propensity Score Matching algorithms Finally, we show that our results are not sensitive to the type of matching algorithm used. In Table 1.15 below, we report the estimates for sales rank using nearest-neighbor and local linear regression matching algorithms, an obtain results consistent with those reported in column 5 of Table 1.7.

[24]For example, at the median *log Purge Distance*$_i$ which is 3.89 (about 48 days), the increase in cumulative reviews is about 22% ($p < 0.01$) and the decrease is sales rank is about 15% ($p < 0.01$).

[25]We do not use PSM in this exercise to further reinforce the fact that potential differences between treated and control products are not driving the sales effects reported in Table 1.7 (however, we obtain qualitatively similar results when we apply PSM). In addition, using the full data sample and the real purge, we obtain results consistent with those reported in Table 1.7.

Table 1.14: Estimates using placebo review purges

	(1) 4 weeks before	(2) 4 weeks after
After	−0.166*	−0.142*
	(0.079)	(0.060)
After × Treated	0.027	0.001
	(0.086)	(0.065)
N	15077	15077
R^2	0.87	0.87

Note: All specifications include product and year-week FE. Cluster-robust standard errors (at the product level) in parentheses.

Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Table 1.15: Estimates using alternative matching approaches

	(1) NN	(2) LLR
After	0.213*	0.188
	(0.100)	(0.098)
After × Treated	−0.368**	−0.370**
	(0.120)	(0.117)
N	7288	11489
R^2	0.88	0.87

Note: In column1 we report the results using the nearest-neighbor algorithm for matching with $n = 20$, and in column 2 we report the results using the local linear regression algorithm with a bandwidth of 0.005. All specifications include product and year-week FE. Cluster-robust standard errors (at the product level) in parentheses.

Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Chapter 2

Optimizing Rating Systems for Innovation

2.1 Introduction

It is well-understood that innovation drives economic growth, firm performance, and consumer welfare (Hauser et al., 2006). As a result, policymakers and researchers strive to identify and understand the factors that influence innovation incentives. A large body of literature focuses on the relationship among innovation and market structure (Scherer, 1967; Goettler and Gordon, 2011), monetary policies (Atkeson et al., 2018), and legal protections (Levin et al., 1987). Taking a different approach, I look into an understudied factor driving innovation: information design. Specifically, I ask, how does the way firms convey their quality to consumers impact firm innovation incentives?

I study this innovation force in the context of rating systems. Firms often rely on rating systems to reach consumers, and many consumers use rating systems on a daily basis. Firms convey the quality of their products through rating displayed. As ratings impact consumer

choice and sales, firms are motivated to engage in activities that keep ratings high. That is, ratings are likely to influence firm innovation, especially for the many industries that have a time-varying product/service quality: hotels, restaurants, software, and so forth. Because of a rating's impact on sales, the rating system motivates firms to innovate and improve their product offerings. However, this innovation incentive might be limited given that most rating systems use simple averages to display ratings. When its product has accumulated many good ratings, a firm may not be motivated to continue innovating and maintain existing quality, because a marginal review will not change its average display rating. In response to such rating inflexibility, firms may even shrink their investment or go so far as to sacrifice the product's quality to increase profitability.

If firms use displayed ratings to guide their investments in improving products, then platform rating aggregation policies can play a key role in increasing or decreasing their innovation incentives. However, many rating platforms use a simple average, for example, Freelancer.com, TripAdvisor, Bookings.com, and Google Business.[1] A better understanding of how ratings motivate firms could enable designing a rating aggregation policy that better motivates innovation and quality refinement in firm offerings. In this paper, I conduct an in-depth study of the impact of information disclosure on innovation incentives in the context of rating platforms and, more specifically, of the corresponding implications for the design of the rating aggregation policies. I seek to answer the following questions: in a scenario in which quality varies, do display ratings have an impact on innovation decisions, and if so, when and how do they influence innovation decisions? What rating aggregation policy is optimal for motivating innovation?

[1]Other platforms either use a variation of a simple average or choose not to disclose the specific calculations used in their rating systems, with the exception of Google Play Store. Vrbo calculates a simple average from all the reviews in the past 365 days. Airbnb and Yelp do not disclose their specific rating calculation. After a policy change in 2017, the iOS App Store gave developers the option of resetting ratings after a version update. If a developer chooses not to reset, the system assigns a simple average rating to the app. Google Play Store weighs ratings from the current app version instead of using the simple average system in place before 2019.

I collect a firm-level dataset from a mobile game app platform, Tap.io. Like the Apple App Store and Google Play Store, Tap.io is an app distribution channel with a sizable market share. Tap.io uses a simple average rating system, similar to those of many review platforms, and the insights and counterfactuals derived from this context can therefore be widely applied. I define innovation as the creation and subsequent introduction of a good or service that is either novel or an improved version of existing goods or services. The setting permits me to use app updates as a measure of innovation decisions, which is consistent with the body of literature that uses app updates to capture digital innovation (Boudreau, 2012; Wen and Zhu, 2019). Besides the observable innovation measurement, this dataset contains the number of app installs, which allows me to approximate firm sales.

My empirical strategy combines a reduced-form analysis and a structural model. The reduced-form analysis describes the relationship between rating systems and innovation decisions. I start by showing that innovation has positive impacts on all key rating system metrics: number of installs, number of reviews, and displayed ratings. Next, I propose a new metric, rating agility, which is defined as how quickly displayed ratings can be changed by the addition of recent reviews. This metric captures the key intuition behind this project and allows me to further quantify the impact of innovation. The analysis shows that innovation has three benefits: direct benefit to consumer demand, indirect benefit to consumer demand via improved ratings, and improved rating agility to allow change in the displayed rating.

Empirical patterns suggest that ratings impact innovation decisions and that firms consider trade-offs between innovation costs and potential profits through innovation. To evaluate how different rating aggregation policies affect innovation incentives, I develop a dynamic structural model that explicitly shows the decisions of a forward-looking firm. I measure revenue and innovation cost by the number of installs. I estimate the innovation cost as the equivalent of 1,520 installs for each update. Then I evaluate the impact of alternative rating aggregation policies on innovation incentives. The preliminary counterfactual analysis

investigates the impact of two platform information design policies: weighting recent ratings more and highlighting innovation. This analysis shows that weighting recent ratings more can increase the innovation rate by 4.87%.

This study has direct managerial implications. Review platforms with varying quality settings are prevalent. By understanding how firms react to reviews and ratings, managers of review platforms can design better rating display policies to motivate firms to continue innovating and improving their product offerings. Such policies might increase reviewer engagement and possibly build more profit revenue for the review platforms through high consumer dependence and high consumer engagement. This paper also provides a model framework for practitioners to evaluate the trade-off of innovation decisions. Thus, this study not only enriches the literature on online reviews but also offers insights into several important sectors for marketing practitioners.

Literature Review- While a great deal of work has focused on how consumers respond to reviews and ratings, the literature on the supply-side response is sparse, but growing. The extant literature indicates that when firms know the sales impact of ratings, they react by adjusting their advertising strategy (Hollenbeck et al., 2019) or by increasing prices to enjoy the benefits of a good reputation (Lewis and Zervas, 2016). Hunter (2020) empirically shows that ratings incentivize firms to take strategic, but costly short-run actions to improve their ratings. Firms also enhance quality based on the consumer reviews in the hotel industry (Ananthakrishnan et al., 2019). Contributing to this stream of scholarship, this paper quantifies the impact of ratings on innovation incentives, showcasing another type of firms' response to ratings. As such, it goes toward redressing the dearth of literature pertaining to the design of rating aggregation policies, despite the significant impact of displayed ratings. In one of the few studies on this topic, Jin et al. (2018) use a structural model to propose an adjusted average rating to improve information efficiency for consumers.

The current paper also contributes to the discourse around the impact of information de-

sign on supplier behavior. Information design can complement monetary levers for platform owners to manage supply-side decisions. Bimpikis et al. (2020) shows theoretically that information design can influence supply-side entry, exit, and pricing decisions and that platform owners can leverage information to increase profitability. Harbaugh and Rasmusen (2018) establish that quality certifiers can increase information revealed to the public by employing coarser quality grades that result in increased supplier participation. Empirically, Ershov (2018) shows that the introduction of new product categories lowers search costs, leading to increased entry and welfare increases in the context of the app store. Comino et al. (2019) argue that the level of quality control by a platform, in the context of app stores, can affect the returns to product updating and therefore affect the incentive to engage in product innovation. Hui et al. (2022) demonstrate that a quality certificate with a higher bar motivates some sellers to incur costs for quality improvements, while other sellers give up on the badge and reduce effort. Nosko and Tadelis (2015a) show that buyers may draw conclusions about the quality of the platform from single transactions, causing a reputational externality across sellers, which further emphasizes the importance of communicating product quality and optimizing information design for platform owners.

By explicitly modeling firms' endogenous investment decisions, this paper also relates to the theoretical literature on reputation and firms' incentives for investment. Board and Meyer-ter Vehn (2013) propose a model of firm reputation in which a firm can invest or disinvest in product quality and market learned firm reputation through the news. Similarly, in this study, the firm's investment is an endogenous decision with long-lasting effects. However, this study's context (i.e., review platforms) and the measure of market belief (i.e., displayed ratings) are more common and applicable. Horner and Lambert (2016) argue that rating systems motivate rated agents and suggest that optimal rating systems focus on recent ratings, maintaining that this approach prevents sellers shrinking. Following a similar rationale, the current paper applies a different modeling framework and explicitly provides counterfactual analysis evidence of real-world data.

2.2 Data

I collected data from Tap.io, a mobile app game distribution and review platform. The platform is headquartered is in China, and the website mainly serves East Asian consumers (Chinese, Korean, and Japanese). By May 2018, its penetration rate among active users who use mobile game review platforms was 20.48% in the Chinese market.[2] Because this platform functions similarly to other major mobile application platforms (e.g., the iOS store and Android App store), data from this platform are likely representative.

I scraped daily data from October 16, 2019, to August 11, 2020. The main variables are game name, game description, developer name, developer ID, number of installs, what is new, all version update details, file size, current version, update time, and total ratings.[3] I also scraped the reviews themselves, including user names, ratings, and written content.

Specific features of the data allow me to answer my research questions. First, I observe the number of installs. This information permits me to control for demand when I measure the impact of ratings on innovation decisions and to detect the impact of those decisions. As firms are likely to care about ratings owing to their effect on demand, observing this variable is crucial. Second, I observe update content, which I can classify through text analysis. Third, Tap.io displays simple average ratings, along the same lines as other rating systems such as TripAdvisor and Google Business, among others. From this, I can derive generalizable implications from counterfactual analysis and make interpretations about managerial implications.

[2] http://www.woshipm.com/evaluating/1070766.html

[3] I use "number of installs" and "number of downloads" interchangeably in this paper.

2.2.1 Data Description

Table 2.1 presents summary statistics for active games. I define "active games" as those games that have more than five new reviews and 10 new installs over the course of the 44 weeks of the data-scraping period. My data include a total of 2,270 active games. The summary statistics indicate significant variation in the number of installs across games, with 75% of the games averaging only 50 installs per week. Less than 3% of consumers leave reviews; the weekly average number of installs is 313, while the weekly average number of reviews is 8.6. On Tap.io, consumers can leave a rating ranging from one to five stars, but the displayed rating on the platform is based on a scale from 1 to 10. The average update probability is 17.4%, and the first and third quantiles are 7.3% and 23.3%, respectively. The mean age of games is 82.7 weeks, but 25% of the games are younger than 27.5 weeks, which means I observe a substantial number of new games.

Table 2.1: Summary Statistics for Active Games

	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Weekly avg number of installs	2,270	313.0	3,142.4	0.08	1.9	50.7	90,967.9
Weekly avg number of reviews	2,270	8.6	25.2	0.04	0.3	4.8	356.0
Weekly avg ratings	2,270	3.8	0.8	1.0	3.3	4.5	5.0
Weekly avg displayed ratings	2,270	7.9	1.4	2.0	7.2	9.0	10.0
Avg update probability	2,270	0.174	0.130	0.0	0.073	0.233	0.889
Median age (weeks)	2,270	82.7	60.8	0.5	27.5	135.5	207.5

2.2.2 Data Features for Identification

To pin down the causal impact of ratings on updates or the number of installs, I would ideally like to observe two identical games receiving different ratings and see how firms update games differently or how the number of installs changes correspondingly. With Tap.io, it is not uncommon for the same game to be released in different regions (e.g., mainland China, Japan, Korea). For each region, the game has a corresponding Tap.io game introduction page, where reviews for the specific regional release appear. Correspondingly, firms might update

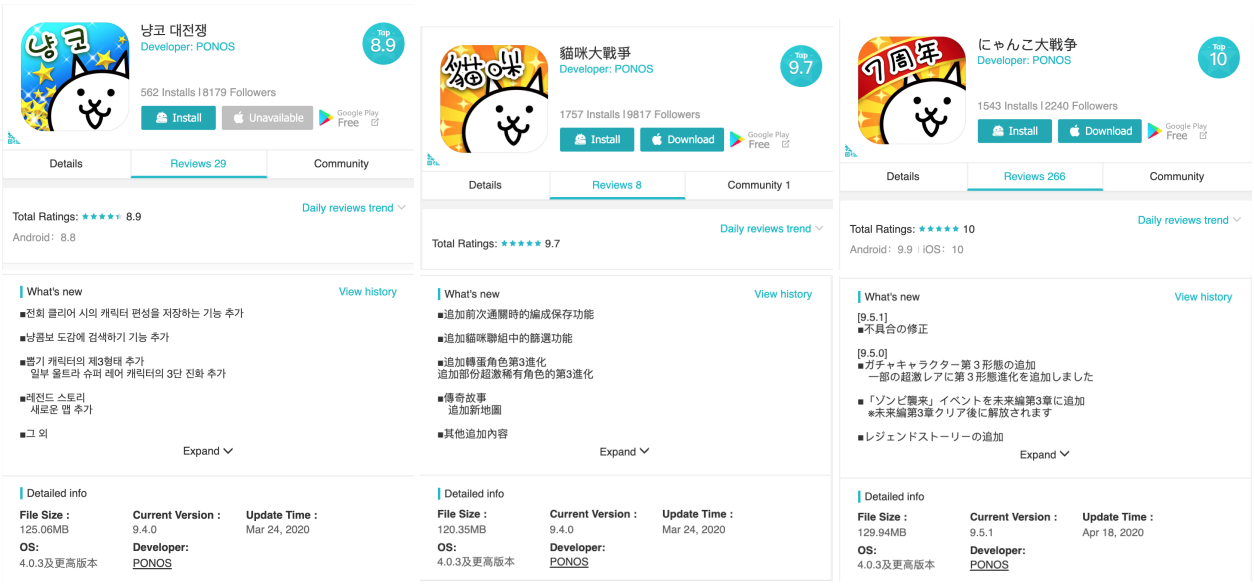


Figure 2.1: The Same Game in Three Regions

the game differently across regions, with different regions receiving different installation numbers. In my data, 400 games have two or more regional releases with the same version update behavior and 65 games have two or more regional releases with different version update behavior. This feature is helpful to identify the causal impact of displayed ratings on number of installs and to identify the impact of displayed ratings on update decisions.

Figure 1 shows an example of a game with three regional releases. The game "The Battle of Cats" was released in South Korea, China, and Japan. The Japanese version has more updates than the versions in the two other regions, by version number and update time. This example shows that the regional versions of the same game can receive the same or different version updates, different displayed ratings, and different numbers of installs.

2.2.3 Text Analysis for Observed Update Heterogeneity

The main objective of the text analysis is to classify product update content to enrich the structural model and provide more insights into the context of innovation efforts. Product

update content contains information regarding the nature of innovation. I define three types of updates: Bug Fix, Additional Content, and Balance Patch. In a Bug Fix, the developers correct back-end code errors to make the game experience smoother. Additional Content refers to the developer adding new features, new characters, new events, and/or new content to the game. A Balance Patch involves the developer changing numbers or probabilities in the game to, for example, make some enemies more difficult to battle, weaken a particularly powerful weapon, or adjust the chance of receiving a rare material. Note that one update can include multiple update types. For example, update content could be "Fixed some bugs, Implementing new Multi-Mission (Beta) feature, Increasing Parts Box limit." This example includes all three update types. The first part "fixed some bugs" corresponds to the Bug Fix type. The second part "implementing new Multi-Mission (Beta) feature" corresponds to the Additional Content type. The last part "Increasing Parts Box limit" implies a number change in the game, which corresponds to a Balance Patch; the "Parts Box" feature is pre-existing and there is no bug to fix.

Classifying the updates is important because they represent different types of innovation efforts. Additional Content represents innovation effort directly because the developers worked to create something completely novel. Bug Fix represents quality refinement efforts. Even if there is nothing new by definition, the developer still made an effort to improve the quality of their product. Conceptually, Balance Patch falls between Bug Fix and New Content. The developer did not invent something completely new, but improved the balance of the game by changing statistics and probability in the back end to make the overall game experience better.

To classify the update content, I first manually labeled the content of 1,019 unique version updates out of 6,479 updates. I then trained a LogitBoost model and classified the rest. Specifically, I converted update content into a vector representing the words within the update content. Then I tested a few supervised machine learning models with 10-fold cross-

validation to pick the best model. I ultimately selected a LogitBoost model and trained the model with preclassified update content. I classified the rest of the update content given the trained model. The accuracy of classification is between 0.86 and 0.92.

Table 2.3 shows the summary statistics for the update type classification. More than 80% of updates involve adding new content, and almost 20% of the updates include all three types of updates. Balance Patch is the least common type of update. The results of update type classification show that the version updates include innovation efforts and quality refinement efforts. Table 2.4 shows the precision and recall for the classification.

Table 2.2: Summary Statistics of Update Type Categorization

Type	Count	Percent
Bug	652	3.74
Balance	255	1.46
Content	10,612	60.80
Bug, Balance	429	2.46
Balance, Content	512	2.93
Bug, Content	867	4.97
Bug, Balance, Content	3,212	18.40
Not Categorized	914	5.24

Table 2.3: Summary Statistics of Update Type Categorization

Type	Count	Percent
Content	10,612	60.80
Bug, Balance, Content	3,212	18.40
Bug, Content	867	4.97
Bug	652	3.74
Balance, Content	512	2.93
Bug, Balance	429	2.46
Balance	255	1.46

Table 2.4: Update Classification Precision and Recall

Update Type, Language	Precision	Recall	N
Content, English	0.48	0.56	100
Bug, English	0.96	0.83	100
Balance, English	0.95	0.73	100
Content, Chinese	0.82	0.82	203
Bug, Chinese	0.96	0.97	203
Balance, Chinese	0.89	0.76	203
Average	0.86	0.80	303

Note: precision indicates the ratio of the number of correctly predicted cases for a given update type to the total number of cases predicted to be of that type. Recall is the ratio of the number of correctly predicted cases for a given update type to the total number of cases of that type.

2.3 Descriptive Evidence

This section describes how developers benefit from updates and how ratings influence developers' update decisions.

2.3.1 Benefits of Updates

In this section, I examine how game updates affect the number of installs, the number of reviews, and displayed ratings. I show that updates increase the number of installs, the number of reviews, and the average displayed rating. Without an update, the number of installs, number of reviews, and display ratings slowly trend downward. Firms' investment in innovation do tend to have impacts on the metrics pertaining to game popularity. The descriptive results also serve as a motivation for structural model choices in the later sections.

In Figure 2.2, I show average data patterns 14 days before and after an update. The figure shows that an update boosts the number of installs, the number of reviews, and the displayed rating.

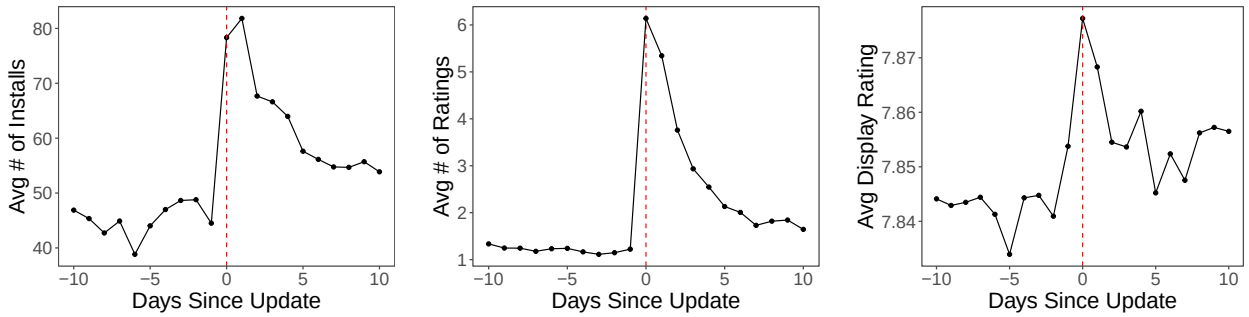


Figure 2.2: Daily Installs, Reviews and Ratings Data Pattern Before and After An Update

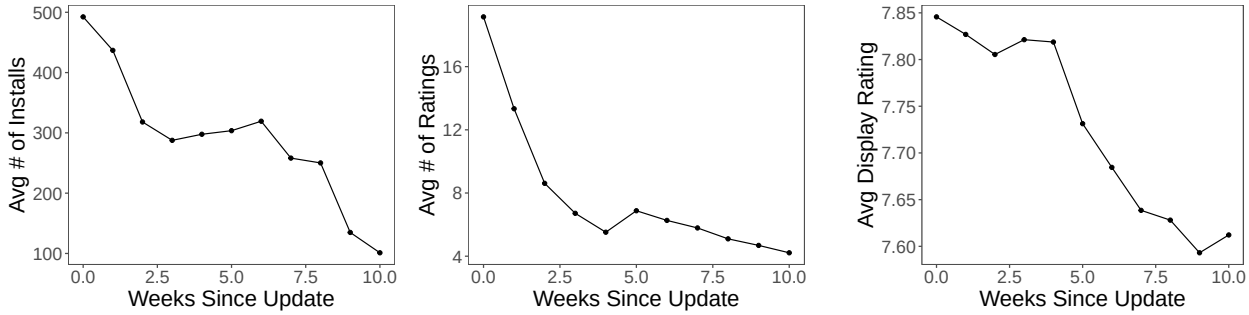


Figure 2.3: Weekly Installs, Reviews and Ratings After an Update

The effect of firms' investment may last some time, but it does not last forever. Figure 2.3 shows how the average number of installs, the average number of reviews, and the average displayed ratings change week by week for 10 weeks after an update. These three key metrics generally trend downward over the course of a few weeks if there is no update. This pattern creates incentives for firms to periodically invest in updating to increase demand or to use innovation to increase ratings and reviews.

The above patterns indicate that innovation yields three benefits in the context of the rating systems. First, innovation has a direct benefit on consumer demand. Second, innovation has a direct benefit on ratings. Given that a higher rating is associated with a higher demand, innovation also indirectly boosts consumer demand via ratings.[4] The third benefit of innovation is allowing an increase in new ratings, due to the addition of new reviews, which will improve displayed ratings. However, the displayed rating is unlikely to change substantially

[4]See Appendix B for the analysis of the positive impact of ratings on demand.

if the game has accumulated a lot of reviews under the current rating aggregation policy. Because different games have different numbers of existing reviews, the number of new reviews is not directly comparable across games. Thus, I define a new variable in the next subsection to advance the discussion of the third benefit of innovation.

2.3.2 Rating Agility is a Key Metrics

To capture the intuition of how quickly displayed ratings can changed by the influx of new reviews, I define a new variable: rating agility. In an simple average rating system, rating agility in each time period for each firm is defined as $\omega = \frac{R}{TR}$ where R represents the new number of reviews written in a week and TR represents the total number of reviews. This variable captures a key aspect of rating system aggregation policy: if part of innovation motivation comes from the desire to improve the displayed ratings, then the speed with which a displayed rating can be changed plays a key role in motivation. A redesign of the rating system aggregating policy will have an impact on this aspect.

In an simple average rating system, rating agility decreases over time, as shown in Figure 2.4, and it decreases quickly. It approaches zero 12 weeks after the game has been launched.

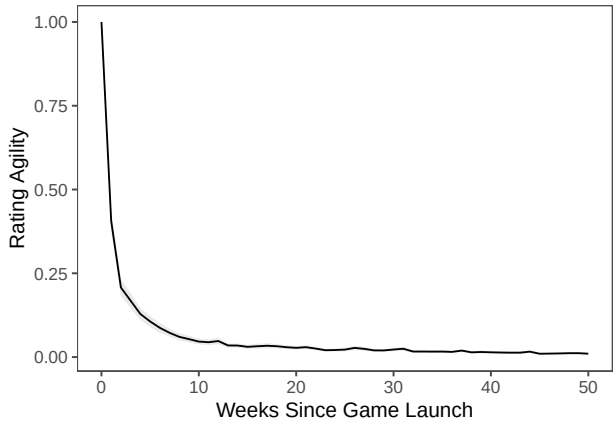


Figure 2.4: Rating Agility over Time

If developers are motivated by a desire to improve ratings via updates, then the higher the

rating agility, the more likely they are to push an update to improve the ratings. Such a pattern is shown in Figure 2.5. The upward trend shows that when the rating agility is higher, developers are more likely to push out an update.

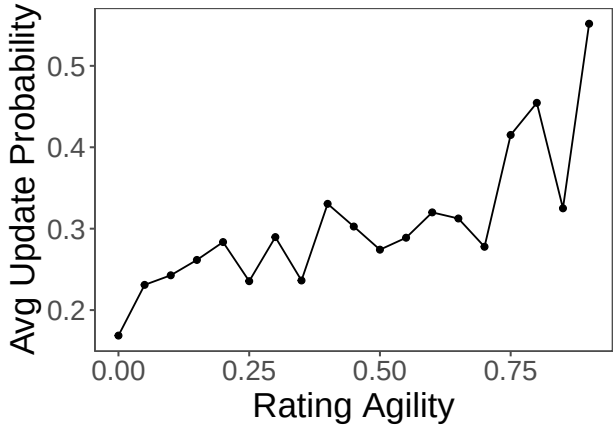


Figure 2.5: Update Probability v.s. Rating Agility

As discussed previously, number of reviews is not a directly comparable variable across products. Rating agility as a concept can capture the essence of the key motivation of this paper as well as mimic the movement of number of reviews. Table 2.5 shows summary statistics of the correlation between log rating agility and log number of reviews for each game. It provides evidence that log rating agility is highly correlated with log number of reviews. Interestingly, log number of reviews and log number of installs are not highly correlated for each game. It means number of reviews cannot be modeled as a fraction of number of installs and these two processes must be modeled separately.

Table 2.5: Correlation Summary

	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Correlation(log rating agility, log reviews)	1,432	0.98	0.09	−1.00	0.99	1.00	1.00
Correlation(log installs, log reviews)	1,432	0.25	0.48	−1.00	−0.04	0.61	1.00

2.4 A Dynamic Model of Firm Innovation Decisions

The goal of this paper is to determine the optimal rating aggregation to maximize firm innovation behavior in a dynamic quality environment. My reduced form analysis indicates that when firms just launch a game, a lower rating will lead to a larger likelihood of updating it. On average, if the displayed rating is high, it indicates a promising future, and the developers are more likely to update the game. The results of the analysis indicate that ratings affect innovation decisions and that firms consider trade-offs between innovation costs and potential profits through innovation. To evaluate how different rating aggregation policies affect innovation incentives, I develop a structural model that explicitly models the firm's trade-offs. This is a single-agent dynamic model in which the firms make weekly decisions about whether to update their game. I assume that firms are rational and forward-looking, with an objective to maximize their total discounted revenue by making optimal choices.

2.4.1 Setup

I first assume that each firm only cares about its own state and make decisions correspondingly, without considering the competitors' information. The corresponding solution concept is oblivious equilibrium (Weintraub et al., 2008). Oblivious equilibrium assumes that each firm makes decisions based on its own state and knowledge of the long-run average industry state, and firms ignore current information about competitors' states. This solution concept is suitable in this case given that there are hundreds, if not thousands, of firms in the market. It is impossible for a firm to take in each competitor's state information and derive a strategy correspondingly.

The timing of the events for my model is shown in Figure 2.6. At the beginning of each

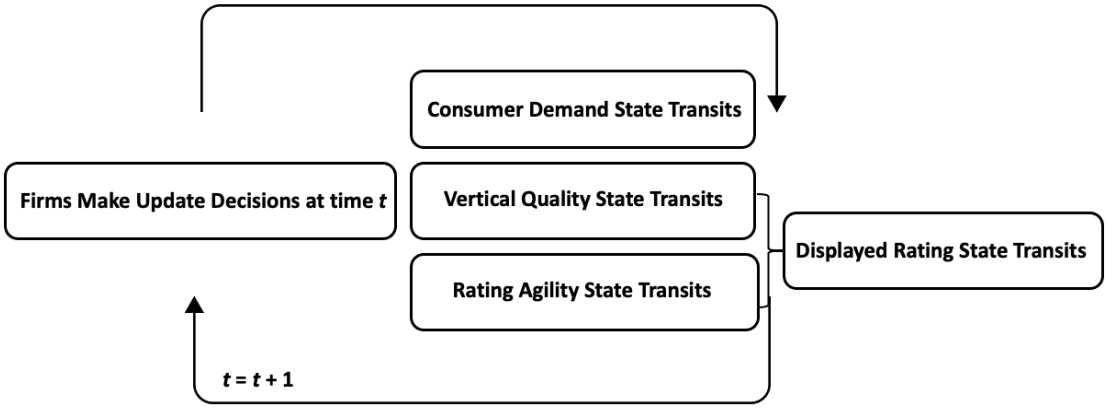


Figure 2.6: Timeline of the Structural Model

time period, the firm decides whether to pay a fixed cost C to update the game or not given the states and how states transit. Then, the state variables evolve and the firm obtains its realized utility and observes the change in the state variables at the end of each time period.

Firms make innovation decisions via their update decision: Action $A = \{0, 1\}$. The decision $A = 1$ if the firm updates the game and $A = 0$ if it does not update the game. There are multiple types of updates (bug fixes, balance improvements, and new content) and a firm can have multiple games. In the simplest model, I assume that each firm has only one game and the firm can only choose to update or not to update. In later estimations, I incorporate heterogeneity in the cost regarding update cost, elaborated in Section 5.

I assume that the firm pays a fixed cost C with each update and there is no cost of operating. As many inactive apps exist on the market, the assumption of no variable cost is reasonable.

State space S has four state variables: the number of installs D, rating change agile state ω (derived from the current number of reviews divided by the total number of reviews), rating R, and game quality Q. Firms observe the first three state variables, but not game quality. Firms infer quality through weekly average ratings. In addition, firms know how states transit empirically. In the later specification, the state transition matrix may contain stochastic information because some state transition processes may not be deterministic.

They are modeled as a Markov process.

2.4.2 State Transition

In this subsection, I outline how the four state variables transit. Note that the correlation between the number of downloads and the number of reviews is close to zero, as shown in Table 2.5. Therefore, the number of downloads state transition is independent of how rating agility state ω transits.

In equation 2, number of downloads D_t is a function of D_{t-1}, past rating R_{t-1}, and action A_t. Empirical data patterns show that demand is highly correlated over time. The potential reason could be that search rank or consumer awareness of the game is highly correlated over time. Consistent with previous literature and data patterns, the higher the rating, the higher the consumer demand. In addition, empirical data patterns show that whenever an update occurs, a spike occurs in the number of downloads. To incorporate these two patterns, I include variables R_{t-1} and A_t. In previous research (Hunter, 2020; Chevalier and Mayzlin, 2006), rating is modeled as having a first-order effect on demand, which further justifies this linear relationship. I specify that $logD_t$ has a linear relationship with $logD_{t-1}$, R_{t-1} and A_t. Based on this specification, I interpret ρ^D as the impact of exposure and consumer awareness from the past time periods, while γ evaluates the impact of displayed rating on consumer choice and γ^D represents the update promotional effect. Furthermore, the equation implies that these three impacts are independent of each other.

$$logD_{j,t} = \rho^D logD_{j,t-1} + \gamma R_{j,t-1} + \delta^D A_{j,t} + \varepsilon^D_{A,j,t} \tag{2.1}$$

where $\varepsilon^D_{A,j,t} = \xi_j + \epsilon_{A,t}$, ξ_j is the fixed effects of firms and $\epsilon_{A,t}$ represents the demand state

transit uncertainty in each time period. Different firms might carry a different reputation, which in turn produces a differential impact on the number of downloads. Thus I include firm fixed effect in the equation as well. Equation 2 can also justify the empirical pattern of a high rating and a high demand shock quickly pushing a newly launched game to have high demand. On the other hand, a popular game might not gain as much benefit from high ratings, and a downtrend pattern in demand is mainly observed.

In equation 3, perceived quality Q depreciates over time with the discount factor ρ^Q as the app's content becomes outdated and consumer tastes evolve.[5] If the firm chooses to innovate, the perceived quality will get a boost by a fixed amount δ. Perceived quality is not an observable parameter. Firms can only infer the perceived game quality through weekly average ratings $\tilde{Q}_t$.

$$Q_t = \rho^Q Q_{t-1} + \delta^Q A_t \tag{2.2}$$

While true quality Q evolves in a deterministic way, $\tilde{Q}_t$, the weekly average rating reflecting the true quality, can be seen as a noisy signal of true quality. Equation 4 captures the relationship between weekly average rating and actual quality.

[5]Perceived quality and true quality are different from objective quality. Previous literature (Li and Hitt, 2008; Godes and Silva, 2012) documents a pattern whereby ratings change systematically over both order and time and shows that this pattern is caused by self-selection bias and increasing difficulty in accessing ratings. In either case, it is possible for firms to innovate the product so that it works well for a larger audience, thereby mitigating the downward trend. This paper also only models vertical quality, not horizontal quality.

$$\tilde{Q}_t = \begin{cases} Q_t + \epsilon^Q_{A,t} & \text{otherwise} \\ \overline{Q} & \text{if } Q_t + \epsilon^Q_{A,t} > \overline{Q} \\ \underline{Q} & \text{if } Q_t + \epsilon^Q_{A,t} < \underline{Q} \end{cases} \tag{2.3}$$

where $\overline{Q}$ represents the highest rating the consumers can give, and $\underline{Q}$ represents the lowest rating the consumers can give. $\epsilon^Q_{A,t}$ is the random part in this quality signaling process.

The rating agility state ω_t is a function of the past rating agility state ω_{t-1} and action A_t, as specified in equation 5. The rationale is that, with an update, it becomes easier for the firm to change the current rating state, as the firm usually gets an increase in reviews right after the update. However, this boost decreases over time.

$$log\omega_t = \rho^\omega log\omega_{t-1} + \delta^\omega A_t + \epsilon^\omega_{A,t} \tag{2.4}$$

where $\epsilon^\omega_{A,t}$ represents the rating agile state transit uncertainty in each time period.

The rating transition is specified in equation 6, where ω is rating agility. This rating state transition represents the simple average rating aggregation system. The details in the derivation are shown in Appendix A. This specification simplifies the need to separately specify the number of reviews and the total number of reviews, and it captures the core of the rating averaging algorithm by centering on how quickly the current rating is affected by the current quality and the past rating.

$$R_t = (1 - \omega_t)R_{t-1} + \omega_t \tilde{Q}_t \tag{2.5}$$

2.4.3 Revenue Function Calibration

In my data, I do not observe the actual revenue/profit; instead, I observe the number of downloads. Therefore, with the assumption that each download contributes the same revenue to the firm, I calibrate the realized revenue function in the following:[6]

$$Rev_t(S_t, A_t) = D_t - C \cdot A_t \tag{2.6}$$

Note that the stochastic component is in the number of downloads state transition, and the firm makes a decision based on the expected current time period payoff and the total future discounted payoff. With this specification, cost C is measured by the number of downloads.

2.4.4 The Dynamic Optimization Problem and Firm Trade-off

We can now state the complete optimization problem facing each firm. Each firm chooses an infinite sequence of innovation decisions to maximize the expected total discounted revenue:

[6]From the data, the vast majority of games are free to download. This "free-to-download" mode is the norm for the mobile game market in the East Asian region. The developers mainly profit from in-app advertisement and/or in-app purchases. Given the research questions and data limitations, I abstract away developers' revenue choices.

$$\max_{\{A_{jt}\}_{t=0}^{\infty}} E\{\sum_{t=0}^{\infty} \beta_j^t Rev(S_{jt}, A_{jt})|S_{j0}\},$$

where

$$Rev(S_{jt}, A_{jt}) = -C \cdot A_t + \int exp\{\rho^D logD_{j,t-1} + \gamma R_{j,t-1} + \delta^D A_{j,t} + \epsilon_{Ajt}^D\} dF(\epsilon)_{Ajt}^D$$

Given state variables, the firm obtains realized payoff $Rev(S, A, \epsilon_A, \theta)$, where R represents revenue, ϵ_A represents the stochastic process given action A, and θ represents a set of parameters.

However, given the timeline, the firm can only make update decisions given expected revenue and expected total future discounted utility. Therefore, the value function can be expressed in the following:

$$V(S_{jt}) \equiv \max_{\{A_{jt}, A_{j,t+1}, \ldots\}} E[\sum_{\tau=t}^{\infty} \beta^{\tau} Rev(S_{j\tau}, A_{j\tau})|S_{j,t-1}],$$

where β is the discount factor. This value function is known to be the unique solution to the Bellman equation below:

$$V(S_t) = \max_{\{A_{jt}, A_{j,t+1}, \ldots\}} \{E[Rev(S_{jt}, A_{jt})|S_{j,t-1}] + \beta E_{S_{jt}}[V(S_{jt}, A_{jt})|S_{j,t-1}]\} \tag{2.7}$$

In the infinite horizon dynamic programming problem, the policy function does not depend on time. We can thus eliminate the time subscript.

In summary, firms' intertemporal trade-offs are associated with cost-benefit trade-off considerations. For each time period, the firm will compare the expected revenue with the innovation decision and without innovation decision, and see whether the difference would justify the innovation cost.

2.5 Identification and Estimation

2.5.1 Identification

The set of structural parameters θ includes the innovation cost C, as well as state transition parameters ρ^D, γ, δ^D, σ^D, ρ^Q, δ^Q, σ^Q, ρ^ω, δ^ω, and σ^ω. State transition parameters can be estimated using panel data regressions, which are elaborated below.

To identify the impact of past installs, displayed rating, and update dummy on future installs, I estimated equation 2 with Game fixed effects and Region fixed effects on the games with multiple regional releases and the same update schedule. When a game has multiple regional releases in different regions, it can be perceived that the same game is distributed on multiple platforms . I apply game fixed effects to control for game quality and region fixed effects to control for consumer taste. The identification strategy is similar to Chevalier and Mayzlin (2006). More robustness checks for the demand transition process appear in Appendix B. Rating agility transition process (equation 5) is identified and estimated in the same way.

The identification for equation 3 comes from the imposed assumption of a perceived pattern of decreasing quality. A similar assumption is commonly seen in the literature, such as Goettler and Gordon (2014) and Allon et al. (2021). Equation 3 is estimated using a Tobit

model. I used weekly average rating to approximate quality. Consumers are only allowed to give a rating ranging from 1 to 5 while the ratings are displayed out of a scale of 10. To make scale consistent, I double the weekly average ratings to be at a scale of 10. In addition, the quality can be higher than the possible scale (10 out of 10) or lower than possible scale (2 out of 10), thus I applied the Tobit model to accommodate such a possibility.

2.5.2 Likelihood

The full likelihood function is

$$Likelihood = L(\{\{S_{jt}|\widehat{S_{jt-1}}, A_{jt}\}_{t=1}^{T})\}_{j=1}^{J}) \cdot L(\{\{A_{jt}|\widehat{S_{jt-1}}\}_{t=1}^{T})\}_{j=1}^{J})$$

where $\widehat{S_{jt}} = \{D_{jt}, R_{jt}, Q_{jt}, \omega_{jt}\}$. Because the likelihood for the optimal choice and for the state transition process are additively separable when we apply a log transformation to the likelihood function, we can first estimate the state transition process from the data and then maximize the likelihood for the optimal choice. The likelihood for the optimal choice is

$$L(\{\{A_{jt}|\widehat{S_{jt-1}}\}_{t=1}^{T})\}_{j=1}^{J}) = \prod_{j=1}^{J}\prod_{t=1}^{T} L(A_{jt}|\widehat{S_{jt-1}}) = \prod_{j=1}^{J}\prod_{t=1}^{T} Pr(A_{jt}|\widehat{S_{jt-1}}),$$

where $Pr(A_{jt}|\widehat{S_{jt-1}})$ can be written as

$$Pr(A_{jt}|\widehat{S_{jt-1}}) = EV_{A_{jt}}(D_{jt}, R_{jt}, Q_{jt}, \omega_{jt}).$$

2.5.3 State Transition Parameters

Table 2.6 shows an initial estimation of equations 2, 3 and 5. Each additional increase in rating (on a scale of 10) will increase installs in the next time period by 6%. An update will promote installs by 6% as well. Rating agility decreases quickly and updates offer a significant boost in terms of rating agility. The results indicate that quality depreciates slowly. In each time period, the quality will decrease by 1.5%. Each update will increase quality by 0.7 on of a scale out of 10. More robustness checks are in Appendix C.2.

Table 2.6: Transition Parameter Estimation

	Installs	Quality	Agility
Variables			
ρ	0.689***	0.986***	0.687***
	(0.021)	(0.002)	(0.014)
γ	0.055***		
	(0.019)		
δ	0.135***	0.066***	0.319***
	(0.016)	(0.026)	(0.027)
Fit statistics			
Observations	12,082	23,371	7,989
R^2	0.927	0.340	0.759

Note: The first and the third specifications include game and region FE. Cluster-robust standard errors (at the game level) in parentheses.
Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

2.5.4 Innovation Cost

Innovation cost is identified by the revealed preference argument. I assume that firms observe the innovation cost and can predict how states will evolve. The firms make the rational choice that they will only invest in innovation if the total discounted future benefit is worth the

innovation cost. Further, I assume discount factor β is 0.95, given the identification argument in Rust (1987).

To estimate the model, I adopt the IJC algorithm (Imai et al., 2009). I choose to apply the IJC algorithm owing to the large total number of states; IJC saves a massive amount of computation time. For the estimation, I discretize the state variables. There are 14 install states, nine rating states, 12 quality states, and four agility states. There are more quality states than rating states because perceived quality can be higher or lower than the allowed rating range. In total, there are 6,048 states.

I estimate specifications both without heterogeneity and with observed heterogeneity. In the latter case, the observed heterogeneity comes from the Update types, as classified by update text content. The revenue function I estimate becomes the following:

$$Rev_{jt}(S_{jt}, A_{jt}) = D_{jt} - \sum_{h=1}^{h=H} C_h \cdot A_{hjt} \tag{2.8}$$

where h represents the observed heterogeneity type. If the firm's type or update type is h at time t and the firm chooses to update, then $A_{hjt} = 1$. Otherwise $A_{hjt} = 0$.

The estimation results show that, on average, the cost of each update is 1.52 thousand installs, with a standard deviation of 0.02 thousand installs. Among all types, the New Content type costs the most, with 1.28 thousand installs with a standard deviation of 0.03 thousand installs. The Bug Fix type costs 0.49 thousand installs with a standard deviation of 0.06 thousand installs, and the Balance Patch type costs 0.09 thousand installs with a standard deviation of 0.10 thousand installs. Note that each update can have multiple types of updates. If an update includes all three types, then the total cost would be 1.86 thousand installs.

2.5.5 Model Analysis

In this subsection, I examine how well the model fits the data by comparing model prediction to actual data. Given the complex nature of the model, it is useful to perform a prediction exercise to show how well the model can approximate reality, despite the lack of formal tests. I solve the model computationally. Given the estimated parameters, one can adopt a fix-point approach to determine the expected value for each firm in each state. Given this expected value for each state space and a firm-specific state transition process, I can evaluate the probability of choosing innovation in each state space, and forward simulate the innovation decision and the next state given the transition probability matrix accordingly. If the model can simulate state transitions and innovation choices comparable to reality, then the model can predict reality well, and the counterfactual analysis holds more validation.

I proceed by using the estimated parameters to obtain the estimated value for each state. I used state data from 474 games in week 1, forward simulate 40 time periods for 500 times for each game, then compare innovation choices, state space distribution for the same set of games in week 10 in Table 2.7. Table 2.7 shows that the model performs well, and the prediction overall matches the reality quite closely.

Table 2.7: Comparison of Model Simulations to Data

State	True Week 1 mean (sd)	True Week 40 mean (sd)	Simulated Week 40 mean (sd)
Install	3.51 (2.49)	3.33 (2.32)	3.54 (2.16)
Rating	7.28 (1.56)	7.18 (1.45)	6.32 (1.41)
Agility	2.23 (1.09)	1.83 (0.89)	1.95 (0.80)
Quality	9.73 (1.95)	6.22 (1.21)	6.11(1.00)
Innovation Rate	0.285 (0.452)	0.171 (0.377)	0.183 (0.039)

Notes: The left two columns show the distribution of discrete states in the first week and in the 40th week of the data. The right column shows the results of the model simulation. For each game, the model is started at the true week 1 distribution and forward simulated 500 times.

2.6 Counterfactuals

The structural model allows me to examine counterfactual studies that consider the impact of alternative platform design policies on the innovation rate. This section evaluates different platform design policies, aiming to provide insights into actual business practice. Specifically, I discuss the implications of weighting recent ratings differently, investigate alternative rating aggregation policies and compare them to simple average policy and innovation highlight policy. I conclude by comparing the promotional effect of innovation and the impact of ratings on demand.

2.6.1 The implication of the weighting trade-off

Some platforms may opt to do a simple average over a certain time period (e.g., Opentable, Vrbo), while others may weight current ratings more heavily (e.g., Google Play Store). However, what the optimal weight is for motivating innovation remains unclear. In Figure 2.7, I explain one possible drawback of giving too much weight to recent ratings. In the figure, the example firm has a true quality of 4.8 out of 5 across time periods. Given the true quality, some consumers may see the firm as having a quality of 3.4 while other consumers may see the firm as having a quality of 6.5. As I model consumer ratings as a noisy signal of true quality, the actual consumer ratings will form a normal distribution around the true quality, with the mean equal to the true quality. However, consumers can only offer a maximum rating of 5, a rating cap that results in an asymmetric truncated distribution. Consumers who perceive the game with a quality of 6.5 can only give a rating of 5. Suppose the platform will only display the average rating from the past week, then this average rating will have more noise than the average taken over a longer time period. Because this noisier distribution is truncated asymmetrically, the average rating has a lower expected value than the less noisy distribution would. Therefore, weighting recent ratings

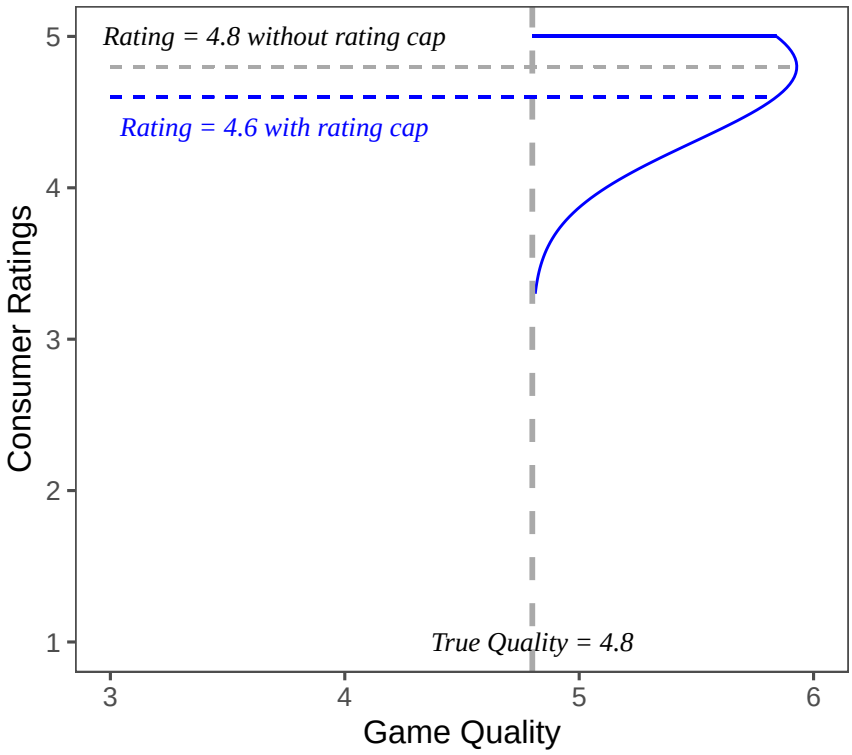


Figure 2.7: Weighting Tradeoff Illustration

too heavily will discourage high-quality firms from investing in innovation.

On the other side, if we weight recent ratings too little, the displayed ratings are unlikely to change, and firms are less incentivized to invest in innovation because they are less likely to see the investment reflected in the displayed ratings. Ultimately, firms' innovation incentives depend on the future payoffs and how four state variables transit. How to weight ratings is an empirical question, as the state transition parameters are given within the data context.

2.6.2 Alternative Platform Design Policies

In business practice, some rating systems use a simple average (e.g., Freelancer.com, TripAdvisor, Bookings.com, Meta Quest App Store, and Google Business). This will be the baseline policy. Some rating systems use a simple average of recent reviews. For example,

Vrbo calculates a simple average from all the reviews in the previous 365 days. OpenTable calculates ratings based on recent reviews from the previous 120 days without specifying the aggregation method. The first alternative rating aggregation policy I investigate is the simple average within a range of recent time periods. Before 2017, the iOS App Store would reset ratings when the developers pushed an update. After 2017, developers can choose whether they want to reset ratings after a version update. If developers choose not to reset, the system is essentially a simple average. These two policies will be the second and third policies I evaluate. I will compare the results with Leyden (2020). Leyden (2020) used apps from the Education, Productivity and Utilities categories. Hence the effect size might be different from the Game category. The general finding in Leyden (2020) is that the self-select version-reset results in more innovation than the mandatory version-reset. The fourth policy I evaluate corresponds to a recent Google Play Store policy change. Starting in 2019, the Google Play Store weighs ratings from the current app version more, instead of a simple average system. In the fifth counterfactual policy, I will investigate weighting recent ratings more by number of reviews, without consideration of version change. In the last counterfactual scenario, I investigate the impact of highlighting innovation. For example, platforms can choose to display games with major updates on the front page or prioritize apps with major updates in the search ranking algorithm. However, it is hard for platforms to gauge the quality of each innovation, and it is unclear how prominently the platforms should highlight the apps with updates. Thus, this scenario serves to compare the effectiveness of alternative rating aggregation policies and highlighting innovation policies.

The above alternative policies are summarized Table 2.8. I make one key assumption, which is that consumers will not react to the alternative rating policies by leaving more reviews or rating the games differently. Hence, the state transition process stays the same. In the counterfactual policies, I forward simulate each firm's reaction and their innovation decisions for 26 weeks (6 months). The counterfactual policies will result in a change in the innovation rate and hence a change in the market composition. Thus, forward simulation is necessary

to evaluate the impact of counterfactuals.

Table 2.8: Counterfactual Scenario

Case Number	Counterfactual Name	Corresponding Industry Practice
C0-Baseline	Simple Average	Tap.io, TripAdvisor, Google Business, freelancer.com, Oculus App Store, etc.
C1	Simple Average with Recent Reviews	Vrbo, Opentable, Lyft, Uber
C2	Mandatory Version Reset	iOS store before 2017
C3	Self-select Version Reset	iOS store after 2017
C4	More Weight on Current Version	Google Play store after 2019
C5	More Weight on Recent Ratings	N/A
C6	Highlight Innovation	iOS store, Google Play store

Some counterfactual scenarios involve a weight component (counterfactual scenarios 1, 4, and 5). To derive the optimal weight, I run the simulation over a grid of different weights and compare innovation rates. I found out that the optimal weight for counterfactual scenario 1 is to do a simple average of the last 6 weeks (counterfactual scenario 1). The optimal weight for counterfactual scenario 4 is to weight ratings from the current version 2.71 times more than the ratings from past versions, and the optimal weight for counterfactual scenario 5 is to weight recent ratings 50% more.

The results of counterfactual analysis are shown in Table 2.9. I show the impact of alternative policies on innovation rate, average product quality, average percentage change in revenue (calculated by average number of installs under new policies and compared with a simple average scenario), and average percentage change in profit (calculated by average revenue minus innovation rate times innovation cost). I conduct counterfactual analysis by varying the rating agility state transition process and how the displayed rating is calculated based on past ratings and current ratings. The baseline strategy is the simple average policy (C0). The other counterfactual I will use for comparison is C6 highlight innovation, where I double the return of innovation on demand. As expected, doubling the return of innovation on demand will increase innovation rate, product quality, firm revenue, and firm profit.

Counterfactual scenarios 2 and 3 evaluate the impact of iOS store policy change on the innovation rate. Interestingly, both the mandatory rating reset and self-select rating reset will

cause a loss of innovation and, correspondingly, a loss of average quality and average revenue. The comparison between mandatory version-reset and self-select version-reset results is consistent with (Leyden, 2020), self-select rating reset scenario leads to a 3.36% higher innovation rate than mandatory version-reset. However, both scenarios lead to a lower innovation rate than simple average policy. It is possible that, right after the rating reset, the high rating agility is will lead to a lower expected displayed rating and thus a lower future revenue, as Figure 2.7 suggests. Under such policies, when a game receives a negative shock to its reputation, it is harder to build a good reputation online compared to other policies.

There are three ways to weight recent ratings more: simple average with recent ratings (C1), weighting ratings from current versions more (C4), and weighting ratings from recent ratings more (C5). The counterfactual results indicate that the best counterfactual scenario is the simple average over the last six weeks. This counterfactual policy leads to a 4.87% innovation rate increase, a 0.70% quality increase and a 12.94% revenue increase.

Table 2.9: Welfare Analysis Under Alternative Rating Aggregation Policies

Impact	C0	C1	C2	C3	C4	C5	C6
	simple average	simple average w/ recent 6 wks reviews	mandatory version reset	self-select version reset	2.71 times more weight on current version	50% more weight on recent ratings	highlight innovation (double return)
Innovation Rate	17.56%	18.42%	16.09%	16.63%	17.48%	17.75%	18.06%
	(0.30%)	(0.27%)	(0.33%)	(0.32%)	(0.30%)	(0.33%)	(0.31%)
Δ Innovation Rate %	0.00%	4.87%	-8.40%	-5.33%	-0.46%	1.09%	2.86%
Avg Quality	6.844	6.892	6.818	6.824	6.844	6.853	6.852
	(0.030)	(0.032)	(0.027)	(0.029)	(0.030)	(0.030)	(0.030)
Δ Avg Quality %	0.00%	0.70%	-0.38%	-0.29%	0.00%	0.14%	0.12%
Δ Revenue %	0.00%	12.94%	-1.33%	-1.07%	-1.66%	0.36%	6.03%

Note: The forward simulation is performed on the entire sample across time, as the distribution of market composition is representative. The differences in percentages are performed by comparing the metrics of interest in counterfactual scenarios to those in the benchmark/simple average scenario. Revenue change is calculated by the change in average number of installs.

2.6.3 Source of Motivation

Two sources of motivation affect how rating systems influence the innovation choices: the promotional effects from the innovation decision itself (i.e. δ^d), and the rating effect (i.e. γ), which in turn affects future utility flow. Following the decomposition method in Amano and Simonov (2022), I investigate the sources of motivation by decomposing the demand transition process. I shut down the impact of rating on demand, and both rating impact and the promotional effect, respectively, and then evaluate the proportion of rating effect. I define two scenarios by their demand transition process in the following. All other parts of the model remain the same.

No Promotional Effect (NP): $logD_{j,t} = \rho^D logD_{j,t-1} + \gamma r_{j,t-1} + \varepsilon^D_{a,j,t}$

No Promotional Effect and No Rating Effect (NE): $logD_{j,t} = \rho^D logD_{j,t-1} + \varepsilon^D_{a,j,t}$

The impact of rating is expressed as

$$\frac{EV_{NP} - EV_{NE}}{EV_{baseline} - EV_{NE}}.$$

I use this setup and calculate the expected value for each state using the fix-point method. I calculate the expected value using the first observed state of each game to avoid repeat observations. I sum the expected values up respectively and calculate the above ratio. I find that rating effect is responsible for the majority of the utility difference, 83%. In other words, the promotion effect is 17%.

2.7 Conclusion

With the increasing influence that review platforms carry, understanding how the design of rating systems can affect supply-side behavior is important. Specifically, in this paper, I investigate the impact of rating systems on innovation incentives and how to optimize rating aggregation policies for innovations. I find that displayed ratings motivate or demotivate firms from investing in innovation, given the trade-off between innovation costs and potential benefits from a better rating. I conceptualize four key dimensions in motivating innovation: consumer demand, product quality, displayed ratings, and rating agility. Using a structural model, I explicitly model this trade-off and conduct a counterfactual analysis to show that review platforms can motivate more supply-side innovation behavior by weighting recent ratings more or by increasing the return from innovation. The counterfactual analysis shows that giving more weight to recent ratings can significantly increase the rate of innovation.

Several limitations of the current study call for future work. First, I do not observe consumer retention data. Suppose a firm mainly pushes out updates to engage with existing consumers and increase monetization among existing consumers, then I would underestimate the promotional effect of innovation. This concern is mitigated by two factors: the short product life-cycle in the game category as well as the relatively low impact of the promotional effect of demand in Section 6.3. Second, the study context involves product quality varying over time with the presence of observable measures such as updates, which means that the product quality is updated periodically. A parallel context might be renovations at a hotel, new menu options at a restaurant, and so forth. However, in some scenarios, the product quality can change from day to day given the efforts put forth. For example, an auto shop could be extra nice to its customers on some days. This model can be potentially extended to study this context by including variable cost and different supply-side quality decision rules. Third, given the specific design elements of the studied context, I mainly study two counterfactual scenarios: highlighting information as well as rating aggregation

policies. Given the study context, this model can be extended to incorporate more conceptual constructs and study more platform design policies. For example, one important factor I did not consider is advertisements on the firm's side. Advertising can be viewed as a tool for increasing consumer awareness of a product. It will lead to an increase in the number of installs but is unlikely to influence ratings, given the assumption that ratings reveal the true quality of a product relatively accurately. We can use the model to study how much firms choose to advertise and compare the return of advertising to the return of innovation.

2.A Derivation of Equations 5 and 6

The derivation of equation 5 and equation 6 is in the following:

I use R to represents the number of reviews received at the current time period, TR represents the total number of reviews at the (end of) current time period, "$\prime$" represents the next time period. I define rating agility ω as the number of reviews divided by the total number of reviews in the current time period, thus $\omega = \frac{R}{TR}$ and $\omega' = \frac{R'}{TR'}$. Based on the definition, $TR' = TR + R'$.

In historical average rating aggregation algorithm, the rating in the next time period can be written in the following:

$$\begin{aligned} r' &= \frac{r \cdot TR + \tilde{q} \cdot R}{TR'} \\ &= r \cdot \frac{TR' - R'}{TR'} + \tilde{q} \cdot \frac{R'}{TR'} \\ &= r \cdot (1 - \omega') + \tilde{q} \cdot \omega' \end{aligned}$$

To derive equation 11, I make an assumption in the data generation process: $logR' = \rho^R \cdot logR + \delta^R a + \epsilon$.

There is no direct derivation between R and ω given the log form and the necessary existence of ρ^R. However, because of how I defined rating agility, if I compute the correlation between log number of reviews and log rating agility for each game, we can see that these two variables are highly correlated (shown in Table 2.5). It is reasonable to model ω in the same fashion.

Therefore, I model the state transition ω as $log\omega' = \rho^\omega \cdot log\omega + \delta^\omega a + \epsilon^\omega$.

2.B Causal Impact of Displayed Ratings on Consumer Choice

For firms to be incentivized to strategically invest in innovation in response to ratings, a higher rating must carry benefits for firms, such as receiving a correspondingly higher number of installs. I verify that this is indeed the case in my study context. To pin down the causal impact of ratings on consumer choice, the primary endogeneity concern is how game quality potentially influences both ratings and number of installs. In addition, given the version update, game quality might vary from time to time. I apply my analysis on a set of games with multiple regional versions with the same update behavior. This method is parallel to using a difference-in-difference methodology via the use of data from two platforms as seen in Chevalier and Mayzlin (2006). In addition, I also included region fixed effects and time fixed effects to account for regional-level differences and seasonal impacts on the number of installs.

I aggregate data at the weekly level. I select games with multiple regional versions that all update at the same time, allowing me to apply a difference-in-difference identification strategy. Table 2.10 shows the regression results when I control for game quality through a diff-in-diff strategy. Column 1 is the baseline estimation. Given that the game introduction page also displays the total number of installs and total number of reviews, and the fact that consumers might use these two numbers to infer the game quality, I include these two variables in Column 2-4 for control. The results show that each additional star of the rating causally increases the number of installs in the next time period by 2%-7%. In addition, update behavior will increase the next time period's number of installs by 5%-6%. The table replicates the causal impact of ratings on product adoptions documented in the previous literature and verifies firms are incentivized to care about ratings due to rating's impact on sales.

Table 2.10: Linear Regression of Rating on Total Installs

	Log(# of installs at week$_{t+1}$)			
	(1)	(2)	(3)	(4)
Variables				
Log(# of installs at week$_t$)	0.68***	0.67***	0.56***	0.56***
	(0.02)	(0.02)	(0.02)	(0.02)
Displayed rating	0.05***	0.07***	0.02*	0.03*
	(0.02)	(0.02)	(0.01)	(0.01)
Update dummy	0.05***	0.05***	0.06***	0.06***
	(0.02)	(0.02)	(0.01)	(0.01)
Log(# of reviews by week$_t$)		0.06***		0.01
		(0.01)		(0.009)
Log(# of installs by week$_t$)			0.25***	0.25***
			(0.02)	(0.02)
Fit statistics				
Observations	12,082	12,082	12,082	12,082
R^2	0.93	0.94	0.94	0.94

Note: All specifications include game, region and week FE. Cluster-robust standard errors (at the game level) in parentheses.
Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

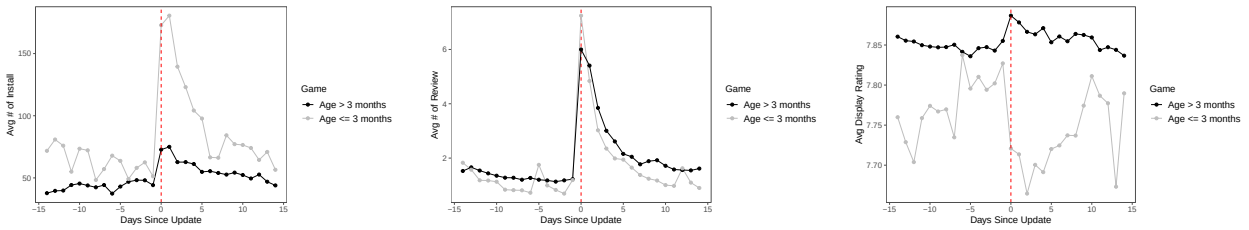


Figure 2.8: Daily Installs, Reviews and Ratings Data Pattern Before and After An Update

2.C Robustness Checks and Heterogeneous Effects

2.C.1 Benefits of Updates

Figure 2.2 shows updates have direct benefits on number of installs, number of reviews, and displayed ratings. Do these data pattern hold for both newly launched games and established games? In Figure 2.8, I show daily installs, reviews and ratings data pattern before and after an update for established games (defined as games that have launched for more than three months), and newly launched games (defined as games launched within three months). The same data patterns hold for number of installs and number of reviews. However, for newly launched games, displayed ratings fluctuate substantially. An update might bring down the displayed rating, highlighting a high chance of failure rate among new games.

2.C.2 Robustness Checks in Quality Transition Process

As a robustness check, I estimate the quality transition process using a sample with more than one weekly number of reviews to reduce potential noises. In addition, I also evaluate the impact of updates on new games. Table 2.11 shows that updates have a robust positive impact on established games while updates bring the average subjective quality down for new games, consistent with the data patterns showing in Figure 2.8.

Table 2.11: Quality Transition Process Robustness Checks

	Dependent variable: Quality			
	(1)	(2)	(3)	(4)
ρ	0.996***	0.986***	0.996***	0.986***
	(0.002)	(0.002)	(0.002)	(0.002)
δ	0.152***	0.066**	0.173***	0.091***
	(0.035)	(0.026)	(0.036)	(0.027)
$\delta \cdot I_{new}$			−0.217**	−0.259***
			(0.104)	(0.078)
Weekly # of Reviews ≥ 2	NO	YES	NO	YES
Observations	33,326	23,371	33,326	23,371
Log Likelihood	−68,105.010	−42,966.270	−68,102.860	−42,960.730

Note: I_{new} is a dummy variable indicating whether the game is newly launched.
Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

2.C.3 Rating Dynamics

A key question is, does the displayed rating change come from first-time reviewers or existing reviewers? In other words, if the new ratings come from existing reviewers, they might comment and rate the update content alone, not the product as a whole. This scenario will tell a different story from how mobile game app developers improve their products to attract new users, since existing reviewers will leave ratings multiple times. To answer this question, I first compute the percentage of ratings coming from existing reviewers. If the same reviewer, identified by the user name, leaves multiple reviews under the same game, I label such reviewers as "repeat reviewers" and all the reviews from such reviewers as "repeat reviews". On average, most reviewers (97%) will only leave a review once. Only 5.5% of the reviews come from reviewers who leave reviews multiple times for the same game.

Figure 2.9 shows how the percentage of repeat reviews changes over the weeks since the initial game launch. The percentage of repeat reviews does not change a lot overtime. Figure 2.10

shows that there is no significant change in terms of the percentage of repeat reviewers before and after an update. The figure indicates that an update does not motivate existing reviewers to leave a review and rate the update itself. The displayed rating changes and the number of rating changes are mainly influenced by the first-time reviewers.

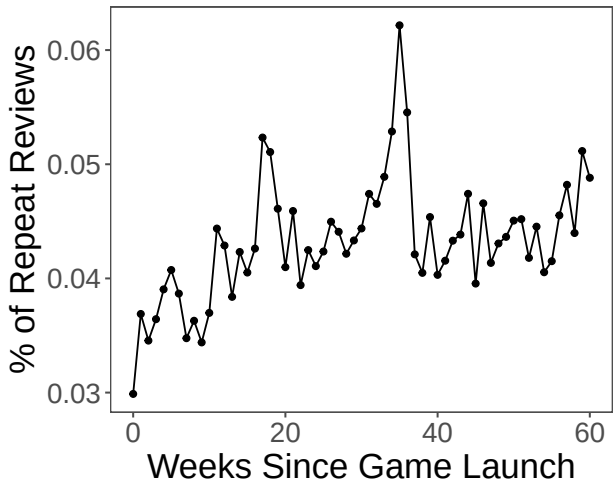


Figure 2.9: Percentage of Repeat Reviews Over Time

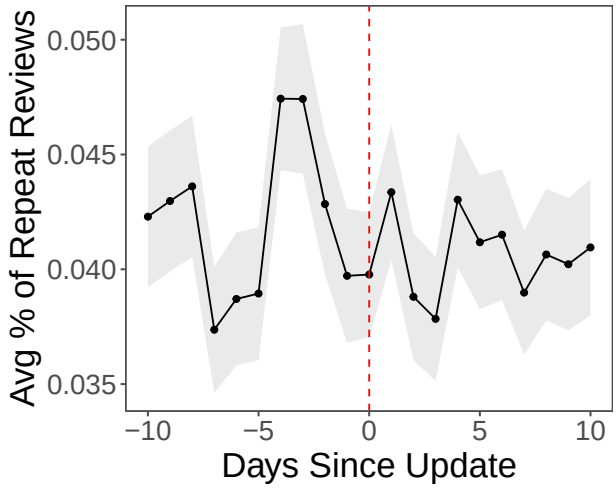


Figure 2.10: Impact of Update on Percentage of Repeat Reviews

2.C.4 Impact of Competition on Innovation

As previous literature indicates (Scherer, 1967; Schumpeter, 1942; Arrow, 1962; Goettler and Gordon, 2011), competition is a primary factor that impacts innovation. In this paper, I abstract away the impact of competition on innovation incentives. To justify this simplified assumption that competition doesn't impact firms' innovation decisions, I conducted the following analysis: I first construct submarkets given the game tags, which indicate the game type and targeted audience. I then show the relationship between the competitiveness in the submarkets and the innovation level.

I utilize the game tag information to construct submarkets. When the firm launches a game, it will use a series of tags to describe the game. The game tags indicate the targeted audience and game type. A game can have multiple game tags. One such example of one particular game is "causal, strategy, puzzle, matching". In this example, the game has four game tags. In total, there are 976 unique game tags, indicating a highly fragmented and competitive market. I do not observe the change of game tags in my data sample, which indicates the games rarely change their targeted audience. I applied k-means clustering methods to cluster the games into submarkets. The goal of the clustering analysis is to discover the submarkets. I use the Silhousettes value to evaluate the clustering performance. Figure 2.11 indicates the market should be split into hundreds of submarkets, another piece of evidence showing that this market is highly fragmented. Given the clustering analysis performance, I choose to cluster the 2270 games into 100 submarkets.

To investigate the relationship between competition and innovation, I aggregate data at the submarket level. I construct HHI to capture competitiveness. HHI is calculated using the market share value of each game in its respective submarket. The market share of each game is derived by its average number of installs divided by the sum of average number of installs from all the games in the same submarket. I measure innovation level by the average update

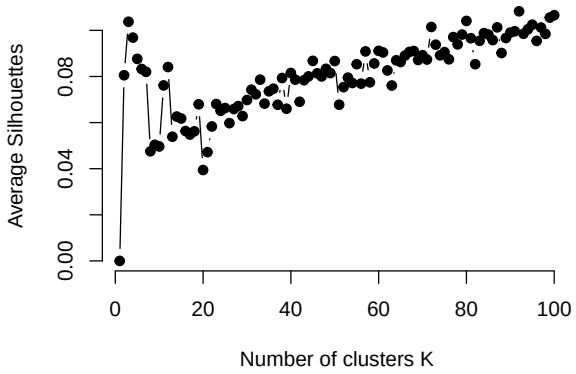


Figure 2.11: Average Silhousttes Value

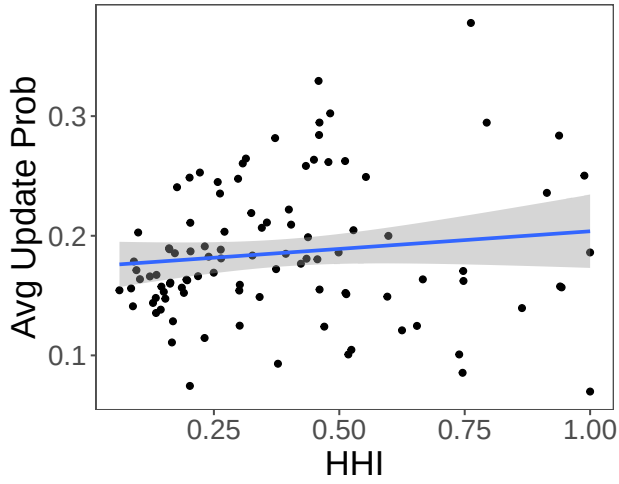


Figure 2.12: Competitiveness v.s. Innovation Level

probability in each submarket, which is the mean of the average update probability of each game in the same submarket.

Figure 2.12 shows the relationship between competitiveness and innovation level with a linear fitted line. The figure indicates that there is no significant relationship between competition level and innovation level.

In conclusion, the mobile game app market is highly fragmented and highly competitive in nature. While competition can influence game firms' decisions, it is unlikely that in such

a setting, the game developers strategically respond to each decision from hundreds, if not thousands, of app developers. Even in the unlikely case that there are close competitors within the submarket, there is no significant relationship between competitiveness and innovation level.

2.D Impact of Review Content on Update Content

In Section 3.3, I looked into the impact of displayed rating on version update choice. While quantitative data is easily accessible, it does not provide us insights into how consumer reviews motivate firm innovation. In this section, I aim to shed some light into how consumer reviews content motivates firm innovation content. I combined a novel NLP-API tool with manual data labeling and then used logit regressions to describe how review content connects to update content. As an initial analysis, I answer the following three exploratory questions: 1. Do more reviews motivate the addition of new content? 2. Do negative reviews motivate firms to fix bugs and balance issues? 3. When consumers complain about micro-transaction markets, how do firms respond?

There are three types of mobile game update: Bug Fix, Additional Content, and Balance Patch. Correspondingly, I define four review content types: Bug Fix, Additional Content, Balance Patch and Money Complaints. The definition of the first three types are consistent with update content classification. The last type is related to the monetization feature of the game, or micro-transaction markets. It is possible that with an update, the developers tune up the monetization features and consumers will complain about how much of a "rip-off" the micro-transaction is in the review content. This "Money Complaint" type also belongs to the "Balance Patch" type, because the game developer could lower the game difficulty, or increase the chance of getting a rarer fragment to avoid consumers complaining about how the game "forces" them to participate in the microtransaction market to get a better game experience. I add this additional type because it directly speaks to the possibility that innovation can lead to subjective quality decrease and looking into this type might shed some light on how microtransaction markets relate to firm innovation and consumer welfare.

To classify review content, I applied a novel NLP-API tool provided by Google to identify frequently mentioned entities, such as "event", "problems", "gameplot", etc, for each review

Figure 2.13: Word Cloud of Entities

and update content. Then I manually categorize whether the entity is closely related to one of the four review types. If it is unclear which update type the entity is closely related to or the entity might be an indication of multiple update types, I discard the entity. I only label the most frequently-mentioned entities. After this manual labeling step, for the reviews in Chinese, I translate Chinese entities to English. The word cloud plot is shown in Figure 2.13.

I selected 52 game IDs which have versions in other regions with the intention to include fixed effects and control for unobservable factors in the future analysis. Among these 52 game IDs, there are 387 version updates in total. 39.5% of the time the version update is about bug fixes, 61.2% of the time the version update contains additional content, and 35.1% of the time the version update content mentions balance improving. The mean and median of the updating period is 29.57 and 24 days, respectively. There are 120540 reviews for these 52 games. 15.65% of the time the review mentioned bug-related entities, 29.85% of the time the review mentioned content-related entities, 22.13% of the time the review mentioned

balance-related entities, and 7.67% of the time the review mentioned money-related entities.

Table 2.12 shows that the number of newly gained reviews is statistically positively correlated with updates that add additional content, but not the two other update types. It indicates that generally when consumers leave more feedback, game developer are more motivated to add content to hold consumer interest. Table 2.13 shows that, when consumers complain about content and balance issues, the game developer is more likely to address these issues in the form of a content update. Negative reviews are defined as 1-star or 2-star reviews. However, the percentage of negative reviews is negatively correlated with bug fix updates and in column 5, the percentage of bug-related negative reviews is not statistically correlated with bug fix updates. It could be that bug fixing is less important than balance improvement and additional content and thus is not highlighted in the update content. Table 2.14 shows that, when consumers complain about micro-transaction markets, the game developer will choose to add new content, instead of improving balance to address those concerns. It is possible that the games with more engaging content will attract more consumers and tend to gate some content behind microtransactions. This result might indicate that even as consumers complain about micro-transaction markets, these additional profits from microtransaction markets will instead motivate game developers to add more content to maintain consumer engagement.

In summary, this section outlines the initial descriptive evidence about how review content is connected to update content. Review content indeed connects to firm's update content and both the number of reviews and microtransaction market can drive firm innovation significantly.

Table 2.12: Impact of # of Reviews on Update Type

	Dependent variable: Update Type Choice		
	Content	Bug	Balance
# of reviews	0.003^{***}	-0.001	0.0002
	(0.001)	(0.001)	(0.0004)
Constant	0.307^{***}	-0.375^{***}	-0.625^{***}
	(0.112)	(0.108)	(0.110)
Observations	387	387	387
Log Likelihood	−250.427	−258.334	−250.811
Akaike Inf. Crit.	504.853	520.668	505.622

Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table 2.13: Impact of Negative Reviews on Update Type

	Dependent variable: Update Type Choice					
	Content	Bug	Balance	Content	Bug	Balance
# of reviews	0.003^{**}	-0.001	0.0003			
	(0.001)	(0.001)	(0.0004)			
% of negative reviews	1.172^{**}	-1.296^{***}	1.481^{***}			
	(0.487)	(0.479)	(0.441)			
% of content-related negative reviews				0.520^{***}		
				(0.161)		
% of bug-related negative reviews					−0.014	
					(0.013)	
% of balance-related negative reviews						0.012^{**}
						(0.006)
Constant	0.048	−0.029	-1.109^{***}	0.163	-0.393^{***}	-0.673^{***}
	(0.180)	(0.174)	(0.190)	(0.113)	(0.107)	(0.110)
Observations	251	251	251	387	387	387
Log Likelihood	−154.486	−163.238	−154.050	−233.300	−259.021	−248.051
Akaike Inf. Crit.	314.971	332.477	314.101	470.601	522.043	500.102

Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table 2.14: Impact of Negative Money-related Reviews on Update Type

	Dependent variable: Update Type Choice		
	Content	Bug	Balance
# of reviews	0.002**	−0.001	0.0003
	(0.001)	(0.001)	(0.0004)
% of money-related negative reviews	28.197**	−5.214	−0.114
	(11.954)	(3.439)	(1.603)
% of negative reviews	0.474	−0.902*	1.495***
	(0.506)	(0.512)	(0.480)
Constant	0.048	−0.038	−1.110***
	(0.181)	(0.173)	(0.190)
Observations	251	251	251
Log Likelihood	−146.002	−161.585	−154.048
Akaike Inf. Crit.	300.003	331.169	316.096

Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

2.E Association between Displayed Ratings and Version Update

In this section, I show the association between displayed ratings and update decision, controlling for as many endogeneity concerns as possible. I first discuss the challenges and potential remedies in identifying the impact of displayed ratings on innovation incentives. Then I discuss the regression results and their implications.

Examining the impact of displayed ratings presents a number of challenges. First, a higher quality firm might lead to a higher rating, and it might have a higher update frequency. To control for firm characteristics, I include game or ID fixed effects.[7] Second, if firms used an

[7]ID fixed effects and game fixed effects differ from one another. Specifically, some games have multiple regional releases, and each regional release has its own ID and receives installs, reviews, and ratings independently. Therefore, some games might have multiple IDs. Games that are released in a single region only

update to improve ratings, then ratings might be affected by the update. For example, if an update contains a lot of bugs, a lot of the reviews may complain about bugs, which would motivate firms to push another update to fix the bug. The argument is that it is not the ratings that affect innovation, but rather the last innovation. However, even if firms intend to use updates to positively influence key metrics in rating systems, it is impossible for firms to perfectly predict what will happen. If firms only use updates to affect key metrics and displayed ratings do not have any impact on the update decision in the next time period, then firm fixed effects and the variable "weeks from last update" should explain all the variations in update behavior. Hence, in the same example, the negative reviews motivate firms to push an update to fix bugs. If firms knew about the existence of bugs, they would be unlikely to release the update in the first place. Lastly, there might be unobserved demand shocks and thus firms might update to address them. In this case, I include control variables to mitigate concerns.

In the regression analysis, I aggregate data at a weekly level. The dependent variable is an update dummy representing whether the firm updates the product in the next week. Table 2.15 shows the regression results when I include fixed effects. Errors are clustered at the ID level, and I include the same set of control variables. Column 1 applies the regression on the full data sample with ID fixed effects. The results show that the displayed rating does not have a significant impact on update probability, while rating agility has a positive impact on update probability. Column 2 applies the regression on the newly launched games with ID fixed effects. The newly launched game sample is defined as observations associated with games within 12 weeks of launch. Current displayed rating has a negative impact on update probability. It indicates that the higher the current displayed rating, the less likely an update will occur. Column 3 applies the regression on the difference-in-difference (DID) sample, consists of games that have multiple regional releases with different update schedules. Current displayed rating has a positive association with update probability. It

have one ID.

means that when developers are working on multiple regional releases, they are more likely to update the game with a relatively higher displayed rating. The table indicates that rating reflects consumer preference. Game developers are more likely to update the game when it is well received by consumers, or when changing the game quality and the displayed rating could provide a more promising future for the game.

Table 2.15: Impacts of Ratings on Update Probability

	Update Probability in Next Week		
	(1) All Games	(2) Newly Launched	(3) DID Sample
Variables			
Displayed rating at week t	0.006	-0.07*	0.02**
	(0.007)	(0.04)	(0.01)
Rating agility	0.09*	0.04	0.07
	(0.05)	(0.08)	(0.37)
Log(# of reviews by week t)	-0.06***	-0.08	0.008
	(0.02)	(0.08)	(0.008)
Log(# of installs at week t)	0.01***	0.04***	0.04
	(0.002)	(0.01)	(0.02)
Log(# of installs by week t)	-0.07***	-0.15***	-0.04
	(0.007)	(0.05)	(0.03)
# of weeks since last update	0.002***	0.04***	0.002
	(0.0003)	(0.004)	(0.002)
Observations	58,952	2,616	1,988
R^2	0.16	0.36	0.13

Note: The first two specifications include game-id FE. The third specification include game and region FE. The samples in the third specification are games with multiple regional releases and with different update schedule. Cluster-robust standard errors (at the game-id level) in parentheses.
Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

In summary, the displayed rating is significantly associated with firms' innovation decisions, but the story is more than "the lower the rating, the more likely the innovation." Ratings serve as a feedback system. The lower the rating for a game, the more incentive firms have to update it, which represents product innovation efforts. However, if the rating is too low for an extended period of time, developers are likely to give up working on the game, which stifles innovation altogether. Displayed ratings continue to serve as a form of feedback but

the results differ depending on the period within the game's lifespan.

Bibliography

Allon, G., Askalidis, G., Berry, R., Immorlica, N., Moon, K., and Singh, A. (2021). When to be agile: Ratings and version updates in mobile apps. *Management Science*.

Amano, T. and Simonov, A. (2022). Gaming or gambling? An empirical investigation of the role of Loot boxes addiction in video games. *working paper*.

Ananthakrishnan, U., Li, B., and Smith, M. (2020). A tangled web: Should online review portals display fraudulent reviews? *Information Systems Research*, forthcoming.

Ananthakrishnan, U. M., Proserpio, D., and Sharma, S. (2019). I hear you: Does quality improve with customer voice?

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors*, pages 609–626. Princeton University Press.

Atkeson, A., Burstein, A., and Chatzikonstantinou, M. (2018). Transitional dynamics in aggregate models of innovative investment.

Bimpikis, K., Papanastasiou, Y., and Zhang, W. (2020). Information provision in two-sided platforms: Optimizing for supply.

Board, S. and Meyer-ter Vehn, M. (2013). Reputation for quality. *Econometrica*, 81(6):2381–2462.

Boudreau, K. J. (2012). Let a thousand flowers bloom? An early look at large numbers of software app developers and patterns of innovation. *Organization Science*, 23(5):1409–1427.

Cabral, L. and Hortacsu, A. (2010). The Dynamics Of Seller Reputation: Evidence From Ebay. *Journal of Industrial Economics*, 58(1):54–78.

Chevalier, J. and Goolsbee, A. (2003). Measuring Prices and Price Competition Online: Amazon.com and BarnesandNoble.com. *Quantitative Marketing and Economics*, 1(2).

Chevalier, J. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43:345–354.

Chiou, L. and Tucker, C. (2018). Fake news and advertising on social media: A study of the anti-vaccination movement.

Comino, S., Manenti, F. M., and Mariuzzo, F. (2019). Updates management in mobile applications: iTunes versus Google Play. *Journal of Economics & Management Strategy*, 28(3):392–419.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management science*, 52(10):1577–1593.

Einav, L., Farronato, C., and Levin, J. (2016). Peer-to-peer markets. *Annual Review of Economics*, 8(1):615–635.

Ershov, D. (2018). Competing with superstars in the mobile app market.

Glazer, J., Herrera, H., and Perry, M. (2020). Fake reviews. *The Economic Journal*.

Godes, D. and Silva, J. C. (2012). Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3):448–473.

Goettler, R. and Gordon, B. (2014). Competition and product innovation in dynamic oligopoly. *Quantitative Marketing and Economics*, 12:1–42.

Goettler, R. L. and Gordon, B. R. (2011). Does AMD spur Intel to innovate more? *Journal of Political Economy*, 119(6):1141–1200.

Gordon, B., Jerath, K., Katona, Z., Narayanan, S., Shin, J., and Wilbur, K. (2021). Inefficiencies in digital advertising markets. *Journal of Marketing*, 85(1):7–25.

Harbaugh, R. and Rasmusen, E. (2018). Coarse grades: Informing the public by withholding information. *American Economic Journal: Microeconomics*, 10(1):210–35.

Hauser, J., Tellis, G. J., and Griffin, A. (2006). Research on innovation: A review and agenda for Marketing Science. *Marketing science*, 25(6):687–717.

He, S. and Hollenbeck, B. (2020). Sales and rank on amazon.com.

Hollenbeck, B. (2018). Online reputation mechanisms and the decreasing value of chain affiliation. *Journal of Marketing Research*, 55(5):636–654.

Hollenbeck, B., Moorthy, S., and Proserpio, D. (2019). Advertising strategy in the presence of reviews: An empirical analysis. *Marketing Science*, 38(5):793–811.

Horner, J. and Lambert, N. S. (2016). Motivational ratings.

Hui, X., Jin, G. Z., and Liu, M. (2022). Designing quality certificates: Insights from eBay.

Hunter, M. (2020). Chasing stars: Firms' strategic responses to online consumer ratings.

Imai, S., Jain, N., and Ching, A. (2009). Bayesian estimation of dynamic discrete choice models. *Econometrica*, 77(6):1865–1899.

Jin, G., Lee, J., Luca, M., et al. (2018). Aggregation of consumer ratings: An application to Yelp.com. *Quantitative Marketing and Economics*, 16(3):289–339.

Levin, R. C., Klevorick, A. K., Nelson, R. R., Winter, S. G., Gilbert, R., and Griliches, Z. (1987). Appropriating the returns from industrial research and development. *Brookings papers on economic activity*, 1987(3):783–831.

Lewis, G. and Zervas, G. (2016). The welfare impact of consumer reviews: A case study of the hotel industry.

Leyden, B. T. (2020). Platform design and innovation incentives: Evidence from the product ratings system on apple's app store.

Li, X., Bresnahan, T. F., and Yin, P.-L. (2016). Paying incumbents and customers to enter an industry: Buying downloads.

Li, X. and Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Mayzlin, D., Y., D., and Chevalier, J. (2014). Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *The American Economic Review*, 104:2421–2455.

Milgrom, P. and Roberts, J. (1986). Prices and Advertising Signals of Product Quality. *Journal of Political Economy*, 94:297–310.

Nelson, P. (1970). Information and consumer behavior. *Journal of political economy*, 78(2):311–329.

Nosko, C. and Tadelis, S. (2015a). The limits of reputation in platform markets: An empirical analysis and field experiment.

Nosko, C. and Tadelis, S. (2015b). The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment. *NBER Working Paper 20830*.

Proserpio, D. and Zervas, G. (2016). Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews. *Marketing Science*.

Rao, A. (2018). Deceptive claims using fake news marketing: The impact on consumers.

Rao, A. and Wang, E. (2017). Demand for "healthy" products: False claims and ftc regulation. *Journal of Marketing Research*, 54.

Rhodes, A. and Wilson, C. M. (2018). False advertising. *The RAND Journal of Economics*, 49(2):348–369.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, 55(5):999–1033.

Scherer, F. M. (1967). Market structure and the employment of scientists and engineers. *The American Economic Review*, 57(3):524–531.

Schumpeter, J. (1942). Capitalism, socialism and democracy.

Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8(1):321–340.

Weintraub, G. Y., Benkard, C. L., and Van Roy, B. (2008). Markov perfect industry dynamics with many firms. *Econometrica*, 76(6):1375–1411.

Wen, W. and Zhu, F. (2019). Threat of platform-owner entry and complementor responses: Evidence from the mobile app market. *Strategic Management Journal*, 40(9):1336–1367.

Wilbur, K. and Zhu, Y. (2009). Click fraud. *Marketing Science*, 28(2):293–308.

Wu, Y. and Geylani, T. (2020). Regulating deceptive advertising: False claims and skeptical consumers. *Marketing Science*.

Yasui, Y. (2020). Controlling fake reviews.