THE UNIVERSITY OF CHICAGO

THE ROLE OF SALES INFORMATION IN ONLINE CONSUMER SEARCH

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

KENNETH C. GRIFFIN DEPARTMENT OF ECONOMICS

BY

FANG FU

CHICAGO, ILLINOIS

JUNE 2020

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am deeply grateful to my advisors Ali Hortaçsu, Pradeep Chintagunta, Michael Dinerstein, Jean-Pierre Dubé, and Anita Rao for their support and advice. I want to thank the participants of the University of Chicago IO and Marketing working groups for their valuable comments. I also appreciate the helpful suggestions from Yuehao Bai, Stéphane Bonhomme, Juanna Schrøter Joensen, Kyeongbae Kim, Karthik Nagarajan, and Wenji Xu.

ABSTRACT

To help consumers make choices from a large number of alternatives, e-commerce platforms provide them with information about historical sales. However, the effects of disclosing sales information on consumers' search decisions and welfare are unclear. Furthermore, revealing sales information can make popular products more attractive to subsequent consumers due to the revealed popularity. This positive feedback loop leads to concerns about the fairness of competition in online marketplaces when some sellers are able to influence the short-term popularity of their products, particularly if a seller is also the marketplace operator. I use novel click-stream data on consumers that search for specific books on a large online book-selling platform and find that popular products attract more clicks. Conditional on clicking, sales information does not affect consumers' purchasing decisions significantly. This finding confirms that sales information serves as a proxy signal for product-page attributes that consumers do not see without a costly search of the product. In counterfactual experiments, I first show that disclosing sales information has uncertain effects on consumer welfare. Making the initial group of consumers search under random rankings benefits subsequent consumers by disclosing sales information. Second, I find that the initial position of a product affects its sales volume in subsequent periods when sales information is disclosed. This confirms that the feedback loop exists if sales information is available. Then, I specifically focus on products sold by the first-party seller and find that compared to ranking products by their mean utilities, assigning these products top positions in the initial periods leads to persistently more sales for the first-party seller even though products are fairly ranked by sales volume after the initial periods. These findings provide platforms with managerial tools to disclose sales information that benefits consumers. The findings also give insights into the recent concerns about the fairness of competition on e-commerce platforms between first-party and third-party sellers.

CHAPTER 1

INTRODUCTION

In this paper, I study the effect of disclosing historical sales information on consumers' search decisions when they search across vertically differentiated products with incomplete information about their qualities (à la Stigler 1961). I further explore how the availability of sales information affects consumer welfare and potentially leads to a positive feedback loop that asymmetrically benefits popular products due to the disclosed popularity. The Internet provides consumers with more choices of products than ever, yet the research has shown that too many alternatives can make search decisions difficult when the search is costly (Kuksov and Villas-Boas 2010). Thus, as intermediaries, e-commerce platforms have developed various tools to make consumers search more efficiently. A leading example of such tools is information about sales volume. This metric is not only widely used to determine default rankings (e.g., Amazon, JD.com) but it is information that is also directly available to consumers since it can be easily generated by platforms and understood by consumers.[1] For example, eBay selectively discloses the sales volume of certain products. Amazon, instead, allows consumers to check the sales rank within each category. However, the effects of disclosing sales information on consumers' search decisions and welfare are not well understood. For consumers, they gain additional information to infer product quality that they do not observe before clicking on the products. But this information may also make consumers fail to choose the best product (Salganik et al. 2006) because the information cascades in such a way that they may copy the choices of earlier consumers that can lead to further inefficient outcomes (Banerjee 1992; Bikhchandani et al. 1992; Smith and Sørensen 2000; Zhang 2010). Thus, whether the disclosure of sales infor-

1. On platforms that sales information is not directly disclosed, consumers are still likely to infer the popularity of products from the default ranking. In this case, the position of a product on the list also contains sales information that affects a consumer's belief over product quality (Athey and Ellision 2010).

mation benefits consumers or not is unclear. Furthermore, disclosing sales information can asymmetrically benefits high-selling products since they are more attractive to consumers due to their revealed popularity. Also, popular products are usually assigned prominent positions that further leads to lower search costs (Chen and Yao 2017; Ursu 2018). This positive feedback loop potentially allows short-run popularity of a product to persist in the long run that leads to concerns about the fairness of competition in the online market when some sellers can influence the temporary popularity of their products. Especially, when a platform is both the marketplace operator and the first-party seller, it has various exclusive methods to temporarily influence the popularity of its products, for example, by assigning them top positions in sponsored ads as shown in Figure 1.1. In this manner, they can gain long-run advantages in competition against other sellers even if the ranking algorithm fairly ranks all products using criteria such as the popularity.[2] However, the effect of such a feedback mechanism is also not obvious since the search costs are generally lower in online marketplaces. These lower costs lead to the lower probability of "bad herding" in which subsequent consumers fail to search for high-quality products (Hendricks et al. 2012).

I first show that sales information only affects consumer's clicking decisions but not purchasing decisions conditional on clicking. Based on this finding, I then develop and estimate a structural model that provides a general framework to study the effect of disclosing sales information[3] on both consumer welfare and the existence of a feedback loop. My model assumes each consumer sequentially searches across vertically differentiated

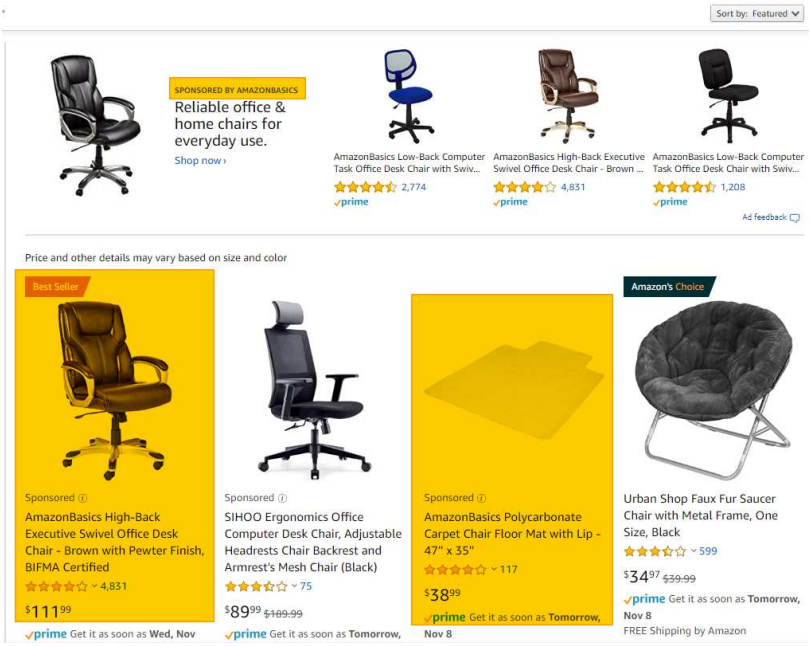
2. There are recent concerns about how Amazon tries to make itself more profitable by listing its own brand in prominent positions (Wall Street Journal, Sept 16, 2019). Also, Amazon has removed the option that allowed consumers to sort products by relevance in the search results. Later, Amazon denied these conjectures and claimed all products, including the private-brand products, were fairly listed under an algorithm that relied on their popularity (Wall Street Journal, Oct 24, 2019).

3. Sales information could be sales volume, sales rank, market share, and other metrics. In this paper, I specifically study the case of sales volume. However, the framework I provided can be used to study other cases as well.

products (see Weitzman 1979) and infers unrevealed product quality using disclosed sales information. In the counterfactuals, I first show that the disclosure of sales information has uncertain effects on consumer welfare that depends on whether the sales information correctly reflects the overall quality of each product. Randomly ranking products in the initial trail period and disclosing sales information to subsequent consumers leads to higher consumer welfare. Second, I find that in general, initial positions affect sales in subsequent periods through the initial popularity, and the persistence of this effect varies with the search cost, the number of consumers, the degree of the initial popularity advantage, as well as the credibility of popularity information. This finding confirms the existence of the feedback loop. In addition, a more salient feedback loop of the popularity comes along with a larger search cost, more consumers shopping for the products per period, a more significant initial popularity advantage, and more credible sales information. Third, I specifically investigate whether a first-party seller can achieve persistently more sales in competition with other sellers by influencing the short-run popularity of its products, such as manipulating their initial positions. I find that initially placing products sold by the first-party seller in top positions and fairly ranking all products by sales volume afterward leads to more sales of its products in the long run compared to a utility-based ranking (see Ghose, Ipeirotis, et al. (2012) and Ghose, Ipeirotis, et al. (2014)), which fairly ranks all products by their mean utilities. These results affect most consumers who shop online and provide insights into the recent concerns about the fairness of competition on e-commerce platforms in which there are both first-party and with third-party sellers competing in the marketplace.

I begin by studying how consumers use the disclosed sales information of products in their online shopping decisions. The failure to consider sales information may lead to a biased estimate of the search cost since default rankings are usually correlated with previous sales records, and consumers usually perceive this correlation. I use a novel data

Figure 1.1. Private-Brand Products in Sponsored Positions



set that contains consumers' individual-level clicks and purchases on a large e-commerce platform in China that is known for its bookselling. In addition, this platform not only has third-party sellers listing their products but also sells similar products as the first-party seller. Using this data, I analyze how sales information affects consumers in both clicking and purchasing decisions. The reduced form analysis finds that the sales volume primarily affects consumers' decisions in the clicking stage, but not their choices in the purchasing stage that are conditional on clicking. This finding indicates that sales information is a signal that consumers use to infer the unrevealed product characteristics that they do not observe without clicking on the product. Consumers are more likely to click on high-selling products in the search results, since they expect the mean utility of those products to be positively correlated with their popularity. However, among clicked products, consumers' conditional purchasing decisions no longer depend on sales information. Therefore, sales information is not part of the consumer's utility. Instead, it exclusively serves as a signal for the unrevealed characteristics that consumers costly search for.

I then estimate a structural model that incorporates sales information into the consumers' sequential-search decisions and quantifies the value of sales information to different sellers. The model is related to Gu (2016)'s that assumes consumers have empirical expectations about the unrevealed product characteristics on the product pages after observing characteristics in the search-result pages. My model also includes a learning mechanism that allows consumers to infer the characteristics on the product page from the disclosed sales information. In the model, consumers observe some product characteristics, the product position, and sales information in the search results without any cost. Using the observed information, consumers infer additional characteristics on the product page, which they observe by costly searching that page. In this setting, sales information is a signal for those characteristics unrevealed in the search results and helps consumers make clicking decisions. Thus, sales information only enters the clicking decision by affecting the consumer's expected utility before sampling the product. The estimation result shows consumers respond actively to the product characteristics they observe. Further, the value of sales information to each product depends on its position and market share. I find that disclosing sales information benefits the higher selling products and harms products with low market shares.

Finally, the estimates of the structural model facilitate the study of how the availability of sales information affects the welfare of consumers and the fairness of competition in e-commerce platforms when some sellers are able to influence the short-run popularity of products. Using counterfactual experiments, I first find that the disclosure of sales information has uncertain effects on consumer welfare. When sales information of products reflects well their true quality (i.e., mean utilities), consumers benefit from observing these signals. Otherwise, consumers are more likely to make more clicks and choose inferior products. I show that by randomly ranking products for the initial group of consumers, disclosing sales information to subsequent consumers leads to higher consumer welfare. Then, I find

that initial positions do affect the sales volume of products in subsequent periods through their initial popularity if sales information is disclosed. In particular, initially lower ranked products have lower sales volume on average in each of subsequent periods. This effect drops to a stable level in the long run, which is after around 35 days based on my counterfactual settings. This finding confirms the existence of the feedback loop. Furthermore, the persistence of the popularity under the feedback loop depends on the degree of correlation between sales information and the overall product quality, as well as the portion of initial popularity advantage in the disclosed cumulative sales information in the later periods. In addition, I specifically focus on the private-label products and study whether assigning them initially prominent positions has a sustained effect on their sales. In my data, the first-party seller (i.e., the platform) competes with other sellers by selling similar books, so I use products sold by the first-party seller as the private-label products in my analysis. In initial periods, I place the first-party seller's products in top positions, and the others are ranked by their mean utilities. Later, I rank all products by their sales volume (i.e., similar to ranking by the popularity as claimed by platforms such as Amazon). Using a utility-based ranking as the benchmark ranking method, the results indicate that consumers are worse off in the short run under the ranking method proposed above, while their welfare converges to the outcome under the benchmark ranking. Meanwhile, the first-party seller gains persistently more sales in the long run compared to the utility-based ranking.

This paper makes two contributions to the empirical search literature. First, in this paper, I shows how disclosed sales information affects consumers' search decisions. This further provides a general framework to study the role of sales information in empirical search settings by incorporating a learning mechanism into consumers' sequential search decisions. Hendricks et al. (2012) provide a theoretical analysis of observational learning in a different search setting such that consumers decide whether to search each product in the list independently (i.e., no substitution effect across products) with two types of quality.

I, instead, study the effect of disclosing sales information in an empirical setting using a sequential search model, such that every consumer can choose at most one item from a list of vertically differentiated products. Second, this paper provides some insights into the recent concerns about the fairness of competition on e-commerce platforms with both first-party and third-party sellers by studying the existence and properties of a feedback loop. Especially, how platforms with the ability to temporarily influence the popularity of products can potentially bring advantages to their own products when competing with other sellers, even if products are "fairly" rank by their popularity.

The structure of the paper is as follows. In Chapter 2, I discuss the related literature. Chapter 3 presents the data and some background information about the online book market. Chapter 4 provides the reduced form evidence of the effect of sales information on consumers' search and purchasing decisions. Chapter 5 introduces a structural model that incorporates the sales information into the traditional sequential search model. Chapter 6 presents the estimation strategy, and the estimation results are shown in Chapter 7. Then, I show counterfactual experiments in Chapter 8 that is followed by the conclusion.

CHAPTER 2

LITERATURE

This paper mainly relates to the empirical search literature. There are various settings for what consumers search for under the framework of sequential search Weitzman (1979). For example, a consumer can search for prices (Honka and Chintagunta 2017), or an *iid* utility shock (i.e., a match value) that is not observed without costly searching the product (Kim et al. 2010; Kim et al. 2017). The setting of this paper is close to (Gu 2016) such that consumers search for not only match values but also some additional product characteristics in the product page. However, I do not assume consumers have identical rational expectations about product-page characteristics given characteristics that they observe in search results. Instead, I model consumers as inferring those unrevealed product-page characteristics given both sales information and revealed characteristics using the Bayesian framework. The identification strategy of parameters is similar to Chen and Yao (2017), Kim et al. (2010), Kim et al. (2017), and Ursu (2018). For the price endogeneity issue, I use an approach similar to De los Santos and Koulayev 2017.

The literature proposes various data sets such as hotels (Ghose, Ipeirotis, et al. 2012; Ghose, Ipeirotis, et al. 2014) and electronics (Kim et al. 2010; Kim et al. 2017). The book shopping data in De Los Santos et al. (2012) is the closest one to ours. In their paper, each consumer searches across different platforms for a specific book, while I study consumers that shop across different sellers in one platform for the same book.

One of my findings about the role of sales information provides some insights into the position effect in the consumer search problem. Ursu (2018) uses an unique data set to confirm that the position affects the search cost. Based on my findings, if consumers perceive the default ranking as having a correlation with the historical sales performance of

products, the position also likely affects the expected utility (Athey and Ellision 2010)when the sales information is not available to consumers or/and econometricians.

In addition to the empirical search papers, this paper also relates to the observational learning literature (Banerjee 1992; Bikhchandani et al. 1992; Smith and Sørensen 2000). Salganik et al. (2006) study the effect of disclosing sales information on the market structure in an experimental setting in which they find that the sales performance of products does not just depend on their qualities when consumers search with imperfect information also observe aggregate histories of past choices. Some previous papers (Hendricks et al. 2012; Mueller-Frank and Pai 2016) study cases that consumers use publicly available information to make search decisions and derive some important theoretical results. My model also incorporates the idea of inferring unrevealed product characteristics from sales information but by using a sequential search model to fit an empirical setting.

CHAPTER 3

DATA

3.1 Data description

The data contain consumers who search on a Chinese online e-commerce platform. The e-commerce industry usually recognizes this platform as one of the largest bookselling platforms in China with around one-third of the market share in bookselling in 2017 with sales were around 40 billion RMB as reported by the platform.[1] Besides books, there are various other categories of products sold on this platform, but I specifically focus on consumers who search for books. Studying book searches has a couple of advantages as compared to other products. First, books have a high conversion rate due to their relatively low prices in China that creates more purchase records. In my data, 6.7% of clicks and 10.34% of search sessions end up with a purchase. Second, the chance of repeated purchases of the same book is smaller than other products, which alleviates the concern that consumers know the products and making a purchase without a search.

The data consist of three parts. The first part is the anonymous url-level click-stream data on the click and browsing histories of consumers on the platform from August 13, 2018 to October 7, 2018. The click-stream data includes detailed information of search queries[2], which includes keywords and refinements, clicking and purchasing with the corresponding product IDs and time stamps, and unique identifiers of consumers who perform each action.

1. The official data on *The Yearbook of Publication in China (*2017*)* indicates that the total book sales in China were 85 billion RMB in 2017. Some third-party sources show the platform owns around 35% of the market share.

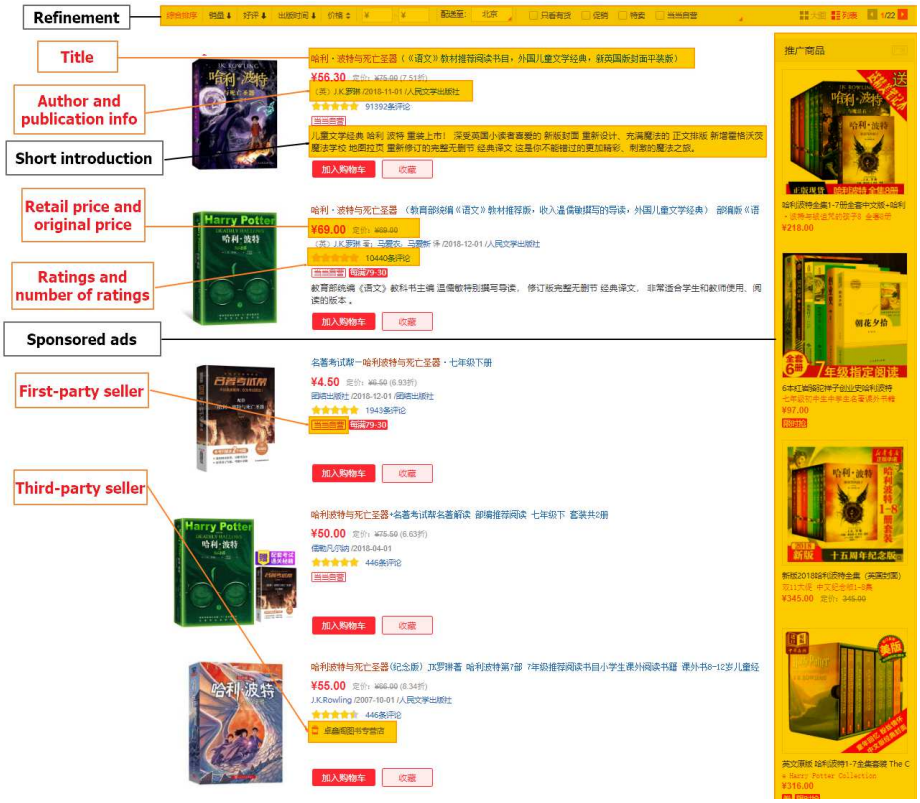
2. A search query is defined as one consumer searching by entering a keyword. In this paper, I use search queries and search sessions interchangeably. I merge two search sessions by the same consumer if they share the same search parameters (e.g., keywords and refinements) and the searches occur closely together.

Since the data is anonymous, I have no demographic information except the type of device that the consumers used, i.e., a PC or a mobile device.

In addition to the click-stream data, I collected the search-result page data by using search-query parameters (e.g., keywords and refinements) that I observed in the click-stream data set. A consumer sees the search results after entering the keywords in the search bar (Figure 3.1 is an example of a page of search results). This data set contains product characteristics in search results, and the position of each product. Some search sessions contain hundreds of products listed in the results, but it is not realistic for consumers to be aware of them all. Defining consumers' choice sets differently can influence the empirical results significantly (Honka, Hortaçsu, et al. 2017). Ideally, knowing the consumer awareness of listed products could solve this issue. Since such data isn't available, I assume consumers are aware of the products up to the 60^{th} in the list. Consumers are less likely to check products after the 60^{th}, as a typical 24-inch screen only shows around six products and even less on a mobile device. I collected these data within 24 hours after consumers do a search. To check the potential variation in the position and availability of products in this period, I track the page layouts of $3,785$ search sessions across a 24-hours period and find no change in the ranks in all search sessions.

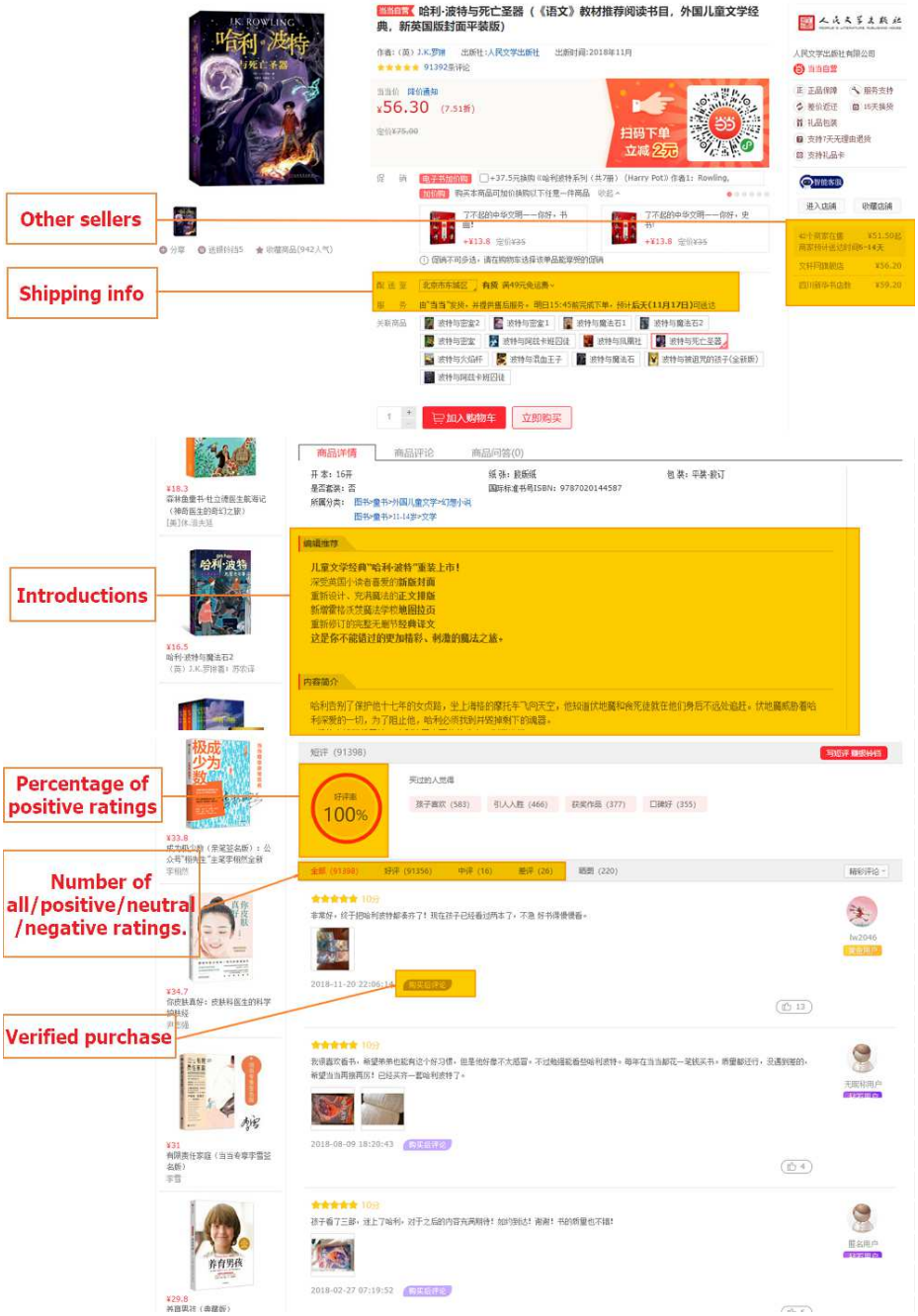
The third data set is the information on the product page of each product in addition to the characteristics in the search results. These characteristics can include the shipping cost, the introductions, details of consumers' reviews, etc. They are not available to consumers without clicking the product. Figure 3.2 shows a typical product page that consumers observe after clicking on it. Further, in the search results, consumers observe the summarized star rating, but not individual-level ratings. In the product page, they can observe not only the content of each review but also the number of positive, negative, and neutral ratings of the product. Due to the special rating system, some of the rating information is not informative, but others are.

Figure 3.1. Search-Result Page



Keyword: *Harry Potter and the Deathly Hallows*. Names in red color are included in my analysis.

Figure 3.2. Product Page



3.2 The rating system of the platform

Both consumers and econometricians do not observe the sales volume directly. Instead, I construct a proxy for sales volume from the number of ratings. There are two potential issues with assuming consumers perceive the number of ratings as a proxy for sales volume. First, the number of buyers who rate their purchases is usually different from the actual sales. Second, the number of ratings also affects the consumer's belief about the overall quality of products in conjunction with the rating valence (Etzion and Awad 2007). For example, consumers are more confident about the reliability of the rating with 1000 ratings than the one with only 10 ratings.

To address the first issue, I rely on the unique rating system of the platform. In this platform, there are two types of ratings for each product: default ratings and non-default. Different from other platforms, in which many buyers do not rate their purchases, this platform automatically assigns a default five-star rating to the product if a buyer does not rate it within ten days after the purchase. In my data, there are three types of rating-related information that consumers observe during their search process: the number of ratings, the number of stars, and the number of the tine (positive, neutral, or negative) of ratings. Some of them only contain the non-default ratings, while others contain both. Table 3.2 summarizes the components of different rating-related variables. Since the number of ratings includes both default and non-default ratings, it ensures that the number of ratings is no less than the actual purchases. I cannot entirely rule out the possibility that the ratings are by consumers who did not purchase the product or are fake, as all registered consumers of the platform can rate any book on the platform. So, the actual sales volume is less than or equal to the number of ratings. By studying the review data, I find that around 95.92% of ratings are rated by verified purchasers. So, I believe the number of ratings should be a good proxy for the actual sales on this platform, although it may contain certain "noise".

Table 3.1: Rating System of the Platform

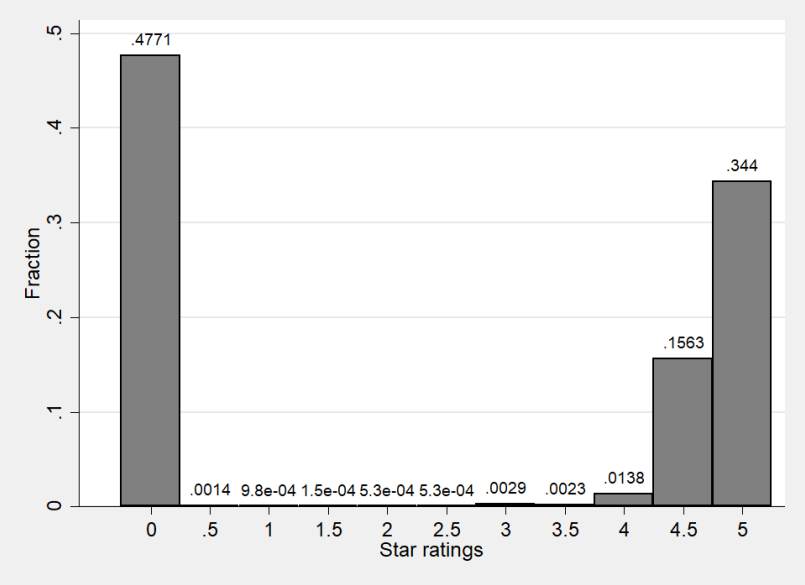
	Buyer-written	System-default
Search-result page		
Star ratings	$\surd$	
Number of ratings	$\surd$	$\surd$
Product page		
Number of positive ratings	$\surd$	$\surd$
Number of non-positive ratings	$\surd$	

Buyers can rate products within ten days of purchases. Otherwise, the system will assign a default five-star rating after ten days.

The second issue is alleviated by the fact that there is very little variation in the star ratings. The overall star rating only summarizes those non-default ratings from purchasers. However, according to some consumers, the platform intensively censors the negative reviews. This is an extraordinarily small portion - less than 0.1% on average - as negative ratings leads to very high overall star ratings. Figure 3.3 shows that in the distribution of star ratings, which has very little variations, i.e., more than 99% of the products with ratings are above four stars. This variation in the data is much less than what Chevalier and Mayzlin (2006) found on Amazon.com, where the fraction of four- and five-star reviews is 73% (86.5% on Barnes&Nobel). When consumers observe almost all products in the search results have more than 4.5 stars, it is doubtful that the star rating itself is informative to the consumer. So, I assume the effect of the number of ratings on the valence of star ratings is negligible. Due to these two features of the rating system, I argue that assuming consumers perceive the number of ratings as a signal of sales volume only is reasonable.

While the summarized star rating in the search-result pages is not informative, consumers can gain useful information from the valence of each rating, which consumers observe in the product pages. After clicking on the product pages, consumers observe both the number of positive/neutral/negative ratings and user comments. These pieces of information contain both default and non-default ratings, so the number of positive ratings is

Figure 3.3. Distribution of Star Ratings



Note, over 99% of ratings are above four stars potentially due to the censorship of the negative reviews.

not informative. However, the negative reviews will affect consumer's decisions, as they are written by buyers instead of being generated by the system. Furthermore, since the platform is likely censoring negative reviews, the negative reviews are more likely to have a stronger effect than on the other platforms.

3.3 Two types of sellers

Similar to Amazon.com, on this platform, there are both a first-party seller (i.e., sold and shipped by the platform) and third-party sellers. For books sold and shipped by the platform, the profits of the platform come from the difference between the price the platform pays to publishers (or intermediaries) and the retail price.[3] For the third-party sellers, the

3. Retailers usually buy books at around $60\% - 70\%$ of the original price set by publishers. Large retailers with bargaining powers may buy books from publishers directly in lower discounts. Smaller retailers usually buy books from intermediaries at higher prices.

platform receives a service fee, which consists of a fixed portion plus a commission proportional to the actual sales. The fixed part ranges from $6,000$ RMB to $30,000$ RMB per year (around 850 USD to 4200 USD) that depends on the category of products, and the books have the highest. The commission on bookselling varies across different sellers and is determined through the bargain between each seller and the platform. A reasonable approximation is around 5%, which is based on the commission of other products on the platform ($1.5\% \sim 5\%$). Usually, the platform gains higher profits if the same book is sold by the first-party seller rather than by third-party sellers. Appendix A.5 provides some evidence that the platform may be influencing the sales of products sold by itself.

Different from other platforms such as eBay and Taobao.com, all third-party sellers on this platform are required to own business licenses from the government that means they either own physical stores or companies. This requirement leads sellers on this platform to be relatively more homogeneous in terms of service qualities, such as the authenticity of products and return process. Furthermore, the platform requires sellers to make a large deposit. Similar to eBay's money-back guarantee, the platform will refund the buyer directly from the deposit in case of a dispute if a seller does not fulfill the consumer's request or cannot provide evidence in their favor. One main difference across sellers is the shipping methods. The platform has its own logistics service, while third-party sellers usually use commercial delivery services. Their shipping costs may be different due to their locations and contracts with carriers, but the overall delivery speed should be close.[4] So, I believe the seller service qualities are homogeneous, and sellers are differentiated in the shipping costs and designs of product pages. Appendix A.2 provides more information about third-party sellers' services and ratings.

4. The delivery usually takes less than two days from Shanghai to Beijing that is around 750 miles.

3.4 Limitation of data

Although books have advantages such as high conversion rates and less likely to be repeatedly purchased, there are a couple of limitations to using books in this study. First, books are heterogeneous in their content, and I do not have information about the quality of the book's content. For example, a consumer searches for fantasy novels may bring up *The Lord of the Rings* and *Happy Potter*, but I cannot characterize and compare the content quality of these two books. In particular, most of consumers in the data search for different keywords. One possible solution is to have a book-content or ISBN fixed effect. But I do not have any identifier for the content of books, and it is infeasible to identify the content of so many books in the data manually. Further, since consumers search different books in the data, some ISBNs[5] appear only once in the data, that makes it is not feasible to include the book fixed effect in the analysis. If I exclude searches in which books exist only once, the sample size would be much smaller.

To avoid this issue, I only focus on consumers who each searches for a specific book in terms of the contents. And, these consumers search for different versions (older or newer editions, international edition, hard cover, etc) across different sellers. To rule out consumers who search across books with different contents, I perform the following cleaning process. First, I drop all search sessions with no clicks, as the click-stream data also records partial keyword before consumers finish typing. Next, I check the score of keyword-title string similarity of the clicked products. If a consumer clicked a product with zero similarity between the keyword and its title, I assume he or she searches across different books, then I drops that search session. For the remaining searches, if a consumer clicked more than one product, I check if the clicked products share the same title, author, or ISBN. This criterion further rules out the possibility that consumers search across different books. For

5. Every book has a unique ISBN. Different editions of the same book have different ISBNs.

example, both *Microeconomics Theory* by M.W.G. and *Advanced Microeconomics Theory* by J.R. have non-zero similarity with the keyword "microeconomics theory". But the consumers who clicked both should be considered as a non-specific searcher, and this criterion ensures they are excluded from my analysis. This rule cannot be applied to consumers who clicked only one product. So, instead, I check if there is more than one product in the search results that share the same ISBN or title to the product that the consumer clicked. If this criterion is not met, I exclude the search session from my analysis.

These filtering criteria may not completely rule out consumers who search across different books. However, they do significantly reduce the chance of non-specific searches, and the remaining consumers are looking for the same book across different editions or sellers. Also, consumers in my data are more likely to search for specific book content since the data I collected contains the period at the beginning of the school year. Figure 3.4 shows around one-third of the books that consumers search for are textbooks or reference books. These consumers are less likely to search across different books, as textbooks and reference books are usually assigned by schools.

Another limitation is that platforms usually over-list products with titles that only have little similarity with the keywords, so some books listed in a consumer's search results do not have the content that the consumer searched for. These books would not be in the search list[6] that consumer would potentially click and search. For example, a consumer enters keyword "microeconomic theory" may be specifically looking for *Microeconomic Theory* by M.W.G. , but the platform is likely to list other microeconomic textbooks that share similarity with the title or topic. To deal with this issue, I assume a consumer's search list only includes books that either have the same title, or the same ISBN as the books that he or she clicked or purchased. The reason my criteria includes both ISBN and the title

6. Throughout this paper, I define the search list in the same way as I describe here.

is that sometimes the title of a product contains extra information, such as the author's name or the edition in addition to the name of the book, that makes the product titles of the same book (ISBN) sold by two sellers to be different. Also, different editions of a book have different ISBNs, but they both have the content that the consumer searched for. So, I include these books to avoid over excluding products.

Table 3.2 shows that the average number of products in consumer's search list is around 4, and there are only 1.5 different books after the data cleaning processes. I find that in most search sessions (74%), there is only one ISBN, that is, multiple sellers sell the same book. So, I think the limitations stated above are alleviated.

Due to the selection of search sessions, I do not try to claim that the findings in this paper fit all consumer search cases. One reason is that I focus on consumers who only search across different sellers, editions, and bundles, but not different book contents. Consumers who shop across different book contents may be uncertain about the quality of the book content after browsing the product pages when books are experience goods (Nelson 1970). So, previous consumers' choices may affect their beliefs about these uncertain qualities that further affects the purchasing decisions among clicked books. An observational learning model may fit this case better.

The third limitation is that consumers may click a product directly from another product page instead of the search-result page, but I cannot identify where consumers click from. I identify these potential product-page clicks by comparing the positions of two consecutive clicks and making adjustments to their positions. If the difference between the positions of two successive clicks is too large, the second click is likely made from the first product page. Then I adjust the position of the second product to the position of the first product plus one.

Table 3.3 summarizes the consumer's utility-related characteristics in the data set after the cleaning process. The first panel of the table shows the characteristics that consumers

observe in the search results free of cost. These product characteristics include price, ratings, first-party seller icon, the suggested price set by the publisher, and some other product information. In addition to these intrinsic characteristics of products, consumers also observe the position and the sales volume, which is constructed from the number of ratings. In most cases, the number of ratings is not an ideal proxy for the sales volume, but due to the unique rating system of this platform, this is a reliable measure of the sales volume for consumers. The keyword-title similarity score is a measure of the string similarity between the keyword and the title of products constructed using the idea of edit distance, which is a standard method used in NLP and other applications. The detailed steps of constructing this measure are shown in Appendix A.1. After clicking on a product, consumers observe additional choice-related information on the product page, such as shipping cost, the sales rank of the book in its category (normalized to 0-1 range), the more detailed information about ratings and reviews, and the introduction of the product. Although sellers are relatively homogeneous in terms of overall service qualities, there could be consumers who received defective items or had bad shopping experiences leaving negative reviews. Consumers are likely to respond to these ratings and reviews even if sellers' overall qualities are similar. The indicators of the bundle, paper quality, and premium edition are listed on the product page, along with other detailed book information such as the ISBN and publisher. But since consumers can recover this information from the picture and title of products in the search results, I consider them as characteristics in the search-result pages.

Table 3.2: Summary Statistics: Session-Level Information

	All			3-5 Products		
VARIABLES	Mean	Std. Dev.	Median	Mean	Std. Dev.	Median
Consumer actions						
Number of clicks	1.554	1.156	1	1.551	0.940	1
Purchase dummy	0.103	0.304	0	0.108	0.310	0
Session-level information						
Number of different ISBN	1.501	1.150	1	1.537	0.928	1
Number of books	3.942	2.877	3	3.727	0.770	4
Top ranked product sold by the platform(%)	0.819	0.385	1	0.825	0.380	1
Number of prev. consuemrs	34,771	133,147	1,602	41,452	156,126	2,810
Months since books published	57.07	61.46	40	59.64	60.30	43
Observations	3,379			1,354		

Figure 3.4. Category Weight of Clicked Books

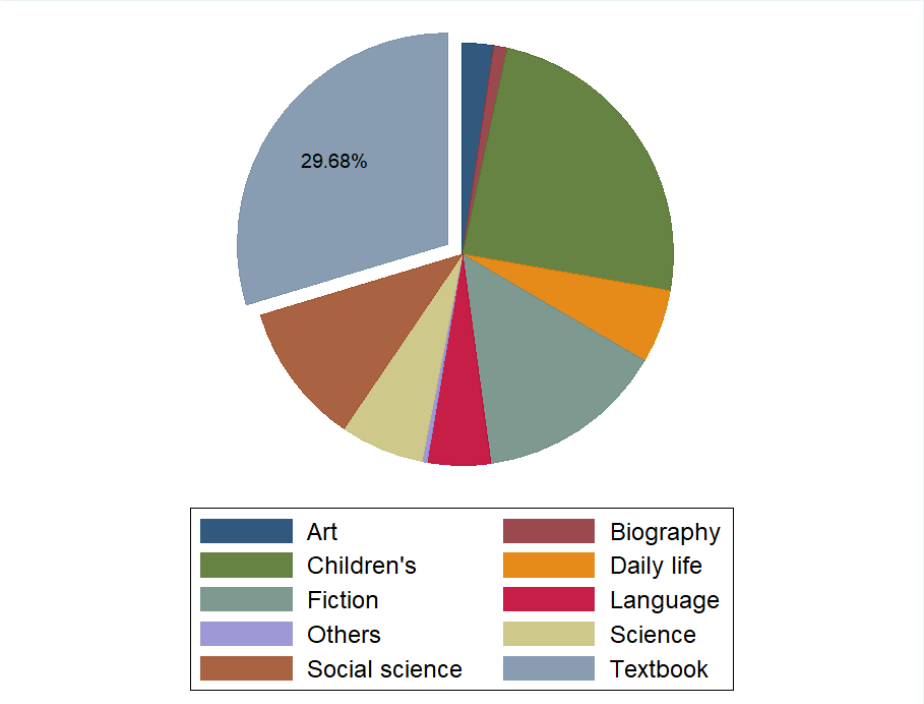


Table 3.3: Summary Statistics: Product Characteristics

	All		Clicked		Purchased	
VARIABLES	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Search-result page						
S&S by platform	0.379	0.485	0.760	0.427	0.742	0.438
Star rating	4.781	0.430	4.791	0.355	4.806	0.261
No rating dummy	0.477	0.499	0.173	0.379	0.149	0.357
Price	58.21	65.62	50.78	59.59	51.20	55.95
Original price(MSRP)	78.56	83.70	67.33	76.42	71.55	75.02
Bundle	0.552	0.497	0.442	0.497	0.553	0.498
Regular edition	0.892	0.311	0.849	0.358	0.862	0.345
Years since published	3.530	4.288	3.263	3.927	2.751	3.657
Position	9.452	8.481	4.473	5.592	3.997	5.107
Market share	0.245	0.376	0.505	0.419	0.608	0.421
Keyword-title match score	0.856	0.233	0.859	0.224	0.868	0.206
Product page						
Shipping cost			4.726	4.031	4.335	4.031
Sales rank in the category			159.7	142.0	163.7	135.9
Number of bad reivew (in 1000)			0.875	5.918	0.902	8.929
Length of introduction (100 words)			19.16	26.72	20.67	26.44
Observations	13,307		5,243		349	

CHAPTER 4

REDUCED FORM EVIDENCE

In this section, I show how sales information affects consumers' decisions at two different stages in the product search. The reduced form evidence shows that consumers are more likely to click products that have higher sales volume. However, conditional on clicking, consumers' purchasing choices are not significantly affected by sales information.

4.1 Effect of sales information on clicks

In search results, consumers observe sales information and some characteristics of products without costly sampling. Let index (i, j, s) denote consumer i and product j in the search session s. X_j^r is the characteristics of product j that the consumer observed in the search results for free. Furthermore, sales information I_{js} and position R_{js} are also observed in this page before clicking any product. Here, I include the subscript of session s in I_{js} and R_{js} to consider the case that the same product j can have different sales information and positions in different search sessions.

Let $SL(i)$ be the set of products in i's search list that is the list of books with consumer i's target book content. Here, I use the sales volume of product j as its sales information. After observing $(I_{js}, R_{js}, X_j^r)_{j \in SL(i)}$, consumer i decides which product(s) to click on. Consumers are allowed to click multiple products in their search lists. To simplify the analysis, I only check how sales information affect consumer's first click. I use a multinomial logit model to show how I_{js} influences the clicking decision that is conditional on controlling for the position and other characteristics X_j^r. Further, the identification of sales information effect relies on the exogenous variation of sales information I_{js} given other observed controls.

The main challenge of the causal inference of the sale information is the potential endogeneity issue of the price p_j, the sales volume I_{js}, and the position R_{js}. To address these issues, I use the control function approach (Petrin and Train 2010) for the sales volume and position as well as adding a proxy control for the unobserved product characteristics.

Sales volume A potential source of endogeneity is some decision-relevant product characteristics are omitted in the data. This is not very likely as most product characteristics that are available to consumers in the search results (Figure 3.1) are collected in the data through web scraping. Potential omitted characteristics are the pictures and/or the titles of products, which may persistently affect the past consumers' choices when considering overall consumers. However, this may not be the case when only considering consumers search for specific books. Another potential unobserved characteristic is the seller quality. But, as I discussed in Section 3, the seller's services are relatively more homogeneous on this platform than other platforms such as eBay. Also, since there is only one nationwide bookstore in China[1], the branding effect is almost negligible. I find that seller quality has little variation (see Appendix A.2) and does not affect consumers' choices among clicked products in general (see Table A.2).

While sales information is not very likely to be endogenous in my case, I conduct a robustness check using the control function approach when exploring the effect of sales information on clicking decisions. I use product-page characteristics that consumers observe only after clicking to the product page, such as the shipping cost, the length of product-page introduction, and the portion of negative reviews, as exogenous variables to construct a control function for the sales volume. These characteristics are exclusively correlated

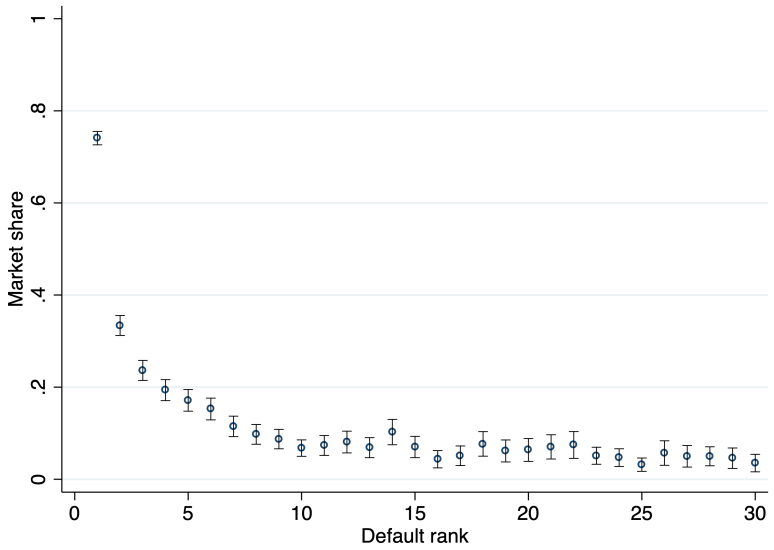
1. The Xinhua bookstore is the largest and only nationwide bookstore in China.

with the sales volume as they are observed by past consumers who purchase the product but do not affect the clicking decision of the current consumer.

Position The default ranking, which is usually generated by algorithms, is a common source of endogeneity in the empirical search literature (Ursu 2018). Since the majority of search sessions in my data use the platform's default rank, the position potentially suffers the same issue. However, the default position of a product generated by a platform is usually determined by the past sales or popularity of the product (De los Santos and Koulayev 2017), so I assume that given sales information, the remaining variations in positions are exogenous. This doesn't mean the position is randomly determined given the sales volume. For example, the relevance between the searching keyword and book titles could be a key determinant of the position. However, this factor does not affect consumers' choices in my case as consumers are assumed to search books with specific content. In the case when sales information is endogenous, the default position also suffers the endogeneity issue. So, I again use the control function to check the robustness of estimations.

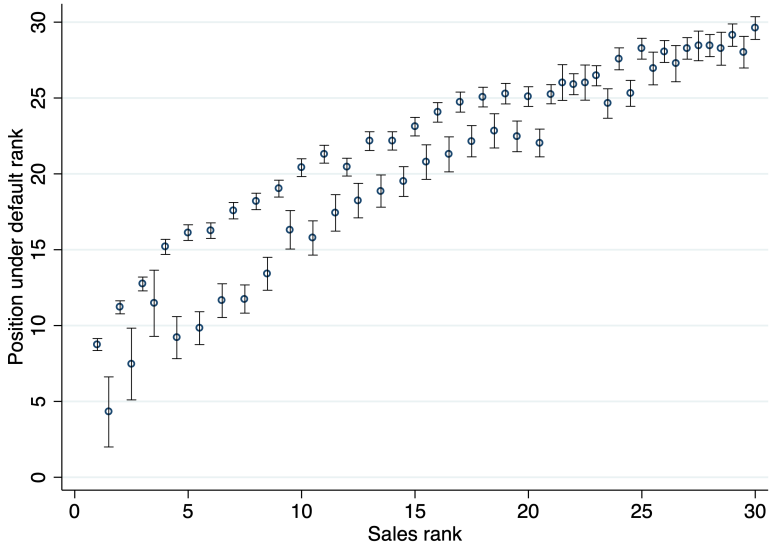
One potential challenge for the identification is the position may suffer from the collinearity with sales information. From the data, I do observe the correlation between the rank of the sales volume and the position, but there are still independent variations between the market share and the position. Figure 4.1 shows the market share at each given position. I find that given the position, the market share varies across different search sessions. This is partially due to my data containing different consumers' searches of various books, and if I consider each search session as a separate market, the market structure is different across different markets. Furthermore, Figure 4.2 shows that the sales rank does not wholly determine the default rank. Instead, there are some "experimental" variations in the default position given the sales rank by the platform.

Figure 4.1. Average Market Share at Given Default Position



Market share has some independent variations at each position.

Figure 4.2. Default Rank and Sales rank



The system default rank and the sales rank have strong correlation. But there are potential "experimental" variations by the platform.

Price Since each product is defined as a seller-book pair, in which a specific book is associated with a unique ISBN, a product indexed by j is equivalent to a (s,k) pair. Thus, book k that is sold by two different sellers s is considered as two different products. One main source of the price endogeneity comes from the unobserved qualities at the *ISBN* levels. Ideally, including the product fixed effect (or ISBN fixed effect) can solve this issue. However, since some ISBNs and books only exist once in the data, this approach is not feasible. Instead, I use the similar approach as the one used in De los Santos and Koulayev 2017 to address the price endogeneity issue.

Potential ISBN-level unobserved characteristics include the version and bundle contents of the book. For example, when a consumer searches a required textbook for school, such as *Microeconomics Theory* by M.W.G., the search results may contain books with the same content but in different versions, such as hardcover, paperback, and international version. For some books, it is more common to find the same book to be printed at different versions, such as the normal edition, collector's edition, and limited edition, that have different quality of printing and packaging.[2] Furthermore, a seller may sell different books as a bundle, e.g., bundle a textbook with a solution manual or other reference books. In the data, I do not observe the exact version and the bundle content. These unobserved qualities are both correlated with the price and persistently affect the consumer's utility.

To address this concern, I include the original price, that is, MSRP, as the control for the ISBN-level unobserved characteristics. According to some reliable sources, the marginal cost of printing a book is a relatively fixed portion of the MSRP in the industry and slightly varies across different categories of books and publishers.[3] So, it is reasonable to assume

2. Different versions and editions of a book usually have different ISBNs

3. In *the Pricing Mechanism of Chinese Book Market* written by the president of a major publisher in China, it claims that the cost of each book was around $35\% - 40\%$ of the MSRP in 2011. This pricing pattern

that the MSRP has a strong correlation with these unobserved book-level qualities or even linear in those characteristics. A regression of the retail price on the MSRP, the book category indicator, and other observed characteristics obtains an adjusted R^2 of 0.9, which suggests observed characteristics explain most of the price variations. The remaining 10% variations in the retail price are assumed to be exogenous variations that are independent of the unobservables. One main source of these variations potentially comes from sellers' different buying prices of the same book from the publisher or intermediaries. The identification of the price effect relies on the joint variation of clicking/purchasing decision and the retail price given the MSRP.

The remaining variation in the retail price may depend on unobserved seller characteristics since large sellers may have lower buying prices from publishers directly or they strategically set prices given the market shares or rankings. An ideal solution is to include the seller fixed effect. But some sellers are only observed once in my data, so this approach is not feasible. Instead, I argue that consumers have limited knowledge about sellers' sizes as almost all bookstores operate locally. Table A.2 also suggests that that seller size doesn't affect consumers' search decisions significantly. Also, strategically setting prices given the market shares or positions in search rankings isn't feasible, as sellers do not have access to others' sales data, and they do not know which other books they are competing with. It is common for one book to be listed in different positions with different sets of other books since search results are both keyword- and time-sensitive. Appendix A.3 provides further evidence and discussions about the price endogeneity caused by sellers' pricing strategies. These evidence should alleviate the concern about the endogeneity caused by unobserved qualities of the seller.

---

is also confirmed by the current staff of a publisher in China, but the cost has decreased to $20\% - 30\%$ in recent years.

To exam how consumers respond to sales information in the clicking stage, I check the effect of sales information on each consumer's first click using a multinomial logit regression. Column (1) of Table 4.1 reports the estimation result, which suggests consumers are more likely to click popular products in their search lists first. In column (2), I use some product-page characteristics as exogenous variables to construct the control functions for the potentially endogenous covariates $log(sales\,volume)$ and position. The results still shows that higher selling products are more attractive. In addition, the insignificant estimates of the controls for endogeneity of sales and position alleviate the concerns about their endogeneity.

4.2 Effect of sales information on transactions conditional on clicking

Conditional on clicking a product, the consumer observes additional characteristics X_j^u on the product page. I then study how consumers respond to sales information when they choose among clicked products. Since I only consider the case when consumers purchase at most one product, there is a substitution effect in choosing among the clicked products. To take this effect into consideration, I again use a multinomial logit model to study the effect of sales information. The control variables include the characteristics X_j^u, that consumers observe on the product page in addition to the information that consumers observed in the clicking stage.

Given the data of each consumer's choice decision, Table 4.1 is the result of the estimation of the choice stage when using maximum likelihood estimation. In the estimation, I exclude the consumers who choose the outside option; those who did not purchase any product after clicking. To correct the potential selection bias in the estimation, I also include the estimation that uses the two-step method developed by Heckman (1979) for comparison. The additional exogenous variable included in the first-stage Probit estimation is a

Table 4.1: Reduced Form Estimation: Effect of Sales Information on Clicks and Purchases Conditional on Clicks

	First clicked product		Purchase conditional on clicked	
	(1)	(2)	(3)	(4)
Search-result page				
log(Sales volume)	0.144***	0.132***	0.00553	-0.0376
	(0.0299)	(0.0484)	(0.0621)	(0.0919)
Position	-0.161***	-0.196***	-0.000450	0.0304
	(0.0154)	(0.0721)	(0.0241)	(0.0578)
Price	-0.0104***	-0.0118***		
	(0.00311)	(0.00410)		
Price plus shipping			-0.0141**	-0.0140**
			(0.00599)	(0.00588)
S&S by platform	0.722***	0.568	0.247	-0.317
	(0.169)	(0.362)	(0.424)	(1.116)
Bundle	-0.106	-0.100	0.0675	0.224
	(0.182)	(0.182)	(0.342)	(0.413)
Regular edition	-0.116	-0.118	-0.694	-0.728
	(0.183)	(0.183)	(0.535)	(0.548)
Years since published	-0.0189	-0.0181	0.00537	0.0190
	(0.0172)	(0.0174)	(0.0565)	(0.0584)
Original price	0.00903***	0.00992***	0.0123**	0.0119**
	(0.00237)	(0.00293)	(0.00542)	(0.00541)
Product page				
Number of bad reivew (in 1000)			-0.417***	-0.422***
			(0.140)	(0.139)
Sales rank in a category			0.0958	0.115
			(0.103)	(0.108)
log(Length of introduction)			-0.152	-0.150
			(0.125)	(0.124)
Endogeneity controls				
Control function - Sales		0.0104		
		(0.0471)		
Control function - Position		0.0356		
		(0.0714)		
Mills ratio				-0.807
				(1.392)
Observations	4,282	4,282	465	465

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note: Multinomial logit model of the first click (Colume (1) and (2)) and purchase conditional on clicked products (Colume (3) and (4)) as a function of product characteristics, sales information, and position. For the clicking stage, I restricted to sessions with no product having zero previous transaction. For the purchasing stage that are conditional on clicking, I restricted to sessions ending in a transaction.

string similarity score between the product title and entered keywords. This score exclusively affects consumer's clicking decisions, but among clicked products, the purchasing decision is less likely affected by it. The estimation results in columns (3) and (4) of Table 4.1 show that both with and without the correction of the selection bias, sales information no longer affects consumers' purchasing decisions among clicked products. One of the main determinants of consumer choice is the total price, which is the sum of the retail price and the shipping cost. This finding means that the price is the main factor that consumers consider when they search for specific books given sellers are relatively homogeneous in the quality of their service.

Again, my findings above may not apply to all cases, such as consumers search for experience goods that have quality that are not observed until using them. For these consumers, I think they are more likely to choose the higher selling products among the sampled products due to observational learning. These consumers are not in the scope of this paper, and I leave them for future study.

CHAPTER 5

STRUCTURAL MODEL

In the previous section, I show that sales information has different effects on consumers at two different stages of the consumer search. In the clicking stage, sales information mainly affects the consumer's belief over the unrevealed characteristics that they search for. Among the clicked products, sales information has no role in the purchase decision. However, I need counterfactual analysis to study how the disclosure of sales information affects consumer welfare and the fairness of competition. Here, I introduce a sequential search model, that incorporates the consumers' learning about the product page characteristics from sales information, which helps them make clicking decisions. I first set up the model and demonstrate the variables that enter the consumer's search decisions. Then I briefly explain the consumer's search process. Finally, I characterize consumers' optimal search decisions.

5.1 Utility

The setup of the layers of product characteristics follows Gu (2016) in which the characteristics of product j contain two components: $X_j = (X_j^r, X_j^u)$. Consumers observe $X_j^r \in \mathbb{R}^{K_1}$ in the search-result pages without search cost, and $X_j^u \in \mathbb{R}^{K_2}$ along with match value ε_{ij} are additional information that is unrevealed to consumers until they costly click on product j. For different consumers, the same product may have different characteristics. For example, the price may vary across periods or the rating may change. So, I treat a product that appears in different consumers' search results as different products, since the product identity itself does not enter my analysis. With this assumption, I can focus on how consumers respond to the displayed product characteristics but not the product itself.

Consumer i's utility from product $j \in \{1, ..., J_i\}$ is

$$u_{ij} = X_j^r \beta_i^r + X_j^u \beta_i^u + \varepsilon_{ij}$$

where J_i is the total number of books in consumer i's search results with the content he or she specifically searches for, that is, in the search list $SL(i)$. The match value of product j to consumer i, ε_{ij}, is observed by consumers after clicking product j, but econometricians cannot observe it. Furthermore, I assume it follows the standard normal distribution for the purpose of identification and be *iid* across consumer-product pairs, that is, $\varepsilon_{ij} \overset{i.i.d.}{\sim} N(0,1)$. So, consumers pay search costs to find both the X_j^u and the ε_{ij} of product j.

As discussed in the Section 3, I narrow down i's searching list by keeping products in the search results that share the same ISBN, author, or the title to the product(s) that the consumer clicked. $\beta_i = (\beta_i^r, \beta_i^u)$ represents the heterogeneity in consumer tastes. I assume that consumers are only heterogeneous in the price tastes to reduce the complexity of computations. So, for all non-price characteristics k,

$$\beta_{ik} = \beta_k \ \forall i$$

that assumes consumer tastes for non-price qualities are homogeneous. For the price coefficient, I follow Kim et al. (2010) and Kim et al. (2017) to assume a log-normal distribution to ensure a negative value

$$\log(-\beta_i^p) \sim N(\gamma, \sigma_\beta^2)$$

The outside option is available to consumers free of search cost, and its utility is assumed to be

$$u_{i0} = V_0 + \varepsilon_{i0}$$

where $\varepsilon_{i0} \overset{i.i.d.}{\sim} N(0,1)$ as other ε_{ij}'s, and V_0 is the mean utility of outside options. Consumers who clicked without purchasing anything are assumed to choose the outside option.

5.2 Search cost

A consumer's search cost depends on the position of the product and the device that consumers used to browse the platform. Let R_j denote the position of product j in search session s, and I simplify the notation by dropping index s. $d_i = 1\{mobile_i\}$ is the indicator of the device type that consumer i used for the search. Consumer i's search cost over product j is

$$c_{ij} = \exp(\alpha_0 + \alpha_1 R_j + \alpha_2 d_i)$$

and ensures the cost is positive. The search cost is assumed to depend on the position of the product and further varies across different devices used to browse the products. On the mobile device, the screen can show around three products at the same time as compared to around five products on a 24-inch computer screen. If consumers choose to show smaller icons, the PC sites can display even more products on one page. Ghose, Goldfarb, et al. (2011) find that search cost is generally higher in mobile devices due to smaller screen. So, I also check if the position effect varies across the types of devices to reflect the idea that the access to products is the key factor in the search cost.

5.3 Learning the unrevealed characteristics X_j^u

Consumers observe X_j^r in the search results free of search costs, then they have to costly search for X_j^u and ε_{ij}. Different from previous empirical search literature, such as Gu (2016) who assumes consumers have rational expectations of X_j^u given X_j^r from the empir-

ical distribution, I assume consumer i additionally receives sales information I_j of product j in the search results that helps him or her to learn the distribution of X_j^u.

Formally, let $\mathbf{X}_i^u = \{X_j^u\}$, $\mathbf{X}_i^r = \{X_j^r\}$, and $\mathbf{I}_i = \{I_j\}$ for $j \in SL(i)$, that is, the characteristics and sales information of all products in consumer i's search list. Suppose consumer i knows the support of each X_j^u. I assume that all products have the same support of product-page characteristics, and the support includes all distinct product-page characteristics in his or her search list, i.e., $X_j^u \in \{x_1^u, ..., x_{K_i}^u\}$. If all products have different unrevealed characteristics X^u, then $K_i = J_i$, where J_i is the number of products in i's search list. I assume that consumer i has a uniform prior over $\mathbf{X}^u$, that is,

$$h_0(\mathbf{X}^u|\mathbf{X}^r) = \prod_j h_0(X_j^u|X_j^r) = (\frac{1}{K_i})^{J_i}$$

where K_i is the size of the support of X_j^u.

Given the characteristics and sales information $(\mathbf{X}^r, \mathbf{I})$ in the search results, the consumer forms a belief about product-page characteristics over all products in his or her search list jointly,

$$p(\mathbf{X}^u|\mathbf{X}^r, \mathbf{I}, F_\beta(\Theta)) \propto \pi(\mathbf{I}|\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta)) h_0(\mathbf{X}^u|\mathbf{X}^r) \tag{5.1}$$

where $(\mathbf{I}|\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta))$ is the likelihood that consumers believe, given the actual characteristics $(\mathbf{X}^r, \mathbf{X}^u)$, the sales information be $\mathbf{I}$. Here, I assume that each consumer has a rational expectation of other consumers' tastes, that is, he or she knows the population distribution of the consumer tastes $F_\beta(\Theta)$.

Because of the assumption on the prior, I can rewrite equation 5.1 as

$$p(\mathbf{X}_k^u|\mathbf{X}^r, \mathbf{I}, F_\beta(\Theta)) = \frac{\pi(\mathbf{I}|\mathbf{X}^r, \mathbf{X}_k^u, F_\beta(\Theta))}{\sum_{l=1}^K \pi(\mathbf{I}|\mathbf{X}^r, \mathbf{X}_l^u, F_\beta(\Theta))}$$

where $\mathbf{X}_k^u$ is a possible realization of $(X_1^u, ...X_{J_i}^u)$ and the number of possible $\mathbf{X}_k^u$ is $K = K_i^{J_i}$.

Here, I use the market share $\mathbf{s} = (s_1, ..., s_J)$ as sales information $\mathbf{I}$, where the market is defined as all products in i's search list $SL(i)$. So,

$$s_j = \frac{sales_j}{\sum_{k=1}^{J_i} sales_k}$$

To model the likelihood that observing the market share $\mathbf{s}$ gives the characteristics $(\mathbf{X}^r, \mathbf{X}^u)$, the observed market share is the weighed sum of a consumer's perceived market share plus an error term $\varepsilon \in \triangle^{J_i}$ that follows the Dirichlet distribution, that is,

$$\mathbf{s} = \mu P(\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta)) + (1-\mu)\varepsilon \tag{5.2}$$

where $P(\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta)) \in \triangle^{J_i}$ is the consumer's perceived market shares of J_i products given $(\mathbf{X}^r, \mathbf{X}^u)$ and the parameters of consumers' tastes.

$\varepsilon \sim Dir(\overrightarrow{\alpha})$ is an error term *iid* across different search sessions that captures the ratings from those who did not buy the products. I have assumed that consumers use the number of ratings as the proxy for sales volume due to the special rating system of the platform, which ensures every purchase ends up with a rating. However, not all ratings are made by consumers purchasing the books. I find that around 95% of ratings are made for verified purchases. So, ε is the share of non-buyer ratings across products in each search session. I assume $\overrightarrow{\alpha} = (\alpha, ..., \alpha) \in \mathbb{R}^{J_i}$, that is, the chance of receiving a non-buyer rating is equally likely for all products on the list.

The weight $\mu \sim Beta(\gamma_1, \gamma_2)$ represents the portion of the ratings made by actual buyers, and I recover the parameter γ_1 and γ_2 from the actual data that indicates ratings made by verified buyers. So, I assume consumers have a rational expectation about the portion of

fake reviews and reviews from non-buyers. Let

$$P_j(\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta)) = \int \frac{\exp(X_j\beta_i)}{\sum_{k=1}^{J_i}\exp(X_k\beta_i)} dF(\beta_i;\Theta)$$

to be a Logit model with a random coefficient for the price, and $X_j = (X_j^r, X_j^u)$. I can compute the consumer's perceived market share by assuming all previous consumers follow the sequential search model that I have proposed in a dynamic setting, but use the logit model for the following reasons. First, the product rankings may vary across different periods, and a new consumer does not know the position of each product in previous consumers' search results, so he or she cannot accurately recover the search cost nor the choice probability by following the sequential search rules. Also, even if the rankings remain unchanged across all periods, due to a large number of previous transactions (about $39,000$ per search session on average), it is computationally infeasible to compute all consumers' choices under the sequential search model. Furthermore, due to the fact a book can exist in the search results of different searched keywords, the previous consumers who purchased a book in consumer i's list very likely have different books in their search results. So, the assumption that each consumer i can fully anticipate all these possible variations in the search results of previous consumers who actually purchased the books listed in i's search results in not computationally feasible.

Rewriting equation 5.2 allows me to write out the likelihood

$$\pi(\mathbf{s}|\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta)) = \int_0^{\overline{\mu}} h_\varepsilon(\frac{1}{1-\mu}(\mathbf{s} - \mu P(\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta))))dG(\mu;\gamma_1,\gamma_2)$$

where $\overline{\mu} = \min_j \frac{s_j}{P_j}$ to ensure $\mathbf{s} - \mu P(\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta)) > 0$ for all entries j, that is, $\varepsilon \in \triangle^{J_i}$ has non-negative entries.

From 5.1, I can compute product j 's marginal probability of X_j^u, that is,

$$P_j(x;\mathbf{X}^r,\mathbf{s},F_\beta(\Theta)) \quad = \quad \Pr(X_j^u = x|\mathbf{X}^r,\mathbf{s},F_\beta(\Theta)) \tag{5.3}$$

for all $x \in \{x_1^u,...,x_{K_i}^u\}$.

5.4 How do consumers search?

Before formally define the consumer's searching strategy, I give an example to illustrate the search process. When a consumer decides to search for a book with a specific content on this platform, he or she enters a keyword into the search box. The keyword may not match the name of the book entirely, but it should at least have certain similarities with that book. After the consumer sends his or her search request to the platform, the platform responds with search results that contain all related products listed in a specific ranking that its algorithm determines. The consumer observes some product characteristics in the search results, such as the title of the product, the overall rating, number of ratings, and the retail price. Then the consumer picks a product in his or her search list, which includes all books in the search results that have the content he or she is looking for. Due to the unique rating system of the platform, consumers use the number of ratings as a signal of the sales volume and infer the product-page characteristics before incurring the cost of clicking on the product(s). After that, the consumer observes all remaining information about the product. He or she decides whether to continue searching or not. The consumer will continue clicking on books until he or she decides to stop searching and to select the best alternative from all sampled products, which includes the outside option.

5.5 Optimal search

In this subsection, I model the consumer's optimal search decisions following Weitzman (1979)'s rules of the sequential search framework. These three rules rationalize the consumer's decision on the clicking order, when to stop searching, and which product to choose. A consumer decides the clicking order after seeing the search-result pages. The optimal clicking order is a descending order of the reservation utility of each product. The reservation utility of a product is defined as the utility level that makes the expected gain of sampling this product equal to its search cost. Denote z_{ij} as consumer i's reservation utility of product j, then it satisfies:

$$\int_{z_{ij}}^{\infty}(u_{ij}-z_{ij})f(u_{ij}|\mathbf{X}^r,\mathbf{s},F_\beta(\Theta))du_{ij}=c_{ij} \tag{5.4}$$

where $\mathbf{X}^r=\{X_j^r\}$ is the characteristics of products $j=1,...,J_i$ that are revealed in the search results, and J_i is the number of products in consumer i's search results that is in his or her potential search list. $\mathbf{s}=\{s_j\}$ is the corresponding displayed market share of product $j=\{1,...,J_i\}$. Here, I assume consumers have the same tastes over non-price characteristics and knowing the population distribution of the price coefficient, that is, $F_\beta(\Theta)$. Equation 5.4 can be rewritten as

$$\int_{z_{ij}}^{\infty}(u_{ij}-z_{ij})(\sum_{k=1}^{K_i}f_\varepsilon(u_{ij}-V_{ij}|X_j^u=x_k,X_j^r)\Pr(X_j^u=x_k|\mathbf{X}^r,\mathbf{s},F_\beta(\Theta)))du_{ij}=c_{ij} \tag{5.5}$$

where K_i is the size of the support of X_j^u, and the marginal probability of $X_j^u=x_k$ given the characteristics and information in the search results is derived in Equation 5.1 and 5.3. As in Kim et al. (2010), equation 5.5 can be further rewritten as

$$\sum_{k=1}^{K_i}p(V_{ijk}|\mathbf{X}^r,\mathbf{s},F_\beta(\Theta))\cdot[\phi(\frac{z_{ij}-V_{ijk}}{\sigma_\varepsilon})-\frac{z_{ij}-V_{ijk}}{\sigma_\varepsilon}(1-\Phi(\frac{z_{ij}-V_{ijk}}{\sigma_\varepsilon}))]=\frac{c_{ij}}{\sigma_\varepsilon} \tag{5.6}$$

where $V_{ijk} = X_j^r \beta_i^r + x_k \beta_i^u$. Since the left-hand side of equation 5.6 is monotonically decreasing in z_{ij}, there is a unique z_{ij} that ensures the equality holds. There is no closed-form solution of the reservation utility. Thus, I solve it numerically. After consumer i computes z_{ij} for $j = 1, ..., J_i$, he or she starts sampling products in the decreasing order ofz_{ij}'s.

Given z_{ij} and u_{ij} for all available products including the outside option, consumers stop searching when the best alternative among the searched products, S, exceed the highest reservation utility in the remaining products $\bar{S}$:

$$\max_{j \in S} u_{ij} \geq \max_{l \in \bar{S}} z_{il}$$

And he or she chooses $j^* = \arg\max_{j \in S} u_{ij}$ among all searched products. Any consumer's action, including clicks and purchases, is fully rationalized by these three rules. In sum, the optimal searching actions are:

- Selection rule: The next product to sample should be the one with the highest reservation utility among the remaining products.
- Stopping rule: Stop searching if the best sampled product has higher utility than the highest reservation utility among the remaining products.
- Choice rule: The best alternative among sampled products, including the outside option, should be chosen.

In Chen and Yao (2017) and De los Santos and Koulayev (2017), consumers have the option to choose different sorting methods and filtering options. In my case, since the majority of the consumers (over 98%) in the data search with the default refinement, I do not model the choice of refinements. Gu (2016) also consider the consumer's choice of switching pages. Since the platform that I use automatically loads the next page when a consumer reaches the bottom of the current page, the position effect in the search cost will fully characterize

the accessibility of products. So, my model does not incorporate the action of switching the page either.

CHAPTER 6

ESTIMATION

In this section, I use the model discussed in the previous section to develop an empirical strategy to estimate the parameters of interest. Then I explain the identification strategies and show that parameters can be recovered in a relatively small sample using a Monte Carlo simulation.

6.1 Likelihood

For each product j in consumer i's search list, I observe the product characteristics X_j^r and X_j^u, as well as its position R_j and market share s_j. Let a_i be consumer i's action that includes the clicking and the purchasing records. Let $j(l)$ be the l^{th} product clicked by consumer i. Here I simplify the notation by dropping the index i in the notation of clicking order $j_i(l)$. If consumer i clicks m products in total, then the set $S = \{j(0), ...j(m)\}$ represents i's searched products, in which $j(0)$ is the outside option that is available to all consumers without a search. By Weitzman's rules, for $i = 1, ..., N$, I can construct the likelihood of observing the action a_i given data $(X_j^r, X_j^u, R_j, s_j)_{j=1,...,J_i}$ and parameters θ.

First, by the selection rule, the clicking order follows the order of reservation utilities. That is,

$$z_{ij(l)} \geq \max_{k=l+1}^{J-1} z_{ij(k)} \tag{6.1}$$

for all $l = 1, ..., m$. The stopping rule characterizes the action of continuing or stopping the search. Thus, if the consumer conducts the l^{th} search, then

$$z_{ij(l)} \geq \max_{k=0}^{l-1} u_{ij(k)} \tag{6.2}$$

for $l = 1, ..., m$, and

$$\max_{l=0}^{m} u_{ij(l)} \geq \max_{j \in \bar{S}} z_{ij} \tag{6.3}$$

when consumer i stops searching after the m^{th} click. Here, $\bar{S} = \{1, ..., J_i\} \backslash S$ is the set of remaining products not viewed. Finally, if product $j^* \in S$ is purchased among all clicked products, it must be the best alternative in S:

$$u_{ij^*} \geq \max_{j \in S} u_{ij} \tag{6.4}$$

The probability that consumer i chooses search action a_i given $data_i = (X_j^r, X_j^u, R_j, s_j)_{j=1,...,J_i}$ and parameters θ is the probability that condition 6.1-6.4 are met jointly. Let $P_i(\theta) = \Pr(a_i; data_i, \theta)$ denote this probability. Since consumers are assumed to have heterogeneous tastes for the price, which follows the log-normal distribution, I use the Gauss–Hermite quadrature with 25 draws of the price coefficient given the mean and variance of its distribution to perform the numerical integration. Suppose in θ, the mean and standard deviation of the price coefficient is μ_p and σ_p, and $(v_m, w_m)_{m=1}^{25}$ are the corresponding draws and associated weights of the Gauss–Hermite quadrature. Then the price coefficient for each node m is

$$\beta_p^m = -\exp(\sqrt{\pi}\sigma_p v_m + \mu_p)$$

Let the corresponding probability of the consumer's observed action given the data and parameters (θ, β_p^m) satisfy condition 6.1-6.4 to be $P_i(\theta, \beta_p^m)$, then the numerical integration of the price coefficient is

$$P_i(\theta) = \frac{1}{\sqrt{\pi}} \sum_{m=1}^{25} w_m P_i(\theta, \beta_p^m)$$

And the corresponding log-likelihood function of all consumers is

$$LL(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log P_i(\theta) \tag{6.5}$$

The main challenge of maximizing equation 6.5 with respect to parameters θ is the non-smoothness of the objective function. Equation 5.4- 5.6 shows that the reservation utility has no randomness given a candidate of parameters. Then by the selection rule, the clicking order given $data_i$ and a parameter θ for each consumer i is deterministic. This rule means, if any consumer i's observed action a_i does not match the clicking order predicted by the model at point $\tilde{\theta}$, then $P_i(\tilde{\theta}) = 0$. And the log-likelihood function $LL(\tilde{\theta})$ is not defined at $\tilde{\theta}$ that causes the objective function to be non-smoothing at some points in the domain of θ. This non-smoothing object function leads to difficulties in the optimization process as the optimizer in the MATLAB requires a smooth objective function. To overcome this issue, I used the logit-smoothed AR simulator, which was first introduced in McFadden (1989), to smooth out the probability $P_i(\theta, \beta_p^m)$. Honka (2014) and Honka and Chintagunta (2017) and other empirical search papers have used this method to solve the discontinuity of the likelihood function. In practice, I draw 50 random utility shocks ε_{ij} for each consumer-product pair from its distribution with the scaling parameter $w = 13$. Theoretically, the simulated loglikelihood function is closer to the original $LL(\theta)$ when w is larger. However, the optimizer has more difficulty finding the optimal point due to the non-smoothness, and the results highly depend on the initial point. So, I test multiple values of w by using the Monte Carlo simulation, and using the value that recovers the parameters best, which is $w = 13$.

6.2 Identification

In this subsection, I discuss how I identify the parameters of interest with consumers' clicking and purchasing behaviors, following the idea similar to Chen and Yao (2017), Kim et al. (2010), and Kim et al. (2017). There are two sets of parameters: tastes for product characteristics and search costs. For the purpose of identification, I normalize the variance of utility shock to 1, that is, $\sigma_{ij} = 1$, as in the empirical search literature.

Consumers' tastes for product characteristics make up the mean utility that can be identified from conditional choice probability among clicked products similar to the multinomial discrete choice environment. Since most consumers search for different products, there are enough variations in the product characteristics. Furthermore, clicking actions can also help identify these characteristics more efficiently. The price heterogeneity is identified if I observe consumers have different price elasticity given other characteristics.

For the set of parameters related to the search cost, Ursu (2018) shows that each consumer's clicking decision on products with different positions identifies the position effect, and comparing the number of clicks made across PC and mobile devices determines the device effect. The constant term in the search cost is pinned down by equation 5.4, which computes the reservation utility.

Furthermore, I normalize the mean utility of the outside option to zero in the estimation. Theoretically, this value can be identified since I observe a group of consumers choosing the outside option after sampling products, and I can pin down this mean utility by comparing the clicked products that lead consumers to choose or not choose the outside option. However, in practice, when the constant term in the search cost is very small, say at -3, the search cost becomes almost linear in the constant. This linearity makes the identification of both the outside option and the constant term in the search cost to be numerically challenging and to cause further issues in other parameter estimates as the constant of the

search cost is pinned down through equation 5.4. In Appendix A.3.1, I show more details about how normalizing the mean utility of the outside option to zero affects the estimation.

For price endogeneity, I use the original price (MSRP) as a control for unobserved book characteristics (see section 4 for details). The reasoning is that the remaining variations in the retail price may either consist of exogenous variations or depend on some qualities of the seller's service. But the latter term does not affect the consumer utility due to the seller type on this platform. The bias caused by position endogeneity is alleviated by controlling for sales information as De los Santos and Koulayev (2017) argues that the primary determinant of the ranking algorithm is past sales. The estimate of the position effect in section 7 is close to Ursu (2018), in which ranking is fully randomized.

6.3 Monte Carlo Simulation

Table 6.3 shows the results of the Monte Carlo simulation with 800 search sessions that uses the information from the actual data that I used for the actual estimation. In each search session, I simulate a consumer by randomly drawing the utility shock ε_{ij} and price coefficient, and then simulating the clicking and purchasing decisions by following optimal sequential search model. The results show that the parameters can be recovered using this finite sample. Especially, the use of $w = 13$ provides an overall better estimation.

Table 6.1: Estimation Results Using Simulated Data

		$w = 13$		$w = 9$	
	True values	Estimates	SE	Estimates	SE
Search-result page					
No rating dummy	-1.000	-0.851***	0.055	-0.866***	0.065
S&S by platform	1.000	0.802***	0.054	0.745***	0.060
Bundle dummy	1.000	0.833***	0.057	0.776***	0.069
Non-premium edition	-1.000	-0.829***	0.065	-0.790***	0.068
log(Years since published)	-1.000	-0.858***	0.034	-0.813***	0.040
Original price	0.010	0.007***	0.001	0.005***	0.001
Product page					
% of bad rating (in 1%)	-5.000	-4.513***	0.726	-4.337***	0.793
Sales rank in category	1.000	0.779***	0.035	0.687***	0.033
log(length of introduction)	0.200	0.197***	0.014	0.170***	0.016
No introduction dummy	-1.000	-0.809***	0.097	-0.770***	0.115
Price					
Mean	-5.000	-4.866***	0.100	-6.335***	0.298
Heterogeneity	2.000	1.948***	0.072	2.606***	0.205
Search cost					
Constant	-3.000	-2.390***	0.143	-2.217***	0.151
Position	0.050	0.039***	0.005	0.037***	0.005
PC device dummy	-1.000	-0.965***	0.152	-0.998***	0.163
Number of search sessions	800				

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

Monte Carlo simulation with sessions from the data with number of products ranges from three to five.

CHAPTER 7

RESULTS

7.1 Estimation results

In this subsection, I show the estimation results of the sequential search model proposed in Section 5 that uses the simulated maximum likelihood estimation. The integration over the random price coefficient is done numerically by using the Gauss–Hermite quadrature with 25 nodes. And for each node, I draw 50 *iid* utility shocks ε_{ij} to construct the logit-smoothed AR simulator of the log-likelihood function. The data I use in the estimation includes searches by consumers who each searches for a specific book across different editions and sellers. To ensure it is computationally feasible, I only include the search sessions with the number of products in the search list at less than six but more than two. Search sessions without any clicks are not included in my analysis as the data record partial keywords before consumers finish entering the entire keyword. And I cannot determine if a consumer searches for a specific book without observing the product he or she clicked. Furthermore, I do not include search sessions in which the consumer uses the same keyword to avoid the potential bias in the estimation caused by these repeated searches.

Table 7.2 shows the structural estimation results. Due to the insignificant estimates of the price heterogeneity, I also include the homogeneous case in Column (2). In general, most of the characteristics I am interested in play statistically significant roles in consumer's decisions on online searches. In terms of sellers, consumers prefer the products sold and shipped by the platform, which may be due to the branding effect. Further, the control for unobserved book characteristics *MSRP* and ISBN-level characteristics, such as bundle dummy and the premium edition dummy, may not be individually identified due to the potential collinearity. The dummy variable of no rating in the search results indicate there are

no previous ratings of the product that were written by purchasers, and I find consumers are less likely to click and purchase a product if it has no ratings. The characteristics in the product page also affect consumer choices. The negative ratings in the product reduce the consumer utility but not significantly, probably because most products have very little or zero negative ratings due to the review censorship by the platform. Further, products with no introduction in the product page are less preferred. However, for products with an introduction, a longer introduction leads to a lower utility. This is possibly due to the information overload of reading a long introduction. The price coefficient is significant and allows me to convert the effect of other characteristics into the monetary value. The estimates of the search cost also confirm the position effect found in the previous literature, such that a lower ranking leads to a higher search cost. Since the search cost is exponential to the position, the marginal effect of the position depends on the position itself. On average, the search cost increases by 19.89 RMB (~2.79 USD) if a product is moved from the top of the list to the 10^{th} position, which is around 1.99 RMB (~0.28 USD) per position. If I consider the cost of the position as the consumer's effort to find the product, then converting 0.28 USD by the difference of real GDP per capita between the US and China (~6.74) leads to 1.89 USD, which is close to the 1.92 USD in Ursu (2018). The device (PC and mobile) does not have a significant effect on the search cost, potentially due to the similar number of products shown on each page.

7.2 Value of sales information

One of my main research questions is the effect of sales information on consumer actions. One challenge of quantifying this effect is that sales information does not enter the consumer's utility function, and there is no coefficient that directly quantifies its effect. Instead, sales information affects consumer's clicking decisions through the belief in the

characteristics in the product page. Here, I use the position as the measure to quantify this effect indirectly. Equation 5.4 shows the only channel in which sales information affects the clicking decision is through the reservation utility. To quantify the effect of sales information, I calculate the equivalent adjustments in the search cost that keep the reservation utility unchanged, that is, clicking decision remains the same, when the availability of sales information varies. I can further represent the changes in search cost by the change in the position leads to finding the equivalent changes in the position as the value of sales information.

Using the search sessions from the data, Figure 7.1 shows the equivalent changes in the position at the given market share and position in the list when sales information is removed, and the clicking decision remains unchanged. I find that disclosing sales information mainly benefits those high-selling products but harms those that are unpopular. For example, if a product ranked at the top of the list with a market share above 0.5, removing sales information is equivalent to reduce its rank by around 0.8. A naive monetary value of this equivalent shift in position for the top-ranked products is about 1.7 RMB (~ 0.25 USD).

Another way to visualize the effect of sales information is the elasticity of the position with respect to sales information with reservation utility fixed:

$$-\frac{\%\text{ position changes in product } j}{\%\text{ market share changes in product } j}$$

Figure 7.2 shows this ratio at each rank and market share in the search sessions from the data. Especially, the effect of sales information decreases with the current market share and the position. For a product with a low market share but at the top of the list, a one percent change in the market share can lead to equivalent changes of 0.08% in the position. For example, if a product in the 5^{th} position has a market share that increases from 5% to

Table 7.1: Structural Estimation Results

	Heterogeneous		Homogeneous	
	Estimates	SE	Estimates	SE
Search-result page				
No rating dummy	-0.042**	0.016	-0.042**	0.016
S&S by platform	0.077***	0.016	0.078***	0.016
Bundle dummy	-0.028	0.024	-0.028	0.024
Non-premium edition	-0.042***	0.021	-0.042**	0.021
log(Years since published)	-0.016	0.011	-0.016	0.011
Original price	0.000	0.000	0.000	0.000
Product page				
% of bad rating (in 1%)	-0.137	0.152	-0.121	0.145
Sales rank in category	0.113***	0.017	0.105***	0.018
log(length of introduction)	-0.221***	0.006	-0.221***	0.006
No introduction dummy	-1.427***	0.069	-1.437***	0.069
Price				
Mean	-7.617***	0.947	-7.526***	0.852
Heterogeneity	0.023	10.285		
Search cost				
Constant	-2.767***	0.081	-2.770***	0.081
Position	0.016***	0.002	0.016***	0.002
PC device dummy	0.004	0.093	0.008	0.093
Observations	4096			
Number of search sessions	1097			
Loglikelihood	-2637.649		-2637.856	

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

Restrict to sessions in which number of products ranges from three to five.

10%, then that increase is equivalent to a move in its position up by one in the list, which is equivalent to 2.2 RMB (~0.3 RMB).

Figure 7.1. Equivalent Changes in Position if Hiding Sales Information

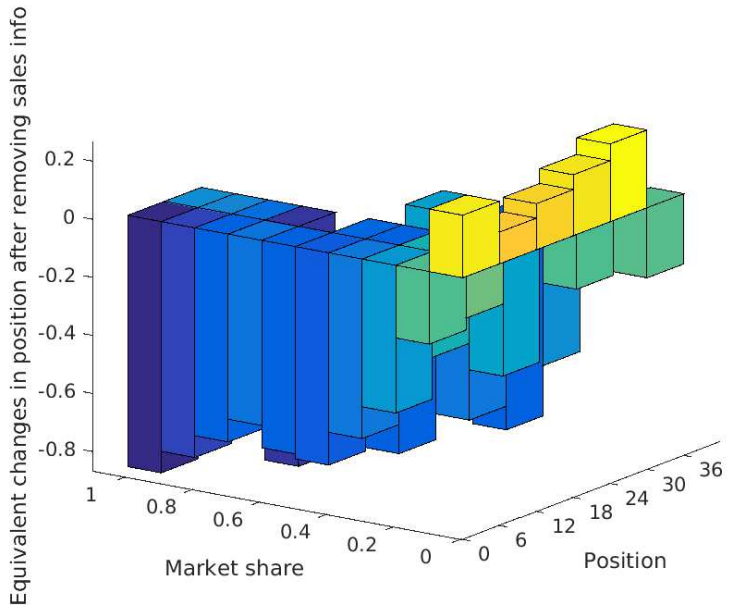
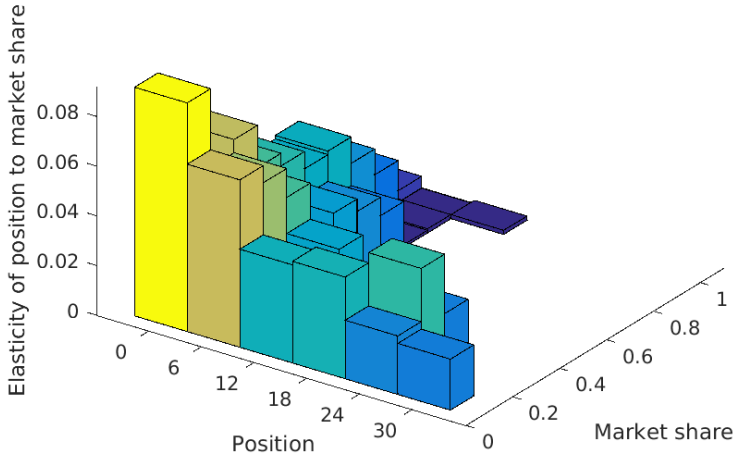


Figure 7.2. Elasticity of Position with Respect to Sales Information



CHAPTER 8

COUNTERFACTUAL

In this section, I use the counterfactual experiments to show the effect of disclosing sales information on consumer welfare, and whether some sellers can gain persistently more sales by influencing temporary popularity of their products. In particular, does disclosing sales information benefit consumers or not? Does the feedback loop, i.e., a popular product is more attractive to future consumers due to its popularity, exist if sales information is available to consumers? Specifically, do first-party sellers gain persistently more sales if they can influence the popularity of their products in the short run? To perform the counterfactual analysis, I use the search sessions from the data with five products in each consumer's search list. I then simulate the consumer's clicking and purchasing decisions under different availabilities of sales information and ranking methods with parameters estimated in the section 7.

8.1 Does disclosing sales information benefit consumers?

Sales information is commonly available to consumers in an online search, but the disclosure of sales information has an unclear effect on the welfare of consumers. Since consumers have imperfect information when searching across alternatives, the disclosure of sales information may benefit consumers by providing them with additional information to infer unrevealed product characteristics. However, previous studies have shown that disclosing information about past consumers' choices may lead to inefficient outcomes (Banerjee 1992; Bikhchandani et al. 1992; Smith and Sørensen 2000; Zhang 2010). In this subsection, I use the observational data from the platform to analyze how the disclosure of sales information affects consumer welfare.

To run the counterfactual, I pick 248 search sessions from the data that have five products in the consumer's search list. For each session, I randomly generate 250 consumers whose idiosyncratic utility component and price coefficient are drawn *iid* from the corresponding distributions. The mean utility and search cost of each product are computed using estimated parameters and the information of products (e.g., characteristics, positions, and sales information) from the data. Since the rating system of this platform assigns a default positive rating if a consumer does not rate the purchase after 10 days, I assume those 250 consumers who arrive at the platform closely together observe the same sales information. Since the default ranking is correlated with the sale volume, I randomize the ranking of the products for each consumer to separately identify the effect of sales information net of the position effect. I simulate consumers' clicking and purchasing decisions as in the optimal search strategy shown in Section 5 under different availabilities of sales information.

Table 8.1 compares the consumer welfare with and without the disclosure of sales information. Here, consumer welfare is defined in the same way as the literature as the utility gain from the chosen product (including the outside option) minus the total search cost. The results show that the disclosure of sales information has no significant effect on consumer surplus in general, and its effect varies across different search sessions. Thus, I group the search sessions by checking whether sales information leads to higher consumer welfare. Separately studying the effect of sales information in two different groups shows that consumers can find better matched products with fewer clicks in sessions where they benefit from the disclosure of sales information. In the other group, disclosing sales information leads consumers to make more clicks and to choose products with lower mean utilities. By comparing the portion of bestsellers also having the highest mean utility across two groups, I find that this number is almost doubled (39.22% vs. 20.18%) in sessions in

which sales information benefits consumers. Therefore, consumers are likely misled by sales information in some sessions that lead to a welfare loss.

I further test a series of procedures with light computations that can potentially alleviate the existence of sales information that misleads consumers. Using the same 248 sessions as above, I simulate a group of 250 consumers in each session per period for 15 periods in total. In the first period, the rankings are randomly determined, and there is no sales information available to consumers to mimic the case when products are newly listed on the platform. The purpose of this initial period is similar to Google assigning a Quality Score to each website that reflects its overall quality. In the subsequent periods, consumers observe the sales volumes of products and make optimal choices. In these periods, I test two popular ranking methods: sales volume and utility-based. Table 8.1 shows that under both ranking methods, the disclosure of sales information benefits consumers. The increment in consumer welfare comes from both fewer total clicks and the choice of better matched products. Meanwhile, the platform also achieves higher total sales, and the sales of the first-party seller remain about the same. In particular, under these procedures, the portion of bestsellers also with the highest mean utility (75.28% and 63.35%) is significantly higher than I find in the data. This finding provides platforms with an easy managerial method to achieve Pareto improvement.

These counterfactual results do not take the seller's pricing strategy into consideration. Sellers may respond to the positions of their products by adjusting their price levels. In Appendix A.4, I conduct a similar counterfactual experiment in which sellers adjusting their price discounts according to the positions. I find that the results stated above still hold.

Table 8.1: Counterfactual I: Effect of Disclosing Sales Information on Consumer Welfare

	Mean difference per session w/ 250 consumers (I - noI)	SE	p-value
All sessions			
Consumer net surplus	-0.139	0.088	0.114
(In RMB)	(-282.378 RMB)		
Number of clicks	-0.754	0.496	0.130
Utility from product	-0.131	0.066	0.048
(In RMB)	(-266.503 RMB)		
Bestselling products have the highest mean utility	31.45%		
Number of sessions	248		
Sessions w/ information harms consumers			
Consumer net surplus	-0.857	0.087	0.000
(In RMB)	(-1740.682 RMB)		
Number of clicks	2.114	0.411	0.001
Utility from product	-0.618	0.074	0.000
(In RMB)	(-1254.838 RMB)		
Bestselling products have the highest mean utility	20.18%		
Number of sessions	114		
Sessions w/ information benefits consumers			
Consumer net surplus	0.620	0.075	0.000
(In RMB)	(1258.901 RMB)		
Number of clicks	-4.196	0.572	0.000
Utility from product	0.371	0.049	0.000
(In RMB)	(754.497 RMB)		
Bestselling products have the highest mean utility	39.22%		
Number of sessions	102		

The search sessions are restricted to those with five products in the consumers' searching lists. Each session has 250 consumers. The ranking is randomized for each consumer. Note: I = disclosing sales information; noI = not disclosing sales information. t-test is used to obtain the significance level.

Table 8.2: Counterfactual I: Simple Method that Makes Sales Information Benefit Consumers

	Utility gain-based ranking			Sales ranking		
	Mean (I - noI)	SE	p-value	Mean (I - noI)	SE	p-value
Consumers (per consumer)						
Surplus	0.002	0.000	0.000	0.002	0.000	0.000
(In RMB)	(4.004 RMB)			(3.801 RMB)		
Number of clicks	-0.012	0.002	0.000	-0.011	0.002	0.000
Utility from product	0.001	0.000	0.001	0.001	0.000	0.001
(In RMB)	(1.781 RMB)			(1.631 RMB)		
Platform (per session w/ 250 *consumers)*						
Total sales (RMB)	23.984	9.063	0.009	23.387	10.998	0.034
Sales of first-party seller (RMB)	-3.758	3.167	0.237	-5.084	2.259	0.025
Bestselling products have the highest mean utility	75.28%			63.35%		
Number of sessions	248					

Using the proposed procedures, consumer surplus and total sales are higher when sales information is available under both utility-based and sales-based rankings. Note: I = disclosing sales information; noI = not disclosing sales information. t-test is used to the obtain significance level.

8.2 The feedback loop of popularity and lock-in effect

Consumers' active response to sales information can lead to a feedback loop in e-commerce platforms, that is, popular products are likely to be more attractive to subsequent consumers due to its disclosed popularity. To confirm the existence of the feedback loop, I test whether the initial ranking of products affects their sales volume in each of subsequent periods through the initial popularity when sales information is available to consumers. If initial popularity influences subsequent sales volume, then it indicates that popularity information, in addition to the intrinsic quality of products, can affect their long-run market shares. And this will be an evidence for the existence of the feedback loop. Then I further discuss different properties of the feedback loop.

I again use 248 search sessions with five products in consumers' search lists from the data and simulate consumer decisions in a dynamic setting. For each session, I generate 30 different random rankings of products in the initial period in which no sales information is available to any consumers. These rankings remain unchanged for the first five periods that I call them the initial treatment periods. In the subsequent periods, I simulate consumers' decisions in four different hypothetical scenarios: random ranking with sales information, random ranking without sales information, sales ranking with sales information, and sales ranking without sales information. In each period and search session, a group of 20 simulated consumers make clicking and purchasing decisions in each of these four scenarios. Since platforms usually do not update sales information and rankings continuously, I assume consumers in the same period observe the same rankings and sales information. A total of 75 periods are simulated after the initial five treatment periods to study the effect of initial positions in a dynamic setting. Each period is approximately one calendar

day, according to the summary statistics shown in Table 3.2.[1] Then for each period $T = 1, ..., 75$, I run a regression of $\triangle sales\,volume_T (= sales\,volume_T^{Info} - sales\,volume_T^{No\ Info})$ on $log(initial\,position)$. Here, $\triangle sales\,volume_T$ is the difference in the number of net sales at period T with and without sales information disclosed. The scenario without disclosing sales information can be seen as a benchmark of no feedback loop. Note that the sales information is cumulative, but the net sales volume in each period is noncumulative.

As shown in Table 8.2, the result of this counterfactual experiment suggests that initial positions have a (statistically) significant impact on the sale volume in subsequent periods when disclosing sales information to consumers. For example, with sales information disclosed and randomly shuffling the ranking after the initial treatment periods, in period $T = 1 - 5$, every 1% increase in the initial position (equivalent to decrease in the ranking) reduces the sales volume by around 0.014 comparing to the scenario with no sales information. This implies that if a product is ranked in the tenth position instead of in the top position, its sales volume will drop by around 14 per period on average during period $T = 1 - 5$. Notice that the effect of initial position on the sales volume decreases gradually and converges to a relatively stable level in the long run. Also, ranking products by sales volume makes the effect of the initial position converges faster than random rankings. One reason is that in the case of this platform, the ordinal ranking of sales volume converges to the quality ranking faster than the cardinal of sales volume, which is cumulative past sales, and the feedback loop is weaker when products are sorted by sales in this case.

To further study the properties of the feedback loop in general cases, Figure 8.1 explores how various factors influence the persistency of feedback loop by plotting the OLS estimator of $\triangle sales\,volume_T$ on $log(initial\,position)$ at different levels of search cost, numbers

1. In search sessions with three to five products in consumers' search lists, the median number of the sales volume is around 3000 in 44 months. Considering the fact that around 11% of searches end up with a purchase, this median leads to approximately 20 consumers per day that search each session.

of consumers per period, and length of initial treatment periods. The result shows that a larger search cost leads to an overall more persistent effect of the initial popularity, i.e., a stronger feedback loop. The lock-in effect of popularity caused by the feedback loop is almost doubled when the search cost increase from $\times 1$ to $\times 5$. Intuitively, consumers with higher search costs are less willing to make more clicks to discover the actual quality of other products. Therefore, the advantage of popular products persists longer with higher search costs as they are more likely to be clicked first. Furthermore, more consumers per period causes a more salient feedback loop. This is potentially due to the more purchases are affected by the initial position in the treatment periods, and it reinforces subsequent consumers' beliefs about the quality of products with initial advantages. Also, longer initial treatment periods lead to a more persistent lock-in effect. Intuitively, longer initial treatment periods let products gain more initial advantage/disadvantage of popularity, which takes longer for its weight in the cumulative sales information to drop through subsequent consumers' choices.

Figure 8.1 also shows that the effect of the initial position (or popularity) may or may not converge to 0 in the long run. For example, when there are five initial treatment periods with 100 consumers per period and $\times 10$ search cost, the initial-position effect maintains at around -0.12 level in the long run. In some other cases (e.g., Initial Period $= 1$, $n_{consumer} = 100$, and $\times 1$ search cost), this effect vanishes. One reason is that sales information is cumulative, so the initial advantage of popularity remains in this information across all periods. Based on the model assumption (equation 5.2), consumers perceive the observed sales information as a weighted sum of a quality-dependent choice probability and a random noise, so they usually hold higher expectations of the unrevealed product quality on high-selling products. When the search cost is low, or the initial popularity advantage is small, the weight of the initial popularity in sales information drops quickly across periods and cause the effect of initial popularity vanishing.

Another factor that affects the long-run steady state of the initial-popularity effect is the portion of actual buyers' reviews (i.e., μ) in equation 5.2, as it characterizes how well the observed sales information can represent the unrevealed product qualities. Figure 8.2 compares the long-run effect of initial popularity under different values of $\mu_{buyer} =$ E[portion of real buyers], which is the mean of a beta distribution characterizes the portion of real buyers' choices in sales information. Comparing to $\mu_{buyer} = 0.96$, which is the average portion of real buyers from the data, we find that lowering μ_{buyer} to 0.6 leads to a stronger and more persistent feedback loop. This may sound counterintuitive as a lower μ_{buyer} means a weaker correlation between sales information and unrevealed qualities. However, as consumers are assumed to have discrete supports over the unrevealed product qualities, it is likely that the observed market share of a high-selling product cannot be achieved by any possible value of product qualities in the discrete support through $P(\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta)) \in \triangle^{J_i}$, which is consumer's perceived choice probability of past consumers given product features $(\mathbf{X}^r, \mathbf{X}^u)$. Thus, if μ_{buyer} is too high, one's conditional expectation of unrevealed quality may not monotonically increase with the observed market share.[2] This property may chance if different $P(\mathbf{X}^r, \mathbf{X}^u, F_\beta(\Theta))$ is used. On the other hand, if μ_{buyer} is further reduced to 0.2, the effect of initial popularity converges to 0 with a higher chance due to the very weak correlation between sales information and actual quality of products. In sum, the persistence of the feedback loop depends on 1) consumers' beliefs about the portion of sales information actually depending on the product quality, and 2) the portion of initial sales advantage in the cumulative sales information in the later periods.

The counterfactual results indicate that the initial positions (or popularity) affect the sales volume in the subsequent periods. Thus, disclosing sales information leads to the

2. See Appendix A.6 for details.

existence of the feedback loop. Consumers who arrive later can discover the true quality of products through costly searching, and the lock-in effect decreases to a stable level. The persistence of the popularity varies with the search cost, the number of consumers, the level of initial advantages of popularity, and the correlation between sales information and actual product characteristics. Higher search costs make consumers less willing to further sample products, and further leads to a stronger feedback loop. Similarly, more consumers per period and more significant advantages of initial popularity also lead to a more salient lock-in effect. Here, I assume that consumers arrive at a uniform rate across all periods. In reality, a larger group of consumers is more likely to search for a product when it was initially released. In this case, the initial position may have a stronger effect.

8.3 Temporary popularity and long-run fairness of competition

Some sellers are more capable of influencing their temporary popularity than the other. For example, many e-commerce platforms not only allow third-party sellers to list their products but also sell products as first-party sellers or even sell their own brands. As consumers may notice, first-party sellers' products are usually at the top of search results even though most platforms claim that their ranking algorithms fairly display products using criteria such as the popularity. However, due to the existence of the feedback loop as shown in the last counterfactual, even if platforms indeed fairly rank products by their popularity (e.g., sales volume), their own products can still gain persistently more sales by influencing their temporary (or initial) popularity. In this counterfactual experiment, I investigate how assigning an initial (or temporary) prominent position to a product sold by the first-party seller affects its long-run sales and consumer welfare if all products are ranked by popularity afterward.

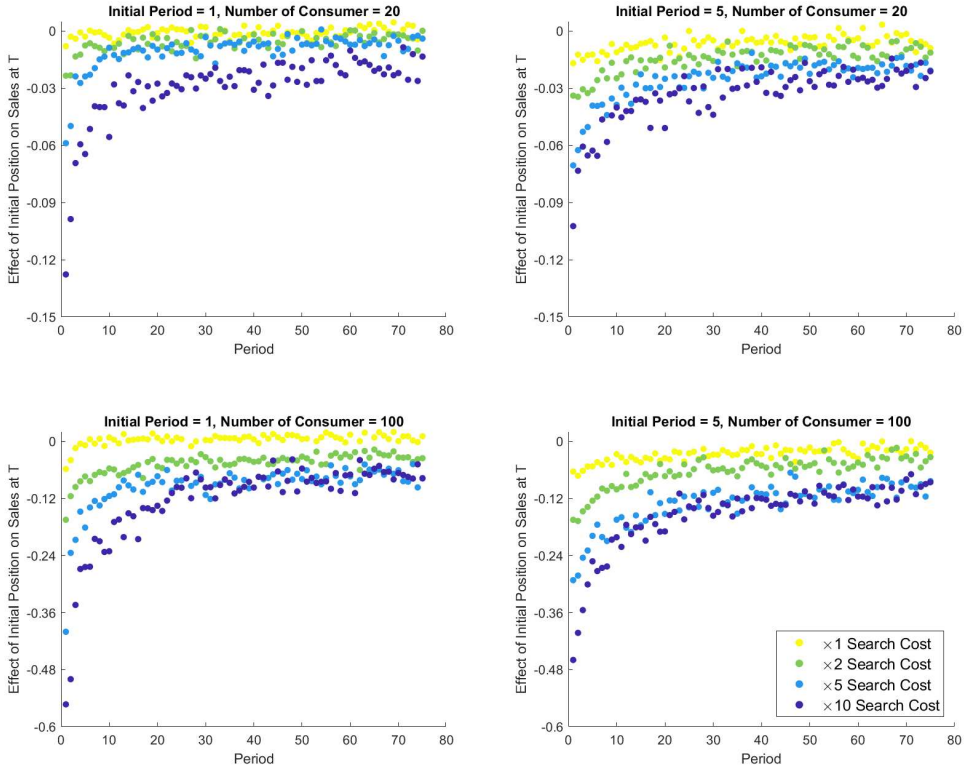
Table 8.3: Counterfactual II: Effect of Initial Positions on Long-Run Sales

	$T=1-5$	$T=6-10$	$T=11-15$	$T=16-20$	$T=21-25$
Random ranking after initial five periods					
	-0.0138***	-0.0114***	-0.0081***	-0.0075***	-0.0082***
	(0.0013)	(0.0012)	(0.0012)	(0.0013)	(0.0013)
Rank by sales volume after initial five periods					
	-0.0072***	-0.0085***	-0.0068***	-0.0054***	-0.0098***
	(0.0010)	(0.0010)	(0.0010)	(0.0011)	(0.0011)
	$T=26-30$	$T=31-35$	$T=36-40$	$T=41-45$	$T=46-50$
Random ranking after initial five periods					
	-0.0052***	-0.0072***	-0.0041***	-0.0047***	-0.0054***
	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0012)
Rank by sales volume after initial five periods					
	-0.0059***	-0.0063***	-0.0046***	-0.0031***	-0.0069***
	(0.0011)	(0.0011)	(0.0010)	(0.0011)	(0.0011)
	$T=51-55$	$T=56-60$	$T=61-65$	$T=66-70$	$T=71-75$
Random ranking after initial five periods					
	-0.0052***	-0.0026**	-0.0032***	-0.0027**	-0.0075***
	(0.0013)	(0.0012)	(0.0012)	(0.0013)	(0.0013)
Rank by sales volume after initial five periods					
	-0.0051***	-0.0036***	-0.0043***	-0.0056***	-0.0057***
	(0.0011)	(0.0010)	(0.0010)	(0.0012)	(0.0011)
Number of sessions	248				
Number of random initial positions per session	30				

$^{***}p<0.01,^{**}p<0.05,^{*}p<0.1$

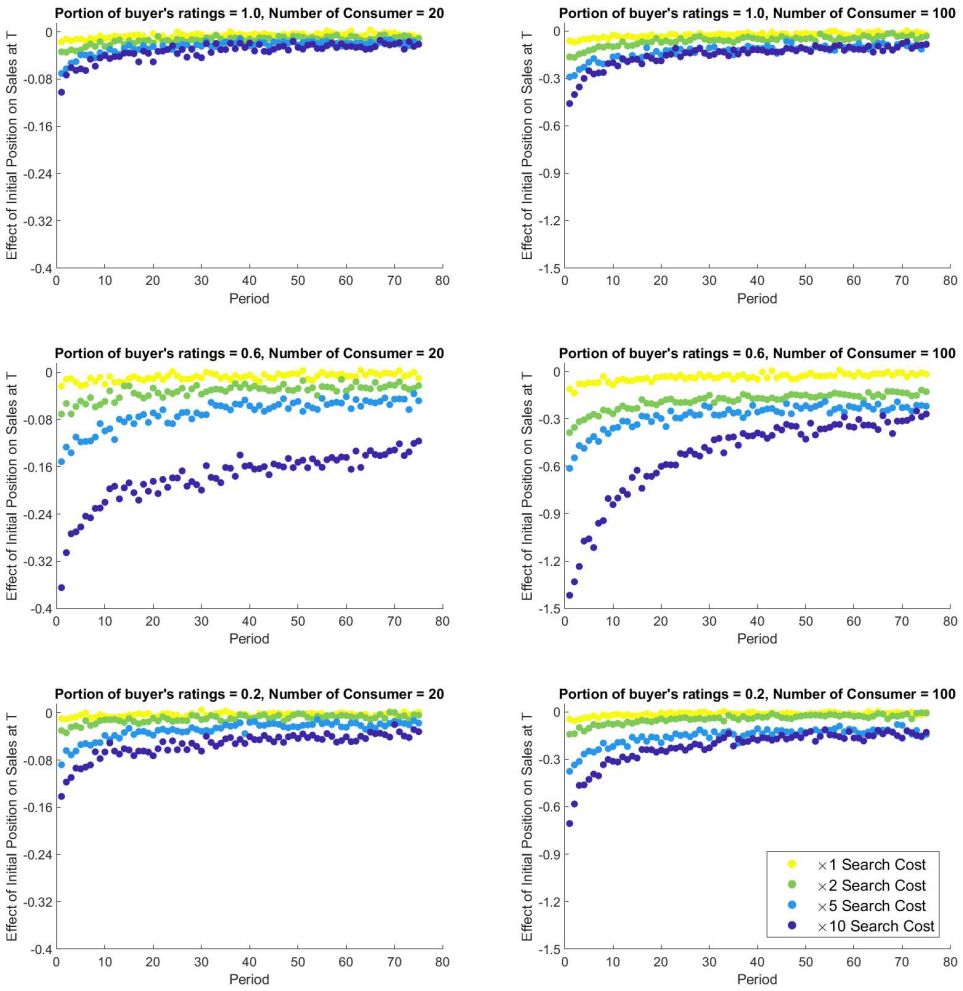
The standard errors are in parentheses. The OLS estimators of $\triangle sales volume_T$ on $log(position\ at\ T=1)$ across period T. A negative estimate implies that a larger initial position (i.e., lower ranking) leads to lower net sales at period T. Products are randomly ranked in the initial period and maintain these positions in the initial five control periods $T=0$. The sessions are restricted to five products in consumers' search lists. Randomly generate 30 initial positions and 20 consumers per initial position for each session and period.

Figure 8.1. Feedback Loop under Different Search Costs



The plot of OLS estimators of $\triangle sales volume_T$ on $log(position\ at\ T = 0)$ with different levels of the search cost, numbers of consumers per period, and numbers of periods maintaining the initial positions. Larger search costs leads to more persistent lock-in effects of the popularity.

Figure 8.2. Feedback Loop under Different Portions of Actual Buyers



The plot of OLS estimators of $\triangle sales\,volume_T$ on $log(position\ at\ T=0)$ with different levels of the search cost, numbers of consumers per period, and the mean portion of sales information generated by actual buyers, μ_{buyer}. There are five initial treatment periods in simulations.

Similar to the previous counterfactual, I use search sessions with five products in consumers' search lists to run the simulation. In each session and period, I simulate 20 consumers using estimated parameters. In the first five periods, products sold by the first-party seller are placed in top positions, and other sellers' products are ranked by their mean utilities. Further, sales information is unavailable to all consumers in the first period. In the subsequent periods, I compare outcomes in a utility-based ranking to two different ranking methods: sort products by sales volume, and maintain first-party sellers' products in top positions. I use a utility-based ranking as the benchmark because it generally leads to more efficient outcomes than other popular rankings for both consumers and platforms (Ghose, Ipeirotis, et al. 2012; Ghose, Ipeirotis, et al. 2014) if not considering private brands.

The results in Table 8.3 indicate that keeping the first-party seller's products in top positions for all periods leads to a significant loss in consumer welfare of around 5 to 6 RMB (~ 0.7 to 0.85 USD) per consumer compared to the benchmark ranking. While this ranking gives the first-party seller more sales ($\sim$ 5 RMB), it leads to significantly less total sales of around 10 to 12 RMB (~ 1.4 to 1.7 USD) for each session in each period. Thus, whether this ranking method makes the platform more profitable is questionable. On the other hand, if products are fairly ranked by their sales volume after the initial periods, consumers suffer from welfare loss in the short run in around 20 periods. In addition, the first-party seller has persistently more sales compared to the utility-based ranking: an increase of 2.6 RMB to 4.9 RMB per session. This increase means that the first-party seller on this platform can give an advantage to its own products in the competition even if products are fairly ranked by sales volume after the initial periods. Appendix A.7 shows the result of a similar counterfactual experiment when assuming sellers respond to their assigned positions by adjusting the prices of their products, and the result is similar to what I find here.

Table 8.4: Counterfactual III: Effect of Assigning Prominent Positions to Private Brands in Initial Periods

	$T=1-5$	$T=6-10$	$T=11-15$	$T=16-20$	$T=21-25$
Private brands at top for all periods					
Consumer surplus (per consumer)	0.003***	0.003***	0.003***	0.003***	0.003***
(in RMB)	5.307	5.634	6.492	5.913	6.747
Total sales per session (RMB)	12.177***	14.953***	13.868***	12.077***	13.138***
First-party seller's sales per session (RMB)	-4.830***	-5.194***	-4.671***	-4.625***	-5.036***
Private brands at top in T=1-5, then sales rank					
Consumer surplus (per consumer)	0.003***	0.002***	0.002***	0.002**	0.000
(in RMB)	5.307	4.757	4.122	3.126	0.802
Total sales per session (RMB)	12.177***	14.164***	5.796***	5.206**	5.583**
First-party seller's sales per session (RMB)	-4.830***	-3.182***	-2.587***	-4.138***	-4.405***
	$T=26-30$	$T=31-35$	$T=36-40$	$T=41-45$	$T=46-50$
Private brands at top for all periods					
Consumer surplus (per consumer)	0.003***	0.002***	0.003***	0.003***	0.002***
(in RMB)	6.377	4.219	5.621	5.118	4.979
Total sales per session (RMB)	12.756***	9.215***	10.261***	10.219***	9.893***
First-party seller's sales per session (RMB)	-4.707***	-4.710***	-4.586***	-5.222***	-5.049***
Private brands at top in T=1-5, then sales rank					
Consumer surplus (per consumer)	0.001**	0.000	0.001	0.001	0.001
(in RMB)	2.531	0.408	1.819	1.482	1.614
Total sales per session (RMB)	8.215***	4.875**	4.575**	7.246***	7.867***
First-party seller's sales per session (RMB)	-2.707***	-2.987***	-2.779***	-2.821***	-3.285***
Number of sessions	248				

$^{***}p<0.01,^{**}p<0.05,^{*}p<0.1$

Twenty consumers are simulated each session per period for 50 periods in total. Using utility-based ranking as the reference. All numbers are computed as the mean difference of the utility-based ranking minus the specific ranking method, i.e., utility-based ranking - proposed ranking method. Total sales and first-party seller's sales are computed for each session per period. t-test is used to obtain the significance level.

CHAPTER 9

CONCLUSION

Sales information is commonly available in online shopping and helps consumers make search decisions. In this paper, I study the role of sales information in consumer search decisions by using a data set from which consumers gain information about previous sales volume due to the unique rating system of the platform. Then, I show that consumers' responses to the available sales information along with the platforms' default ranking algorithms create a feedback loop in online platforms. This mechanism potentially leads to a long-lasting effect of having a temporary leading position in sales volume that raises concerns about the fairness of competition in the online marketplace with both first-party and third-party sellers.

I first show that the sales information serves as a signal of unrevealed product characteristics at the clicking stage, such that consumers are more willing to click on high-selling products. However, conditional on the clicking, the purchasing decision is independent of the popularity of products. Since I only focus on consumers who search for characteristics, such as sellers, prices, and versions, but not the book's content, all demand-related product characteristics are observed after clicking into the product pages. Given these findings, I estimate a sequential search model in which consumers infer the characteristics in the product page using sales information. This model provides a general framework to study the effect of sales information on consumer welfare. With it I can also investigate whether sellers can achieve unfair competitive advantages persistently in online marketplaces by influencing the short-run popularity of their products. A counterfactual experiment indicates that displaying sales information has heterogeneous effects on consumer welfare. I find that when sales information can correctly reflect the overall quality of products, disclosing this information achieves a higher total surplus. I show that simply let the initial group of

consumers search under a random ranking without disclosing sales information can make sales information better represent the quality of products in the subsequent periods and benefit those subsequent consumers. Then, I test the effect of initial positions on the sales volume of overall products in each subsequent period and find that a better initial position brings more sales in subsequent periods. This finding confirms the existence of the feedback loop. In particular, the persistence of the popularity due to the feedback loop depends on both consumers' beliefs about the correlation between observed sales information and the actual product characteristics, and the weight of initial popularity advantage in the cumulative sales information in the later periods. Using this featured data set that contains both first-party and third-party sellers, I specifically focus on the first-party seller and show that placing private-brand products at the top of lists in the initial periods increases their sales in the long run even if all products are fairly ranked by sales volume in subsequent periods.

In this paper, I do not study the case when consumers still have uncertainties about product quality after making clicks, which makes my findings more restrictive. In the case when uncertainties are still unsolved after clicking, one conjecture is that consumers' purchasing decisions will also be affected by the previous consumers' choices due to observational learning. Thus, the effect of disclosing sales information will be more salient as it affects not only the clicking decision but also purchasing decisions. I expect the effects of manipulating temporary popularity will be more persistent in this case. Another limitation of this paper is when I model the posterior belief of unrevealed characteristics, the previous consumer's choice probability is not theory-rooted. One difficulty in modeling the learning component under an equilibrium framework is the computational feasibility. In my data, the aggregate historical sales information contains around $40,000$ consumers on average for each search session. Further, consumers only observe the aggregate past choice histories but not the individual choices, and previous consumers may search with different

keywords and observe different search-result pages. All these factors make the modeling of learning under an equilibrium framework infeasible. Also, in my current analysis, I still face potential endogeneity issues due to unobserved book-level and seller-level characteristics. I may be able to solve this issue by adding fixed effects, if I have enough observations of the same book sold by different sellers and different books sold by the same sellers, or if I observe plenty of different consumers search with the same keyword and there are variations in their search results.

One way to study this question under an equilibrium framework is by focusing on products newly listed for around 10 days on this platform. Since the default rating is not assigned until 10 days after the purchase, the consumers who arrive at day 11 will know that the previous purchases were made without sales information, and the ranking will be less likely to vary. Furthermore, I can extend my findings in this paper to study the threshold of initial advantages that a seller needs to gain to change its market share in a long-run equilibrium under a feedback loop. With knowing this threshold, I can derive the optimal (minimum) amount of time the seller needs to pay for the sponsored ads or other promotions to maximize its long-run profits.

REFERENCES

Athey, S., & Ellision, G. (2010). Position Auction with Consumer Search. *Quarterly Journal of Economics*, (2004), 1–61.

Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*.

Chen, Y., & Yao, S. (2017). Sequential search with refinement: Model and application with click-stream data. *Management Science*.

Chevalier, J. A., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*.

De Los Santos, B., Hortacsu, A., & Wildenbeest, M. R. (2012). *Testing models of consumer search using data on web browsing and purchasing behavior* (tech. rep. No. 6).

De los Santos, B., & Koulayev, S. (2017). Optimizing Click-Through in Online Rankings with Endogenous Search Refinement. *Marketing Science*.

Etzion, H., & Awad, N. F. (2007). Pump up the Volume? Examining the Relationship between Number of Online Reviews and Sales: Is More Necessarily Better? In *Icis*.

Ghose, A., Goldfarb, A., & Han, S. P. (2011). How is the mobile internet different? Search costs and local activities. In *International conference on information systems 2011, icis 2011*.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*, *31*(3), 493–520.

Ghose, A., Ipeirotis, P. G., & Li, B. (2014). Examining the Impact of Ranking on Consumer Behavior and Search Engine Revenue. *Management Science*.

Gu, N. (2016). Consumer online search with partially revealed information. *Working Paper*.

Heckman, J. (1979). Sample Specification Bias as a Selection Error. *Econometrica*.

Hendricks, K., Sorensen, A., & Wiseman, T. (2012). Observational learning and demand for search goods. *American Economic Journal: Microeconomics*.

Honka, E. (2014). Quantifying search and switching costs in the US auto insurance industry. *The RAND Journal of Economics*, *45*(4), 847–884.

Honka, E., & Chintagunta, P. (2017). Simultaneous or Sequential? Search Strategies in the U.S. Auto Insurance Industry. *Marketing Science*, *36*(1), 21–42.

Honka, E., Hortaçsu, A., & Vitorino, M. A. (2017). Advertising, consumer awareness, and choice: evidence from the U.S. banking industry. *RAND Journal of Economics*, *48*(3), 611–646.

Kim, J. B., Albuquerque, P., & Bronnenberg, B. J. (2010). Online Demand Under Limited Consumer Search. *Marketing Science*, *29*(6), 1001–1023.

Kim, J. B., Albuquerque, P., & Bronnenberg, B. J. (2017). The probit choice model under sequential search with an application to online retailing. *Management Science*.

Kuksov, D., & Villas-Boas, J. M. (2010). When more alternatives lead to less choice. *Marketing Science*.

McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*, *57*(5), 995.

Mueller-Frank, M., & Pai, M. M. (2016). Social learning with costly search. *American Economic Journal: Microeconomics*.

Nelson, P. (1970). Information and Consumer Behavior. *Journal of Political Economy*.

Petrin, A., & Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*. arXiv: 0410550 [cond-mat]

Smith, L., & Sørensen, P. (2000). Pathological outcomes of observational learning. *Econometrica*.

Stigler, G. J. (1961). The Economics of Information. *Source Journal of Political Economy*, *63*(6), 213–225.

Ursu, R. M. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*.

Weitzman, M. (1979). Optimal Search for the Best Alternative. *Econometrica*, *47*, 641–654.

Zhang, J. (2010). The Sound of Silence: Observational Learning in the U.S. Kidney Market. *Marketing Science*, *29*(2), 315–335.

APPENDIX A

APPENDIX

A.1 Keyword-title similarity score

In this subsection, I illustrate how I construct the keyword-title similarity score using the Levenshtein distance, which is a special case of edit distance. The Levenshtein distance between two strings is defined as the minimum number of edit operations, which includes insertion, deletion, and substitutions, required to convert one string into another. For example, if I have two strings, "Harry" and "Happy", the minimum number of operations is two. I further develop a series of steps to apply this measure to compare search keywords and book titles in Chinese. Formally,

1. Split the keyword into sub-keywords if there are non-word characters, such as space, $+$, and $-$.

    - For example, if the keyword is "Harry Potter + J.K. Rowling", then there are two sub-keywords, "Harry Potter" and "J.K. Rowling".
    - Note, in Chinese, a phrase or word consists of successive letters, instead of being split by spaces as in the English. So, when there is a space, it is usually considered as a divider between two sub-keywords.

2. Using a word segmentation tool for Chinese to determine the number of segments of sub-keyword and book title

    - For example, if the keyword "Harry Potter written by J.K. Rowling" contains word segments "Harry Potter", "written by", and "J.K. Rowling", then there are three segments of words in the keyword.

3. Suppose there are n_{kw} segments in a sub-keyword and n_t in the title. $[\max\{n_{kw}-1,1\},\min\{n_{kw}+1,n_t\}]$ is the range of sliding windows that is the number of segments in the title I use to compute the Levenshtein distance with the keyword each time.

    - For example, if the keyword is "Harry Potter" and the book title is "Harry Potter and the Deathly Hallows". And if $n_{kw}=2$ and $n_t=6$, then the range of sliding window is 1 to 3. When the sliding window is 2, I compare the keyword "Harry Potter" to each of {"Harry Potter", "Potter and", "and the", "the Deathly", "Deathly Hallows"} . And when it is 3, I compare the keyword to each of {"Harry Potter and", "Potter and the", "and the Deathly", "the Deathly Hallows"}.

4. For a sub-keyword kw_j and each size of sliding window, I compute the Levenshtein distance between the sub-keyword kw_j and each segment i of the title. Suppose the distance is D_{ji}, then the corresponding similarity score is $S_{ji}=1-\frac{D_{JI}}{\max\{\text{length of } kw_j,\text{length of segment} i\}}$. Given the size of a sliding window, the similarity score between kw_j and the entire title is the maximum over all segment i's, that is, $S_j=\max_i S_{ji}$.

    - For example, when the size of the sliding window is 2, comparing the keyword "Harry potter" to each of {"Harry Potter", "Potter and", "and the", "the Deathly", "Deathly Hallows"}, the similarity score $S_j=1$.

5. The similarity score between a sub-keyword and the title is the maximum similarity score over all sizes of sliding window.

6. The similarity score between the keyword and the title is the average similarity score over sub-keywords.

A.2 More discussion about third-party sellers

Here, I provide more information about services provided by third-party sellers, as well as the distributions of their quality scores. Third-party sellers list services that they guarantee to offer in the product pages. Consumers also observe third-party sellers' quality scores in the same page. The data shows that third-party sellers have very similar guaranteed services and quality scores.

Table A.1: Summary of Third-Party Sellers' Services (in %)

	(1)	(2)
VARIABLES	Mean	Std. Dev.
Authentic guarantee	0.935	0.237
Shipping in 24 hours	0.763	0.417
Unconditional return in 7 days	0.992	0.0642
Number of sellers	815	

Figure A.1. Distribution of Third-Party Sellers' Quality Scores

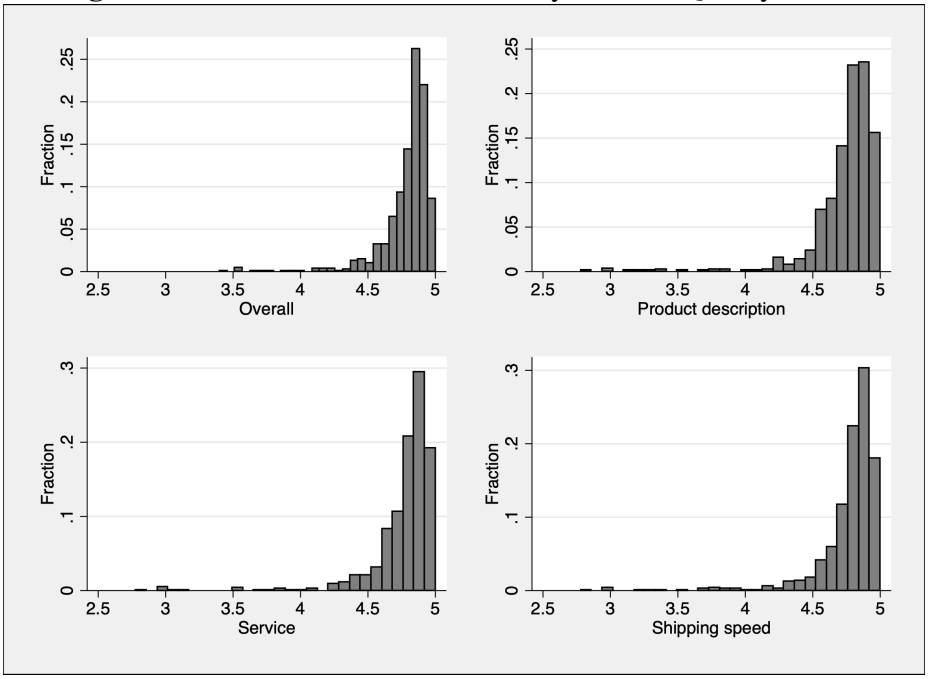


Table A.2: Effect of Third-Party Seller's Quality Scores and Sizes on Consumer Choices

VARIABLES	(1) Purchased	(2) Purchased	(3) Purchased	(4) Purchased
Search-result page				
log(sales volume)	0.00553	-0.0376	0.0143	-0.0435
	(0.0621)	(0.0919)	(0.0629)	(0.0929)
Position	-0.000450	0.0304	-0.00289	0.0393
	(0.0241)	(0.0578)	(0.0242)	(0.0584)
Price plus shipping	-0.0141**	-0.0140**	-0.0136**	-0.0136**
	(0.00599)	(0.00588)	(0.00593)	(0.00572)
SS by platform	0.247	-0.317	0.265	-0.394
	(0.424)	(1.116)	(0.581)	(1.077)
Bundle	0.0675	0.224	0.0970	0.314
	(0.342)	(0.413)	(0.344)	(0.412)
Regular edition	-0.694	-0.728	-0.728	-0.777
	(0.535)	(0.548)	(0.552)	(0.566)
Years since published	0.00537	0.0190	0.00616	0.0246
	(0.0565)	(0.0584)	(0.0558)	(0.0579)
Original price(MSRP)	0.0123**	0.0119**	0.0115**	0.0111**
	(0.00542)	(0.00541)	(0.00527)	(0.00510)
Product page				
Number of bad reivew (in 1000)	-0.417***	-0.422***	-0.443***	-0.453***
	(0.140)	(0.139)	(0.139)	(0.139)
Sales rank in the category	0.0958	0.115	0.0879	0.114
	(0.103)	(0.108)	(0.103)	(0.108)
log(Length of introduction)	-0.152	-0.150	-0.161	-0.159
	(0.125)	(0.124)	(0.126)	(0.124)
Seller score			-2.504	-2.367
			(1.866)	(1.932)
Large 3rd-party seller			-0.0793	0.0791
			(0.485)	(0.525)
Endogeneity controls				
Mills ratio		-0.807		-1.102
		(1.392)		(1.419)
Observations	465	465	465	465

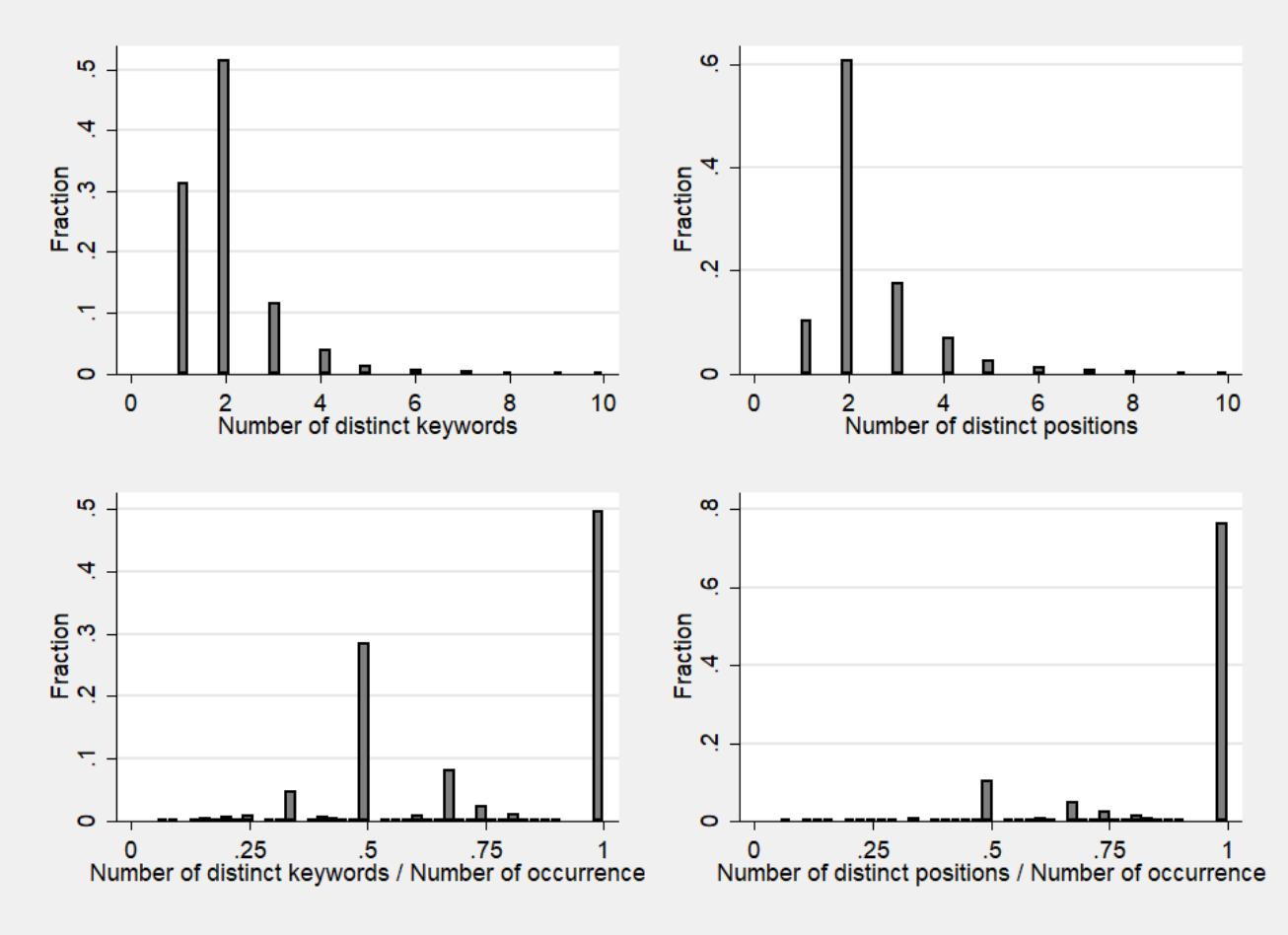
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Consumers observe four scores of third-party seller quality in the product page. The purchase decision isn't affected by seller qualities significantly. Consumers also do not respond to seller's size.

A.3 Further evidence: price endogeneity

In addition to the possible causes of price endogeneity discussed in subsection 4.1, sellers may also strategically pricing their books given their market shares. However, strategically setting prices given sales volumes is no easy on this platform as it is quite common for a product to be listed under different search results with different keywords and at different positions on this platform. In Figure A.2, I select those products exist more than once in the data, and plot the distribution the occurrences of these products displayed in different searched keywords and positions. It suggests that a product is likely to be listed with different sets of other products at different positions, so it is overall a difficult task for sellers to set price strategically given the market share and positions.

Figure A.2. Distribution of the Occurrences of Same Product in Different Keywords and Positions



A.3.1 Normalization of the outside option

The mean utility level of the outside option can be identified as I observe some consumers who search for products without purchasing any, that is, choosing the outside option. However, in practice, due to the functional form of the search cost, the identification can be very weak. This is mainly because the exponential form of the search cost becomes almost linear in α_0 when it is tiny, and affected by the mean utility of outside option. The difficulty in pinning down the search cost parameters further affects the estimates of other utility parameters.

Tables A.3.1 shows the Monte Carlo simulation under two different model specifications, with the true specification, i.e., the nonzero mean utility of outside goods with $V_0 = 2$. I find that for most parameters, normalizing V_0 to zero in the estimation leads to better estimates.

Table A.3: Test Identification under a Different Specification of the Outside Option

		Nonzero V_0		Zero V_0	
	True values	Estimates	SE	Estimates	SE
Search-result page					
No rating dummy	-1	-0.924***	0.095	-1.028***	0.086
S&S by platform	1	0.831***	0.063	0.858***	0.062
MSRP	0.01	0.008***	0.001	0.008***	0.001
Bundle dummy	1	0.900***	0.075	0.949***	0.074
Non-premium edition	-1	-0.932***	0.078	-1.068***	0.079
log(Years since published)	-1	-0.944***	0.046	-1.008***	0.046
Product page					
% of bad rating (in 0.1%)	-5	-2.934**	1.061	-4.075***	0.850
Sales rank in category	1	0.843***	0.036	0.877***	0.035
log(length of introduction)	0.2	0.178***	0.027	-0.007***	0.015
No introduction dummy	-1	-1.468***	0.341	-2.511***	0.210
Price					
Mean	-5	-5.159***	0.361	-5.008***	0.332
Heterogeneity	1	0.813***	0.199	0.811***	0.194
Search cost					
Constant	-3	-2.375***	0.174	-2.491***	0.154
Position	0.05	0.040***	0.006	0.041***	0.005
PC device dummy	-1	-1.167***	0.186	-1.003***	0.178
Outside option					
V_0	2	1.503***	0.214		
Number of search sessions	800				

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

Estimation using simulated data results with $w = 13$. The correct specification has the mean utility of the outside option as $V_0 = 2$. Comparing the parameter recovery under two different specifications in the estimation indicates that normalizing the mean outside option to zero may lead to better estimates.

A.4 Counterfactual I'

In this subsection, I show the counterfactual results following the same procedures as section 8.2 but assuming sellers adjust price level according to the positions. Thus, this is closer to the case of a competitive equilibrium that sellers also respond to their positions. Here, I use a simple method to incorporate the effect of position on the level of the price discount, in which I first run a regression of the discount level on the seller's type (e.g., first-party seller, and large seller) and the position, then generate the predicted price discount at the new position. Table A.4 shows the results, and I find that the overall conclusions are the same as in subsection 8.2.

Table A.4: Counterfactual I' (when sellers respond to positions by adjusting prices)

	Utility gain-based ranking			Sales ranking		
	Mean	SE	p-value	Mean	SE	p-value
Consumers (per consumer)						
Surplus	0.002	0.000	0.000	0.002	0.000	0.000
(In RMB)	(3.884 RMB)			(3.738 RMB)		
Number of clicks	-0.012	0.002	0.000	-0.011	0.002	0.000
Utility from product	0.001	0.000	0.001	0.001	0.000	0.000
(In RMB)	(1.714 RMB)			(1.641 RMB)		
Platform (per session w/ 250 *consumers)*						
Total sales (RMB)	27.513	9.364	0.004	27.235	9.787	0.006
Sales of first-party seller (RMB)	-1.320	2.524	0.602	-2.646	2.267	0.244
Bestselling products has highest mean utility	74.35%			63.31%		
Number of sessions	248					

Assuming the price of a product adjust with its position by empirically computing sellers' pricing strategy at different positions.

A.5 Default ranking, rank of sales, and rank of mean utilities

This appendix compares the default ranking, the sales rank, and the rank of mean utilities. Table A.5 shows the portion of products in the top position of the default ranking with the highest sales volume (and the highest mean utility). It suggests the sales rank has strong correlation with the system default ranking. Products sold by the first-party seller are more likely to gain high market share that does not match their qualities.

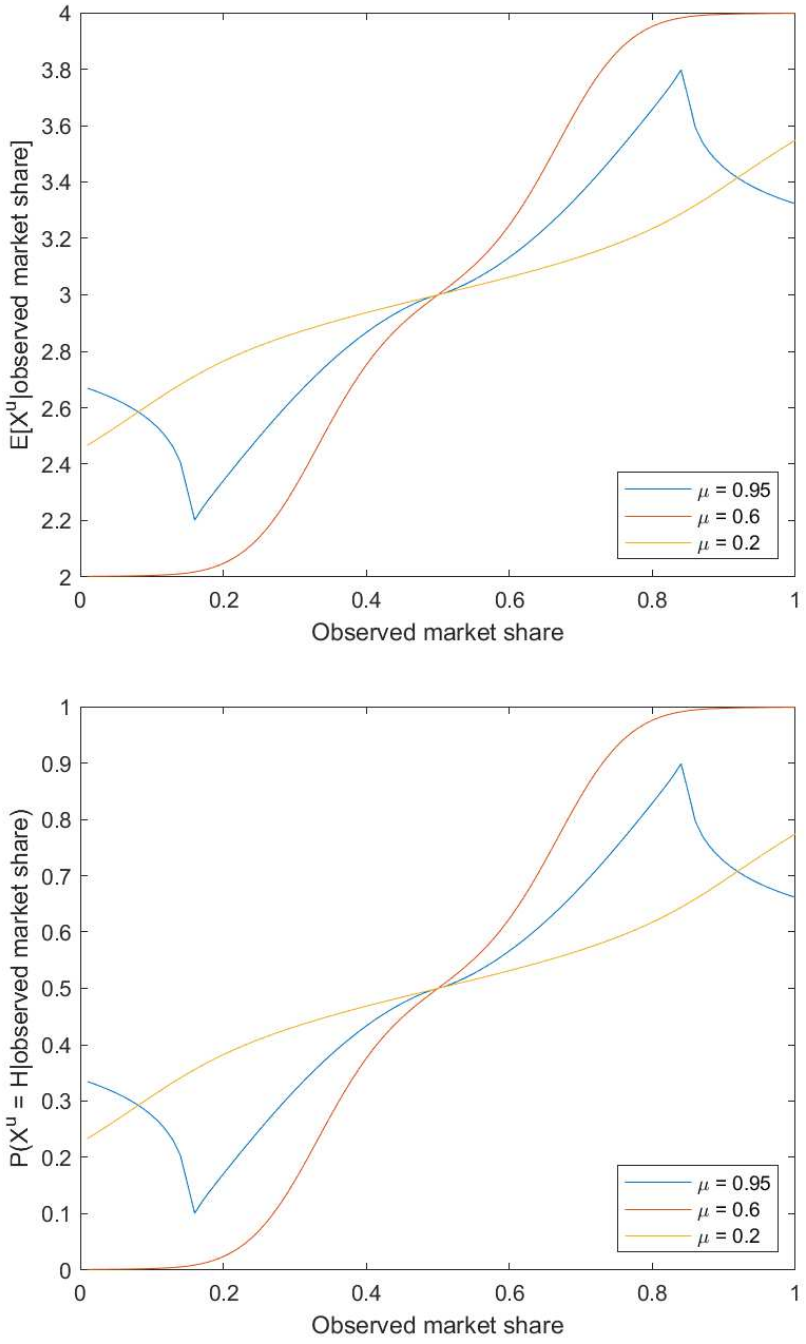
Table A.5: Sales Volumes and Mean Utilities of Top-Ranked Products under the Default Ranking

	w/ the highest sales volume	w/ the highest mean utility
All	73.11%	38.29%
Exclude the first-party seller	77.45%	52.61%

A.6 Sales information, consumers' expectation of unrevealed product quality, and portion of sales from real buyers

Using a toy example, this appendix shows how consumer's expectation of unrevealed product quality varies with sales information and the mean portion of sales from real buyers. As shown in Section 5, a consumer form a posterior belief over the unrevealed product quality given sales information (market share) of each product. Here, I use the same model setup as in Section 5, and create a hypothetical scenario with two products, which have the same observed product characteristics, i.e., $X_1^r = X_2^r = 3$. Suppose the unrevealed product quality $X_j^u \in \{H, L\}$ with $H = 4$ and $L = 2$ in this toy example. Let I_j denote the observed market share of product j. Figure A.3 compares $\mathrm{E}[X_j^u | I_j]$ and $\Pr(X_j^u = H | I_j)$ under different μ's, i.e., which is the mean of a beta distribution that characterizes the portion of sales information from real buyers. The result suggests under some extreme values of μ's, $\mathrm{E}[X_j^u | I_j]$ and $\Pr(X_j^u = H | I_j)$ does not monotonically increase with the observed market share I_j.

Figure A.3. Unrevealed Product Quality and Sales Information



A.7 Counterfactual III'

Table A.6: Counterfactual III' (when sellers respond to positions by adjusting prices)

	$T=1-5$	$T=6-10$	$T=11-15$	$T=16-20$	$T=21-25$
Private brands at top for all periods					
Consumer surplus (per consumer)	0.002***	0.003***	0.003***	0.003***	0.003***
(in RMB)	4.589	5.365	6.152	5.318	6.803
Total sales per session (RMB)	12.478***	15.074***	15.015***	12.500***	14.966***
First-party seller's sales per session (RMB)	-5.110***	-5.611***	-4.731***	-5.026***	-5.371***
Private brands at top in T=1-5, then sales rank					
Consumer surplus (per consumer)	0.002***	0.001**	0.003***	0.002***	0.000
(in RMB)	4.589	2.988	5.215	3.721	-0.021
Total sales per session (RMB)	12.478***	12.047***	9.328***	7.852***	4.985**
First-party seller's sales per session (RMB)	-5.110***	-2.521*	-2.526***	-4.441***	-4.634***
	$T=26-30$	$T=31-35$	$T=36-40$	$T=41-45$	$T=46-50$
Private brands at top for all periods					
Consumer surplus (per consumer)	0.003***	0.002***	0.002***	0.002***	0.002***
(in RMB)	6.330	4.823	4.878	4.957	3.948
Total sales per session (RMB)	13.997***	10.549***	12.526***	12.164***	10.603***
First-party seller's sales per session (RMB)	-4.926***	-5.077***	-5.178***	-5.197***	-5.147***
Private brands at top in T=1-5, then sales rank					
Consumer surplus (per consumer)	0.001	0.000	0.001	0.001	0.000
(in RMB)	3.011	0.516	1.874	1.792	0.745
Total sales per session (RMB)	5.611***	4.364**	4.492***	4.783***	4.908***
First-party seller's sales per session (RMB)	-3.355***	-3.713***	-3.585***	-2.979***	-3.317***
Number of sessions	248				

$^{***}p<0.01,^{**}p<0.05,^{*}p<0.1$

Assuming the price of a product adjust with its position by empirically computing sellers' pricing strategy at different positions.