Platform Information Provision and Consumer Search: A Field Experiment*

Lu Fang
Zhejiang University

Yanyou Chen
University of Toronto

Chiara Farronato
Harvard University, CEPR, NBER

Zhe Yuan
Zhejiang University

Yitong Wang
Alibaba Group

This Version: *December 4, 2024*

Abstract

Despite substantial efforts to help consumers search in more intuitive ways, text search remains the predominant tool for product discovery online. In this paper, we explore the effects of visual and textual cues for search refinement on consumer search and purchasing behavior. We collaborate with one of the largest e-commerce platforms in China and study its roll out of a new search tool. When a consumer searches for a general term (e.g., "headphones"), the tool suggests refined queries (e.g., "bluetooth headphones" or "noise-canceling headphones") with the help of images and text. The search tool was rolled out with a long-run experiment, which allows us to measure its short-run and long-run effects. We find that, although there was no immediate effect on orders or spending, in the long-run the search tool changed consumers' search and purchasing behavior. In the six months following entry into the experiment, consumers with access to the new tool substantially increased orders and spending compared to those in the control group, especially for non top-selling products. The purchase increase comes from more effective searches, rather than an increase in activity on the platform. We also find that the effect is not only driven by the direct value of suggested searches, but also by consumers indirectly learning to conduct more specific searches on their own.

Keywords: E-commerce, Digital Platforms, Consumer Search, Learning, Search Recommendations

*Contact: Chen: yanyou.chen@utoronto.ca; Fang: fl_fanglu@zju.edu.cn; Farronato: cfarronato@hbs.edu; Yuan: corresponding author, yyyuanzhe@gmail.com. Any opinions, findings, conclusions and recommendations expressed are those of the authors and do not necessarily represent the views of the focal platform. We are grateful to Avi Goldfarb, Joshua Gans, Heski Bar-Isaac, Elisabeth Honka, and seminar participants at the University of Toronto and Zhejiang University for helpful discussions and comments. All errors are our own. The authors confirm equal contribution among Chen, Fang, Farronato, and Yuan. We thank Yitong Wang for helpful business insights and for facilitating the data collaboration. Yuan would like to acknowledge financial support from the NSFC (Grants 72203202, 72192803 and 72141305). Fang would like to acknowledge financial support from the NSFC (Grant 72192803). Wang is an employee of the company that shared the data for this research.

1 Introduction

One of the most important roles of digital platforms is to facilitate matches between a wide range of buyers and sellers. By developing increasingly sophisticated ranking algorithms, platforms have invested substantial effort into making search results as relevant as possible. These functionalities operate most effectively when consumers know what they are looking for and how to describe their needs. Less emphasis, however, has been placed on helping consumers effectively identify and articulate their preferences.

In this paper, we study the role of search recommendations in helping consumers express and develop their preferences. We examine the launch of a search tool that recommends more refined searches through text and images on one of the world's largest e-commerce platforms. Our goal is to evaluate whether and how much enhancing search with textual and visual suggestions can effectively assist consumers in finding products online.

Our partner platform, like many other e-commerce sites, allows consumers to input their search queries in text form. Two potential challenges arise with the commonly used text-based search process. First, consumers may have a good understanding of their needs but lack knowledge of the corresponding search terms (Liu and Toubia, 2020). For example, a user may know they want cordless headphones, but may not know that *bluetooth* is the typical technology to connect such headphones to their electronic devices. This challenge, known as *demand expression*, arises when consumers struggle to articulate their requirements effectively while conducting searches.[1]

Second, consumers might have a general idea of what they want, but lack specific information about the characteristics of the products available, and hence of the products they ultimately prefer. For example, a user may know they want headphones, but do not know that they can choose between over-ear or in-ear headphones. This challenge is often referred to as *demand formation*.[2]

To address both demand expression and formation challenges, some platforms have adopted auto-complete technology, offering consumers query suggestions based on their past searches or aggregate search behavior (Hagiu and Wright, 2023). What is unique about our context is the integration of both visual and textual features to guide consumers, providing

[1]Demand expression seems to be an important challenge in online search, at least judging from the number of websites with tips for more efficient search strategies (Markey, 2019). For example, see the tips on Lifewire, TechRepublic, Indeed, or MediaSmarts. Yet, the existing literature remains limited (Lazonder, 2005).

[2]Prior research has demonstrated that recommendation systems can influence consumers' consideration sets, help them identify what they want (Häubl and Murray, 2003; Fong, 2017; Wan et al., 2023; Yuan et al., 2024) and how much they are willing to pay (Adomavicius et al., 2013, 2018, 2019). Such results support the hypothesis that consumers may develop their demand while searching, rather than searching for what they already know they want.

a more direct and intuitive way to address these challenges.

We collaborate with one of the largest e-commerce platforms, which we keep anonymous as part of our research agreement. Given the vast information the platform possesses, it can leverage historical order data and viewership patterns to help consumers search more effectively. Our focus is on a new search tool that, for a subset of query words, suggests refined queries by combining pictures and text, thus enabling consumers to narrow down their choices. When defining those refined queries, the platform utilizes both collaborative filtering models and human curation, taking into account factors such as query popularity and common patterns in search behavior across related terms.

We exploit the experimental roll out of the search tool, which has two helpful features for our analyses. First, the new search tool was randomly made available to a subset of the platform's consumers, allowing us to estimate the causal effects of the tool. Second, the experimental period lasted for approximately ten months, allowing us to quantify both the short-run and long-run effects of the search tool. As we emphasize in the next section, there is remarkably limited empirical work identifying the effects of improved search functionalities, with a few exceptions including Lee et al. (2020), Lei et al. (2023), and Zheng et al. (2023). Even less evidence exists on the long-term effects of such functionalities and the learning value they provide to consumers, who can increase the effectiveness of their searches by imitating the tool, even when those search functionalities are not available.

Our results show that the search tool was immediately effective at changing consumers' search behavior. On the first day they entered the experiment, consumers in the treatment group searched 495% more for queries suggested by the search tool compared to consumers in the control group. This substantial percent increase highlights how rarely consumers perform narrow searches at baseline, without recommendations. Despite the change in search behavior, however, the tool had no immediate effect on consumer transactions, measured as either the number of orders placed or total expenditures.

The long-run effects paint a very different picture. In the following 24 weeks since entering the experiment, consumers in the treatment group spent 3.2% more and completed 1.6% more orders compared to the control group. These are large improvements, especially given the fact that by the end of the experimental period, only 15% of spending is directly linked to searches supported by the tool. These results are not explained by increased activity on the platform such as conducting more searches or viewing and clicking on more products. Instead, the introduction of the search tool altered the distribution of products viewed, raising the expected match quality. This shift resulted in an increase in orders and spending, along with higher ratings and lower returns.

We find evidence that the increase in consumer spending does not only come from searches

directly affected by the new search tool, but also spills over to other searches on the platform. In particular, we confirm that consumers learn to perform more specific searches. Indeed, treated consumers use longer search terms and adopt words suggested by the recommendation tool in their own independent searches.

We highlight important heterogeneous effects across different demand and supply segments. On the demand side, the tool proves particularly beneficial for older consumers, who are likely more used to offline search, and for heavy platform users, whose search frequency leads them to encounter the recommendation tool more often.

On the supply side, the tool enables more specific, targeted searches, which particularly benefit tail products and smaller sellers. Additionally, the tool is especially effective in product categories where search costs are higher, i.e., categories where each purchase requires multiple searches, and categories where sales are not concentrated among a few limited products.

Our results have important implications for search and experiment design. While consumers have private information over what they want to search online, they can benefit from recommendations that help refine their searches and inspire their interests towards products they may not ex-ante know they want or how to describe. Our results suggest that the current design of search mechanisms may still overly rely on consumers' prompts, despite platforms having extensive knowledge about consumer preferences, in the aggregate as well as at the individual level. Our paper highlights the significant opportunities for platforms to leverage vast data on consumer preferences to help consumers more effectively express and identify their needs.

Second, our findings suggest that platforms need to exercise patience in measuring the effects of search refinement tools, because it takes time for their effects to emerge. Yet, platform companies typically run experiments (or A/B tests) for short periods. Our results — showing null effects on purchasing behavior in the short run and large positive effects in the long run — highlight the risk of drawing conclusions from short-run experiments, especially in contexts like ours, where consumer behavior may change significantly but slowly due to the treatment.

The heterogeneous effects can help platforms identify which users and product categories to target with recommendation tools like the one we study. In addition, the fact that niche products benefit from this tool has broader platform implications. On one hand, platforms want to present users with relevant results for their searches, which tends to favor products and sellers that have been popular so far. On the other hand, allowing for smaller and new sellers to be discovered is key to ensure that the platform offerings follow the evolution of consumer preferences. Our findings demonstrate that the search refinement

tool not only boosts sales for small sellers and niche products, but also enhance consumer satisfaction. Therefore, such tools may help foster healthier platform dynamics and growth without sacrificing consumer experience. We expand on these and other implications of our results in the concluding section.

The paper is structured as follows. Section 2 describes the existing literature to which our paper contributes. Section 3 presents the institutional setting, the experiment, and the data available. Section 4 focuses on our empirical approach and results, which are divided into short-run and long-run results. Finally, Section 5 concludes the paper, highlighting the managerial implications of our results.

2 Literature

Our paper contributes to the literature on consumer search behavior, on platform design, and on field experiments. Since at least Stigler (1961), McCall (1965), Gardner (1970), Mortensen (1970), Weitzman (1979), and Rothschild (1974), researchers have been interested in understanding how people search (Greminger et al., 2023). The advent of search engines and digital platforms have allowed empirical tests of the theories (Santos et al., 2012; Honka and Chintagunta, 2017; Ursu, 2018), as well as quantifications of search frictions (Ellison and Ellison, 2009; Lee and Musolff, 2021; Greminger, 2022; Honka et al., 2024). More recently, Bronnenberg et al. (2016) describe online consumers' search behavior while Seiler (2013) shows that search frictions significantly impact purchases. Choi et al. (2018) focus on the unexpected consequences of lowering search frictions. In this paper, we identify consumers' inability to express in words what they are looking for as a source of frictions, and how visual and textual cues, designed to help consumers refine their searches, can be an effective solution. Although we cannot separate the role of visual and textual refinements, existing work has demonstrated that pictures play an important role in facilitating consumers' information acquisition and processing (Blanco et al., 2010; Wu et al., 2021).

There is recent growing research on platform design (Zhang et al., 2023), specifically on how platforms present relevant options and what type of information they disclose about them. In the context of eBay, Dinerstein et al. (2018) is one of the earliest works looking at how search ranking algorithms play a critical role in reducing search frictions and changing competition, ultimately determining market outcomes and welfare. More recent studies on the effects of changing how products are presented to consumers include Chen et al., 2023 and Yang et al., 2023. Several papers focus on identifying what type of information platforms should disclose (Filippas et al., 2022). For example, Fradkin (2017) and Filippas et al. (2023) highlight the importance of disclosing providers' availability, whereas Hui and Liu (2022),

Moravec et al. (2023), and Pu et al. (2023) emphasize the role of platform-managed quality certificates and platform-incentivized online ratings. Chen and Yao (2017) use click-stream data on hotel bookings to find sizable consumer benefits from search refinement tools, such as sorting and filtering. A crucial assumption underlies the ample work estimating the effects of disclosing information about products and services and the effects of changing the order of search results: consumers are assumed to know what they want and how to describe it. The behavioral literature has however found limitations to this assumption (e.g., Kamenica et al., 2011). Our paper confirms that consumers sometimes find it difficult to describe what they want. The search refinement tool we study can also be seen as a more convenient implementation of search filters, which have existed for a long time but are not used very frequently by consumers (Chen and Yao, 2017).

There is more limited work on estimating the effects of tools that help consumers better discover and express their preferences (Lee et al., 2020). Lei et al. (2023) is one of the few papers quantifying the positive effects of auto-complete on consumer search. They leverage an experiment with a small search engine platform, which removes access to search recommendations from the API of a larger competitor. The authors find large benefits of the API in helping consumers find what they want. Similarly, Zheng et al. (2023) leverage an experiment on a food delivery platform to show that query recommender systems increase the probability that consumers place a food order, at least in the short-run. Häubl and Trifts (2000) conduct a controlled experiment using a simulated online store to show that interactive tools designed to assist consumer search have strong positive effects on purchase decisions. Our study adds to this body of work by revealing the impact of offering a search refinement tool on consumer search and purchasing decisions, not just in the short-run, but over an extended period of time.

The majority of the research on the value of search recommendations, such as Sun et al. (2023), Peukert et al. (2023), and Chiou and Tucker (2017), focuses on the role of consumer data to help offer personalized results (Zhang et al., 2019). But consumer data can help refine searches in other ways, by for example, identifying new search filters or search tools (Jiang and Zou, 2020). Our paper contributes to this latter line of research by highlighting the role of visual and textual suggestions in guiding consumer search. Our ability to observe the entire search and purchasing funnel allows us to shed light on the mechanisms through which search refinement tools benefit consumers, by increasing the likelihood that consumers find what they want for any individual search, and by teaching consumers to more effectively search on their own even when those tools are not available.

Our results that tail and niche products gain more from the introduction of search refinement tools relates to the extensive literature exploring how the Internet reshapes market

structure and concentration, particularly when comparing sales of popular products versus niche products (Elberse and Oberholzer-Gee, 2006; Bar-Isaac et al., 2012). Fleder and Hosanagar (2009) employ a theoretical model to investigate the effect of recommendation systems on sales diversity, and predict that recommendation systems based on sales and ratings tend to promote product concentration at the expense of diversity. Although that may be true of baseline recommender systems, our findings suggest that search tools like the one we study may actually provide a correcting mechanism against sales concentration. Our findings align more closely with work by Brynjolfsson et al. (2011). They demonstrate that e-commerce and online search technologies allow consumers to discover and buy products that better match their preferences. Notably, this effect is not solely attributed to the expansion of product variety, but also to platforms' efforts to help consumers navigate through products and find better-matched options more effectively, as our paper showcases.

Finally, our paper also highlights the importance of running long-term experiments to identify the equilibrium effects of product changes (Gupta et al., 2019). In doing that, we relate to the literature on long-term experiments (Goli et al., 2021; Huang et al., 2018) and approaches to infer long-term outcomes from short-term proxies (Athey et al., 2019). We find that short-term results may be very different from long-term results. The typical risk of experiments is that one may find positive short-term effects, but null or negative effects in the long-run (Kohavi et al., 2012). Our specific case highlights the opposite risk, i.e., improvements in platform design that take time to emerge.

3 Data and Institutional Details

We collaborate with one of the largest e-commerce platforms in the world. Given the large variety of products available, search tools on e-commerce platforms like our partner play a crucial role in helping consumers find products that match their needs. In this section, we describe the search tool that our collaborating platform created, how they experimentally launched it, and the data we have available to study its effects.

3.1 The Picture-Text Search Tool and Its Experimental Roll-Out

The collaborating platform has millions of sellers and hundreds of millions of consumers active on any given month, and billions of products listed on any given day. Consumer search plays a crucial role on this platform. Like in the majority of e-commerce platforms, search is the largest channel through which consumers purchase products.[3]

[3]Other channels include, for example, platform recommendations and live streaming.

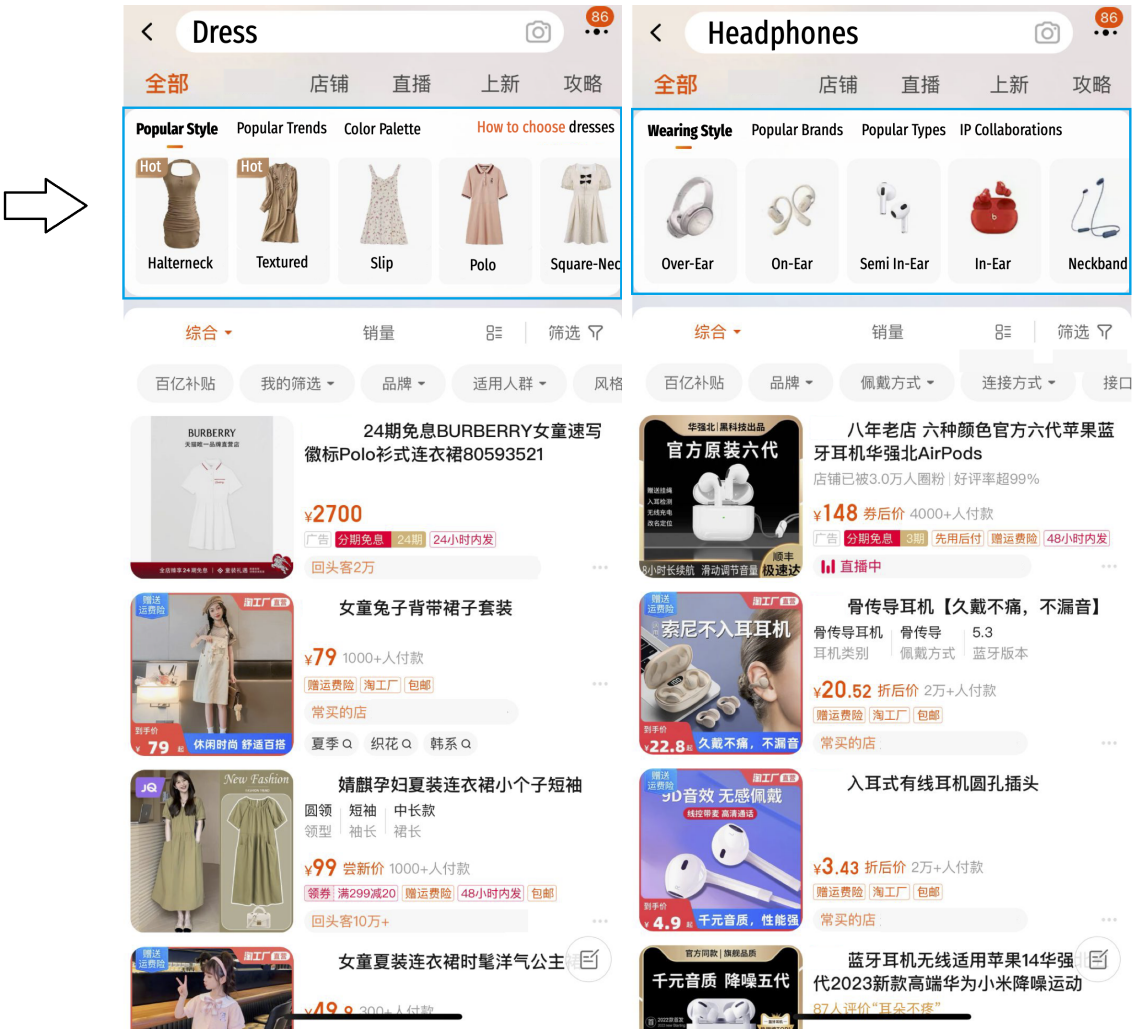
The focus of our study is a new search tool that suggests a combination of picture and textual recommendations for the consumer to refine their searches. We call this tool *Picture-Text Guidance* (PTG henceforth) in the rest of the paper. Figure 1 illustrates how PTG works. When a consumer enters a query that is a candidate for PTG, such as "Dress" on the left panel of Figure 1, the platform's search engine presents the consumer with two levels of sub-categorization of products related to the general query. The first level presents broad dimensions for classifying relevant products. In the dress example, the picture shows "Popular Style," "Popular Trends," and "Color Palette." The second grouping level is presented as a series of pictures with the corresponding descriptive words. In the figure, the pictures correspond to dresses grouped by "Popular Style": halterneck, textured, slip, polo, and square-neck. The right panel of Figure 1 provides an analogous example for headphones.

Consumers can click on any of the PTG elements to refine their search. When they click on one of those elements, the search engine will automatically refine the search query to reflect the finer subset of relevant products. For example, if the consumer clicks on the picture for the halterneck dress, the word "Halterneck" is added to the search box at the top. Instead of returning results matching the query "Dress," the engine will thus return results matching the query "Dress Halterneck." The PTG search refinement tool is not personalized, so for the same query, such as 'Dress,' treatment consumers will see the same PTG elements.[4]

Although we cannot disclose the details of the proprietary algorithms, the set of queries that are candidates for this refinement tool were identified by the product team based on the popularity of consumer searches, and the possibility to break down those searches into narrower queries. Query popularity is determined by factors such as the number of consumers who have searched for that particular query. For example, "Dress" is among the top 1% queries in terms of cumulative searches conducted by consumers in 2021. Queries with higher popularity tend to represent a consumer's initial idea or a general expression of their needs. Therefore, identifying refined recommendations to these popular search queries has the potential to assist consumers in expressing and forming their demands more effectively. The candidates for these finer and more specific queries are identified by a combination of a collaborative filtering model and human curation. For instance, "Halterneck" is chosen as a suggested query associated with "Dress" because it is often used by consumers in conjunction with "Dress."

Because of the substantial effort in identifying finer categories for the many search queries that consumers search on the platform, and because some searches cannot be further broken down into subcategories, not all search queries are candidates for PTG. We thus categorize

[4]Appendix Figure A.1 presents more details about PTG.



(a) Dress (b) Headphones

Figure 1: Illustration of the Picture-Text Guidance (PTG) Search Tool, denoted by the arrow and surrounded by a blue frame.

search queries into three types:

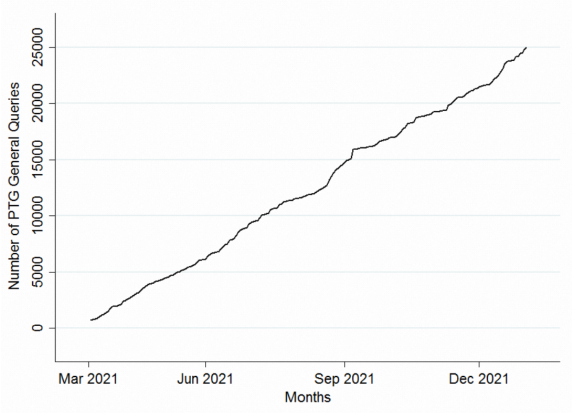
- *PTG general* queries refer to search queries that have been augmented with PTG. Examples of PTG general queries include "Dress" and "Headphones" as in Figure 1.

- *PTG specific* queries refer to search queries generated when a consumer clicks on the pictures following a search for a PTG general query. Examples of PTG specific queries include "Dress Halterneck" and "Dress Textured" on the left panel of Figure 1, and "Headphones Over-Ear" and "Headphones In-Ear" on the right panel. Note that consumers can search for a PTG specific query by directly typing the words in the search box, not just by clicking on the picture provided by PTG. Our data will not be able

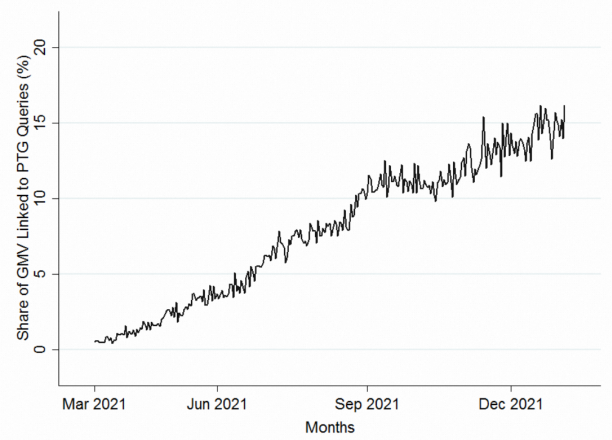
to distinguish whether consumers click on the PTG picture or type the query on their own.

- *Non-PTG* queries refer to search queries that do not qualify for the PTG feature, such as "Squash Racket."

The platform launched PTG on the mobile app in March 2021,[5] with a small number of PTG general and specific queries. Over the course of the following months, it progressively increased both the number of search terms classified as PTG general queries and the number of search terms classified as PTG specific queries. The left panel of Figure 2 shows that by December 2021, around 25,000 queries were classified as PTG general queries. The right panel of Figure 2 shows that PTG queries went from representing 0 to 15% of the gross merchandise volume (GMV henceforth) directly associated to a search query.



(a) Number of Queries Classified as PTG General Queries.



(b) Share of GMV Linked to PTG Queries (out of total GMV directly linked to searches).

Figure 2: Expansion of PTG Between March and December 2021.

During the roll-out of PTG between March and December 2021, the platform conducted a randomized field experiment to measure the effectiveness of the new search tool. All consumers using the mobile platform were ex-ante randomly allocated to a control and a treatment groups with equal probability.[6] Upon entering a PTG general query, treatment and control consumers saw different displays. Treatment consumers saw the PTG search

[5] Over 95% of consumers use the platform on mobile.

[6] Consumers who were not logged in when searching for products are not included in our experiment. This affects a small number of searches on mobile. Similarly, because the tool was only implemented on the mobile app, consumers who searched on the Web were not included in the experiment. Less than 5% of consumers use the platform on a web browser.

tool (picture and text suggestions in the blue rectangle in Figure 1) and could click on any of the search recommendations to refine their searches.[7] The control group did not have access to the PTG tool, and hence would not see the rectangle from Figure 1.

It is worth making two remarks. First, when consumers searched for non-PTG queries, they would face the same standard search experience without the rectangle in Figure 1, regardless of whether they were in the treatment or control group. Second, because not all consumers searched for PTG general queries, in our analysis we only include consumers who searched for PTG general queries during the experimental period. The ex-ante random allocation ensures that focusing on this subset of users does not undermine our causal analyses.

The experiment lasted for ten months, from mid March until end of December 2021. Since there is variation in the timing when consumers first search for PTG general queries, we say that a consumer *enters the experiment* on the first day during the experimental period when they search for a PTG general query.

This experiment proves very valuable for our goal of understanding whether and how search guidance tools help consumers better identify and describe what they want. The long experiment duration was driven by the fact that the search tool was progressively increasing its reach as new queries were included in PTG, but it provides a unique opportunity for us to measure both short-term and long-term effects of the new search tool, and evaluate the validity of the conclusions that would have been drawn if we had had access only to a short-run experiment.

3.2 Data

We obtain proprietary data from the platform. Although the roll-out of PTG continued past the end of 2021, we have access to data between mid March and Dec 31, 2021 (the *experimental period*). We restrict attention to treatment and control consumers who reside in China and who performed a PTG general query during the experimental period.

Data are aggregated at the search level. For each search performed by a consumer included in the experiment, information on the search terms allows us to classify the query into PTG general, PTG specific, or non-PTG. For each of the searches, we also have information on the following outcomes of interest: the number of products viewed in the search results (*views* henceforth);[8] the number of clicks on products returned in the search results

[7] Additional information regarding the PTG search tool and the potential responses of treatment group consumers to PTG are provided in Appendix Figure A.1.

[8] The number of product views is both a function of product availability for the specific search query, and of how much the consumer continues to scroll past the initial results. Products are grouped into sets of

(*clicks*); the number of purchases that were directly linked to the search (*orders*); and the total expenditures for those purchases (*GMV*, for gross transaction volume).

We augment the search-level data with product- and seller-level information. Specifically, for each of the products viewed, we obtain the product category and its seller identifier, in order to calculate sales rankings for both products and sellers. This additional information allows us to distinguish between more and less popular products or sellers, and how PTG affects consumer choices for different product and seller groups.

Similarly, we also augment the search-level data with consumer-level information to compare consumers in the treatment and control groups. Our sample includes 505,485 consumers, half in the treatment group and half in the control groups. Appendix Table A.1 confirms that the randomization was effective at allocating comparable consumers into the two groups. On average, consumers are between 25 and 35 years of age (denoted as age tier 2), they reside in large cities (3 denotes the third largest city-tier in China, out of a total of 6 tiers), they have been users of the platform for 6.3 years, and are 55% women. When it comes to consumer behavior on the platform, Table A.1 shows that in the 8 weeks preceding their entry into the experiment, consumers viewed about 2,500 products, clicked on 112 products, purchased 4.5 of them, and spent CNY410-415 (almost $60).[9]

Although Appendix Table A.1 confirms that users in the treatment and control groups are statistically similar, the entry into the experiment is not randomly assigned to all users of the platform. In particular, heavy users will be more likely to enter the experiment earlier because their frequent search behavior will lead them to search for PTG general queries earlier than infrequent users. Appendix Figure A.2 shows the average spending of users by cohort of entry into the experiment. The figure confirms that earlier cohorts spend much more in the 8 weeks preceding their entry into the experiment—around CNY800—compared to later cohorts, who spend less than CNY100. The phenomenon of selective entry is very common in experiments conducted by platforms where a user action triggers entry into the experiment. Such selection creates concerns around generalizing the results to the entire platform population. Given our long experiment duration however, the characteristics of the users tend to stabilize in the second half of the experiment (as Appendix Figure A.2 confirms), allowing us to test whether the experimental results are likely to generalize to the rest of the user population. In Section 4, our robustness checks with respect to different

about a dozen (we cannot disclose the exact number) – the first 12 results, the next 12 results, and so on. When a consumer scrolls past a multiple of 12, an additional 12 products are added to the list of product views, as long as there are relevant products that remain to display.

[9]Note that the statistics in Appendix Table A.1 do not necessarily reflect the usage characteristics of the entire population of platform consumers, given that consumers in our dataset are selected by the fact that they perform a PTG general query during the experimental period.

entry cohorts confirm that our results are generalizable to the entire platform population.

The next section describes our analyses, divided into a short-run and a long-run analyses. For the short-run, we consider all the experimental consumers. For the long-run, we restrict attention to consumers entering the experiment between mid March and mid July 2021, allowing us to track them for 24 weeks, or almost 6 months until the end of 2021.

4 Empirical Approach and Results

We evaluate the effect of the PTG search tool on consumer search behavior and purchase decisions. To do so, we conduct our analysis at the individual consumer level (i.e., the randomization level) and estimate regressions of this form:

$$y_i = \beta \times Treat_i + \alpha_{c(i)} + \epsilon_i, \tag{1}$$

where i denotes a consumer in the experiment. Since consumers enter the experiment when they first type a PTG general query, we control for their day of entry with cohort $c(i)$ fixed effects. $Treat_i$ is an indicator for whether the consumer belongs to the treatment group, so the coefficient β measures the causal effect of giving consumers access to the search tool.

We estimate the regression for several outcomes y_i tracking the consumer behavior from search to purchase. We focus on the number of products each consumer views, the number of clicks they make to those products, the total number of products purchased, and the overall spending linked to those purchases (GMV). These metrics are computed at the individual consumer level and aggregated over a designated time period, depending on whether we focus on the short-run (one day) or the long-run (24 weeks). To identify the mechanisms through which the effects materialize, we explore additional outcomes as needed.

We are interested in estimating the immediate effects of the search tool as well as the longer-term effects, which may include learning to perform more effective searches independently. In the short-run, we aggregate the outcomes of interest over the course of the first day when a consumer enters the experiment. This allows us to use all consumers who joined the experiment between mid March and end of December 2021 (505,485 consumers). In the long-run, we aggregate the outcomes of interest over the course of 24 weeks following a consumer entry in the experiment, which requires us to constrain the analysis to consumers who entered the experiment between mid March and mid July 2021 (346,110 consumers).[10]

Given recent concerns around using log transformations when outcomes can take the

[10]Limiting the analysis to consumers who entered the experiment between mid March and mid July 2021 guarantees that we can track all those consumers for at least 24 weeks.

value zero (Chen and Roth, 2024), we estimate regressions in levels, and present short-run and long-run estimates in the next two sub-sections.

4.1 Short-Run Results

This section focuses on outcomes measured on the day a consumer enters the experiment. First, we show that the PTG search tool had a large and immediate impact on consumers' search behavior. To do this, we estimate the effect on three separate outcomes: total number of searches conducted on the consumer's first day in the experiment,[11] number of PTG general searches, and number of PTG specific searches.

We run regressions of Equation 1, and Table 1 displays the results. Column 1 shows that the search tool leads consumers to perform 0.049 more queries compared to the baseline, which amounts to a 1.04% increase. This increase in searches comes solely from a rise in the number of PTG specific queries (column 3), which increase by 0.053, an almost identical coefficient to the estimate in column 1. Although this seems like a small increase in levels, the search tool effectively grows the propensity of consumers to perform PTG specific queries 5-fold. At least in the short-run, these additional PTG specific searches do not cannibalize PTG general (column 2) or non-PTG searches, which remain fairly constant.

Table 1: Short-Run Impact on Number of Searches

	Number of Searches	Number of PTG General Searches	Number of PTG Specific Searches
	(1)	(2)	(3)
Treat	0.0486***	0.000479	0.0531***
	(0.0152)	(0.00132)	(0.000628)
% Change	1.04%	0.04%	495.22%
Observations	505,485	505,485	505,485
R-squared	0.043	0.019	0.015

Notes: Regression estimates of Equation 1. The dependent variables are in levels. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses.

Table 1 confirms that consumers actively utilize the new search tool to perform narrower searches than in the absence of PTG, so our next step is to evaluate whether this change in search behavior translates into changes downstream, all the way to purchases. We thus estimate regressions as in Equation 1 for views, clicks, purchases, and expenditures on the consumer's first day in the experiment.

[11] The total number of searches are the sum of PTG general searches, PTG specific searches, and non-PTG searches.

Table 2 presents the results. None of the coefficients on views, clicks, orders, and GMV are large nor statistically significant, implying that PTG does not immediately impact how many products consumers view or purchase, nor the price of those purchases. Note that this null effect may be due to at least two separate reasons. First, a consumer navigating to the e-commerce platform may indicate an underlying purchasing intent (say, buy a dress for a special event) that would not be affected by the availability of the PTG search tool. If this hypothesis were true, PTG would simply shift product views and purchases from one type of searches (non PTG or PTG general searches) to another (PTG specific searches). Second, the effects of improving search may take longer than one day to materialize, as consumers learn to use the tool and to perform more effective searches on their own. We tackle the first hypothesis next, and the second hypothesis in the following sub-section.

Table 2: Short-Run Treatment Effects

	Views	Clicks	Orders	GMV
	(1)	(2)	(3)	(4)
Treat	0.666	0.0563	0.00537	0.35
	(1.177)	(0.0465)	(0.0035)	(0.727)
%Change	0.26%	0.58%	1.23%	1.08%
Observations	505,485	505,485	505,485	505,485
R-squared	0.013	0.025	0.008	0.002

Notes: Regression estimates of Equation 1. The dependent variables are in levels. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses.

To evaluate whether PTG shifts consumers' viewing, clicking, and purchasing behavior toward PTG specific queries, we need to separate the effect on aggregate outcomes by query type. We thus allocate product views, clicks, purchases, and expenditures to the three types of searches described in Section 3.1: PTG general, PTG specific, and non-PTG queries. We analyze the treatment effects on consumer behavior for these three queries separately.

Results for PTG general and PTG specific queries are shown in Table 3.[12] The results confirm a significant and sizable decrease in the number of product views and clicks stemming from PTG general queries. Views decrease by 2.3%, and clicks decrease by 3.7%. Consumers shift viewing and clicking to PTG specific queries. Columns 5 and 6 show that consumers in the treatment group view 2.8 and click on 0.1 more products related to PTG specific queries compared to consumers in the control group. Columns 7 and 8 further confirm that the shift in browsing behavior translates into 0.005 more products purchased and CNY0.255 more spent on products showing up in PTG specific queries. In percent terms, all these

[12]Results for non-PTG queries are presented in Appendix Table A.2.

coefficients represent a more than 500% increase in the very small baseline browsing and purchasing behavior related to PTG specific queries.

Table 3: Decomposition of Short-Run Treatment Effects

	PTG General Queries				PTG Specific Queries			
	Views	Clicks	Orders	GMV	Views	Clicks	Orders	GMV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat	-1.402***	-0.0846***	-0.000989	-0.412	2.756***	0.106***	0.00470***	0.255***
	(0.277)	(0.0122)	(0.00117)	(0.269)	(0.0535)	(0.00227)	(0.0002)	(0.0225)
% Change	-2.34%	-3.74%	-0.80%	-4.69%	515.3%	516.59%	568.36%	560.19%
Observations	505,485	505,485	505,485	505,485	505,485	505,485	505,485	505,485
R-squared	0.009	0.002	0.002	0.001	0.007	0.005	0.002	0.001

Notes: Regression estimates of Equation 1. The dependent variables are in levels. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses. Appendix Table A.2 contains the estimates for non-PTG queries.

These estimates suggest that although the PTG search tool is not changing aggregate purchase intent, consumers find the products they want through the help of narrower searches that are suggested by PTG. An important question is whether consumers eventually see the same products, regardless of whether they are directed to those products from specific or general queries. To check that, we can focus on treated users on their first day in the experiment, and measure the share of products appearing in PTG specific searches that also appear in the corresponding PTG general searches. (Recall that users need to conduct a PTG general search to enter the experiment.) A high share would imply a large overlap in the search results returned by the two queries, whereas a low share would imply a small overlap. In our data, only 15.9% of products appearing in PTG specific searches appear in the corresponding PTG general queries, which suggests a sizable change in the products consumers have access to.

Because the search tool was gradually rolled out over the duration of the experiment, one may worry about whether the null average effect masks heterogeneous effects over time, as more queries are supported by the search tool and the composition of users entering the experiment changes. To confirm that the short-run results are stable over time, we interact the treatment dummy with dummies for the week of entry into the experiment. Appendix Figure A.3 plots estimated treatment effects by entry cohort. Most of the coefficients are small and indistinguishable from zero, without a clear upward trend. The more positive coefficient estimates in October are likely driven by holiday events and related promotions, such as Singles Day (similar to Black Friday in the US), suggesting some short-run value of

the tool around the holidays.[13] In general, this robustness check suggests that the gradual roll-out of the search tool is unlikely to be a major confounder behind our null short-run results.

We have identified a relatively stable null effect, but at least another force is changing over time, in addition to the gradual roll-out. Indeed, the composition of users entering into the experiment dynamically changes (Appendix Figure A.2), which may lead to a decrease in the treatment effect for later entry cohorts composed of less frequent users. Combined with the fact that new words are progressively added to PTG,[14] the implications for how the treatment effect will change over time are not obvious. With the data available to us, we are not fully able to determine whether the relatively stable treatment effects result from selective entry and gradual roll-out canceling each other out, or because neither of the dynamics have large impacts on the estimates. However, the composition of entry cohorts stabilizes around August 2021 (Appendix Figure A.2), and the treatment coefficients do not show any obvious trend from August to the end of the year (Appendix Figure A.3). The estimates in the second half of the experiment thus suggest that the null effect is likely to generalize beyond the experiment period and users.

Overall, the results in this sub-section suggest that, while the PTG search tool impacts consumer search behavior (as evidenced by the type of searches they conduct and the resulting purchases), in the immediate short-term it does not impact how many products consumers buy or how much they spend on the platform.[15]

In order to understand whether the short-run results are due to PTG being ineffective at improving search enough to generate market expansion or the consumers needing time to find PTG valuable and learn from it, in the next sub-section we explore the long-run effects of the search tool.

4.2 Long-Run Results

To investigate the long-run effects of introducing the PTG search tool, we restrict attention to 68% of users who entered the experiment early enough to give us about 6 months of experimental data for all of them (346,110 users). Specifically, we focus on consumers who first searched for PTG-related queries between mid March and July 16, 2021. This constraint ensures that we can observe all these consumers for a minimum of 24 weeks by the end of

[13] Singles Day is November 11, but the promotions often start before then (around October 20) and last a while after the main day.

[14] It is possible for searches that are added to the search tool earlier to benefit more or less from the tool compared to later searches.

[15] Appendix Table A.3 shows that, even if consumers buy different products in the treatment group relative to the control group, consumer satisfaction does not change.

2021. Just like Table 1 in the previous sub-section, Appendix Table A.4 (column 3) confirms that PTG was effective at shifting consumers in the treatment group to perform narrower searchers that were recommended by the tool itself: the number of PTG specific searches increases by 450%, up from an otherwise small baseline, whereas the total number of queries does not meaningfully change.

Given the impact of PTG on consumer search behavior, we start by estimating the effects of the search tool on product views, clicks, purchases, and expenditures. We compute those outcomes at the consumer level by aggregating views, clicks, purchases, and expenditures over the course of the 24 weeks following the consumer's entry into the experiment.

Table 4 shows the estimates of Equation 1, where the outcomes are measured in the long-run. Starting from columns 1 and 2, the estimates imply that access to the new search tool does not significantly change the number of products viewed, nor the number of clicks on those products. Both the point estimates and the percentage changes are fairly small in magnitude. Columns 3 and 4 however, indicate that consumers with the PTG search tool purchase on average 0.34 more orders and spend on average CNY62 more compared to consumers without the search tool, an increase of 1.6% in orders and 3.2% in spending compared to the baseline. Together with the null results on product views and clicks (and the null effect on total searches from Appendix Table A.4), this purchase expansion seems primarily due to searches returning better products in the treatment group, rather than consumers dedicating more time to viewing and clicking on more products.

Table 4: Long-Run Treatment Effects

	Views	Clicks	Orders	GMV
	(1)	(2)	(3)	(4)
Treat	-35.78	-0.932	0.336**	62.44**
	(54.09)	(2.215)	(0.140)	(27.60)
% Change	-0.29%	-0.19%	1.57%	3.24%
Observations	346,110	346,110	346,110	346,110
R-squared	0.06	0.08	0.03	0.006

Notes: Regression estimates of Equation 1. The dependent variables are in levels. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses.

A 1.6% increase in orders and a 3.2% increase in spending may seem small, but it is worth putting these improvements into perspective. Digital platforms like the one we study constantly optimize their product offerings, which makes improvements of our magnitude increasingly rare. Coupled with the fact that PTG is directly linked to only 15% of search-related spending by the end of our experiment period (Figure 2), we interpret our results as

providing a sizable improvement in both orders and spending relative to the status quo.

Because a substantial proportion of purchases are not immediately linked to searches, we want to ensure that the increase in orders and spending does not cannibalize purchasing from non-search channels, such as platform recommendations. Appendix Table A.5 tests whether orders and spending from non-search channels over the 24 weeks of analysis are different between treatment and control consumers. The treatment effects are statistically indistinguishable from zero. If anything, we cannot exclude sizable positive effects on spending (column 2). This suggests that the observed positive impact on orders and spending from Table 4 does not come from substitution between search and non-search channels.

It is worth placing our results within the existing theories of consumer search. We find that in the long run, the consumers exposed to the refinement tool bought more with the same level of search effort as consumers in the control group. These results are difficult to rationalize with standard search models (Stigler, 1961; McCall, 1965; Mortensen, 1970), where search effort is determined by equating the marginal benefits of search with marginal costs. In our case, it is unlikely that the search tool changes the costs of inspecting individual products, but, as we have seen in Section 4.1, it changes the distribution of products seen. The direction of change seems to be positive, as evidenced by the increase in orders and spending, and by the increase in consumer satisfaction that we describe later in Table 10. The standard models would thus predict more search, which we do not find. Instead, the results seem consistent with a model where search effort is determined by the opportunity cost of time (Greminger et al., 2023) rather than the marginal benefit of inspecting additional products.

Given the null effects on purchasing behavior in the short-run (same-day) and the large positive effects in the long-run (the following 24 weeks), we want to explore how early these positive effects start emerging. Typically at platform companies, experiments (or A/B tests) are run for a few weeks, so this exercise can help us understand the extent to which short-run experiments can capture the effects of product changes like ours.

To evaluate how early the positive effects on purchasing behavior materialize, we replicate columns 3 and 4 from Table 4 with outcomes aggregated over the first week, the first 2 weeks, the first 3 weeks, and the first 4 weeks since a consumer enters the experiment.[16] Table 5 presents the results. Columns 1 and 2 report treatment effects on orders and expenditures in the first week following entry into the experiment. Columns 3 and 4 do the same for orders and expenditures in the first two weeks, and so on. All coefficients are statistically

[16] Appendix Table A.6 presents analyses for time aggregations beyond the first four weeks. Those other time aggregations are all comparable to the long-run results presented in Table 4, obviously not in magnitudes, but rather in percentage terms.

indistinguishable from 0, except for those in columns 7 and 8, which aggregate consumer activity within the first month since entry into the experiment. In percentage terms, those effects (1.29% and 3.55%) are comparable to the longer-run results from Table 4, providing support for the hypothesis that it takes time for the positive benefits of improved search to materialize.[17] The introduction of PTG progressively influences the purchasing decisions of the treatment group consumers, ultimately leading to a significant increase in product orders and purchases. Despite the benefits, typical durations of A/B tests would not be able to capture these benefits.[18]

Table 5: Treatment Effects Over Different Time Aggregations

	Week 1		Weeks 1-2		Weeks 1-3		Weeks 1-4	
	Orders	GMV	Orders	GMV	Orders	GMV	Orders	GMV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat	0.0157	2.758	0.0201	2.899	0.0310	7.366	0.0538*	13.10**
	(0.0109)	(2.117)	(0.0174)	(3.480)	(0.0238)	(4.757)	(0.0302)	(5.964)
% Change	1.12%	2.34%	0.86%	1.42%	0.95%	2.57%	1.29%	3.55%
Observations	346,110	346,110	346,110	346,110	346,110	346,110	346,110	346,110
R-squared	0.014	0.003	0.019	0.004	0.022	0.004	0.024	0.005

Notes: Regression estimates of Equation 1. The dependent variables are in levels. In column 1, the number of orders placed by a consumer is aggregated over the first week since the consumer enters the experiment. In column 3, the number of orders is aggregated over the first two weeks, then in column 5 it is calculated over the first three weeks, and in columns 7 over the first four weeks. Columns 2, 4, 6, and 8 compute the same aggregations for spending. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses. Appendix Table A.6 presents similar analyses of longer time periods.

As we did in Section 4.1 for the short-run, we check whether there are any obvious time trends in the treatment effects, given selective entry into the experiment and the gradual roll-out of the search tool over an increasingly larger number of queries. Appendix Figure A.4 presents the treatment effects by entry cohort. After the first month, coefficients tend to be positive, although the limited sample size constrains our power to detect effects that are statistically different from zero for most of the weeks.

In the rest of this section, we explore how the effects differ across consumers and products, we offer some evidence that the tool is effective at teaching users to perform better searches,

[17] Appendix Table A.1 shows that, on average, consumers completed 4.5 orders in the eight weeks prior to the experiment, which translates to a shopping frequency of approximately one order every two weeks. Therefore, observing a significant effect of the PTG tool on orders starting around four weeks seems reasonable.

[18] Note that the estimates in Table 5 and Appendix Table A.6 do not exactly resemble the analysis one could conduct with a short-run experiment. Indeed, the number of experimental participants, and hence the statistical power of the tests, is large only because the experiment was run over many months. In a sense, the analysis presented here offers an upper bound of what can be inferred from a short-run experiment.

and finally, we show that consumers are more satisfied with the purchases they make, as evidenced by higher ratings and lower returns.

Heterogeneous Effects. Although PTG increases purchases on average, we are interested in exploring for which types of consumers it is most effective in helping them find what they want. To do this, we examine how the treatment effects on orders and spending differ among different consumer categories. Our data allow us to pick five dimensions. Beyond gender, the other four dimensions can be seen as proxies for how Internet-savvy consumers are: age, city of residence, year of registration on the platform, and frequency of platform use. For each of the four dimensions, we divide the consumers into two separate groups and interact the treatment dummy with a variable denoting one of the two groups.

Results are presented in Table 6. Column 1 interacts the treatment indicator with a dummy for whether the consumer is under 35 years old (58% of the experiment users are under 35 years old). The results indicate that the benefits of the search tool are concentrated among consumers over 35 years old, providing support for the hypothesis that younger users know how to more effectively search for products online. The next three columns do not find support for heterogeneous effects. Column 2 interacts the treatment indicator with a dummy for whether the consumer resides in a large city in China (32% of the experiment users live in a large city). Column 3 interacts the treatment indicator with a dummy for whether the consumer is new to the platform, i.e., they created their account in the last 5 years (40% of the users are considered new). Column 4 interacts the treatment indicator with a female dummy (55% of the consumers self-identify as female). Finally, the last column shows a perhaps unexpected result, that more frequent users, despite their deeper knowledge of the platform, are those that truly benefit from PTG. Here, frequent users are defined as those in the top quartile of spending in the 8 weeks preceding the experiment. A likely explanation for this result is that heavier users are more likely to conduct searches supported by PTG, and thus stand to benefit the most.

Because PTG facilitates more specific and narrower searches and, as shown in Section 4.1, it affects the type of products consumers find, it is likely that PTG allows consumers to find less popular products. To test this hypothesis, we create two classifications of products into more versus less popular.

First, we compute product-level revenues for 2021,[19] which allows us to rank products from best to worst selling within their respective product categories. We then classify the products into five groups: top 10 selling products, the next 10-100 products, the next 100-

[19]The purchases related to the users in our experiment are a small share of the revenues for these products in 2021, so the likelihood that our experimental treatment affects the product sales rank is very low.

Table 6: Heterogeneous Treatment Effects on GMV (Consumers)

	GMV	GMV	GMV	GMV	GMV
	(1)	(2)	(3)	(4)	(5)
Treat	117.5***	50.33	36.49	47.01	22.87
	(41.70)	(32.55)	(34.67)	(40.60)	(30.95)
Treat*Young	-101.0*				
	(54.44)				
Treat*Big City		24.43			
		(57.37)			
Treat*New			53.97		
			(54.60)		
Treat*Female				19.82	
				(54.06)	
Treat*Heavy					141.3**
					(61.90)
Observations	346,110	346,110	346,110	346,110	346,110
R-squared	0.062	0.062	0.062	0.062	0.062

Notes: Regression estimates of Equation 1, where the treatment indicator is interacted with demographic characteristics. In column 1, "Young=1" refers to consumers younger than 35 years old. In column 2, "Big City=1" denotes consumers residing in first and second-tier cities (in China, the cities are categorized into six tiers, and first and second-tier cities typically refer to large cities). In column 3, "New=1" refers to consumers who created their account in the last five years. In column 4, "Female=1" refers to consumers who self-identify as female. In column 5, "Heavy=1" denotes consumers in the top quartile of spending during the 8 weeks prior to entering the experiment. Standard errors are in parentheses. The results of a similar analysis using Orders as the outcome variable are presented in Appendix Table A.7.

1,000 products, the next 1,000-10,000 products, and finally, the products beyond the top 10,000 selling products. We then compare the expenditures of treatment and control consumers in each of these five product groups.

Panel I of Table 7 shows the results. We find that products further down in the sales rank benefit the most because the tool makes it easier for consumers to find them. Columns 4 and 5 confirm that expenditures on products beyond the top 1,000 selling products increase by between 4.5% and 6% when PTG is available. For more popular products, the percentage increase is smaller (e.g., column 3) and sometimes even indistinguishable from zero (columns 1 and 2).

The second approach is to classify products by their seller's overall sales rank. We categorize sellers into five quantiles based on their cumulative annual revenues by the start

Table 7: Heterogeneous Treatment Effects on GMV (Products and Sellers)

	Panel I: GMV Across Products Grouped by Products' GMV Rank				
	Top 10	10-100	100-1000	1000-10000	Beyond 10000
	(1)	(2)	(3)	(4)	(5)
Treat	1.639	5.668	15.25*	20.60**	19.57***
	(7.208)	(6.382)	(9.097)	(8.027)	(5.954)
% Change	0.77%	1.55%	2.89%	4.53%	5.96%
Observations	346,110	346,110	346,110	346,110	346,110
R-squared	0.001	0.005	0.004	0.004	0.004
	Panel II: GMV Across Products Grouped by Sellers' Revenue Quantile				
	Top 20%	Med-high 20%	Medium 20%	Med-low 20%	Tail 20%
	(1)	(2)	(3)	(4)	(5)
Treat	3.120	3.473	15.13	26.05**	14.96**
	(2.857)	(6.906)	(10.27)	(10.23)	(7.078)
% Change	2.03%	1.03%	3.17%	5.00%	3.74%
Observations	346,110	346,110	346,110	346,110	346,110
R-squared	0.005	0.004	0.003	0.003	0.004

Notes: Regression estimates of Equation 1, where GMV is disaggregated by type of products/sellers. In Panel I, products are classified into 5 groups, depending on their GMV rank within their respective product categories. In Panel II, products are classified into another 5 groups, depending on their sellers' revenue rank. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses. The results of a similar analysis using Orders as the outcome variable are presented in Appendix Table A.8.

of the roll out of PTG, and conduct a similar analysis. Panel II of Table 7 shows the results. The findings align closely with Panel I. Expenditures on top sellers show no significant change between treatment and control consumers, perhaps because of the ease with which top sellers can already be found on the platform. In contrast, the two bottom quantiles of sellers experience a significant increase in revenues from treated consumers, by between 3.7% and 5%. Together, the two panels of Table 7 indicate that tail and niche products gain more from the introduction of PTG.

The final dimension of heterogeneity we explore is by product categories. We separate GMV across the ten main product categories defined by the platform,[20] and estimate

[20]The platform assigns product category labels to each product, employing a hierarchical categorization system with four primary layers. The first layer includes broad categories, such as Fast-Moving Consumer Goods (referred to as "categories" in the paper). The second layer categorizes these top layer categories into types including Beauty Products and Maternal and Child Products. The third layer further distinguishes between these second layer types, such as Makeup and Skincare Products. The fourth layer identifies specific types, such as lipsticks and eyeshadow (refer to as "subcategories" in the paper). In our analysis of heterogeneity in product categories, we adhere to the platform's definitions of the product categories at the top layer.

category-specific treatment effects. Appendix Table A.9 indicates that there is substantial heterogeneity in the effect of the search tool across product categories, with Home Furnishing, Healthcare and Medicine, and Apparel and Fashion benefiting the most, whereas Electronic Products and Stationary and Educational Supplies do not benefit at all.

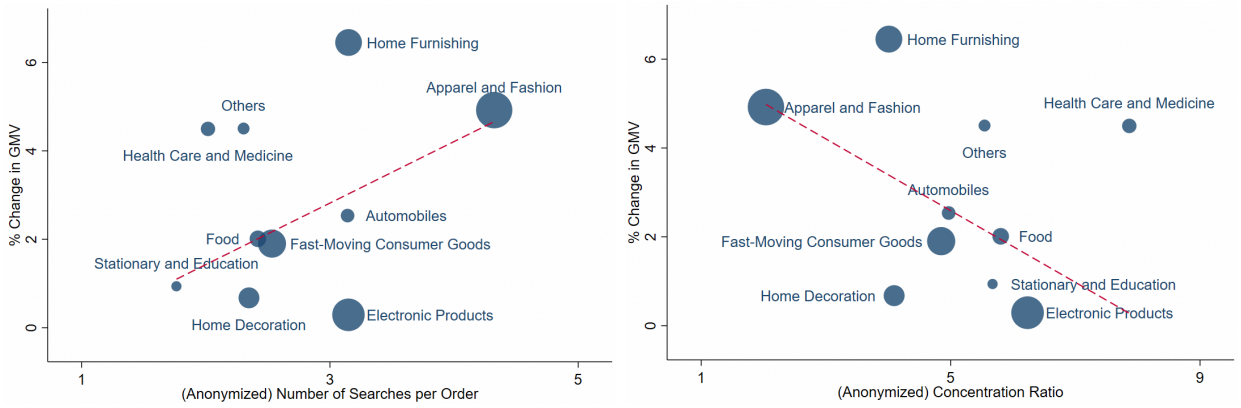
The heterogeneous effects across categories provide an opportunity to test that the mechanism through which the tool helps consumers is by making search more effective. Additionally, identifying correlates of category-level heterogeneity can help provide managerial insights into which types of products will generate the highest return if platforms consider implementing search refinement tools like PTG.

To do this, we characterize each product category using two metrics that proxy for search difficulty: the average number of searches per order, and the concentration ratio, defined as the market share of the top 100 products in each of the underlying subcategories.[21] One would expect the search tool to be most helpful for product categories where users have to perform many searches to find the products they want, and for product categories where sales are dispersed across a larger number of distinct products.

Figure 3 plots the percent change in GMV induced by the search tool against our two proxies for search difficulty. Panel *a* confirms that the search tool has a bigger positive impact in product categories where users search more for each product purchase. Similarly, Panel *b* shows that the search tool has a bigger impact in product categories where sales are less concentrated. Our results thus indicate that Apparel and Fashion, as well as Home Furnishing products, can benefit the most from search refinement tools like PTG. More generally, for product categories where consumers engage in intensive searches and sales are not concentrated among a few big sellers, PTG-like search refinement tools will be most effective.

Mechanisms: Direct Benefits, Frequency of Use, and Consumer Learning. There are a number of possible mechanisms explaining the aggregate improvement in purchases and expenditures given a comparable number of product views and clicks. The first possibility is

[21]The average number of searches per order is calculated by dividing the total number of searches by the total number of product orders for each category. We use our sample data collected prior to the experiment to construct this metric. To determine the concentration ratio for each product category, we first rank products from best to worst selling within their respective subcategories according to their 2021 sales revenues, similar to the methodology used in Table 7. We define the subcategory-level CR100 concentration ratio as the the market share of the top 100 products within each subcategory. Subsequently, the category-level concentration ratio is calculated by taking the average of the CR100 concentration ratios for all subcategories within the specific category. We use CR100 instead of smaller concentration ratios because the sales distribution is more dispersed online compared to offline. Many subcategories have more than tens of thousands of products. Therefore, commonly used measures such as CR4, CR20, or HHI are not appropriate in the online context. Instead, we use the subcategory-level CR100, a measure also adopted by the platform, to assess the concentration level for each category.



(a) (Anonymized) Number of Searches per Order (b) (Anonymized) Concentration Ratio

Figure 3: Heterogeneous Treatment Effects by Product Categories.

Notes: The figures plot the percent change in GMV induced by the treatment (the percent changes presented in Appendix Table A.9) against two proxies for search difficulty: the number of searches per order in Panel *a*, and the market share of the top 100 selling products (averaged across all underlying sub-categories) in Panel *b*. The size of each dot is proportional to the GMV share of each product category prior to the experiment. The red dashed line is a fitted line where each category is weighed by its GMV share prior to the experiment. To comply with the platform's confidentiality requirements, the data on the x-axis has been re-scaled by multiplying it by an arbitrary number.

that the search tool allows consumers to refine their searches whenever the tool is available. This is the most direct benefit of the search tool, which would imply that the improvements are concentrated on searches where the tool is available, i.e., PTG general and specific queries.

To test this, columns 1 and 2 in Table 8 estimate the effect of the search tool on orders and expenditures generated through PTG general and specific queries. The coefficient estimates confirm large and significant effects of the search tool for PTG queries: orders increase by 2.2% and expenditures increase by 5.3% in the 24 weeks since the consumer's entry in the experiment.

The second, more indirect, channel through which the search tool can be helpful is by increasing overall consumer satisfaction with the platform, which in turn can increase consumer loyalty. This hypothesis would imply that consumers in the treatment group use the platform more often than consumers in the control group. To test this, we consider two metrics: the number of days performing searches and the number of categories searched.[22] We regress the two metrics on the treatment indicator to test whether treated consumers

[22] As each query is related to various purchased products, we can calculate each query's number of orders within every product category. We define a query's category as the one with the highest number of orders attributed to it.

Table 8: Treatment Effects for PTG and Non-PTG Queries

	PTG Queries		Non-PTG Queries	
	Orders	GMV	Orders	GMV
	(1)	(2)	(3)	(4)
Treat	0.0453***	7.779***	0.291**	54.67**
	(0.0128)	(2.452)	(0.131)	(26.00)
% Change	2.22%	5.34%	1.50%	3.07%
Observations	346,110	346,110	346,110	346,110
R-squared	0.01	0.001	0.031	0.006

Notes: Regression estimates of Equation 1. The dependent variables are in levels. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses.

use the platform more than control consumers. As shown in Appendix Table A.10, the PTG search tool does not increase the overall usage of the platform.

Finally, the third possibility that we explore is that consumers learn to perform more effective searches from PTG, which they can then apply even to product searches that are not augmented by the search tool. If this were the case, we would expect an increase in orders and spending originating from non-PTG searches, as well as a shift of non-PTG queries towards finer and more specific searches.

We test whether the tool leads to an increase in orders and spending from non-PTG queries. Columns 3 and 4 in Table 8 confirm sizable effects: orders increase by 1.5% and spending increases by 3.1%. The increases estimated in columns 3 and 4 are larger in levels compared to those estimated in columns 1 and 2 (because a bigger share of sales come from non-PTG queries), but the opposite is true in percentage terms. This result confirms that the percent increase in orders and spending directly linked to PTG queries is larger than the indirect effect on non-PTG queries.

Does the increase in orders and spending on non-PTG queries come from consumers learning to perform more effective searches on their own? We start to explore this possibility by focusing on the text length of consumer searches. We would expect that consumers in the treatment group may learn to conduct more specific searches, perhaps using longer descriptions of the items they want. We measure query length as the number of Chinese characters that the consumer types in the search box. We then compute the average query length across all searches performed in the 24 weeks since entry in the experiment, and the average query length across all non-PTG searches.

Columns 1 and 2 of Table 9 present the results. Across all searches, the average query length increases by 0.02 characters, or 0.3%. For non-PTG searches, the result is smaller

in magnitude, as expected given that it is an indirect effect, but statistically different from zero at conventional levels.

Table 9: Tests for Consumer Learning

Outcome:	Avg Query Length	Avg Query Length	Number of Searches	Number of Searches
Query Type:	All Queries	Non-PTG	Matched Non-PTG	Unmatched Non-PTG
	(1)	(2)	(3)	(4)
Treat	0.0191***	0.0126**	0.313**	-0.790
	(0.00479)	(0.00535)	(0.133)	(0.682)
% Change	0.31%	0.20%	2.41%	-0.40%
Observations	346,110	346,110	346,110	346,110
R-squared	0.021	0.045	0.012	0.135

Notes: Regression estimates of Equation 1. The dependent variables are in levels. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses.

As they search for products not included in PTG, consumers might also start using words they learned from PTG queries and integrate them into other searches. For instance, a consumer could discover the term "Halterneck" from a PTG specific query when searching for dresses, and then use the same descriptor in other searches, e.g., "Halterneck Top." To investigate this possibility, we make a list of all the words that the search tool uses in PTG specific queries (e.g., "Halterneck"). We call this list the PTG specific vocabulary. We then categorize non-PTG queries into two groups: queries whose words match at least one word included in the PTG specific vocabulary (*matched non-PTG queries*); and queries whose words do not match any of the words included in the PTG specific vocabulary (*unmatched non-PTG queries*). In the example above, the search for "Halterneck Top" would be classified as a matched non-PTG query. We want to test whether the number of matched non-PTG queries increases in the treatment group relative to the control group.

Columns 3 of Table 9 shows that the number of matched non-PTG queries significantly increases by 0.3 queries, or 2.4% relative to the baseline level. Although the coefficient estimate is negative in column 4, the increase in matched queries does not seem to come at the expense of unmatched queries.

Consumer Satisfaction. So far, we have showed that consumers perform more effective searches through the help of PTG. The benefits arise both from the direct use of the tool, and the indirect learning provided by the tool. Before concluding, we want to ensure that the additional purchases induced by PTG are as good as or better than the purchases consumers

would make in the absence of PTG. Indeed, there are worries that digital platforms may exploit consumer biases to induce their users to overspend (Fletcher et al., 2023; Spencer, 2020), so we want to evaluate whether treated consumers regret their purchases more or less compared to control consumers.

We consider all orders placed in the 24 weeks since the consumer's entry into the experiment. This amounts to 7,479,300 orders placed by the 346,110 consumers for whom we have at least 24 weeks of data, for an average of 22 orders per consumer. We run two linear probability models of the following type:

$$y_{ij} = \beta * Treat_i + \alpha_{c(i)} + \epsilon_{ij}, \tag{2}$$

where i denotes the consumer as in Equation 1, and j denotes a product purchased during the relevant time period. For y_{ij}, we use two proxies for consumer satisfaction: an indicator for whether the consumer submits a positive review for the purchased product (i.e., 4- or 5-star review out of a 1-5 scale); and an indicator for whether the consumer returns the product and requests a refund. We also control for consumers' day of entry into the experiment with cohort $c(i)$ fixed effects.

Table 10: Effects on Consumer Satisfaction: Positive Reviews and Return Rates

	Positive Rating	Return
	(1)	(2)
Treat	0.00636***	-0.00274***
	(0.000201)	(0.0000731)
% Change	3.56%	-3.22%
Observations	7,479,300	7,479,300
R-squared	0.003	0.012

Notes: Regression estimates of Equation 2. An observation is an order placed in the 24 weeks following a consumer's entry in the experiment. In column 1, the dependent variable is an indicator for a positive rating. When the rating is on a 1-5 scale, we define 4- or 5-stars as positive rating. In column 2, the dependent variable is an indicator for whether the consumer initiated a request for return and refund. Standard errors are in parentheses. Similar regressions restricting attention to orders related to PTG queries are displayed in Appendix Table A.11.

Results are presented in Table 10. Purchases from consumers in the treatment group are significantly more likely to be rated positively (0.64 percentage points more likely) and less likely to be returned (0.27 percentage points less likely). The effects are substantial relative to the baseline, with a 3.56% increase in the positive rating probability and a 3.22% decrease in the return probability. The results thus confirm that the purchases made with the support

of PTG are perceived as higher quality.

5 Conclusion

Our research shows that improving search tools with textual and visual suggestions can be an effective way to help consumers express and develop their preferences. Leveraging a long-run experiment linked to the launch of a text- and picture-based search refinement tool on a major e-commerce platform, we find that having access to textual and visual search recommendations increases purchases by 1.57% and spending by 3.24%. The increase does not seem to be driven by consumers viewing or clicking on more products, nor by consumers conducting more searches, suggesting an increase in search effectiveness. The increase is not only driven by searches that are directly affected by the search tool, but rather it spills over to other searches, implying that consumers learn to perform better searches on their own.

The value of the tool is concentrated among a subset of buyers and a subset of sellers. On the buyer side, we find two distinct results. On one hand, the tool helps older consumers find what they need. Younger consumers do not seem to be greatly affected by the search refinement tool, perhaps given their intrinsic ability to search online. On the other hand, we also find that heavier consumers of the platform (rather than lighter consumers) benefit the most from the tool. The latter result suggests that experience with the platform is not enough to reduce search costs related to demand expression and demand formation.

Although the experiment directly affected demand, we find important indirect effects for sellers as well. In particular, the ability of the search tool to narrow down searches to more specific product categories benefits products and sellers outside of the most popular, reducing the concentration of sales among the top sellers.

Importantly, we also find that if we restrict our analysis to the short-run, we are unable to detect the significant benefits of the new search tool. In fact, we find that the immediate effect is a precisely estimated zero. This result likely reflects the fact that people entering the experiment are visiting the platform with a specific intent to buy (or not to buy) something, which the search tool does not immediately impact. It also implies that it takes time for search tools like the one we study to display their beneficial effects on consumers. In this specific case, our analysis reveals that it takes about one month since entry in the experiment for consumers to experience the improvements in search effectiveness driven by the refinement tool.

Our results have important implications for the design of search mechanisms, and for the design of experiments. Related to search mechanisms, our results highlight the importance of addressing demand expression and demand formation challenges. The design of search

mechanisms is evolving away from purely relying on consumer prompts to identify what consumers want, towards increasing the role of machine learning to leverage large data on consumer search and purchasing behavior (Zhang et al., 2020). In a world where platforms can very accurately predict consumer preferences, there would not even be any need for search because platforms would ship products before consumers even have a chance to search (Agrawal et al., 2022). Despite the many promises of product recommendations, our paper highlights that platforms are still far from the frontier of search optimization. There thus remains ample opportunity to invest resources to turn consumer data into insights to guide future searches.

Our paper also emphasizes that search refinement tools have to be used to be useful. One could argue that PTG is nothing more than an additional filter. Yet, data from Farronato et al. (2023) suggest that only about 15% of searches use filters on amazon.com. So, platforms not only need to invest in tools to make search easier, but those tools need to be designed for consumers to effectively leverage them. PTG could have been designed as an additional filter whose options consumers could check. Instead, by suggesting refinements immediately below the search bar, PTG makes it easier for consumers to benefit from it. It remains an open question which combination of search tools is most effective, and which position on the search page each tool should occupy.

When a platform considers introducing search refinement tools with textual and visual suggestions, specific user segments and supply categories should be carefully targeted. In this instance, older consumers and heavier platform users were the best targets, a result likely applicable more broadly to search refinement functionalities. Similarly, product categories such as Apparel and Fashion, or Home Furnishing, are likely to yield the highest returns from such tools because in those categories, it is particularly cumbersome for consumers to find what they want.

The value of search refinement tools for niche products has broader platform implications. Indeed, when presenting search results to consumers, platforms face a main trade-off between rewarding past seller success and promoting smaller sellers without an established reputation. Our paper highlights that effective search refinement tools can benefit small sellers and niche products without harming, and in fact while enhancing customer satisfaction.

Search mechanisms have historically been relying primarily on textual prompts. Only in recent years, some digital platforms have started offering visual recommendations (such as Amazon) or visual-based searches (such as Google Lens). Although our research cannot separate the role of textual versus visual suggestions, it highlights the potential value of visual cues. An important avenue for future research would be to identify the separate and complementary benefits of the two.

A final implication on search design is the fact that our results do not seem to fit standard search models (Stigler, 1961; McCall, 1965; Mortensen, 1970): if PTG improves the match quality of products showed, existing theories cannot explain why search effort (as measured by product views and clicks) stays constant. Our results are instead consistent with a model where search effort is fixed regardless of the outcome (Dinerstein et al., 2018). For example, it could be that search effort is determined by the opportunity cost of time (Greminger et al., 2023), which our experiment does not change. Under this model, platforms need to focus on search tools that allow consumers to quickly reach the products they like, because wasted time, although potentially entertaining, can result in consumers leaving the platform without purchases.

Related to experiment design, our results highlight the risk of drawing conclusions from short-run experiments. Despite recent efforts to identify long-run results from short-term metrics (Athey et al., 2019), there are many contexts, such as ours, where the best approach is simply to run a long-run experiment. In our case, if we had only had access to a few weeks of data, we would have concluded that the search tool did not make consumer search more effective, both because the estimates are too noisy to detect a significant effect and because the short-term impact is quantitatively close to zero. It is thus important to expand research to understand when and how practitioners and researchers can rely on short-run experiments, and when instead longer run approaches may be required.

The paper has a number of limitations. Our analysis is unable to look at how sellers would respond to the roll out of search refinement tools like the one we study. There are two main reasons for our focus on the demand side. Although increasing over the course of the experimental period, the searches qualifying for the refinement tool accounted for only about 15% of GMV linked to searches (Figure 2) by the end of 2021. Additionally, because of the experimental roll-out, not all users had the opportunity to benefit from the search tool recommendations. We leave the important question of how sellers would adjust their product offering in response to changes in search design to future research.

Many recommendation systems run the risk of recommending impulse purchases that consumers may ex-post regret. We have robust analyses indicating that the tool had net benefits on consumers: they purchased more given the same search effort, and those purchases had higher ratings and lower returns. However, it is possible that ratings and return rates may not capture longer-term regret. Similarly, because choices may not reflect true preferences, search recommendation tools risk diverting consumers away from what they truly want towards what the platform can measure, in a potential spiral of error propagation (Fu et al., 2022). In the financial setting, for example, this has been identified as a potential risk of autocomplete tools for stock tickers (Rubin and Rubin, 2021). The combination

of search aid tools, ranking algorithms, and product recommendations, Mik (2016) argues, risks eroding consumer autonomy in online transactions. Although in our setting consumer satisfaction metrics suggest otherwise, we leave this important and under-explored topic to future research.

References

Adomavicius, Gediminas, Jesse Bockstedt, Shawn P Curley, Jingjing Zhang, and Sam Ransbotham, "The hidden side effects of recommendation systems," *MIT Sloan Management Review*, 2019, *60* (2), 1.

_ , Jesse C Bockstedt, Shawn P Curley, and Jingjing Zhang, "Do recommender systems manipulate consumer preferences? A study of anchoring effects," *Information Systems Research*, 2013, *24* (4), 956–975.

_ , _ , _ , and _ , "Effects of online recommendations on consumers' willingness to pay," *Information Systems Research*, 2018, *29* (1), 84–102.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*, Harvard Business Press, 2022.

Athey, Susan, Raj Chetty, Guido W Imbens, and Hyunseung Kang, "The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely," *NBER Working Paper No. 26463*, 2019.

Bar-Isaac, Heski, Guillermo Caruana, and Vicente Cuñat, "Search, design, and market structure," *The American Economic Review*, 2012, *102* (2), 1140–1160.

Blanco, Carlos Flavián, Raquel Gurrea Sarasa, and Carlos Orús Sanclemente, "Effects of visual and textual information in online product presentations: looking for the best combination in website design," *European Journal of Information Systems*, 2010, *19* (6), 668–686.

Bronnenberg, Bart J., Jun B. Kim, and Carl F. Mela, "Zooming in on choice: How do consumers search for cameras online?," *Marketing Science*, 2016, *35* (5), 693–712.

Brynjolfsson, Erik, Yu Hu, and Duncan Simester, "Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales," *Management Science*, 2011, *57* (8), 1373–1386.

Chen, Jiafeng and Jonathan Roth, "Logs with zeros? Some problems and solutions," *The Quarterly Journal of Economics*, 2024, *139* (2), 891–936.

Chen, Yuxin and Song Yao, "Sequential search with refinement: Model and application with click-stream data," *Management Science*, 2017, *63* (12), 4345–4365.

_ , **Zhe Yuan, Tianshu Sun, and AJ Chen**, "Understanding the Impacts of Depersonalization in Search Algorithm on Consumer Behavior: A Field Experiment with a Large Online Retail Platform," *Available at SSRN 4412157*, 2023.

Chiou, Lesley and Catherine Tucker, "Search engines and data retention: Implications for privacy and antitrust," Technical Report, National Bureau of Economic Research 2017.

Choi, Michael, Anovia Yifan Dai, and Kyungmin Kim, "Consumer search and price competition," *Econometrica*, 2018, *86* (4), 1257–1281.

Dinerstein, Michael, Liran Einav, Jonathan Levin, and Neel Sundaresan, "Consumer price search and platform design in internet commerce," *The American Economic Review*, 2018, *108* (7), 1820–1859.

Elberse, Anita and Felix Oberholzer-Gee, "Superstars and underdogs: An examination of the long tail phenomenon in video sales," *Division of Research, Harvard Business School*, 2006, *7*.

Ellison, Glenn and Sara Fisher Ellison, "Search, obfuscation, and price elasticities on the internet," *Econometrica*, 2009, *77* (2), 427–452.

Farronato, Chiara, Andrey Fradkin, and Alexander MacKay, "Self-preferencing at Amazon: evidence from search rankings," in "AEA Papers and Proceedings," Vol. 113 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2023, pp. 239–243.

Filippas, Apostolos, John J Horton, and Diego Urraca, "Advertising as coordination: Evidence from a field experiment," *Working Paper*, 2023.

_ , **John J. Horton, and Joseph M. Golden**, "Reputation inflation," *Marketing Science*, 2022, *41* (4), 733–745.

Fleder, Daniel and Kartik Hosanagar, "Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity," *Management Science*, 2009, *55* (5), 697–712.

Fletcher, Amelia, Gregory S Crawford, Jacques Crémer, David Dinielli, Paul Heidhues, Michael Luca, Tobias Salz, Monika Schnitzer, Fiona M Scott Morton, Katja Seim et al., "Consumer protection for online markets and large digital platforms," *Yale J. on Reg.*, 2023, *40*, 875.

Fong, Nathan M, "How targeting affects customer search: A field experiment," *Management Science*, 2017, *63* (7), 2353–2364.

Fradkin, Andrey, "Search, matching, and the role of digital marketplace design in enabling trade: Evidence from airbnb," *Available at SSRN 2939084*, 2017.

Fu, Runshan, Ginger Zhe Jin, and Meng Liu, "Does human-algorithm feedback loop lead to error propagation? Evidence from Zillow's Zestimate," *NBER Working Paper No. 29880*, 2022.

Gardner, Martin, "Mathematical games," *Scientific American*, 1970, *222* (6), 132–140.

Goli, Ali, David H. Reiley, and Hongkai Zhang, "Personalized Versioning: Product Strategies Constructed from Experiments on Pandora," *Working Paper*, 2021.

Greminger, Rafael P, "Optimal search and discovery," *Management Science*, 2022, *68* (5), 3904–3924.

Greminger, Rafael, Yufeng Huang, and Ilya Morozov, "Make every second count: Time allocation in online shopping," *Available at SSRN*, 2023.

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey et al., "Top challenges from the first practical online controlled experiments summit," *ACM SIGKDD Explorations Newsletter*, 2019, *21* (1), 20–35.

Hagiu, Andrei and Julian Wright, "Data-enabled learning, network effects, and competitive advantage," *The RAND Journal of Economics*, 2023, *54* (4), 638–667.

Häubl, Gerald and Kyle B Murray, "Preference construction and persistence in digital marketplaces: The role of electronic recommendation agents," *Journal of Consumer Psychology*, 2003, *13* (1-2), 75–91.

_ **and Valerie Trifts**, "Consumer decision making in online shopping environments: The effects of interactive decision aids," *Marketing Science*, 2000, *19* (1), 4–21.

Honka, Elisabeth and Pradeep Chintagunta, "Simultaneous or sequential? Search strategies in the US auto insurance industry," *Marketing Science*, 2017, *36* (1), 21–42.

_ **, Stephan Seiler, and Raluca Ursu**, "Consumer search: What can we learn from pre-purchase data?," *Journal of Retailing*, 2024, *100* (1), 114–129.

Huang, Jason, David Reiley, and Nick Riabov, "Measuring consumer sensitivity to audio advertising: A field experiment on pandora internet radio," *Available at SSRN 3166676*, 2018.

Hui, Xiang and Meng Liu, "Quality certificates alleviate consumer aversion to sponsored search advertising," 2022.

Jiang, Baojun and Tianxin Zou, "Consumer search and filtering on online retail platforms," *Journal of Marketing Research*, 2020, *57* (5), 900–916.

Kamenica, Emir, Sendhil Mullainathan, and Richard Thaler, "Helping consumers know themselves," *The American Economic Review*, 2011, *101* (3), 417–422.

Kohavi, Ron, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu, "Trustworthy online controlled experiments: Five puzzling outcomes explained," in "Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining" 2012, pp. 786–794.

Lazonder, Ard W, "Do two heads search better than one? Effects of student collaboration on web search behaviour and search outcomes," *British Journal of Educational Technology*, 2005, *36* (3), 465–475.

Lee, Dongwon, Anandasivam Gopal, and Sung-Hyuk Park, "Different but equal? A field experiment on the impact of recommendation systems on mobile and personal computer channels in retail," *Information Systems Research*, 2020, *31* (3), 892–912.

Lee, Kwok Hao and Leon Musolff, "Entry into two-sided markets shaped by platform-guided search," *Working Paper*, 2021.

Lei, Xiaoxia, Yixing Chen, and Ananya Sen, "The value of external data for digital platforms: Evidence from a field experiment on search suggestions," *Available at SSRN*, 2023.

Liu, Jia and Olivier Toubia, "Search query formation by strategic consumers," *Quantitative Marketing and Economics*, 2020, *18*, 155–194.

los Santos, Babur De, Ali Hortaçsu, and Matthijs R Wildenbeest, "Testing models of consumer search using data on web browsing and purchasing behavior," *The American Economic Review*, 2012, *102* (6), 2955–2980.

Markey, Karen, *Online searching: A guide to finding quality information efficiently and effectively*, Rowman & Littlefield, 2019.

McCall, John J, "The economics of information and optimal stopping rules," *The Journal of Business*, 1965, *38* (3), 300–317.

Mik, Eliza, "The erosion of autonomy in online consumer transactions," *Law, Innovation and Technology*, 2016, *8* (1), 1–38.

Moravec, Patricia L, Avinash Collis, and Nicholas Wolczynski, "Countering state-controlled media propaganda through labeling: Evidence from Facebook," *Information Systems Research*, 2023.

Mortensen, Dale T, "Job search, the duration of unemployment, and the Phillips curve," *The American Economic Review*, 1970, *60* (5), 847–862.

Peukert, Christian, Ananya Sen, and Jörg Claussen, "The editor and the algorithm: Recommendation technology in online news," *Management Science*, 2023.

Pu, Jingchuan, Young Kwark, Sang Pil Han, Qiang Ye, and Bin Gu, "Uncertainty reduction vs. reciprocity: Understanding the effect of a platform-initiated reviewer incentive program on regular ratings," *Information Systems Research*, 2023.

Rothschild, Michael, "Searching for the Lowest Price When the Distribution of Prices Is Unknown," *Journal of Political Economy*, 1974, *82* (4), 689–711.

Rubin, Eran and Amir Rubin, "On the economic effects of the text completion interface: empirical analysis of financial markets," *Electronic Markets*, 2021, *31* (3), 717–735.

Seiler, Stephan, "The impact of search costs on consumer behavior: A dynamic approach," *Quantitative Marketing and Economics*, 2013, *11*, 155–203.

Spencer, Shaun B, "The problem of online manipulation," *U. Ill. L. Rev.*, 2020, p. 959.

Stigler, George J, "The economics of information," *Journal of Political Economy*, 1961, *69* (3), 213–225.

Sun, Tianshu, Zhe Yuan, Chunxiao Li, Kaifu Zhang, and Jun Xu, "The value of personal data in internet commerce: A high-stakes field experiment on data regulation policy," *Management Science*, 2023.

Ursu, Raluca M, "The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions," *Marketing Science*, 2018, *37* (4), 530–552.

Wan, Xiang (Shawn), Anuj Kumar, and Xitong Li, "How Do Product Recommendations Help Consumers Search? Evidence from a Field Experiment," *Management Science*, 2023.

Weitzman, Martin L, "Optimal Search for the Best Alternative," *Econometrica*, 1979, *47* (3), 641–654.

Wu, Ruijuan, Heng-Hui Wu, and Cheng Lu Wang, "Why is a picture 'worth a thousand words'? Pictures as information in perceived helpfulness of online reviews," *International Journal of Consumer Studies*, 2021, *45* (3), 364–378.

Yang, Joonhyuk, Navdeep S Sahni, Harikesh S Nair, and Xi Xiong, "Advertising as information for ranking e-commerce search listings," *Marketing Science*, 2023.

Yuan, Zhe, AJ Chen, Yitong Wang, and Tianshu Sun, "How recommendation affects customer search: A field experiment," *Information Systems Research*, 2024.

Zhang, Xingyue Luna, Raluca Ursu, Elisabeth Honka, and Yuliang Oliver Yao, "Product discovery and consumer search routes: Evidence from a mobile app," *Available at SSRN 4444774*, 2023.

Zhang, Yingjie, Beibei Li, and Ramayya Krishnan, "Learning individual behavior using sensor data: The case of global positioning system traces and taxi drivers," *Information Systems Research*, 2020, *31* (4), 1301–1321.

_ , _ , Xueming Luo, and Xiaoyi Wang, "Personalized mobile targeting with user engagement stages: Combining a structural hidden markov model and field experiment," *Information Systems Research*, 2019, *30* (3), 787–804.

Zheng, Shuang, Siliang Tong, Hyeokkoo Eric Kwon, Gordon Burtch, and Xianneng Li, "Recommending what to search: Sales volume and consumption diversity effects of a query recommender system," *Available at SSRN 4667778*, 2023.

Appendix to "Platform Information Provision and Consumer Search: A Field Experiment"

A Additional Figures and Tables

Table A.1: Covariate Balance

		Control Group $n = 252,737$	Treatment Group $n = 252,748$	P-value $(C = T)$
Age Tier	Mean	2.4853	2.4892	0.2654
	Std Err	0.0025	0.0025	
City Tier	Mean	3.8480	3.8518	0.4463
	Std Err	0.0036	0.0036	
Number of Registered Years	Mean	6.2573	6.2636	0.5044
	Std Err	0.0067	0.0066	
Female	Mean	0.5455	0.5454	0.9707
	Std Err	0.0010	0.0010	
View in the Past 8 Weeks	Mean	2585.4	2563.9	0.0865
	Std Err	8.95	8.80	
Clicks in the Past 8 Weeks	Mean	112.0	111.5	0.3835
	Std Err	0.39	0.39	
Orders in the Past 8 Weeks	Mean	4.53	4.55	0.4997
	Std Err	0.02	0.02	
GMV in the Past 8 Weeks	Mean	410.6	415.4	0.4888
	Std Err	5.73	4.06	

Notes: The table displays characteristics of consumers included in the experiment. Consumer characteristics refer to demographics – age grouping (where age is grouped in 10-year groupings, and 1 is assigned to the youngest 10-year grouping between 15 and 25 years old), city tier (where 1 is assigned to the largest cities in China, such as Beijing and Shanghai, 2 is assigned to cities like Hangzhou and Nanjing, all the way to tier 6, which includes the smallest towns and villages), tenure on the platform in years, the proportion of women – and behavior on the platform in the 8 weeks preceding their entry into the experiment – product views, clicks, orders, and GMV.

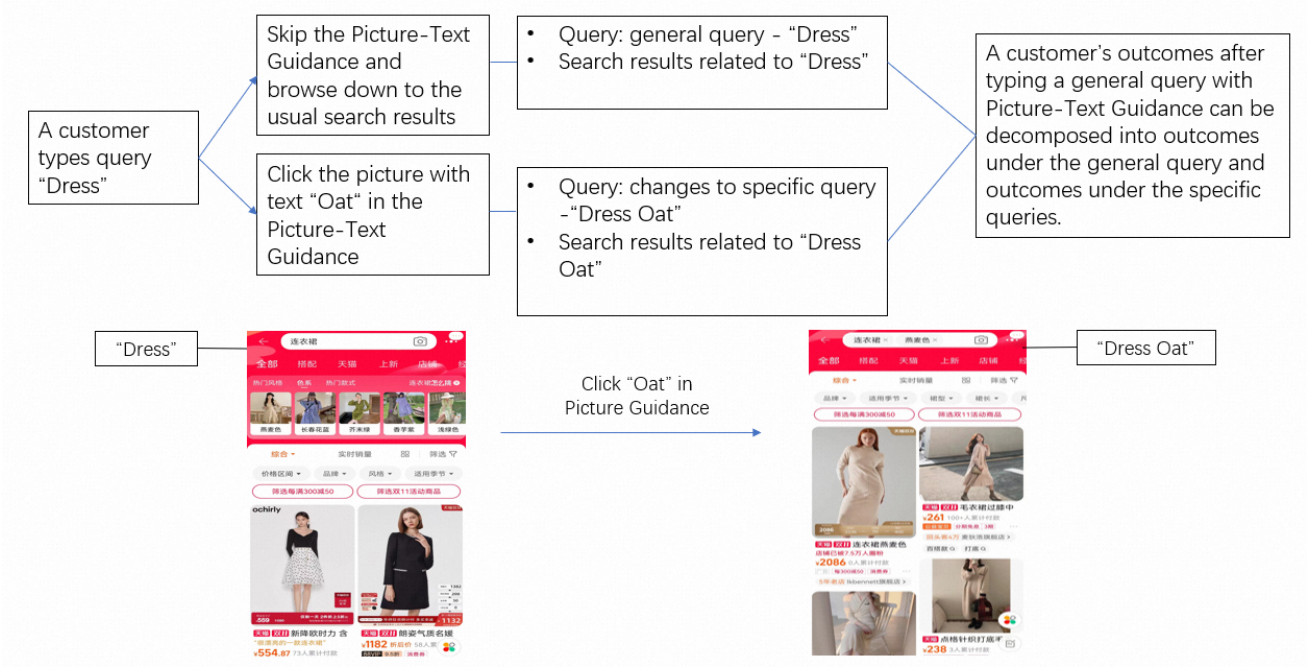


Figure A.1: Consumer Search Process with Picture-Text Guidance

Notes: Treatment group consumers have two options after searching for a PTG general query word. For example, if a consumer types in "Dress" (PTG general query), they can either skip Picture-Text Guidance and search for products related to "Dress," or click on the picture with text "Oat" and search for products related to "Dress Oat" (PTG specific query) instead of just "Dress".

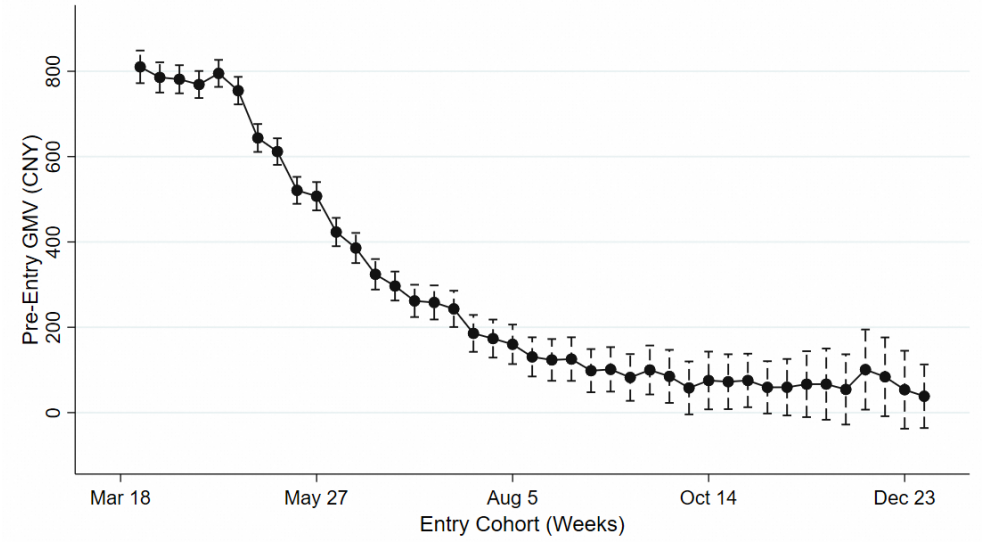


Figure A.2: Spending in the 8 Weeks Preceding Entry into the Experiment.

Notes: The solid dotted line plots the estimated coefficients obtained by regressing consumers' GMV in the 8 weeks prior to entering the experiment on dummies denoting the week of entry into the experiment. Vertical bars denote 95% confidence intervals.

Table A.2: Short-Run Treatment Effects for Non-PTG Queries

	Views	Clicks	Orders	GMV
	(1)	(2)	(3)	(4)
Treat	-0.687	0.0352	0.00166	0.507
	(1.081)	(0.0427)	(0.0031)	(0.638)
% Change	-0.35%	0.47%	0.53%	2.14%
Observations	505,485	505,485	505,485	505,485
R-squared	0.018	0.029	0.01	0.002

Notes: The table presents results similar to Table 3. Instead of focusing on PTG queries, it restrictions attention to views, clicks, orders, and GMV associated to non-PTG queries.

Table A.3: Short-Run Effects on Customer Satisfaction: Positive Reviews and Return Rates

	Positive Rating	Return
	(1)	(2)
Treat	3.82e-07	-0.00115
	(0.00137)	(0.00094)
% Change	0.0003%	-1.37%
Observations	222,517	222,517
R-squared	0.006	0.014

Notes: An observation is a completed purchase on the day a consumer enters the experiment. The dependent variable in column 1 is an indicator for whether the consumer leaves a 4-star or 5-star rating for the purchase. The dependent variable in column 2 is an indicator for whether the consumer requests a return of the item. Standard errors are in parentheses. We include cohort fixed effects.

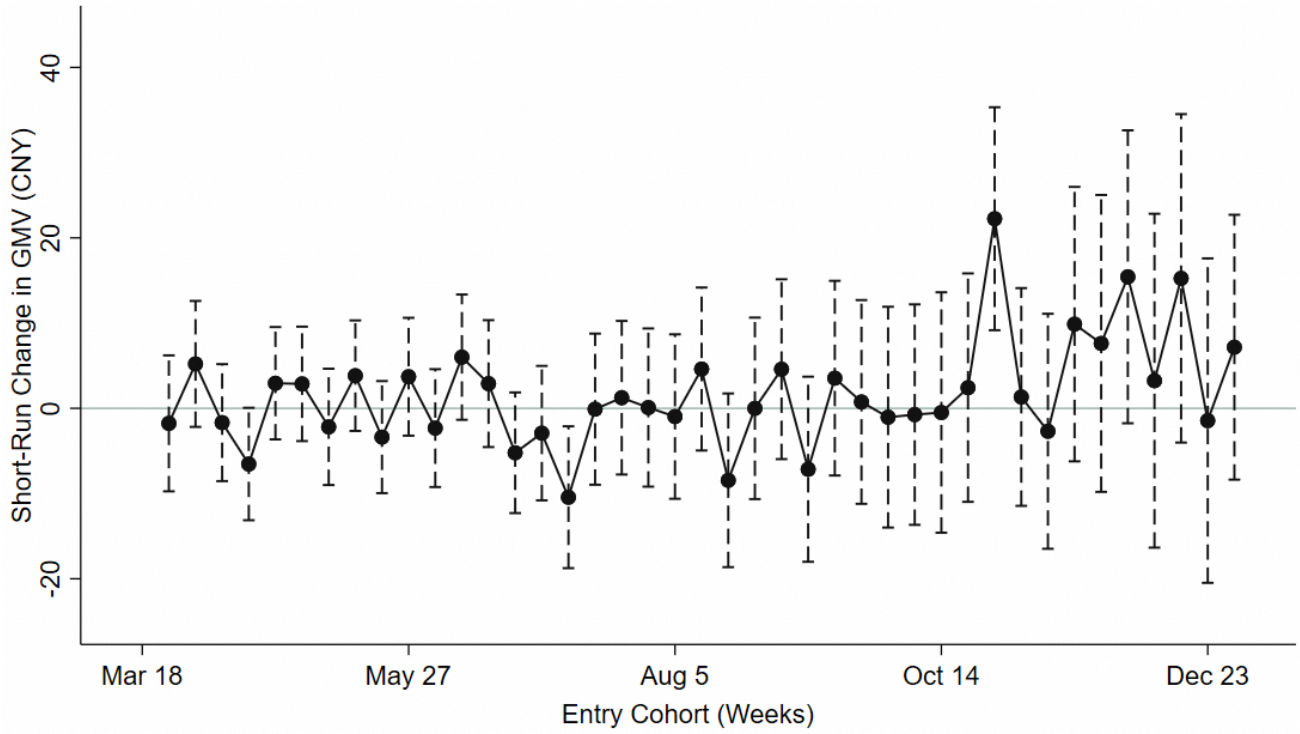


Figure A.3: Short-Run Treatment Effects Across Entry Cohorts

Notes: The figure plots coefficient estimates of Equation 1 where the outcome is GMV on the day a consumer enters the experiment, and the treatment dummy is interacted with each of the entry cohort weeks. Vertical bars denote 95% confidence intervals.

Table A.4: Long-Run Impact on Number of Searches

	Number of Searches	Number of PTG General Searches	Number of PTG Specific Searches
	(1)	(2)	(3)
Treat	0.172	-0.0212	0.670***
	(0.814)	(0.0583)	(0.00523)
% Change	0.08%	-0.14%	450.55%
Observations	346,110	346,110	346,110
R-squared	0.116	0.034	0.048

Notes: The dependent variables are in levels. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses. We include cohort fixed effects according to equation 1.

Table A.5: Long-Run Treatment Effects for Non-Search Channels

	Orders	GMV
	(1)	(2)
Treat	0.0192	0.426
	(0.338)	(0.340)
% Change	0.05%	4.35%
Observations	346,110	346,110
R-squared	0.013	0.001

Notes: The dependent variables are the number of orders placed by consumers and their expenditures within non-search channels, both measured in levels. To comply with the platform's confidentiality guidelines, the data has been normalized by dividing each variable by its respective standard deviation and then multiplying by 100. %Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses. We include cohort fixed effects as per Equation 1.

Table A.6: Treatment Effects Over Different Time Aggregations (for 8, 12, 16, 20 weeks)

	Weeks 1-8		Weeks 1-12		Weeks 1-16		Weeks 1-20	
	Orders	GMV	Orders	GMV	Orders	GMV	Orders	GMV
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat	0.107*	20.59*	0.161**	31.58**	0.202**	41.89**	0.269**	53.32**
	(0.0563)	(10.93)	(0.0791)	(15.55)	(0.101)	(19.79)	(0.122)	(23.92)
% Change	1.38%	2.94%	1.44%	3.11%	1.39%	3.18%	1.50%	3.30%
Observations	346,110	346,110	346,110	346,110	346,110	346,110	346,110	346,110
R-squared	0.028	0.006	0.030	0.006	0.029	0.006	0.029	0.006

Notes: The dependent variables are in levels. In column 1, the number of orders placed by a consumer is aggregated over the first eight weeks since the consumer enters the experiment. In column 3, the number of orders is aggregated over the first twelve weeks, then in column 5 it is calculated over the first sixteen weeks, and in columns 7 over the first twenty weeks. Columns 2, 4, 6, and 8 compute the same aggregations for expenditures. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses. We include cohort fixed effects as per Equation 1.

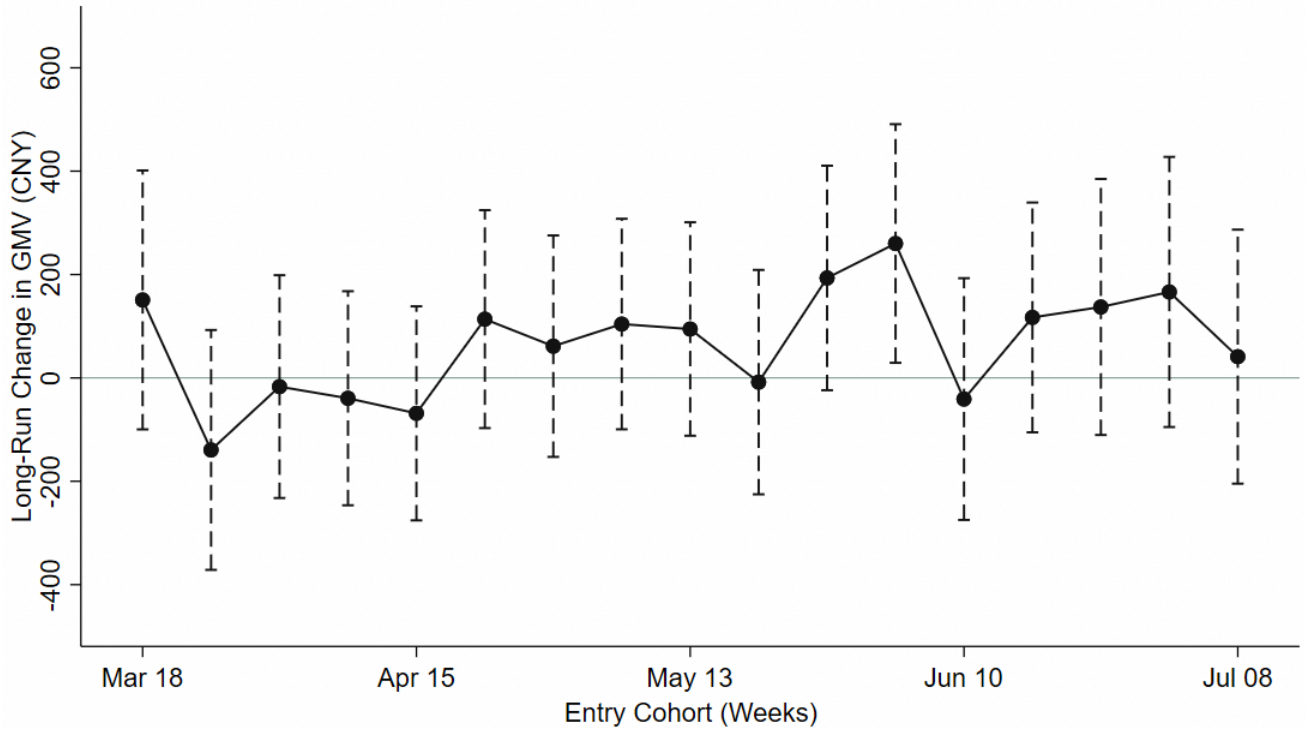


Figure A.4: Long-Run Treatment Effects Across Entry Cohorts

Notes: The figure plots coefficient estimates of Equation 1 where the outcome is GMV in the 24 weeks following a consumer entry into the experiment, and the treatment dummy is interacted with each of the entry cohort weeks. Vertical bars denote 95% confidence intervals.

Table A.7: Heterogeneous Treatment Effects on Orders (Consumers)

	Orders	Orders	Orders	Orders	Orders
	(1)	(2)	(3)	(4)	(5)
Treat	0.281	0.234	0.216	0.392**	0.159
	(0.205)	(0.160)	(0.170)	(0.199)	(0.152)
Treat*Young	0.061				
	(0.267)				
Treat*Big City		0.258			
		(0.281)			
Treat*New			0.251		
			(0.268)		
Treat*Female				-0.134	
				(0.265)	
Treat*Heavy					0.632**
					(0.304)
Observations	346,110	346,110	346,110	346,110	346,110
R-squared	0.146	0.146	0.146	0.146	0.146

Notes: The dependent variables are Orders in levels. Column 1 reports different treatment results by consumer age, where "Young=1" refers to consumers younger than 35 years old. Column 2 reports the results by city tier, where "Big City=1" denotes consumers residing in first and second-tier cities (in China, the cities are categorized into six tiers, and first and second-tier cities typically refer to large cities). Column 3 reports the results by the number of registered years on the platform, where "New=1" refers to consumers who created their account in the last five years. Column 4 reports the results by gender, where "Female=1" refers to consumers who self-identify as female. Column 5 reports the results by spending, where "Heavy=1" denotes consumers in the top quartile of expenditures during the 8 weeks prior to entering the experiment. Standard errors are in parentheses. We include cohort fixed effects as per Equation 1.

Table A.8: Heterogeneous Treatment Effects on Orders (Products and Sellers)

	Panel I: Orders Across Products Grouped by Products' GMV Rank				
	Top 10	10-100	100-1000	1000-10000	Beyond 10000
	(1)	(2)	(3)	(4)	(5)
Treat	0.0182	0.0438**	0.0653	0.0863**	0.119**
	(0.0125)	(0.0216)	(0.0403)	(0.0399)	(0.0527)
% Change	0.8%	1.12%	1.23%	1.78%	2.59%
Observations	346,110	346,110	346,110	346,110	346,110
R-squared	0.047	0.046	0.024	0.018	0.01
	Panel II: Orders Across Products Grouped by Sellers' Revenue Quantile				
	Top 20%	Med-high 20%	Medium 20%	Med-low 20%	Low 20%
Treat	0.0113	0.0176	0.0740**	0.113***	0.118***
	(0.0181)	(0.0441)	(0.0361)	(0.0369)	(0.0452)
% Change	0.48%	0.44%	1.62%	2.31%	2.31%
Observations	346,110	346,110	346,110	346,110	346,110
R-squared	0.029	0.013	0.023	0.022	0.014

Notes: The table is identical to Table 7 except that the outcome variable is number of Orders rather than GMV. The dependent variable is in levels in all columns. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses. We include cohort fixed effects according as per Equation 1.

Table A.9: Heterogeneous Treatment Effects on GMV (by Product Categorises)

	Apparel and Fashion	Electronic Products	Fast-Moving Consumer Goods	Food	Healthcare and Medicine
	(1)	(2)	(3)	(4)	(5)
Treat	24.16***	1.090	5.288*	1.625	2.854**
	(8.222)	(9.297)	(2.965)	(2.164)	(1.113)
% Change	4.92%	0.29%	1.90%	2.01%	4.50%
Observations	346,110	346,110	346,110	346,110	346,110
R-squared	0.004	0.002	0.013	0.002	0.004
	Stationary and Educational Supplies	Home Decoration	Home Furnishing	Automobiles	Others
	(6)	(7)	(8)	(9)	(10)
Treat	0.267	1.064	22.82*	1.623	1.651
	(0.873)	(2.937)	(12.25)	(1.748)	(4.649)
% Change	0.94%	0.68%	6.45%	2.54%	4.51%
Observations	346,110	346,110	346,110	346,110	346,110
R-squared	0.002	0.004	0.001	0.002	0.000

Notes: The dependent variables, always GMV, are in levels. We categorize consumers' total expenditures into ten different product categories. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses. We include cohort fixed effects according as per Equation 1.

Table A.10: Effects on Consumer Frequency of Use of the Platform

	Number of Search Days	Number of Query Categories
	(1)	(2)
Treat	-0.099	-0.039
	(0.120)	(0.290)
% Change	-0.16%	-0.04%
Observations	346,110	346,110
R-squared	0.14	0.14

Notes: The dependent variables are in levels. This table considers two metrics: the number of days performing searches and the number of search query categories. As each query is related to various purchased products, we can calculate each query's number of orders within every product category. We define a query's category as the one with the highest number of orders attributed to it. % Change is calculated by dividing the treatment effect by the control group average. Standard errors are in parentheses. We include cohort fixed effects as per Equation 1.

Table A.11: Positive Reviews and Return Rates for PTG queries

	Positive Rating	Return
	(1)	(2)
Treat	0.00515***	-0.00363***
	(0.000686)	(0.000384)
% Change	3.28%	-4.58%
Observations	713,883	713,883
R-squared	0.002	0.004

Notes: The table is identical to Table 10, except that the observations are restricted to orders directly related to PTG queries.