Advertising as Information for Ranking E-Commerce Search Listings[†]

Joonhyuk Yang

Mendoza College of Business
University of Notre Dame

joonhyuk.yang@nd.edu

Navdeep S. Sahni

Graduate School of Business
Stanford University

navdeep.sahni@stanford.edu

Harikesh S. Nair

Graduate School of Business
Stanford University

harikesh.nair@stanford.edu

Xi Xiong

TikTok

xiongxi9@gmail.com

This version: September 21, 2022

[†]We thank Jack Lin, Xiliang Lin, Bo Long, Yun Xiao, Paul Yan and JD.com's search and ads teams for support with data access and business context. We also thank Brett Gordon, seminar participants at the Stanford Quant Marketing Seminar, Kellogg Quant Marketing Workshop, KAIST College of Business, Korea University Business School, Harvard Business School, University of Notre Dame and 2021 ISMS Marketing Science Conference, and Caio Waisman in particular for helpful comments and suggestions. The randomization involved in the experiment are covered by JD.com's terms of service agreed via opt-in by users at the point of sign-up for the platform's services. Nair served as a consultant to JD.com and Xi was an employee of JD.com during the time the study was conducted. First version: Aug 17, 2021.

Advertising as Information for Ranking E-Commerce Search Listings

Abstract

Search engines and e-commerce platforms have substantial difficulty exposing new products to their users on account of an information problem: new products typically do not have enough sales or other user-engagement that enables platforms to reliably assess product quality. This paper evaluates the role of advertising in providing information to the platform regarding new product quality so as to solve this "cold-start" problem and to engineer higher quality organic listings. Using a large-scale experiment implemented at `JD.com`—a large e-commerce platform in China—we show that using ad propensity information for ranking new products benefits both the platform and consumers. Our findings showcase a new channel by which advertising can potentially improve outcomes for consumers and platforms in e-commerce, through its ability to reveal information that can be used by platforms to improve search ranking algorithms.

Keywords: Advertising as information, Search engines, New products, Platform businesses, E-Commerce, Field-experiments.

JEL Codes: M37, D82, D83, C93

1 Introduction

Sorting products and presenting the best set of options in response to a user's search query is a crucial function of platforms such as online retailers and search engines. A fundamental information problem makes this task difficult; the platform's ability to predict the most relevant product list for the user depends on what the platform knows about the products. For this purpose, platforms spend significant resources and gather a lot of information from historical clicks, reviews, and purchases. Still, this task is hard, because of the massive scale of the problem. For example, retail platforms typically offer millions of products for sale, and cater to millions of consumers with heterogeneous preferences looking for products to buy. Clearly, this challenge is more serious when it comes to ranking new products where platforms suffer from a "cold-start problem" because information about new products is generally scarce.

Speaking to this problem, this paper empirically investigates a novel information channel: i.e., how platforms can engineer better organic listing algorithms by using information on advertising actions (specifically, whether they previously advertised, and how much they spent on advertising) to improve their assessments of new products and address new product discoverability.

Among multitudes of information sources, what is special about advertising? Unlike other sources of data such as clicks and reviews which are outcomes of consumer evaluation, advertising is directly controlled by sellers, and search engines have long worried that giving weightage to advertising can deteriorate the quality of search engine listings (Brin and Page, 1998). Consequently, while many new products are advertised and these decisions are observed by platforms, in practice, the major platforms do not use information on whether a product was advertising for their organic product ranking. For example, Google (2021) clarifies Google's policy regarding the independence of its organic rankings and advertising:

> "Google Ads are paid online advertisements which appear next to relevant searches and other content on the web. Running a Google Ads campaign does not help your SEO [Search Engine Optimization] rankings, despite some myths and claims."

Overall, platforms are known to gather and use vast amounts of data for the purpose of search rankings, but not advertising data.

Theory, however, suggests a possibility that the act of advertising can be informative, and help separate high and lower quality sellers. The literature started by Nelson (1974) suggested that the act of advertising in and of itself may be informative to market participants because firms producing products of higher unobserved quality (unobserved to consumers and other market intermediaries) may have higher incentive to advertise than firms producing products of lower unobserved quality, leading to advertising becoming an informative signal of such quality. Recent research (e.g., Abhishek et al., 2019; Sahni and Nair, 2020) has extended these insights to modern

digital platform settings and have shown the potential "signaling" role of advertising.[1]

However, it remains unclear whether platforms can leverage the information contained in advertising decisions in a way that improves their search algorithms; and whether this can materially affect their total sales and user experience. This paper empirically investigates this possibility. Specifically, we argue the information contained in past seller advertising decisions can be used to improve platforms' organic listing algorithms. We demonstrate this via a field experiment on `JD.com`—an e-commerce company where we helped develop such an algorithm.

There are some complex challenges inherent in such an evaluation that are worth noting. First, assessing whether using advertising information improves search listings is fundamentally a distinct question from whether a signaling equilibrium exists, requiring different evaluation strategies and different empirical methods. This is because the mere prevalence of a signaling equilibrium does not imply that the platform can harness advertising information fruitfully. Incorporating advertising information may not materially change the platform's rankings, or might cause users to substitute one product with another without increasing total platform sales. Also, the impact may depend on the mix of users exposed to advertised and organic listings. Further, a signaling equilibrium implies that by boosting advertised products in the organic listings, the platform would be boosting products that are high quality – in terms of their "experience" attributes – which the users cannot observe from the organic placement. So the users who see such products listed in organic listings may not perceive them to be of high quality, unless they also saw the ads for those products.[2] Hence, boosting such products may even backfire if such users do not try them and obtain disutility from seeing them in organic listings, perceiving them to be lower quality.

All of this makes assessing whether using advertising information improves search listings a fundamentally empirical question. Whether there is an improvement, and the magnitude of any such improvement depends on the sophistication of the platform's existing search algorithm in sorting between good and bad new products; the information content of the data utilized by the platform in training its existing search algorithms; and on user's search and choice behavior on the platform. The question of improvement pertains to whether advertising provides incremental benefit over and above what is *in place*, which requires evaluation *in place* on an actual platform. It is hard to conceive of implementing such a study without close cooperation of a platform.

Second, this evaluation requires comparing outcomes under the existing search algorithm to outcomes under the new search algorithm which leverages the advertising information. This requires

[1] Sahni and Nair (2020) provide field experimental evidence that the act of advertising on search engines on online platforms (via so-called paid "sponsored listings") can serve as a signal of unobserved product quality to searching consumers on the platform's marketplace. Abhishek et al. (2019) and Sahni and Nair (2020) present theoretical models of separating equilibria in which advertising signals quality in such settings. In these models, high quality products are more likely to be advertised when the platform is unable to sort easily between high and low quality products in its organic search listings using its own information and algorithms, and advertising forms a screening device that distinguishes between high and low quality products.

[2] For instance, suppose that advertised products have high quality after sales service (an experience attribute), and users care about this aspect. When users see product A's ad, under a signaling equilibrium, they correctly believe it has high quality service. They do not believe so when they see A only in organic listings.

engineering the new search algorithm and deploying it on the platform, which is challenging because search algorithms on large platforms are complex products, requiring sophisticated data pipe-lining, feature engineering, machine learning model training and inference, and complex user-level behavior tracking, all of which involve large teams, code bases and infrastructure. This is a non-trivial exercise, which nonetheless has to be completed in order to do the evaluation.

Third, a credible evaluation needs to be causally robust. To assess the impact of the algorithm change on user-level outcomes, we need to compare ex-ante identical users exposed to different algorithms while holding everything else fixed. This requires experimentation and persistent user-level randomization, in which users are randomized into platform experiences under one of two algorithms and maintained in their assigned experience for some period of time, so their behavior can be assessed. Maintaining persistent user-level randomization is non-trivial (involving for instance, economic costs for the platform on the users randomized into the inferior algorithm), and again requires the platform's cooperation.

Fourth, experimentation has to be "platform-scale," i.e., involving deployment of the new search algorithm at the platform-level and not changing the search algorithm on some subset of products, or for some specific day parts or times of day. When the new search algorithm is deployed on only some subset of products, or some day parts, users in the treated group are exposed to the new algorithm in some product searches and to the old algorithm in others. Hence, users' responses in this scenario will represent exposure to a mix of the new and old algorithms. In contrast, we need to obtain user response to the new algorithm when the new algorithm is utilized in every platform search (treatment group), which should be compared to user response to the old algorithm when the old algorithm is utilized in every platform search (control group). Hence, proper estimation of the counterfactual improvement of a platform-wide adoption to the new algorithm requires experimentation on *the platform as a whole* for a contiguous, sustained period of time. Such randomization is also non-trivial.

We work with `JD.com` to redesign its search ranking algorithm and to run a field-experiment that addresses each of the concerns discussed above.[3] The `JD` platform confronts a demanding cold-start problem with respect to the discoverability of new products due to a continuous influx of a large number of new products introduced into its marketplace.[4] We refer to the platform's existing search ranking algorithm that does not use ad-information as algorithm "A" and the newly developed algorithm which uses ad-information as algorithm "B." Both algorithms A and B have a model component that boosts new products rankings in search listings. This boosting is done by putting additional weight on new products' search scores that determine their search rankings given a search query. The additional weight, namely the boosting score, is computed based on observable characteristics of the products and sellers (both A and B), consumer-product engagement data

[3] In 2020, `JD`'s e-commerce marketplace generated about 400 billion USD of GMV (Gross Merchandise Value). As benchmarks, the GMV of `Amazon`, `Shopify` and `Ebay` in 2020 was 475, 119 and 100 billion USD, respectively (Statista, 2021).

[4] About one million new SKUs (stock-keeping units) are added to the platform on some days in our data.

(both A and B) and ad-information (B only). That is, the two algorithms are identical in terms of all input features other than that B takes ad-information as additional input. Both algorithms also went through the same development process, involving modeling, training and testing, in order to optimize their performance given the predetermined feature set.

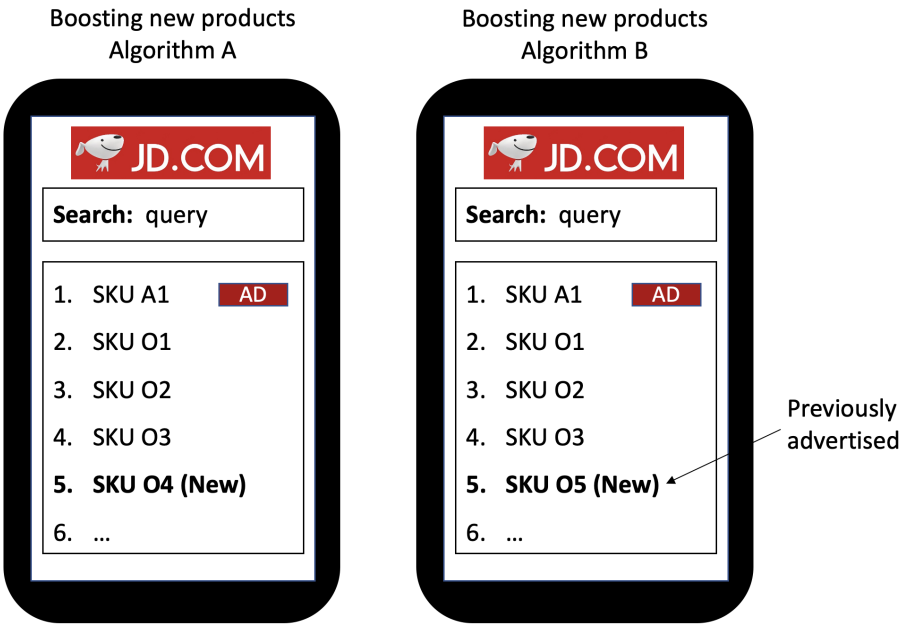
An initial version of algorithm B was first developed in October 2019. Following this, the algorithm was refined over a period of six months, till a final version was developed that was deployed platform wide on JD in March 2020. The field experiment reported in this paper pertains to the comparison of this final version of the new algorithm to the existing algorithm. The experiment involves two treatment conditions to which about 7.7 million users were randomly assigned during 16 days in March 2020. A user who arrives at JD.com and conducts a search during the experimental window is randomly assigned to treatment A or B, which uses algorithm A or B, respectively, for presenting search listings. Once a user is assigned to a treatment condition, she stays in the same treatment group during the entire experiment period. Overall, our experimental treatment's effect on user-behavior is the effect of utilizing new product ad-information in ranking search listings.

Analyzing the data on products that actually got listed in response to user searches, we show that the experimental assignment does impact the organic listings as expected in the algorithm design, and treatment B is more likely to show advertised new products, relative to treatment A. So, going from A to B affects users in two ways. First, on any given search, users are more likely to be exposed to advertised new products in response to their searches in B than A ("advertised new-product boosting" effect). Second, users are less likely to be exposed to non-advertised new products in response to their searches in B than A ("non-advertised new-product displacement" effect). The treatment effect we report measures the platform's overall outcome, which is the sum total of the advertised new-product boosting effect and the non-advertised new-product displacement effect. See Figure 1 for a visual illustration.

We primarily assess treatment effects on three outcome variables (1) whether a user placed at least one order (*conversion*), (2) the number of orders placed by a user (*#Orders*), and (3) the total gross merchandise value generated by a user (*GMV*), all linked to the purchase behavior of the user with respect to the platform as a whole. Our rationale is that if the search listings produced by the algorithm make finding an acceptable product difficult for the user, we would expect conversion to decrease. Similarly, if the algorithm change causes users to buy products that are less useful to them, they would expect lower benefit from shopping on the platform in the future, and we would see an overall decrease in purchases. On the other hand, intermediate outcomes such as clicks can either increase or decrease due an enhancement in search quality: a search engine of higher quality can induce users to engage with more searches and clicks due to the improved platform experience, whereas it can also decrease additional searches and clicks as it is more effective in surfacing what users are looking for.

We find that the use of ad-information in the search engine improves outcomes for the platform as a whole. During the experiment, users treated with algorithm B had, on average, 0.25%, 0.40%

Figure 1: Visual illustration of treatment conditions



Notes: Plots illustrate hypothetical outcomes from two designs of new product boosting algorithms (A and B) used as two treatment conditions in the field-experiment. In the figure, the position of a sponsored product (here SKU A1) is constant across the two conditions. In algorithm A, a new product (SKU O4) may show up in the fifth position among other non-new products (from SKU O1 to SKU O3), and in algorithm B, the new product showing up is more likely to be advertised previously (SKU O5). So, when we compare algorithm A and B, the effect (if any) comes from the appearance of advertised new products (boosting effect) and the disappearance of non-new and/or non-advertised new products (displacement effect).

and 0.57% more conversion, #Orders and GMV, respectively, than those treated with algorithm A. These effects are economically meaningful: the annual GMV of JD in 2020 was 400 billion USD, so a 0.57% increase translates into an incremental 2.28 billion USD in revenue. The results are found to be statistically robust and qualitatively unchanged using alternative estimators based on trimmed means, which are resistant to extreme skew and outliers in the distribution of the outcome data (Wilcox, 2011).

We find that assigning a user to treatment B increases the likelihood of the user returning to the platform in the future, relative to treatment A. Analyzing the composition of the treatment effect on purchases we find that treatment B increases sales not just for advertised new products that are boosted by the algorithm B, but also for other products, referred to as non-boosted products, that are not directly affected by the algorithm change. These findings are consistent with a broader positive effect of the algorithm change – a change in the presented listings caused by algorithm B increases the likelihood of users adopting the platform for their other shopping needs.

Examining the channel through which users reach the products they buy, we find that treatment B shifts boosted product sales from sponsored to organic channel relative to treatment A. For instance, while the overall GMV for boosted products increases by 1.02%, the GMV generated by

purchases occurring through organic clicks (i.e., users reaching the products without clicking on a sponsored listing) increases by 1.23% whereas that from sponsored clicks decreases by 0.15%. We see no such shift in purchase channel for non-boosted products. We find a similar pattern for other outcome variables. This shift is expected; since treatment B places the boosted products higher in the organic listing while keeping the sponsored listings same, relative to treatment A, it makes it more likely that buyers reach the boosted products through organic clicks as opposed to sponsored clicks. At the same time, this shift in revenue channel may be consequential for retail platforms setting mechanisms for both product and advertising markets.[5]

To check for the robustness of our findings we examine the heterogeneity in our effects – we group users on the basis of proxies for treatment intensity to check whether users who get more exposure to the new search algorithm have larger treatment effects. Our proxies are based on (1) users' pre-experimental usage of the platform and (2) their search queries up to the first search during our experiment. Overall, we consistently find the treatment effect to be higher among users who are more likely to experience the treatment B algorithm relative to treatment A.

Implications. These results show that using advertising information is consequential in organic ranking – users who are exposed to a platform search experience that utilizes ad information in assessing new products have more positive outcomes towards the platform. The seller's advertising decisions seem to be positively associated with product quality; if using advertising information led to treatment B boosting worse quality products, assignment to treatment B would have caused poorer user experience relative to treatment A. Taken overall, our findings are consistent with the view that using advertising information in ranking new products not only increases platform revenue in the short term, but it also makes the consumers better off on average by increasing the likelihood of them finding useful products to purchase.

Our proposal and results go against the conventional wisdom in industry which argues for a stark separation of ranking of organic listings from the advertising market. This separation of "church and state" has traditionally been motivated by the need to preserve the sanctity and relevance of search listings. The concern is that search listings could become biased if they boost the rankings of advertised products if it turns out that products of poorer quality tend to advertise more, thereby allowing sellers of poorer quality products to buy rankings (Google, 2021). Clearly, this is an empirical question, and the answer to it depends on the platform's prior information, the quality of its organic listings, and the nature of the prevalent advertising. Our results obtained here suggest that both consumers and platforms may benefit from using information from the advertising market in organic ranking. While we do not claim that this result holds generally across all platforms and settings, our findings do imply a rethinking of imposing strict separation of this sort on platforms by default.

[5]This shift in purchase channel has implications for how the platform reports the effectiveness of sponsored ads and its ad attribution logic. While treatment B increases the overall sales for boosted products, simple last click attribution would show a decrease in ad effectiveness in terms of sales.

The use of advertising information suggested here may be most relevant for e-commerce platforms such as JD. We believe the growing importance of e-commerce in retail, along with the increasing prevalence of search advertising on e-commerce platforms makes it important to develop a deeper understanding of the interaction of the product and advertising markets on such platforms, and to find ways to leverage their interaction to benefit both the platform and consumers. This paper serves as a step in this direction.

A caveat to our results is the short-run nature of the field experiment (16 days). In the "longer-run," several factors can adjust as a result in response to the platform-wide deployment of the new algorithm: a partial list of the possible margins of adjustment include product prices, promotion policies, product and seller entry/exit, products' platform store and product-page design, advertising aspects such as the identity of advertisers, their ad-intensity, ad-content, bids, budgets, frequency, and possible platform redesign that respond to these changes. All of these in turn can affect the overall "long-run" platform experience of users.[6] The current experiment holds fixed these aspects in the comparison between the A and B groups and does not fully reflect user responses to such potential "longer-term" changes. Accounting for all these aspects would require varying algorithms across isolated markets and across long periods of time; both of these have formidable challenges and are beyond the scope of this paper.

Given this, one way to interpret the empirical content of the results from our experiment is as answering the following question. Following long-standing practice in the tech industry, the platform's search ranking and advertising markets are strictly separated, and the platform and its users are in an equilibrium that is induced by that practice. Is this set up optimal for consumers? Can deviating from the current state benefit the platform in the short term? This paper documents that it is. In light of the long-standing nature of current practice of ad and search separation in the tech industry, and due to the fact that the speed of business decision making often requires responding to credible short-run effects, the results documented here have practical value.

Indeed, JD's platform has updated its search ranking algorithm incorporating our findings in the following year. Leveraging the staggered roll-out of the algorithm update, we analyze a few months of the post-experiment period in which the algorithm B was fully adopted in many product categories. We compare within and between adopted and yet-to-adopt product categories using a difference-in-differences analysis framework.[7] Our findings suggest that categories that adopted the algorithm B saw (1) more new products added to the platform and (2) an increase in new products purchases without cannibalizing non-new products sales. We detect no change in the overall product quality as measured by star ratings.

Finally, this study documents evidence of consumer sensitivity to the information shown by platforms – a change in the way product listings are displayed leads to a non-trivial consumer

[6] We write "long-run" in quotes because the unclear time-lines at which these effects will manifest itself, and the ambiguous quantitative impact of these changes and their interactions makes it fuzzy in our view where exactly to draw the line between what we call "short-run" versus "long-run" results.

[7] Eventually, all product categories adopted the use of ad-information in their organic search rankings.

response in terms of their overall spending on the platform (i.e., not just in shifting demand for a particular product, but shifting overall usage of the platform). This shows that consumer demand for the platform as a whole is sensitive to the user experience it provides. We believe this effect, which is rarely estimated, is interesting in its own right because it shows that consumers are able to put competitive pressure on the platform. The extent of such an effect is an input to the policy debates about the market power of e-commerce platforms as well as possible regulations that depend on assessments of such market power.

The remainder of this paper is as follows. The next section briefly reviews the relationship to related literature. We explain our empirical strategy in Section 3 and experimental design in Section 4. In Section 5, we describe the data and report descriptive statistics, experimental variation and randomization checks. We report the results from the experiment as well as from additional analyses and robustness checks in Section 6. In Section 7, we present additional analyses on the post-experimental period. Section 8 concludes.

2 Relationship to the Literature

This paper is broadly related to an empirical literature on online marketplaces that analyzes how marketplaces can be designed to reduce asymmetric information between buyers, sellers and platforms and to improve platform efficiency. A strand of this literature has emphasized the value of reviews and other forms of online feedback mechanisms as sources of information that can be used for resolving uncertainty about product quality. A difficulty with such mechanisms has already been mentioned previously – such information tends to be scarce for new products. Another difficulty is that online feedback can be under-provided due to its public good nature (see Dellarocas (2003)). Further, on account of the fact that it is provided by users, online feedback can sometimes be biased (Nosko and Tadelis, 2015; Filippas et al., 2018; Zervas et al., 2021) or possibly manipulated (Hu et al., 2012; Mukherjee et al., 2012; Mayzlin et al., 2014; Luca and Zervas, 2016). All of this implies that reviews and user generated feedback are not a panacea for product quality revelation for platforms, and the discovery of additional informational provision mechanisms remains useful.[8]

To the best of our knowledge, this study is the first to provide field experimental evidence on the value of leveraging ad-information in improving e-commerce organic search listings. Our results suggest that meaningful informational gaps can exist between sellers and platforms even for very information rich platforms, and that leveraging the informational content of advertising constructively can improve search listings which benefits both platforms and consumers.

Long et al. (2022) present a game theoretic analysis of the broader marketplace design problem of aligning advertising and search engines in e-commerce. They conclude that optimal design in this

[8]Rather than use search engine rankings, platforms can also reveal the information they obtain about product/seller quality via badges displayed to consumers, or implement buyer protection programs directly by say extracting compensatory payments from errant sellers for selling poor quality products; see Hui et al. (2016).

scenario would also involve adjusting the commission rate charged by the platform to its marketplace sellers as a function of how much weight it gives to the bidding information in its search engine algorithm. This study is complementary to their work, and can be seen as a contemporaneous real-world deployed, empirical counterpart.[9]

This study relates to the voluminous empirical literature on the field experimental evaluation of digital advertising effects (see Gordon et al. (2021) for an overview of recent work). A distinguishing feature of this study relative to this stream is its focus on platform-wide outcomes, rather than the outcomes for an individual advertiser or a firm operating on a platform. In this respect, the study is related to a smaller set of recent papers that have assessed via field experiments whether allowing advertising is good or bad for platforms (e.g., Abhishek et al. (2019); Sahni and Zhang (2020) who conclude that advertising improves consumer demand for the platform, and Huang et al. (2018); Moshary (2021) who conclude the opposite, that advertising reduces consumer demand for the platform). The findings from this study are closer to the former set of papers: net-net, our take-away is that having advertising is a net positive for the platform because it allows it to improve its search algorithms. In contrast to these papers, which evaluate the direct effect of advertising, this study focuses on an indirect mechanism by which advertising provides information to the platform that it might be able to leverage to improve overall outcomes.

To the extent that advertising is a seller action that has informational value, this paper is also related to a burgeoning stream of literature that have empirically shown that (costly) actions by sellers on online marketplaces can reveal private information held by sellers about product quality and serve as a signal to market participants. One example is Li et al. (2020) who analyze the "rebate-for-feedback" system implemented by `Taobao.com` that allows sellers to provide rebates to buyers for leaving feedback about product quality. Such rebates are shown to help sellers of high-quality new goods signal their product's quality to consumers, helping alleviate a cold-start problem for new goods. Other examples include studies documenting the signaling role of round-numbered asking prices by sellers on `eBay.com` (Backus et al., 2019) and of interest rates posted by loan seekers on `Prosper.com` (Zhang and Liu, 2012; Kawai et al., 2020). More broadly, this study is related to earlier offline marketing literature on enhancing new products' sales (e.g., Sudhir and Rao, 2006; Narayanan and Manchanda, 2009) and utilizing advertising decisions as information (e.g., Erdem et al., 2008; Luan and Sudhir, 2010).

This study is also related to a stream of empirical work that has assessed the product-market effects of online search engines and search advertising, and evaluated how their design affects downstream outcomes (see Yao and Mela (2011); Dinerstein et al. (2018); Fradkin (2017); Choi and Mela (2019); Bai et al. (2020) for example). Compared to these papers which use a structural econometric approach to platform design, a novelty of the current study is to assess the new search engine design directly via a field experiment implemented on a real platform.

[9]The suggestion to adjust commission rates along with the use of ad-information is not part of the intervention considered in this study, and is left as a topic for future work.

Finally, there is a large and active computer science literature on the algorithmic and engineering implementation of search engines (e.g., Metzler et al., 2021). The contribution of this study to this literature is the use of economic theory to inform the feature engineering of search algorithms, and the evaluation of its effects via a randomized controlled trial on a real e-commerce platform. In this respect, our study fits into a stream of work that uses social science theory to improve the performance of machine learning algorithms.

3 Empirical Strategy

The goal of the empirical analysis is to measure changes in users' responses when a platform adopts a new search algorithm. In our case, the new search algorithm admits sellers' advertising decisions as additional input features while keeping all other features utilized in its existing algorithm.

To understand our empirical strategy, assume the experiment begins at period $t = 1$ and ends at $t = T$. Upon arrival at the platform, users are randomized into one of the two search algorithms and stay within the algorithm until the experiment ends. Consider a user i who arrives at the platform at time $t = \tau_i$ and is randomized into one of the platform search engine treatments. Without loss of generality, let us define the length of each period t such that each user can submit a maximum of one search query in each t to the platform's search engine. We denote the search query by q_t. In response to q_t, the platform's search engine generates an ordered list of products, $\tilde{l}(q_t)$, and displays it to the user. To generate $\tilde{l}(q_t)$, the algorithm incorporated into the search engine assigns a score to every product j and sorts them in descending order based on the scores. That is, $\tilde{l}(q_t) = \{j_{(1)}, j_{(2)}, j_{(3)}, ...|q_t\}$, where $j_{(n)}$ is the product with the n-th highest score given query q_t. If i did not search in t, both q_t and $\tilde{l}(q_t)$ are $\varnothing$. Collect the history of search listings exposed to the user as of t as $\boldsymbol{l}(\boldsymbol{q}_t) = \left\{\tilde{l}(q_{\tau_i}), \tilde{l}(q_{\tau_i+1}), .., \tilde{l}(q_t)\right\}$. The user's action in t, denoted $y_i[\boldsymbol{l}(\boldsymbol{q}_t)]$, is influenced by the search algorithm she is randomized to through her exposure to these listings. The dependence of y_i on the vector of past listings is meant to capture that the user's behavior is potentially influenced by the search algorithm not just by the current listings, but by the entire history of returned listings. Though we focus on purchases, in general, $y_i[\boldsymbol{l}(\boldsymbol{q}_t)]$ can also represent clicks (whether i visits a product description page), or exit (whether i leaves the platform without making any further action), among others.

The causal effect of switching from a search algorithm (say A), to another (B) is,

$$\Delta_i \equiv \sum_{t=\tau_i}^{T} y_i[\boldsymbol{l}^B\left(\boldsymbol{q}_t^B\right)] - y_i[\boldsymbol{l}^A\left(\boldsymbol{q}_t^A\right)]. \tag{1}$$

where, $\boldsymbol{q}_t^A$ and $\boldsymbol{q}_t^B$ indicate the vectors of queries submitted by the user as of time t under the two algorithms, and $\boldsymbol{l}^A\left(\boldsymbol{q}_t^A\right)$ and $\boldsymbol{l}^B\left(\boldsymbol{q}_t^B\right)$ are the vectors of search listings returned by the platform as of time t under the two algorithms. The superscript "A" or "B" on $\boldsymbol{q}_t^A$ and $\boldsymbol{q}_t^B$ are shown to make

explicit that the set of queries submitted by the user (including $\varnothing$) can itself depend on the search algorithm she is randomized to, which can affect her outcomes over the T periods. The estimand in Equation 1 is meant to capture this effect as well.

The estimand in Equation 1 represents an *overall* (unconditional) effect of a search technology on user i rather than its effect in a specific query by the user at a particular time. User responsiveness to the new platform experience is better measured from a comparison of users' actions over a longer horizon than in a specific moment during the experiment. To address this, we measure the overall treatment effect of switching between search algorithms during a finite horizon of time (of T periods). During this horizon, a user can submit as many query terms as they like and we aggregate across the outcomes of each search.

Averaging Δ_i across users, the average treatment effect is:

$$\Delta = \mathbb{E}_i \Delta_i \equiv \mathbb{E}_i \{ \sum_{t=\tau_i}^{T} y_i[\boldsymbol{l}^B \left(\boldsymbol{q}_t^B\right)] - y_i[\boldsymbol{l}^A \left(\boldsymbol{q}_t^A\right)] \}. \tag{2}$$

Its sample analog is simply the mean difference between the outcomes in the two treatment groups:

$$\hat{\Delta} \equiv \frac{1}{N^B} \sum_{i=1}^{N^B} \sum_{\tau_i=1}^{T} y_i[\boldsymbol{l}^B \left(\boldsymbol{q}_t^B\right)] - \frac{1}{N^A} \sum_{i=1}^{N^A} \sum_{\tau_i=1}^{T} y_i[\boldsymbol{l}^A \left(\boldsymbol{q}_t^A\right)], \tag{3}$$

where N^B and N^A indicate the number of users assigned to each treatment condition. Consider an outcome measure y_i for which the platform favors greater value (e.g., sales amount). If we observe $\hat{\Delta} > 0$ for such measure, we conclude that the search algorithm B, on average, generates more favorable outcomes to the platform than algorithm A.

Choice of outcome measures. In response to the exposure to the search listings delivered by the platform's search engine, the user could take several actions including additional search, visitation of detailed product description pages (i.e., click), and place an order (i.e., purchase). We focus mostly on its last stage, purchase decision, as an outcome for analysis. Our rationale is as follows. If the search listings produced by the algorithm make finding an acceptable product difficult for the user, we would expect conversion – the likelihood of the user making a purchase – to decrease. If the algorithm change causes users to buy products that are less useful to them, they would expect lower benefit from shopping on the platform in the future, and we would see an overall decrease in purchases. The effect of a quality-improvement in search results on other search related actions is ambiguous.

Overall, we consider three outcome measures that summarize users' purchase decisions: (1) whether a user placed at least one order (*conversion*, $y_i = 1$ if yes or $y_i = 0$ otherwise); (2) the number of orders placed by a user (*#Orders*, $y_i = 0, 1, 2, ...$); and (3) the total Gross Merchandise Value generated by a user (*GMV*, $y_i \geq 0$). For all these measures, $\hat{\Delta} > 0$ indicates that algorithm

B is more favorable than A for the platform.

4 Experiment Design

The field-experiment was implemented on JD.com's mobile app platform for 16 days between March 4th, 2020 and March 19th, 2020. JD is the second largest e-commerce firm in China and carries over 300 million SKUs from various consumer product categories.[10] The platform charges purchase commissions for both sponsored and organic listings, and the commission rate is higher for sponsored listings.[11] Sponsored listings additionally pay per click, and typically, there is one sponsored listing on the top of every six organic listings.

The experiment was motivated by the firm's continuous efforts to enhance the quality of its search engine, focusing in particular on ensuring sufficient exposure of newly added products. Over a million new SKUs may be added to the platform in one day. Inferring the quality of these new products and presenting those with higher quality to users is important to the firm.

Like most typical e-commerce search engines, the platform's search engine largely relies on a customer's prior engagement with a particular item (e.g., page visits, purchases, reviews and ratings) to determine the product's search ranking for that customer's searches. The search ranking determines the order of products viewed by users in the search results pages. By definition, new products lack such information, which poses an informational challenge to the platform in sorting between high and low quality new products in its rankings. To address this *cold-start* problem, JD.com engineered two versions of search algorithms (namely algorithms A and B) that are designed to boost new products' search rankings, one not using data on sellers' advertising decisions and another using advertising data. The use of ad-information in algorithm B was proposed and evangelized by us to JD's search and ad teams. The two algorithms serve as two treatment conditions in the field-experiment. Below, we elaborate on the details of the field-experiment and the platform's new product boosting algorithms.

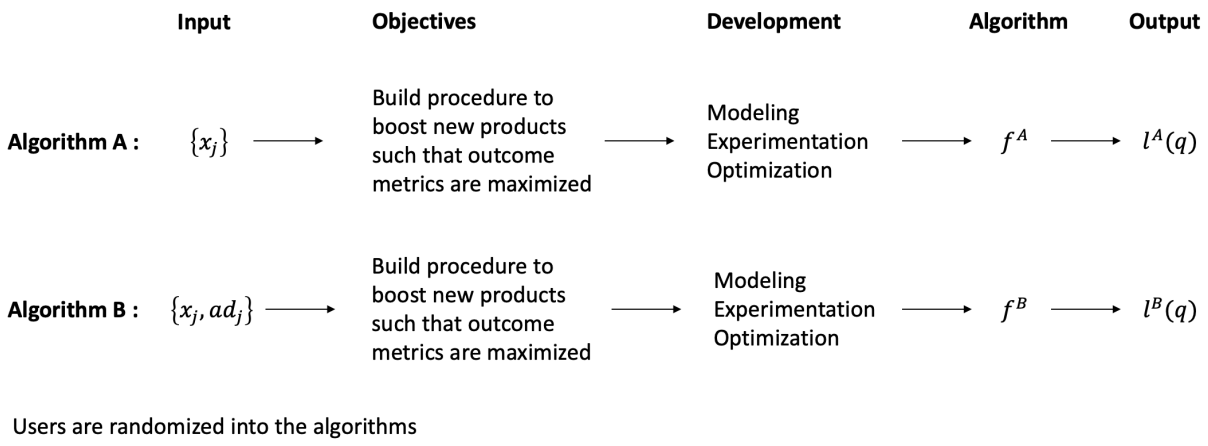
4.1 Randomization

A user is identified by a unique user id, and each user is randomly assigned to one of two treatment conditions at the time s/he firstly submits a query term to the search engine at JD.com during the experiment period. Only a randomly-selected fraction (less than 10%) of the entire user base is assigned to this experiment and randomized into one of the two treatment groups. Once a user is assigned to a treatment condition, the user stays in the treatment group for the entire experiment period.

[10] A SKU (stock-keeping unit) represents a unique combination of a product and a seller.

[11] The commission rate is typically 8% although it varies across product categories. See https://rule.jd.com/rule/ruleDetail.action?ruleId=63820964731198 2592&btype=1 (in Chinese; last accessed July 5, 2022).

Figure 2: Comparison of algorithm A vs. B



4.2 Treatment conditions

The two treatment conditions are:

1. Treatment A (new product boosting with no advertising data): Users in this group see a listing of products, generated by algorithm A, in which the search rankings of new products are boosted using various features other than advertising data.

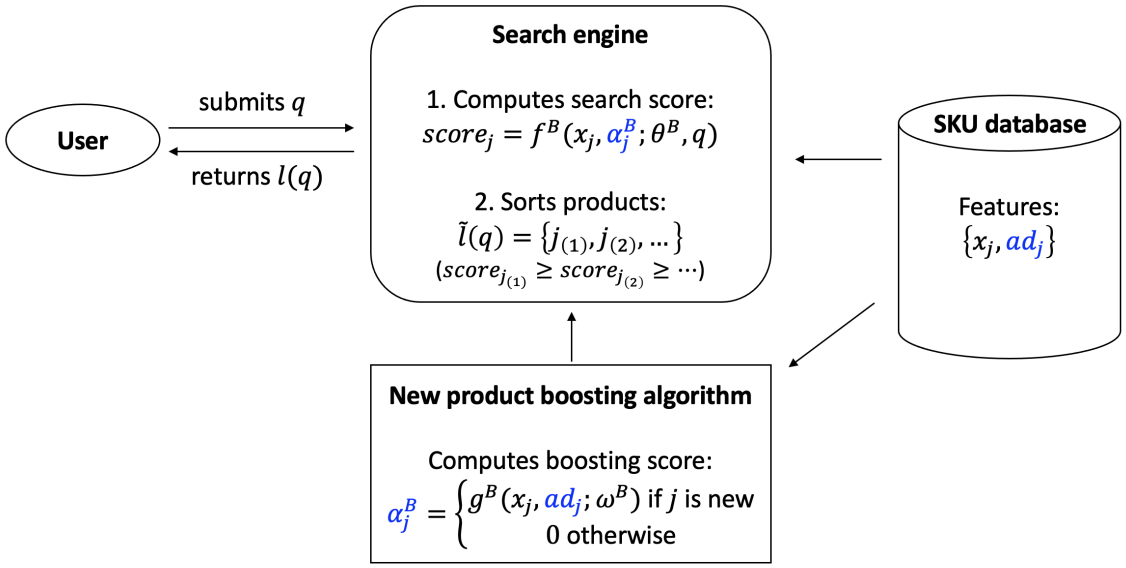
2. Treatment B (new product boosting with advertising data): Users in this group see a listing of products, generated by algorithm B, in which the search rankings of new products are boosted using the same features as in Treatment A *and* advertising data.

4.3 New product boosting algorithms: A vs. B

The format of the two new product boosting algorithms along with their differences are highlighted in Figure 2. As shown, the key difference lies in its input features—algorithm A only takes product attributes (x_j) as input, whereas algorithm B leverages both product attributes and sellers' advertising decisions (x_j, ad_j). Product attributes (x_j) include not only those accumulate over time (e.g., ratings and sales), but also information that is readily available for new products upon its release (e.g., brand score and shop score, which represent respectively a normalized measure of quality of the brand of the new product and the store that sells the new product).[12] Note that x_j also includes metrics such as sales in the past N days; thus, algorithm A does not always over-weigh old listings or under-weigh new listings. Advertising data (ad_j) are represented by two features: ads status (whether the product was advertised in the last 15 days) and ads score (a normalized measure of ad spending in the last 15 days).

[12]Due to business confidentiality reasons, we cannot disclose the full list of input features and/or discuss the exact configuration of each algorithm.

Figure 3: Illustration of algorithm B's architecture



Algorithm B vs. Algorithm A

- Algorithm A: $score_j = f^A(x_j, \alpha_j^A; \theta^A, q)$ and $\alpha_j^A = g^A(x_j; \omega^A)$ if j is new or 0 otherwise

Given the pre-specified set of input features, the platform builds an algorithm to boost new products such that outcome metrics (e.g., GMV) are maximized. Both algorithms A and B went through the same development process before the beginning of our experiment. This process, which typically takes several months, follows the typical engineering test-and-tune practice which involves multiple rounds of optimization using modeling and experimentation. Hence, we view each of the algorithms as the best possible outcome for the platform, given the choice of utilizing advertising data (algorithm B, indexed by f^B) or not (algorithm A, indexed by f^A). Users randomized into either of the algorithms can see different listings of products after submitting the same query term q. Following the notation from section 3, we refer to the two listings of products generated by algorithm A and B as $\tilde{l}^A(q)$ and $\tilde{l}^B(q)$.

Next, we elaborate on how the platform's search engine and its new product boosting algorithms work in each treatment condition, starting with algorithm B.

Algorithm B: boosting new products using advertising data. The architecture of algorithm B is shown in Figure 3. In response to a query term q submitted by a user, the search engine returns a listing of products $\tilde{l}(q)$ to the user.[13] A listing of products $\tilde{l}(q)$ is an ordered list of products $(j = 1, 2, ..., J)$, wherein the product with the highest "search score" (denoted by $score_j$) is placed in the top position, followed by the product with the second highest search score. That is,

[13] For expositional brevity, we suppress time index in our discussion in this section.

$\tilde{l}(q) = \{j_{(1)}, j_{(2)}, ..., j_{(J)}\}$ where $score_{j_{(1)}} \geq score_{j_{(2)}} \geq ... \geq score_{j_{(J)}}$. The search score, $score_j$, is a function of the submitted search query term q, product attributes (x_j), "boosting score" ($\alpha_j^B \geq 0$), and a set of weight parameters (θ^B) that govern the scoring function $f^B(x_j, \alpha_j^B; \theta^B, q) = score_j$.[14] The score function is weakly increasing in α_j^B, so the greater α_j^B is, the search ranking of j is more likely pushed upward.

The boosting score (α_j^B) is computed for every product j by the platform's new product boosting algorithm (the rectangle in the bottom of Figure 3), and is returned to the score function. The term α_j^B is set to be zero for the products not eligible to be considered as new products. Typically, the boosting intensity for a newly released product gradually decreases as time goes on until it becomes zero (i.e., losing its new product status) after a category-specific period of time from release. For eligible new products, the boosting algorithm computes the boosting score, which we denote by $g^B(x_j, ad_j; \omega^B)$, where ω^B is a set of weight parameters for the boosting function g^B and ad_j is a vector of the two features that summarize advertising (ads status and ads score). The boosting score is weakly increasing in both elements of ad_j.

Algorithm A: boosting new products not using advertising data. Algorithm A does not admit advertising data as input. That is, $\alpha_j^A = g^A(x_j; \omega^A)$ if j is new, or 0 otherwise; and $score_j = f^A(x_j, \alpha_j^A; \theta^A, q)$. In other words, algorithm A sets to zero a subset of weight parameters (in ω^A) that are associated with advertising data in the boosting score function.

4.4 Experimental interpretation

For a given listing of products, the number of products viewed by users is likely limited. In the context of mobile shopping, a majority of users browse only the first or up to the second page of the listings after submitting a query. Since the number of products displayed in a single search results page is limited (about 10 products for the typical smartphone), it is unlikely that products ranked low will get user viewership. In other words, the search results page can be viewed as a type of retail real estate with limited physical space, in which both new and non-new products are competing with each other.[15]

This implies that boosting some products' search rankings can displace non-boosted products downward in search listings, which in turn lowers the likelihood of viewership for the non-boosted products. To be specific, going from treatment A to treatment B increases (decreases) the likelihood of users seeing advertised new products (non-advertised new products and non-new products). Thus, the experiment does not explicitly manipulate the extent to which users see a particular product, rather it manipulates the mix of products seen. Users in each of the two treatment

[14] We index θ and f for each algorithm by θ^B and θ^A, and f^B and f^A, where the superscripts stand for algorithm B and A, respectively.

[15] In principle, the type of search engine used can induce users to browse more (or fewer) search results pages. In our experiment, we find empirically that users' tendency to browse more or fewer pages is unaffected by the treatment condition.

Figure 4: The effects of switching from algorithm A to B

Type of listing	Type of products	Implication for the products rank in algorithm B vs. algorithm A
Organic	New products that have advertised previously	↑ Boosted
Organic	New & existing products not advertised previously	↓ Displacement
Ads	New & existing products advertising in t	No change *(but potential substitution in sales)*

The sum of these effects is the treatment effect of B-A on overall purchase behavior

groups see a different product listing for a given query term—users in treatment group B sees more advertised new products than those in group A.

Also, note that we do not have direct control over the position an (advertised) new products is shown at in the search results page. Depending on search query term, it is possible that the two conditions can show exactly same listings of products, at least in the first few pages. Later, a manipulation check shows that algorithm B shows more advertised new products than algorithm A, as expected (Table 2).

The experimental variation should therefore be interpreted as a manipulation of the *overall* quality of search listing through a change in the composition of products (i.e., mixing search ranks). Boosting search rankings may weakly increase the purchase likelihood of the boosted products (boosting effect), while displacing the purchases of other products (displacement effect), and this in turn can spillover into the purchase of other products. In addition, although there is no change in sponsored listings, there is a possibility of substitution in purchases made through sponsored listings versus organic listings because previously advertised new products are more likely to show up in the organic listing of B (substitution effect). The sum of these three effects (boosting, displacement and substitution) translates into a treatment effect of B-A; Figure 4 illustrates. The important thing to emphasize is that the effect of the treatment is measured with respect to overall purchases on the platform, not just on boosted or non-boosted products, or new versus old products, or on advertised versus non-advertised ones. Thus, the interpretation is that the change in the search algorithm generates a change in the platform experience for the user, which changes her behavior with respect to the platform as a whole, and this is the treatment effect we are measuring. The sign and magnitude of the net effect would depend on the change in the quality of search listings from a treatment group to another. It can be either negative or positive, as well as zero.

5 Data

During the 16 days of the field-experiment (between March 4th and March 19th, 2020), we observe a total of 7,687,390 users who were randomized into one of the two treatment conditions with equal probability (3,841,923 users in treatment A and 3,845,467 users in treatment B). These users represent a randomly-selected, undisclosed fraction of all traffic to JD.com during the experiment. The randomization into treatment conditions is done the first time a user visits JD.com and submits a query term, which we define as a "search." We observe all searches by users in our sample during the experiment period ("in-experiment" period) and during the 30 days prior to the experiment ("pre-experiment" period). The date and time each user submits the first search during the experiment and the total number of searches vary across users.

For each search, we observe the following data: user identifier; treatment condition (treatment A or treatment B); date and time of each search; submitted query term(s); listings of products returned by JD.com and viewed by the user for each search; whether the user clicked a product in the listing to see more descriptions about the product; and details of the users purchases at JD. We also observe whether a product is considered a new product by JD.com on a given date.[16] For new products we observe the two advertising metrics mentioned previously (whether advertised in the previous 15 days and advertising intensity in the previous 15 days).

5.1 Descriptive statistics

Table 1 summarizes the search engine usage in the experimental data. We start by looking at 54,347,316 searches made by users in the treatment A group during the experiment period. For a search, we define a product as "viewed" if it is included in a search listings pages that was browsed by the user. For confidentiality, the number of products viewed by users on average per search is normalized to 100. Looking at row 2, on average, 3.09% of products that users view per search are

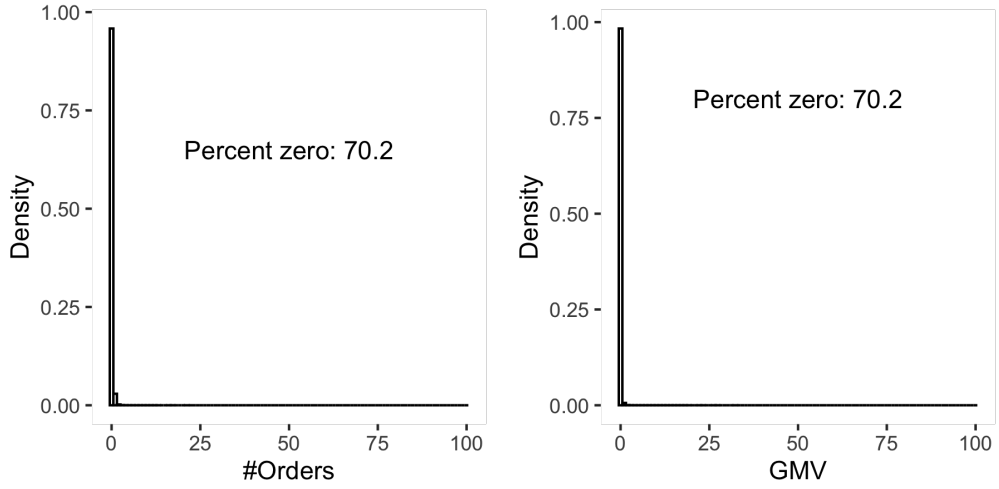
Table 1: Summary statistics: product views, clicks and orders per search by users in treatment A

	Mean	SD	Min	1st	Median	99th Per.	Max	%-nonzero
# products viewed per search†	100.00	167.01	3.73	6.64	49.64	742.85	28,529	100.0
# new products	3.09	22.65	0	0	0	53.22	10,666	15.86
# clicks per search	4.17	6.67	0	0	0	42.48	504	49.32
# new products	0.09	0.85	0	0	0	4.53	187	1.73

Notes: Table reports the summary statistics of product views, clicks and orders per search. The numbers are based on 54,347,316 searches made by users in the treatment A group during the in-experiment period.
† For confidentiality purposes, numbers in all rows (other than those in column '%-nonzero') are reported relative to the number of products viewed per search, which is normalized to 100.00 on average.

[16] At JD, there is a separate team whose job is to tag products coming into the platform as a new product. In order to decide which products are new, the team uses various methods that include image recognition and comparison to the existing product database.

Figure 5: Skewed distribution of the number of orders and GMV



Notes: Figure reports the distribution of the number of orders and GMV for users in the either treatment A or B group. The percentage of observations with zero values are reported for each variable. For confidentiality purposes, the numbers are normalized so that the maximum value of each variable is set at 100.

new; the median is 0. Users view one or more new products in only about 15.86% of search results. Looking at rows 3 and 4, users tend to click about 4.17% of the products from the listings to visit detailed product description pages, although this varies across searches (the standard deviation is 6.67% and the median is zero).

Figure 5 shows the distribution of #Orders and GMV in groups A and B. We observe that both metrics about purchase behavior are skewed and zero-inflated. For both metrics, 70.2% of observations are zero while the maximum value is quite high. This pattern is common for e-commerce purchase data (e.g., Wu et al., 2020). The highly skewed distribution of the outcome measures poses an empirical challenge in precisely estimating average treatment effects, because the sample mean represents a summary statistic that has low power to detect differences in such distributions. In addition, in the presence of such extreme skew and outliers, the mean may not be a good descriptor of the purchase of the typical user in each group, and is not a robust measure of location. Motivated by this, we follow Wilcox (2011) and Wu et al. (2020), and also assess tests of differences in purchases between the experimental groups on the trimmed mean, which is robust to skew and outliers.

5.2 Randomization checks

We check whether the randomization is done properly in three different ways. First, we check whether users are randomly assigned to each of the two treatment groups with equal probability. We observe 49.98% and 50.02% of users who were assigned to the experiment are in treatment A and treatment B groups, respectively. These numbers are statistically not different from $\frac{1}{2}$ (p-value

of 0.201 from a Chi-squared test for equal probability).

Second, we check whether users are randomly assigned at each date and time of arrival. We do this by comparing the quantile values of users' arrival date and time between the two groups. We report the results in Figure B.1 in the Online Appendix. In the figure, each point (x, y) represents the n-th quantile values of the arrival date and time from two groups we compare, where we vary n from 0 to 100. If the arrival date and time is balanced, we should see all the points are on the 45-degree line, which is what the figure shows.

Third, we check whether the two groups of users are balanced in their pre-experimental behaviors. Specifically, we compare the search engine usage behavior during the 30-day period prior to the first date of the experiment on (i) whether a user made at least one search (visit); (ii) number of searches either conditional or unconditional on visit; (iii) average number of products viewed per search; (iv) the average number of clicks per search; (v) average number of orders per search; (vi) total number of new products viewed; (vii) total number of clicks for new products; and (viii) total number of orders for new products. Table B.1 in the Online Appendix reports the results. We do not detect any systematic differences in each of the variable across the treatment groups. Recognizing that these variables are correlated, we conducted a joint chi-square test. We fail to reject that these pre-experimental usage behaviors are equal between users in treatment A and B conditions (p-value = 0.728).

5.3 Experimental variation

As a manipulation check, Table 2 describes the experimental variation generated going from treatment A to B. We report the changes in terms of (a) the number of products viewed, (b) the share of new products among the products viewed, (c) the probability of seeing at least one advertised new product, and (d) the sum of viewed new products' ad-propensity score. We report the numbers only for organic listings using the first search made by each user since the subsequent searches can be endogenous to the treatment condition users were assigned (Sahni and Nair, 2020).

Table 2: Summary statistics: experimental variation

	Treatment A $N = 3,841,923$ Mean (SD)†	Treatment B $N = 3,845,467$ Mean (SD)	t-stat
(a) Number of products viewed (first search)	100.0 (250.8)	99.94 (236.5)	−0.33
(b) Share of new products viewed (first search)	1.000 (3.304)	1.035 (3.363)	14.5*
(c) Probability of seeing advertised new products (first search)	1.000 (0.944)	1.032 (0.973)	11.1*
(d) Sum of viewed new products' ads score (first search)	1.000 (16.57)	1.044 (17.08)	3.62*

Notes: For all comparisons, we report the mean and standard deviation and t-stats from a two-sided t-test for the equality of means with unequal variances assumption (Welch's t-test).

† For confidentiality purposes, all numbers are normalized so that the mean values of the treatment A group are set at either 100 or 1.

* The absolute value of t-statistic is greater than equal to 2.

We find that there is no statistically significant difference in the number of searches between the two treatment conditions (row (a)). However, in treatment B, new products are more likely to be viewed by users than in treatment A (row (b)). On average, users tend to view new products 3.5% more in treatment B and the difference is highly statistically significant with a t-stat of 14.5. Additionally, the probability of seeing at least one advertised new product is 3.2% higher in treatment B (row (c)). These suggest that the platform can boost new products' ranking to a greater extent when utilized advertising data. Lastly, the new products viewed by users in treatment B are more likely to be advertised products than those viewed by users in treatment A (row (d)).

To summarize, the data show our experimental treatment generates variation in the identity of products in search listings. As expected, allowing advertising data to be used for ranking new products leads to users viewing more new products that advertised.

6 Results

This section reports on the results from the field-experiment. Following the discussion in section 3, we focus on three outcome measures: conversion, #Orders and GMV during the experimental time period. Since these measures are highly skewed, we take the following steps in our analysis to improve our statistical power.

First, we log transform #Orders and GMV and compare the sample means and standard errors of $\log(1 + \#Orders)$ and $\log(1 + GMV)$. Second, we check the robustness of our findings by comparing the trimmed means of #Orders and GMV with various level of truncation, following the rationale in Section 5. Third, we expect larger treatment effects, and therefore also, higher statistical power to detect treatment effects among individuals who receive a higher intensity of treatment. So, we also analyze heterogeneous treatment effects along this dimension in addition to comparison of overall means.

6.1 Main results: the incremental value of advertising data

We report the comparison between treatments A and B in the bottom panel of Table 3. Overall, we find the average treatment effect of switching from treatment A to B is positive in all our outcome measures. The magnitude of differences is 0.15%, 0.40%, and 0.57% for conversion, #Orders, and GMV, respectively. These effects are both statistically and economically significant. Statistically, all the three treatment effects are estimated with p-values less than 0.05.[17] A joint chi-square test rejects that these outcomes are equal between treatments A and B (p-value = 0.001). Economically, given the annual GMV of JD in 2020 was 400 billion USD, so a 0.57% increase translates into an

[17] We also conduct a one-side t-test for each of the three outcome variables where the null hypothesis is $Y_A \geq Y_B$ and the alternative hypothesis is $Y_A < Y_B$. We reject the null hypothesis in all outcome variables with greater statistical precision (conversion: t=2.22, p=0.027; log(1+#Orders): t=2.55, p=0.011; and log(1+GMV): t=2.22, p=0.026) than ones reported in Table 3.

Table 3: Main results: average treatment effects: treatment A vs. B

	DV: Conversion		DV: log(1+#Orders)		DV: log(1+GMV)	
	Estimate	Robust SE	Estimate	Robust SE	Estimate	Robust SE
Treatment B	0.002456**	0.001122	0.001678**	0.000661	0.004927**	0.002267
Percent difference†	0.25%		0.40%		0.57%	
95% Confidence Interval	[.03%, .47%]		[.09%, .70%]		[.06%, 1.1%]	
Intercept: treatment A	1.000000**	0.000794	0.549561**	0.000467	1.962719**	0.001602

Notes: Table compares outcomes from the two treatment conditions using three regressions wherein we regress each outcome variable on an intercept and a dummy variable for treatment B condition (N=7,687,390).
† The percent differences for #Orders and GMV are computed after inversing the log transformation. The 95% confidence intervals are computed based on one million random draws for each parameter estimate.
‡ For confidentiality purposes, we mask the actual values of estimates and standard errors. For conversion, masking is done after estimation to take advantage of the properties of binomial approximation of a binary variable. We divide the estimates and standard errors by the estimated intercept. For #Orders and GMV, masking is done before estimation by multiplying each variable by an undisclosed constant. The two constants may or may not be the same.
* p-value < 0.1; ** p-value < 0.05

incremental 2.28 billion USD in revenue.

To check for robustness to outliers, we replicate the results in Table 3 by comparing γ-trimmed means of #Orders and GMV across the treatment groups. An estimate of a γ-trimmed mean is obtained by averaging observations after truncating the sample at its γ and $1-\gamma$ quantiles.[18] Consistent with our main findings, we find that the treatment effects on both #Orders and GMV tend to increase and get more precisely estimated as we remove extreme valued observations. Table B.2 in the Online Appendix reports our results for five values of $\gamma \in \{0, 0.05, 0.10, 0.15, 0.2\}$, since there is no theoretical guideline on choosing γ.[19]

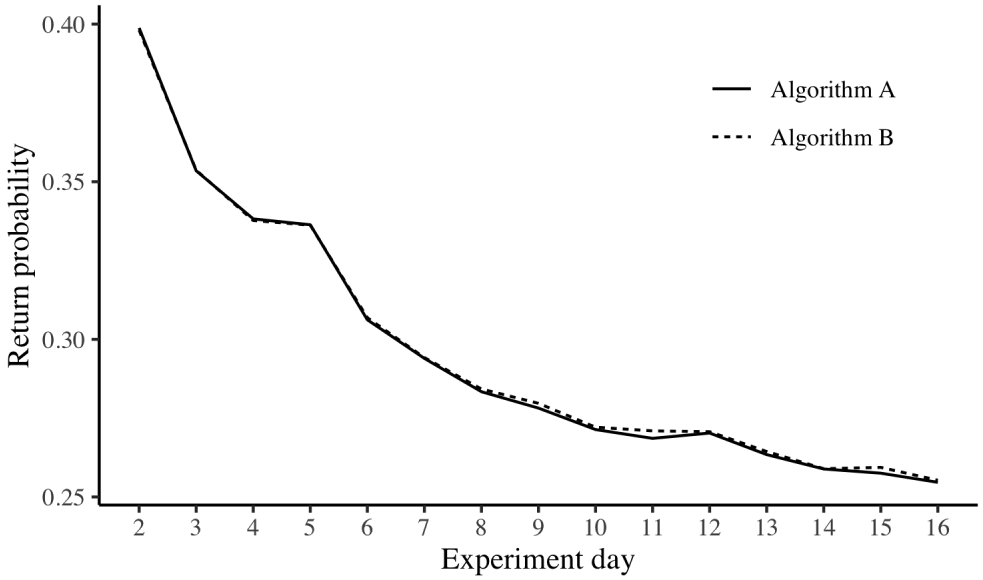
6.2 Broader treatment effect

The entire treatment effect could be driven by purchases of the new products that received a ranking boost in treatment B. However, presenting an improved assortment of the products in treatment B can cause non-trivial broader effects. It is possible that non-boosted products lose sales to the boosted ones implying a substitution effect. Conversely, being able to find an acceptable product, or a more suitable product due to treatment B can increase individual's future likelihood of shopping on the platform which could cause purchases of other products (not necessarily those that are boosted by treatment B). Hence, the broader effect is useful in understanding how boosting of advertised products spills over to the broader marketplace, and whether it helps in market expansion.

[18] One can see from this the mean is obtained by using the minimum possible trimming ($\gamma = 0$), while the median is obtained by using the maximum possible ($\gamma = 0.5$), so the trimmed mean for an intermediate level $\gamma \in (0, 0.5)$ is a compromise between the mean and the median. Analogously, the γ quantile of X is the $\gamma-$trimmed minimum, and the $1-\gamma$ quantile of X is the $\gamma-$trimmed maximum.

[19] The trimmed means and the standard deviations are computed based on 5,000 bootstrap samples of observations in each treatment group, following Wilcox (2011) and Wu et al. (2020). We stop at 0.2 because trimming more than $\gamma = 0.2$ leaves only observations of zero values in one or more treatment group.

Figure 6: Return probabilities of day-1 users



Notes: Figure reports the return probabilities of users who entered the experiment on the first day during the remaining 15 days of the experiment. The solid (dashed) line represents the probabilities for users in algorithm A (B).

Users returning to the platform. Does presenting organic listings generated through treatment B affect platform usage? We focus on users who entered the experiment during the first day, and compare between the treatment groups the probability of users returning to the platform in the remaining 15 days of the experiment. We find that day-1 users assigned to treatment B are 0.23% ($p = .027$) more likely to ever return to the platform during the remaining 15-day period relative to the corresponding treatment A users. Figure 6 further examines the difference in return probabilities of day 1 users across treatment groups during each of the remaining days. The return probabilities of both treatment A and B users are similar and declining in the first few days of the experiment. However, we can detect a relative increase in the likelihood of treatment B users returning in the later days during the experiment. The difference in return probabilities between day 2 and day 8 (i.e., Pr(return on days 2-8|entered on day 1, B)/Pr(return on days 2-8|entered on day 1, A))-1) is 0.08% ($p = 0.339$) whereas the difference for days 9-16 is 0.20% ($p = 0.048$).

Treatment effect on boosted and non-boosted products. Table 4 compares the treatment effects on the three outcomes measures separately for advertised new products (boosted) and non-boosted products. We further decompose the changes into three channels of purchases based on the last clicked listing before the purchase. For example, a purchase that occurred after the buyer reached the product through organic (sponsored) clicks is labeled as coming from the organic (sponsored) listing. For instance, in the first column conversion is counted only if the user placed an order for at least one boosted product during the experiment. Similarly, in the fourth column, conversion is computed based on users' purchase of non-boosted products.

Table 4: Composition of the treatment effect

	Advertised new products (boosted)			Non-boosted products		
	Conversion	log(1+#Orders)	log(1+GMV)	Conversion	log(1+#Orders)	log(1+GMV)
All listings						
Percent difference	1.26%	1.09%	1.02%	0.32%	0.39%	0.34%
p-value†	0.023	0.016	0.020	0.003	0.012	0.017
Organic listings						
Percent difference	1.79%	1.65%	1.23%	0.36%	0.41%	0.40%
p-value	0.023	0.016	0.020	$<.001$	0.001	0.012
Sponsored listings						
Percent difference	-0.92%	-1.22%	-0.15%	0.73%	0.64%	0.57%
p-value	0.212	0.143	0.393	0.041	0.085	0.085

Notes: Table reports the percent difference between B and A in the mean of the three outcome variables for new and non-new products. The numbers are from active users who submitted at least one search query term in the 30-day period prior to the experiment. N(treatment A)=2,925,928; N(treatment B)=2,927,304. For confidentiality purposes, all numbers are presented to relative to the mean values of treatment A, which are normalized to be one, on average.
† The p-values are from a two-sided t-test for the equality of means with unequal variances assumption (null hypothesis: $Y[A] = Y[B]$).

The first set of rows in the table ("All listings") shows that the users under treatment B made more purchases than those under treatment A, not only for boosted products but also for non-boosted products. In terms of magnitude, the percent difference is greater for boosted products than for non-boosted products. All these changes are statistically significant ($p < .05$).

Next, we look at purchases from organic and sponsored listings separately. We detect substitution between the two sources of purchases for advertised new products, although the effect size is greater for organic listings. For instance, conversion from organic listings for boosted products increases by 1.79% while that from sponsored listings decreases by 0.92%. We find similar patterns in #Orders and GMV. The statistical precision of the estimated treatment effects is higher for purchases through organic listings, which may be explained by relatively lower amount of viewership of sponsored listings. For non-boosted products, we do not find evidence for substitution as the effect sizes are all positive.

Overall, our interpretation of this analysis is that treatment B relative to A causes users to return to the platform more often which causes the users to increase their overall purchases. This explains our finding a net positive impact even on purchase of products that are not directly boosted by the algorithm. This finding shows that even the competitors of the promoted products benefit from the algorithm change. These findings indicate that use of advertising data in product ranking may enable the platform to better sort products, which improves consumer outcomes. Since the entire treatment effect does not occur entirely through advertised product purchases, and we see a decrease in sponsored listing attributed purchases, we can say that our treatment effect does not just occur because of repetitive presentation of advertised products in treatment B (e.g., effects documented in Johnson et al. (2016)).

6.3 Heterogeneity across users

We next investigate how the treatment effect varies across user groups with different level of treatment intensity.[20] In doing so, we consider two proxies that capture the user-level treatment intensity: user activeness and users' tendency of submitting certain query terms. First, more active users, by virtue of using the platform more, are more likely to get exposed to our treatment relative to other users. Hence, we expect the estimate for the difference between treatment A and B to be larger and measured more precisely for active users. Second, our treatment intensity varies across user search queries because queries vary in their tendency to show advertised new products in treatment B, relative to A. Hence, we expect a larger effect size among users who tend to submit query terms with higher treatment intensity.

We elaborate on these and report the results in the Online Appendix A. Overall, the findings are consistent with our expectations; user sub populations that experienced the treatment with a higher intensity are more affected by our treatment and show a larger increase in their overall spending on the platform when they are shown results from the algorithm that uses advertising data in its ranking. We also note that the estimated treatment effect is negative for inactive users, though this estimate is statistically insignificant. Hence, the treatment may not be unambiguously good for everyone, even though the effect is positive on average.

7 Analyses of the Post-experiment Period

The advantage of our experiment is that it establishes cleanly the effect of the algorithm change in the short-run – during the duration of the experiment. The experiment shows that treatment B is advantageous to the platform in the short-run. Our take on the interpretation and usefulness of the short-run effects was discussed in detail in the introduction. It is possible the long-run effects of adopting treatment B are different from the short-run effects: if the treatment B algorithm is rolled out globally, there may be complex supply side responses to the intervention, which in turn may induce corresponding consumer responses in the longer run. This section provides an exploratory analysis of these issues.

Post-experimental developments. Based on our experimental results, the platform began to deploy the new search ranking algorithm platform-wide after our study. The platform also communicated to its vendors that the act of advertising can influence organic search rankings.[21] These actions indicate that the platform believed in the long-run benefits of using our findings. However, this global roll-out also rendered the current experimental design unsuitable to pin down

[20] Since the randomization is done at the user-level, we are unable to analyze the role of treatment intensity at the product level.

[21] An example of this communication can be found here `https://helpcenter.jd.com/vender/issue/767-5032.html` (in Chinese; last accessed April 14, 2022).

a valid long-run effect. The platform-wide roll out of treatment B meant that both groups A and B of the experiment were exposed to the same treatment (i.e., B) post-experiment. Therefore, comparing post-experimental behavior of groups A and B post-experiment is not helpful to obtain the long-term impact of treatment B vs. A.

Because of this limitation, we explore an alternative strategy based on difference-in-differences at the aggregate, product-category level to investigate long-run effects. To do this, we exploit the staggered roll out of the treatment B algorithm across product categories.

A difference-in-differences (DID) analysis. The platform adopted algorithm B for 13 out of 23 product categories within three weeks of our experiment. All the remaining categories got updated later in 2021. Leveraging this variation in adoption timing, we conduct a difference-in-differences analysis where our unit-of-observation is category-week. Specifically, we estimate the following:

$$Y_{cw} = \beta \cdot \text{Treated}_{cw} + \mu_c + \tau_w + \varepsilon_{cw}. \tag{4}$$

Here, Y_{cw} represents the outcome variable of interest for product category c in week w; $\text{Treated}_{cw} = 1$ if the algorithm B was adopted by c in w or 0 otherwise; μ_c and τ_w are category and week fixed effects; and ε_{cw} is an error term. We consider three sets of outcome variables: (1) the number of new products added to the platform, (2) average product star ratings, and (3) sales (#Orders and GMV).

Inference. Assuming parallel trends in the outcome variables across categories, the coefficient β estimates the change in the outcome caused by adoption of algorithm B among those categories that actually adopted it. To assess the validity of parallel trends, Figure B.2 in the Online Appendix compares the pre-trends in our outcome variables and finds the pre-trends are similar between treated and non-treated categories. Also, note the product categories with higher share of revenue generated from new products adopted the new algorithm during the time covered by our data. Hence our estimate of β is applicable for these categories.

Results. We have 23 categories $\times$ 20 weeks category-week observations – 9 weeks before the experiment and 11 weeks after 13 categories adopted the updated algorithm. We drop the three-week transient time period between the experiment end and the time when the 13 categories adopted.[22] Tables 5 and 6 report the estimation results of equation 4. We find that sellers responded to the change in the organic algorithm by adding more new products to the platform, while the product quality in terms of average star ratings remained unchanged. We also find that sales of new

[22]We find the results are qualitatively unchanged with different choice of estimating sample. In the Online Appendix, we report the results with all data (Tables B.3 and B.4) and the results only using the first cohort of treated product categories (Tables B.5 and B.6)

Table 5: DID results on seller responses

	log(NewSKU)	AvgRating
β: Treated	0.337**	−0.002
	(0.158)	(0.003)

Notes: Table reports the estimation results of Equation 4. NewSKU is the number of new products added to the platform; AvgRating is the average star rating (1,...,5) of all listed products. Standard errors are clustered at the product category level. $N = 391$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Table 6: DID results on sales

	log(Total)	log(New)	log(Non-new)
	(A) #Orders		
β: Treated	0.100	0.474***	0.082
	(0.071)	(0.160)	(0.073)
	(B) Gross Merchandising Value (GMV)		
β: Treated	0.068	0.686***	0.027
	(0.109)	(0.189)	(0.115)

Notes: Table reports the estimation results of Equation 4 for #Order or GMV. Standard errors are clustered at the product category level. $N = 391$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

products increased after the adoption of the new algorithm while changes in sales from non-new products are statistically insignificant.[23]

In sum, we find an increase in the supply of new products in response to the algorithm change, and there is an overall increase in new products sales. We do not detect a change in other product sales at the aggregate level. Average product quality in terms of star ratings remains unchanged. From the same DID specification, we also find that there is no statistically significant change in the number of clicks for sponsored listings as well as for overall clicks (Table B.7 in the Online Appendix). Overall, these results suggest that the value of advertising information in ranking new products continued to exist in the three months after our experiment period.

[23] As a robustness check, we re-estimate equation 4 under various scenarios of hypothetical adoption timing, gradually pushing the adoption timing sooner, and find that the parameter decreases as expected despite the noise. See Figure B.3 in the Online Appendix.

8 Conclusion

This study has empirically demonstrated that ad-information can be incrementally useful for an e-commerce platform to alleviate the cold-start problem with respect to new products' search rankings. Our premise is that sellers' advertising decisions reflect private information that sellers possess about product quality, which is not observable through other data that platforms typically use in ranking products. Therefore, advertising can play a significant role in improving search algorithms. We provide empirical evidence supporting this idea by redesigning the search engine of a large e-commerce platform and conducting a large-scale field experiment. Overall, the study highlights the usefulness of theory-driven feature engineering for algorithm design and calls for blanket separations between ad and product markets to be re-thought. Finally, our empirical results show that using advertising information to rank products has broader positive effects beyond the sales of products that are directly boosted by the algorithm. This change therefore benefits both the platform and its users, on average.[24]

We caution that the empirical part of this study is based on an experiment that was conducted at a platform in a particular time. Future research is needed to replicate our findings in other settings. Relatedly, future research is also needed to further investigate how the effect vary across different product markets, which may require alternative experimental designs and/or choice of randomization units.

[24] We do not make overall welfare statements because the effect of the algorithm may be heterogeneous and may not benefit all users.

References

Abhishek, V., K. Jerath, and S. Sharma (2019). Advertising on online marketplaces: Information asymmetry and the relevance of sponsored listings. *Available at SSRN 3013468*.

Backus, M., T. Blake, and S. Tadelis (2019). On the empirical content of cheap-talk signaling: An application to bargaining. *Journal of Political Economy 127*(4), 1599–1628.

Bai, J., M. Chen, J. Liu, and D. Y. Xu (2020). Search and information frictions on global e-commerce platforms: Evidence from aliexpress. Technical report, National Bureau of Economic Research.

Brin, S. and L. Page (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems 30*(1-7), 107–117.

Choi, H. and C. F. Mela (2019). Monetizing online marketplaces. *Marketing Science 38*(6), 948–972.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science 49*(10), 1407–1424.

Dinerstein, M., L. Einav, J. Levin, and N. Sundaresan (2018). Consumer price search and platform design in internet commerce. *American Economic Review 108*(7), 1820–59.

Erdem, T., M. P. Keane, and B. Sun (2008). A dynamic model of brand choice when price and advertising signal product quality. *Marketing Science 27*(6), 1111–1125.

Filippas, A., J. J. Horton, and J. Golden (2018). Reputation inflation. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 483–484.

Fradkin, A. (2017). Search, matching, and the role of digital marketplace design in enabling trade: Evidence from airbnb. *Working paper, Boston University*.

Google (2021). Google ads: Seo vs. ppc? `https://ads.google.com/home/resources/seo-vs-ppc/`. [Online; accessed 7/26/2021].

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

Hu, N., I. Bose, N. S. Koh, and L. Liu (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems 52*(3), 674–684.

Huang, J., D. Reiley, and N. Riabov (2018). Measuring consumer sensitivity to audio advertising: A field experiment on pandora internet radio. *Available at SSRN 3166676*.

Hui, X., M. Saeedi, Z. Shen, and N. Sundaresan (2016). Reputation and regulations: Evidence from ebay. *Management Science 62*(12), 3604–3616.

Johnson, G., R. A. Lewis, and D. Reiley (2016). Location, location, location: repetition and proximity increase advertising effectiveness. *Available at SSRN 2268215*.

Kawai, K., K. Onishi, and K. Uetake (2020). Signaling in online credit markets. *Available at SSRN 2188693*.

Li, L., S. Tadelis, and X. Zhou (2020). Buying reputation as a signal of quality: Evidence from an online marketplace. *The RAND Journal of Economics 51*(4), 965–988.

Long, F., K. Jerath, and M. Sarvary (2022). Designing an online retail marketplace: Leveraging information from sponsored advertising. *Marketing Science 41*(1), 115–138.

Luan, Y. J. and K. Sudhir (2010). Forecasting marketing-mix responsiveness for new products. *Journal of Marketing Research 47*(3), 444–457.

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

Mayzlin, D., Y. Dover, and J. Chevalier (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review 104*(8), 2421–55.

Metzler, D., Y. Tay, D. Bahri, and M. Najork (2021). Rethinking search: making domain experts out of dilettantes. In *ACM SIGIR Forum*, Volume 55, pp. 1–27. ACM New York, NY, USA.

Moshary, S. (2021). Sponsored search in equilibrium: Evidence from two experiments. *Working paper, Booth School of Business*.

Mukherjee, A., B. Liu, and N. Glance (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 191–200.

Narayanan, S. and P. Manchanda (2009). Heterogeneous learning and the targeting of marketing communication for new products. *Marketing science 28*(3), 424–441.

Nelson, P. (1974). Advertising as information. *Journal of Political Economy 82*(4), 729–754.

Nosko, C. and S. Tadelis (2015). The limits of reputation in platform markets: An empirical analysis and field experiment. Technical report, National Bureau of Economic Research.

Sahni, N. S. and H. S. Nair (2020). Does advertising serve as a signal? evidence from a field experiment in mobile search. *The Review of Economic Studies 87*(3), 1529–1564.

Sahni, N. S. and C. Zhang (2020). Search advertising and information discovery: Are consumers averse to sponsored messages? *Stanford Graduate School of Business Paper No. 3441786*.

Statista (2021). Most popular online marketplaces in the united states in 2020 based on gross merchandise value. *Digital Commerce 360: https://www.statista.com/statistics/977262/top-us-online-marketplaces-by-gmv/*.

Sudhir, K. and V. R. Rao (2006). Do slotting allowances enhance efficiency or hinder competition? *Journal of Marketing Research 43*(2), 137–155.

Wilcox, R. R. (2011). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press.

Wu, D., H. Nair, and T. Geng (2020). Consumption vouchers during covid-19: Evidence from e-commerce. *Available at SSRN 3679362*.

Yao, S. and C. F. Mela (2011). A dynamic model of sponsored search advertising. *Marketing Science 30*(3), 447–468.

Zervas, G., D. Proserpio, and J. W. Byers (2021). A first look at online reputation on airbnb, where every stay is above average. *Marketing Letters 32*(1), 1–16.

Zhang, J. and P. Liu (2012). Rational herding in microloan markets. *Management Science 58*(5), 892–912.

Online Appendix

for

Advertising as Information for Ranking E-Commerce Search Listings

A Heterogeneous Treatment Effects

A.1 Effects by user activeness

More active users, by virtue of using the platform more, are more likely to get exposed to our treatment relative to other users. Hence, we expect the estimate for the difference between treatment A and B to be measured more precisely for active users.

In implementing the idea, however, we avoid to group users based on their activeness during the experiment period, since the in-experiment usage behavior can be endogenously determined by the treatment conditions. Instead, we use users' search engine usage behavior in the 30-day pre-experiment period with an underlying assumption that users' activeness between pre- and in-experiment periods are positively correlated even in the absence of our treatments.

Specifically, we define *inactive* users as those who made no searches in the 30-day period prior to the experiment. On the contrary, *active* users are those who searched at least once during the same period. Among the 7,687,390 users assigned to either treatment A or B groups, more than half submitted at least one search in the pre-experiment period, constituting the active user group. On average, active users, by our definition, made about 2.35 times more searches than inactive users during the experiment period (p-value $< .001$). Also, active users had about 2.19 (1.89) times more searches that had at least one (advertised) new product than inactive users (p-values $< .001$), which suggests that the treatment intensity was greater for active users.[1]

To check how the differences in our outcome measures between treatment A and B vary by user activeness, we estimate the following regression:

$$Y_i = \beta_0 + \beta_1 \cdot \text{Active}_i + (\beta_2 \cdot \text{Inactive}_i + \beta_3 \cdot \text{Active}_i) \cdot \text{Treatment B}_i + \varepsilon_i, \tag{A.1}$$

where Inactive_i and Active_i are binary variables that indicate user activeness. β_2 and β_3 are the parameters of interest, each of which separately reports the estimated average treatment effect for the two groups of users.

Table A.1 reports the estimation results. We find that the average treatment effects are positive and statistically significant ($p < .05$) for active users. The magnitudes of estimates and the percent difference from baseline are also greater than the numbers we report in Table 3. For instance, conversion is 0.30% greater and GMV is 0.73% greater in treatment B than in treatment A. On the other hand, the estimates are negative but statistically insignificant for inactive users (p-values = 0.705, 0.770, and 0.866, respectively).

[1] For confidentiality purposes, we do not disclose the actual numbers.

Table A.1: Treatment effects by user activeness

	DV: Conversion‡		DV: log(1+#Orders)‡		DV: log(1+GMV)‡	
	Estimate	Robust SE	Estimate	Robust SE	Estimate	Robust SE
β_3: Treatment B						
× Active users	0.00550**	0.00238	0.00201**	0.00077	0.00626**	0.00261
Percent difference†	0.30%		0.47%		0.73%	
95% CI	[.05,.56%]		[.12%,.83%]		[.13%,1.3%]	
β_2: Treatment B						
× Inactive users	−0.00119	0.00342	−0.00041	0.00094	−0.00186	0.00351
Percent difference†	−0.12%		−0.18%		−0.30%	
95% CI	[-.79%,.55%]		[-.97%,.62%]		[-1.4%,.81%]	
β_1: Active user	0.82165**	0.00295	0.26388**	0.00086	0.97238**	0.00309
β_0: Intercept	1.00000**	0.00242	0.26388**	0.00067	0.81397**	0.00248

Notes: Table reports the estimation results of the multiple regressions in Equation 1 for each of three outcome variables. Inactive (active) users indicate those who submitted no (at least one) search query term in the 30-day period prior to the experiment. N=7,687,390.

† The percent differences for #Orders and GMV are computed after inversing the log transformation. The 95% confidence intervals are computed based on one million random draws for each parameter estimate.

‡ For confidentiality purposes, we mask the actual values of estimates and standard errors. For conversion, masking is done after estimation to take advantage of the properties of binomial approximation of a binary variable. We divide the estimates and standard errors by the estimated intercept. For #Orders and GMV, masking is done before estimation by multiplying each variable by an undisclosed constant. The two constants may or may not be the same.

* p-value < 0.1; ** p-value < 0.05

A.2 Effects by user treatment score based on query terms

Our treatment intensity varies across user search queries because queries vary in their tendency to show advertised new products in treatment B, relative to A. Hence, we expect a larger effect size among users who tend to submit query terms with higher treatment intensity.

To implement this idea, we start by splitting users into two groups—one for estimating each query term's treatment intensity, and another for estimating the treatment effect. We randomly select 10% of users for the first task. Using data on this subset of users, we compute for each query term, the difference in the average ad-propensity score of all new products in the search results between treatment B and A conditions. To be more specific, consider a query term q. Suppose a user in our 10% sample and in treatment A condition submitted q and retrieved a search results page. From the search results page, the user sees a certain number of new products, wherein the sum of these new products' ad-propensity scores is s_q^A. This value can vary across time, so we take the average across all occasions in which q is submitted under treatment A during our experiment period. We do the same for the users under treatment B and obtain s_q^B. Then, we take the difference, $\Delta s_q = s_q^B - s_q^A$, which we define as the *query treatment score* for q. A higher value of Δs_q indicates that users who search for q are more likely to see advertised new products under treatment B than under treatment A.[2]

[2]Because we only utilize a small portion of our sample to compute Δs_q, it may not cover all the query terms

We then compute the *user treatment score* for each user in the remaining 90% of the data. The user treatment score of user i, Treatment score$_i$, is defined as the average of s_q's for all query terms submitted by the user during the *pre*-experiment period and only the *first* query term submitted during the in-experiment period.[3] For instance, Treatment score$_i > 0$ indicates that, prior to the experiment, user i submitted queries with higher treatment intensity relative to the average user. On average, users with Treatment score$_i > 0$ made 1.31 times more searches and had about 1.51 (1.35) times more searches that had at least one (advertised) new product than users with Treatment score$_i <= 0$ during the experiment period ($p < .001$).

To see how the treatment effect varies by Treatment score$_i$, we estimate the following regression:

$$Y_i = \beta_0 + \beta_1 \cdot 1[\text{Treatment score}_i > 0] + \{\beta_2 \cdot 1[\text{Treatment score}_i \leq 0] + \beta_3 \cdot 1[\text{Treatment score}_i > 0]\} \cdot \text{Treatment B}_i + \varepsilon_i. \quad \text{(A.2)}$$

Table A.2: Treatment effects by treatment score

	DV: Conversion‡		DV: log(1+#Orders)‡		DV: log(1+GMV)‡	
	Estimate	Robust SE	Estimate	Robust SE	Estimate	Robust SE
β_3: Treatment B						
× 1[Treatment score>0]	0.00476*	0.00261	0.00170*	0.00098	0.00530	0.00327
Percent difference†	0.27%		0.38%		0.60%	
95% CI	[-.02%,.57%]		[-.05%,.80%]		[-.13%,1.3%]	
β_2: Treatment B						
× 1[Treatment score≤0]	0.00041	0.00251	0.00070	0.00086	0.00220	0.00311
Percent difference†	0.04%		0.25%		0.32%	
95% CI	[-.45%,.53%]		[-.35%,.86%]		[-.55%,1.2%]	
β_1: 1[Treatment score>0]	0.73422**	0.00256	0.27933**	0.00092	0.91141**	0.00319
β_0: Intercept	1.00000**	0.00177	0.32467**	0.00061	1.21073**	0.00220

Notes: Table reports the estimation results of the multiple regressions in equation (A.2) for each of three outcome variables using randomly selected 90% of users. Using the data on the remaining 10%, we compute for each search query term the difference in the ad-propensity score of new products in the search results between treatment B and A conditions. A user's treatment score is defined as the average of the differences for all query terms submitted by the user during the pre-experiment period and the first query term submitted during the experiment period. N=6,918,315.

† The percent differences for #Orders and GMV are computed after inversing the log transformation. The 95% confidence intervals are computed based on one million random draws for each parameter estimate.

‡ For confidentiality purposes, we mask the actual values of estimates and standard errors. For conversion, masking is done after estimation to take advantage of the properties of binomial approximation of a binary variable. We divide the estimates and standard errors by the estimated intercept. For #Orders and GMV, masking is done before estimation by multiplying each variable by an undisclosed constant. The two constants may or may not be the same.

* p-value < 0.1; ** p-value < 0.05

and/or differ to the values from the entire sample. In our case, the 10% sample covers 85.2% of searches in all pre-experimental searches and the first searches during the experiment. The correlation in the values of Δs_q between the 10% sample and the full data is 0.904.

[3] By doing so, we try to minimize the concern that users' choice of search query terms during in-experiment period can be affected by treatment conditions.

The parameters of interest are β_2 and β_3, which represent the separate treatment effects for low and high treatment intensity groups.

Estimates reported in Table A.2 show that the estimates are positive across all outcome variables regardless of the user treatment score. However, the treatment effect is greater, and more precise for user groups with strictly positive treatment score. For instance, conversion in treatment B is .27% greater than in treatment A for users with higher treatment score while the difference is about .04% for users with lower treatment score. We detect similar patterns for #Orders and GMV.

B Additional Tables and Figures

Table B.1: Randomization checks: comparing pre-experimental user behavior

Variable[†]	Difference between A and B: p-value[‡]
1[visit]	0.263
#Searches	0.995
#Searches|visit	0.727
#ProdViewedPerSearch	0.646
#ClicksPerSearch	0.521
#OrdersPerSearch	0.294
#TotalNewProdViewed	0.986
#TotalClicksNewProd	0.762
#TotalOrdersNewProd	0.718

Notes: Table compares the pre-experimental usage of users in the three treatment groups during the 30-day period prior to the first date of the experiment. N(treatment A)=3,841,923; N(treatment B)=3,845,467. We do not report the raw means and standard deviations for confidentiality purposes.

[†] 1[visit] is a binary variable that indicates whether a user submitted a search query (1 if submitted or 0 otherwise); #Searches is the number of searches unconditional on 1[visit]; #Search|visit is the number of searches conditional on 1[visit]; #ProdViewedPerSearch is the average number of products viewed per search; #ClicksPerSearch is the average number of clicks per search; #OrdersPerSearch is the average number of orders per search; #TotalNewProdViewed is the total number of new products viewed; #TotalClicksNewProd is the total number of clicks for new products; and #TotalOrdersNewProd is the total number of orders for new products.

[‡] The p-values are from a two-sided t-test for the equality of means with unequal variances assumption.

Table B.2: Results from trimmed mean comparisons: Treatment A versus B

	Trimming				
	$\gamma = 0$	$\gamma = 0.05$	$\gamma = 0.10$	$\gamma = 0.15$	$\gamma = 0.2$
	(A) #Orders				
Treatment A					
Mean	1.0000	0.5804	0.4188	0.3002	0.2272
SD	2.5979	1.0910	0.7946	0.5941	0.5327
Treatment B					
Mean	1.0058	0.5830	0.4213	0.3017	0.2290
SD	2.7474	1.0936	0.7979	0.5952	0.5344
Percent difference	0.58%	0.44%	0.61%	0.51%	0.78%
p-value[†]	0.002	0.020	0.010	0.037	0.001
	(B) Gross Merchandise Value (GMV)				
Treatment A					
Mean	1.0000	0.3312	0.1979	0.1167	0.0583
SD	6.2325	0.7368	0.4327	0.2709	0.1568
Treatment B					
Mean	1.0027	0.3327	0.1994	0.1178	0.0591
SD	6.7274	0.7384	0.4351	0.2729	0.1553
Percent difference	0.27%	0.46%	0.75%	0.99%	1.41%
p-value	0.548	0.084	0.019	0.022	0.032
N	7,687,390	6,918,652	6,149,914	5,381,174	4,612,436

Notes: In each column, the mean and the standard deviation (SD) are computed using 5,000 bootstrap samples of observations in treatment group. For confidentiality purposes, numbers are reported relative to the mean value of each variable at $\gamma = 0$, which is set to one.

[†] The p-values are from the percentile bootstrap method for testing the equality of means.

Table B.3: DID results on seller responses (all data)

	NewSKU	log(NewSKU)	AvgRating	Share45	Share45New
β: Treated	0.236	0.272*	−0.001	−0.001	0.001
	(0.249)	(0.132)	(0.002)	(0.003)	(0.001)

Notes: Table reports the estimation results of Equation 4. NewSKU is the number of new products added to the platform; AvgRating is the average star rating (1,...,5) of all listed products; Share45 is the share of star ratings 4 and 5; Share45New is the share of star ratings 4 and 5 for new products. Standard errors are clustered at the product category level. $N = 460$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Table B.4: DID results on sales (all data)

	Total	New	Non-new	log(Total)	log(New)	log(Non-new)
	(A) #Orders					
β: Treated	0.114	1.045	0.099	0.054	0.391**	0.040
	(0.100)	(0.666)	(0.103)	(0.048)	(0.166)	(0.049)
	(B) Gross Merchandising Value (GMV)					
β: Treated	0.053	1.860*	−0.020	0.025	0.570***	−0.006
	(0.096)	(1.000)	(0.107)	(0.074)	(0.166)	(0.078)

Notes: Table reports the estimation results of Equation 4 for #Order or GMV. Standard errors are clustered at the product category level. $N = 460$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Table B.5: DID results on seller responses (cohort 1)

	NewSKU	log(NewSKU)	AvgRating	Share45	Share45New
β: Treated	0.244	0.137	0.003*	0.004*	0.0004
	(0.202)	(0.199)	(0.002)	(0.002)	(0.002)

Notes: Table reports the estimation results of Equation 4. NewSKU is the number of new products added to the platform; AvgRating is the average star rating (1,...,5) of all listed products; Share45 is the share of star ratings 4 and 5; Share45New is the share of star ratings 4 and 5 for new products. Standard errors are clustered at the product category level. $N = 276$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Table B.6: DID results on sales (cohort 1)

	Total	New	Non-new	log(Total)	log(New)	log(Non-new)
	(A) #Orders					
β: Treated	0.136**	1.932	0.106**	0.011	0.427	−0.006
	(0.052)	(1.507)	(0.048)	(0.063)	(0.375)	(0.066)
	(B) Gross Merchandising Value (GMV)					
β: Treated	0.106	2.851**	−0.005	−0.034	0.694**	−0.079
	(0.091)	(1.223)	(0.119)	(0.077)	(0.306)	(0.084)

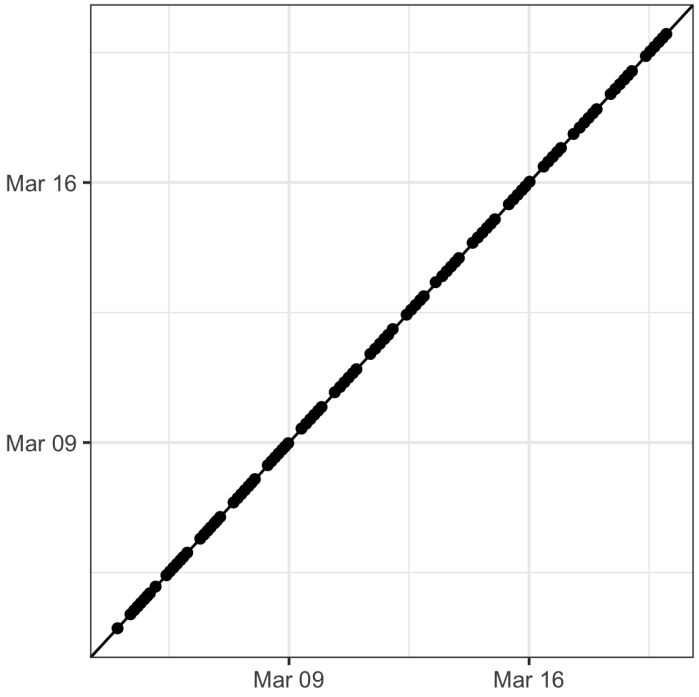
Notes: Table reports the estimation results of Equation 4 for #Order or GMV. Standard errors are clustered at the product category level. $N = 276$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Table B.7: DID results on sponsored clicks

	log(Sponsored clicks)	log(All clicks)
β: Treated	0.110	0.064
	(0.082)	(0.108)

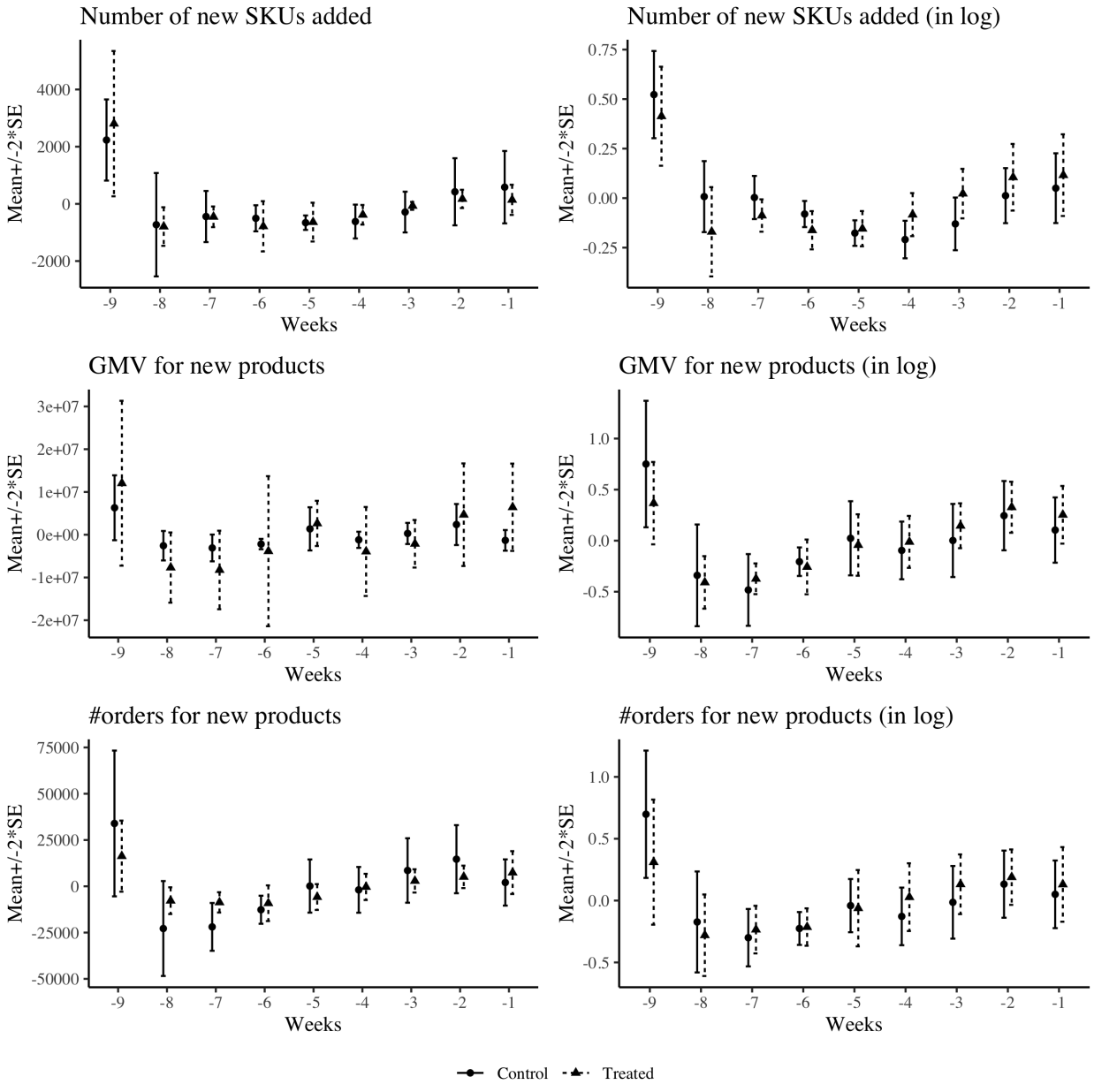
Notes: Table reports the estimation results of Equation 4 for #Sponsored clicks or All clicks. 'Sponsored clicks' is the number of search clicks for sponsored listings and 'All clicks' is the number of search clicks for all listings. $N = 391$. Fixed effects are not reported for brevity.
* p-value < 0.1; ** p-value < 0.05

Figure B.1: Randomization check: comparing user arrival date and time



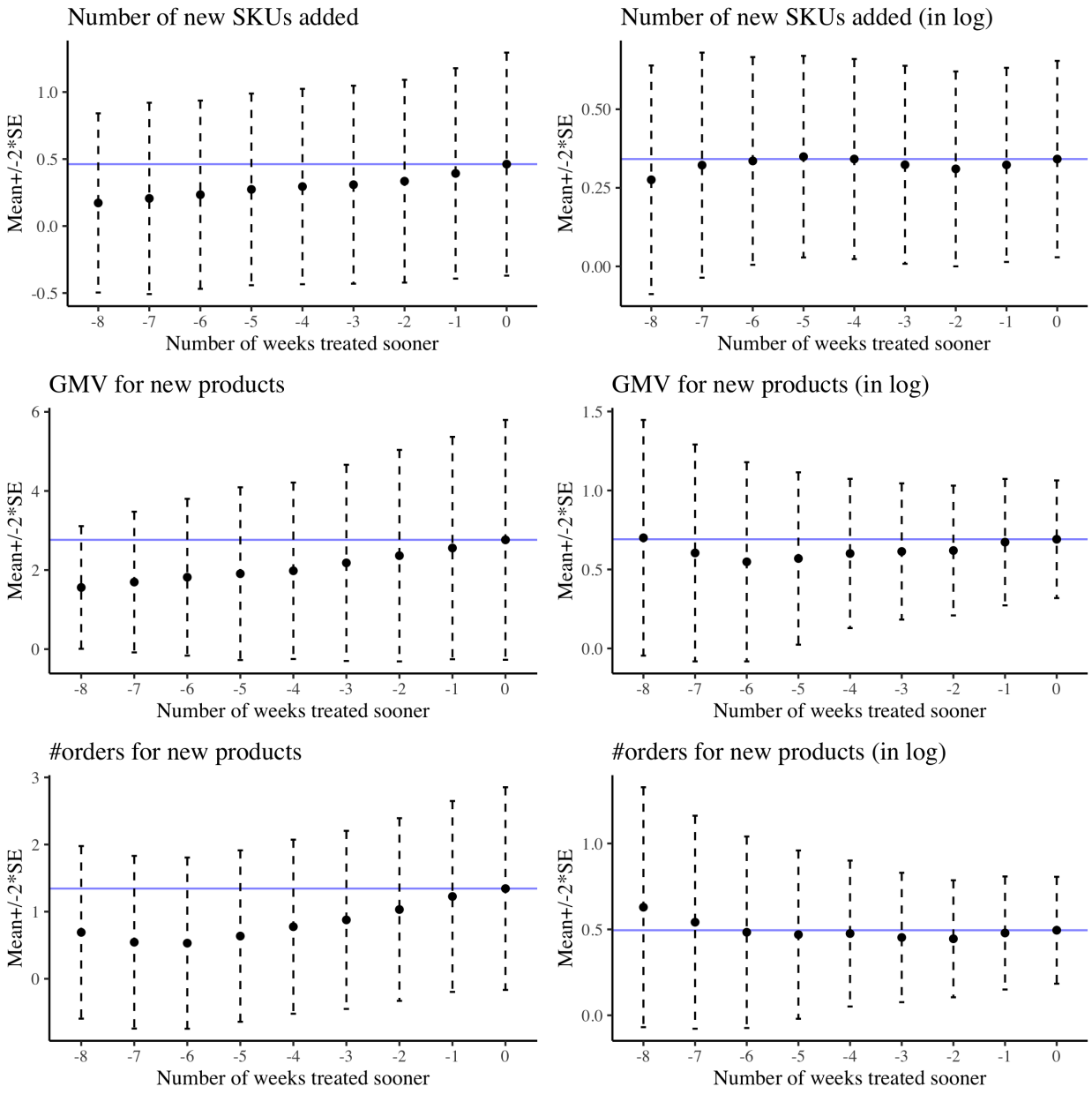
Notes: Figure compares the arrival date and time of users in the the treatment groups by reporting the n-th quantile-quantile values where $n = (0, 1, ..., 100)$. A 45-degree line is overlaid in the plot.

Figure B.2: Pre-trends between treated and non-treated product categories



Notes: Figure compares the pre-trends in each of three outcome variables between treated and non-treated product categories. Each plot reports the means and the confidence intervals of residuals from a regression in which we regress each of the outcome variable on product category fixed effects.

Figure B.3: Placebo test



Notes: Figures report the parameter estimate of β in Equation 4 under various scenarios of alternative adoption timing. The estimate is expected to decrease in the number of weeks pushed sooner.