The Power of Ranking:

Quantifying the Effects of Rankings on Online Consumer Search and Choice*

Raluca M. Ursu†

Current version: May, 2015

PRELIMINARY AND INCOMPLETE

PLEASE DO NOT CITE OR CIRCULATE WITHOUT THE AUTHOR'S PERMISSION

Abstract

When consumers face many options, intermediaries can help by ranking them, which in turn can influence how consumers search and what they ultimately purchase. To understand such influence, it is crucial to separate the role the ranking plays in consumer choices from other characteristics of the firm. As the ranking is endogenous, separately identifying the role of the ranking is challenging. In this paper, I identify the causal effect of rankings by using a data set on hotel searches that includes both a random ranking and the default ranking of a popular online travel agent. I show that rankings affect both clicks and purchases, but that conditional on a click, consumers do not derive any additional utility from purchasing from a higher ranked hotel. In addition, the data permits testing two common modeling assumptions found in the literature: that position mainly affects search costs and that consumer search sequentially, both of which are supported by the data. To quantify the effect of rankings on consumer choices, I estimate a sequential search model that accounts for the order in which consumers search. This model reveals that search cost estimates ignoring the endogeneity bias of position are upward biased. Using this model's search and preference parameter estimates, I construct several counterfactuals of interest comparing the value of the default ranking with the optimal ranking for consumers, hotels and the intermediary.

Keywords: online consumer search, hotel industry, search intermediaries, popularity rankings, endogeneity bias.

*I wish to thank Pradeep Chintagunta, Ali Hortaçsu, Hugo Sonnenschein, Richard Van Weelden, Elisabeth Honka, Chris Nosko, Anita Rao, Bradley Shapiro, Sergei Koulayev, Stephan Seiler, Andrew Rhodes, Alexei Alexandrov and Regis Renault. I thank the participants at the 2015 Workshop on Search and Switching Costs (Groningen) and 2015 IIOC conference (Boston). I thank Kaggle for providing me with the data. The usual disclaimer applies.

†University of Chicago, E-mail: rursu@uchicago.edu.

JEL Classifications: L81, D83.

1 Introduction

Rankings are becoming increasingly popular on the Internet where the abundance of information and products requires search intermediaries to aggregate and order relevant information. Search engines such as Google and Yahoo! rank documents based on relevance, while e-commerce sites like Amazon, Netflix or e-Bay order products by popularity and make product recommendations. Rankings may considerably reduce consumer search costs. They may also divert consumers away from alternatives that they like and toward those that are profitable for the search intermediary. What is the economic value of search intermediaries is an important question in marketing and economics. Clear evidence on this issue is scarce in large part because the rankings that consumers observe are not exogenous, but are rather chosen by intermediaries to maximize the probability of making a sale. The severity of this endogeneity bias is unknown because separating the role that the ranking plays in determining outcomes from other firm characteristics is difficult without random variation in the ranking.

In this paper, I seek to understand how rankings affect consumer choices by studying hotel searches on the world's largest online travel agent, Expedia.[1] Travel spending totaled $458 billion in the U.S. in 2014 alone, of which 43% was sold online, while the rest mostly represents business trips handled by corporate travel agents. Almost 80% of bookings made online are made on online travel agents, which had combined bookings of $157 billion in the U.S. and $278 billion world wide in 2013. OTA's revenues are derived from commission payments on sales. To compete for consumers, OTA's aggregate and rank third-party seller's products. As a result, their rankings are endogenous, making it difficult to separate the effect of the position on consumer choices from other characteristics of the firm.

My paper employs a unique data set to recover the causal effect of rankings on choices. The unique feature of my data set is that only two thirds of the data comes from searches using Expedia's proprietary ranking, while the rest used a ranking that was randomly generated. In constructing the random ranking, a hotel's quality did not affect its probability of being placed in any position, but rather the position of the hotel was randomly determined. Searches with a random ranking are costly for Expedia since they generally lead to fewer purchases. Nevertheless, they are used by Expedia to train their ranking algorithm without the position bias of the existing algorithm. Using this feature of the data set, I show two interesting patterns of the causal effect of rankings on consumer choices. First, I show that higher ranked hotels are clicked and purchased more often under both rankings. The fact that top positions receive more clicks and purchases has been documented of other search intermediaries (e.g. Google) and is thus not surprising of Expedia's curated ranking. However, the fact that the same pattern holds for randomly ranked hotels reveals the importance of rankings in influencing consumer choices. Second, I find that

[1] All figures reported come from three sources: 1. The Economist article: http://www.economist.com/news/business/21604598-market-booking-travel-online-rapidly-consolidating-sun-sea-and-surfing; 2. Forbes article: http://www.forbes.com/sites/greatspeculations/2014/04/08/competitive-landscape\-of-the-u-s-online-travel-market-is-transforming/; 3. Wall Street Journal article: http://www.wsj.com/articles/amazons-new-travel-service-enters-lucrative-online-travel-market-1429623993

conditional on a click, higher ranked hotels receive more purchases only under Expedia's ranking, while under the random ranking the fraction of purchases across positions is constant. This reveals that conditional on a click, consumers do not derive any additional utility from purchasing from a higher ranked hotel. In other words, consumers' realized utility does not depend on the position of the hotel. As a result, for the online travel agent, identifying which hotels consumers wish to purchase and ranking those first becomes paramount.

In addition, I use this unique data set to understand *how* the ranking affects consumer behavior and thus how to properly model rankings in a search model. Two popular assumptions of search models in the hotel industry are that consumers are searching sequentially and that rankings shift consumer search costs. I show evidence for both of these claims. First, I show that the difference in characteristics of the two rankings can be used to test whether consumers are using a sequential or a simultaneous search method and I find evidence for the former. Second, there are several mechanism through which the position of an alternative in a ranking may affect its probability of a click. These mechanisms can be grouped into two main effects: rankings affect consumers' expected utility, either through signaling (Nelson, 1974; Kihlstrom and Riordan, 1984) or consumer learning about the relation between position and the relevance of an alternative (Varian, 2007; Athey and Ellison, 2011), or rankings affect search costs (Ghose et al. 2012b, Chen and Yao, 2014). This great data set allows me to separate these two effects and show that rankings mainly affect search costs not expected utility. To show this, I exploit variation that naturally arises in my data set on the position of the nth displayed hotel, which in some searches appears in position n, while in others is demoted to position $n+1$ by an opaque offer.

To quantify the effect of rankings on choices, I estimate a sequential search model. This model extends Weitzman's (1979) sequential search model to account for the order in which consumers search in estimation. Comparing estimates from the random ranking with those from Expedia's ranking, I show both the direction and the magnitude of the endogeneity bias inherent in the ranking. Because under Expedia's ranking consumers find desirable alternatives faster, a model that does not account for endogeneity bias will attribute this behavior to sizable consumer search costs. I find that an increase in position is equivalent to a increase in price by 53 cents, whereas estimates using Expedia's ranking are upward biased. Finally, I use the preference and search cost parameters to construct counterfactual experiments of interest measuring the welfare of Expedia's current ranking and comparing it to welfare from the consumer, hotel and platform optimal ranking.

The rest of the paper is organized as follows. In the next section I review related work. In Section 3, I describe the data that I use for analysis. In Section 4, I provide reduced form evidence of the effect of rank on consumer choices and describe a test for two common assumptions made in the literature. In Section 5, I introduce the model, discuss identification and provide simulation results that confirm that the coefficients in my model are identified. In Section 6, I estimate the search model proposed on data from consumers searching for hotels on Expedia and compute the net gain of the current ranking as well as evaluate the welfare of a counterfactual ranking. Section 7 describes future research and Section 8 concludes.

2 Related Work

In this section, I describe how my paper relates to several lines of work. On the empirical side of the literature, my paper relates to research on (i) examining how ordered lists of alternatives affect consumers' search behavior, (ii) estimating search costs, (iii) identifying consumers' search method, (iv) controlling for endogeneity of position, and (v) estimating preferences and search costs using the restrictions placed by Weitzman's optimal search rules. On the theoretical side of the literature, my paper is related to work emphasizing consumers' search (i) when some firms are exogenously prominent, (ii) when firms compete for consumers' search order and (iii) when intermediaries, such as the OTA, divert consumer search. I describe each strand of the literature below.

Examining the effect of an ordered list of alternatives on consumer search and purchases in the online hotel industry has been the subject of six recent studies: De los Santos and Koulayev (2014), Ghose et al. (2012a, 2012b, 2013), Koulayev (2014), and Chen and Yao (2014). The paper that is closest to the current study is De los Santos and Koulayev (2014). They estimate a consumer search model and propose a method for ordering hotels that maximizes the click through rate (CTR) at the OTA. They show that expected CTR can be increased almost twofold by replacing the default ranking of the OTA with the ranking proposed in their paper. The paper addresses the endogeneity problem of the ranking by using a control function approach where the residual from a regression of position on past CTR, price and hotel fixed effects is used in estimation. My data set allows me to take a different approach to eliminate the endogeneity bias by using searches from the random ranking. I can then measure the severity of the endogeneity bias by comparing estimates from both types of rankings. The counterfactuals also differ: the counterfactual ranking that I focus on considers the average utility gain of rearranging hotels, while De los Santos and Koulayev (2014) look at maximizing CTR.

Ghose et al. (2012a) was one of the earliest papers to propose a utility based ranking, a method which is closely related to work in online recommender systems (see Ansari et al. 2000, Ansari and Mela 2003). The main difference in my approach is that I model consumer search (their's is a model of consumer discrete choice with no search), and I provide a setting where rankings are exogenous. Ghose et al. (2012b) introduces a model of consumer search, but it does not address the endogeneity of rankings. Ghose et al. (2013) addresses the endogeneity problem using a simultaneous equation model for clicks, purchases, rankings and ratings of the hotel and they show that a utility based ranking would increase the OTA's revenue given their estimates. They model the probability of seeing a hotel in a given position as a function of its past conversion rate and its characteristics. My paper provides a setting where some searches come from a random ranking, thus eliminating the endogeneity bias present in rankings.

The work of Koulayev (2014) and Chen and Yao (2014) is also closely related to my paper and it fits more vastly into the extensive literature on estimating search costs and the marketing literature on consideration set formation (see Mehta et al. 2003; Hortaçsu and Syverson, 2004; Hong and Shum, 2006; Moraga-Gonzalez and Wildenbeest 2008; Moraga-Gonzalez et al. 2010; Kim et

al. 2010; De los Santos et al., 2012; Seiler, 2013; Honka, 2014; Honka and Chintagunta, 2014). Koulayev (2014) proposes a method to identify search costs in the presence of unobserved tastes for consumers searching for differentiated products. He recognizes that the position of a hotel in a ranking may be endogenous in general, but provides evidence that in his data endogeneity may not be a concern. I have data both on searches observing the random ranking and Expedia's ranking, which allows me to quantify the direction and the magnitude of the endogeneity bias. Chen and Yao (2014) also focus on measuring consumer search costs, but in addition explicitly model the consumer's decision to refine their search: filter or sort hotels by different criteria, such as price and number of stars. However, Chen and Yao (2014) does not address the potential endogeneity problem of the ranking.

My paper is also related to studies that model the restrictions placed on preference and search cost parameters by optimal search in a sequential search model (Kim et al. 2010, 2014; Ghose et al. 2013; Chen and Yao, 2014; Honka and Chintagunta, 2014). Kim et al. (2010) introduce the optimal sequential search model of Weitzman (1979) into a model of choice, and in Kim et al. (2014) they extend their model to include purchases. Honka and Chintagunta (2014) is the closest to the current study in that they model the order in which consumers search by restricting reservation utilities. Like them, I do not have data on the exact sequence of searches that consumers make. However, I augment the current data set with evidence that allows me to recover the click order from a companion data set that contains this information. De los Santos et al. (2012) have data on the sequence of searches, but they estimate a simultaneous search model. Chen and Yao (2014) also have access to data on the exact sequence of searches that consumers make, but they do not model the probability of observing a specific click order (although they mention this possibility in an earlier draft).

In this paper, I provide a novel test to identify consumers' search method that uses differences in characteristics between two rankings. As such, my paper is also related to the recent literature on identifying consumers' search method. De los Santos et al. (2012) have data on purchases, consideration sets and the sequence of searches. They provide tests for the search method that consumers use for both homogeneous and differentiated goods and identify through these consumers' search method. Honka and Chintagunta (2014) show that consumers' search method can be identified even with less information: the price pattern in consumers' observed consideration sets can be used to identify consumers' search method. Both papers find evidence of consumers searching simultaneously, while I find suggestive evidence of sequential search.

Understanding how rankings affect consumer search is not a topic that is limited to the online hotel industry. There is an extensive literature on online sponsored search ads that deals with problems of identifying the position effect of a firm from its other characteristics (see Athey and Ellison, 2011; Yao and Mela, 2011; Baye et al., 2014; Jerath et al. 2011; Ghose and Yang, 2009; Yang and Ghose, 2010; Blake et al. 2014; Jeziorski and Segal, 2012; Chan and Park, 2014; Jeziorski and Moorthy, 2014; Narayanan and Kalyanam, 2014). Two papers are most relevant for my work. First, Baye et al. (2014) study search results at Google and Bing to measure the importance of name prominence and position on consumers' clicks. In separately identifying

the two effects, they are worried about the endogeneity bias of position, which they solve by instrumenting for position and ads on Google with position and ads on Bing. They also find that failing to account for the endogeneity in position inflates the position effect and minimizes the effect of name prominence. Unlike them, I also have access to searches from the random ranking allowing me to check how well a method that accounts for endogeneity alleviates the bias. Second, Jeziorski and Moorthy (2014) focus on separating the effect of ad placement and advertiser prominence on click through rates. They find that the two are substitutes: advertisers who are more popular benefit less from a top ad position than less popular ones. They also worry about the potential endogeneity of ad position, but they argue that they can treat the choice sets faced by consumers as exogenous because of the institutional details of their setting. This is similar to my setting where rankings are random, but I also consider the welfare gain of a better ranking and perform counterfactuals to understand how the OTA's default ranking performs compared to the consumer optimal ranking.

On the theoretical side of the literature, my paper is related to the extensive work on modeling consumer search. The most common assumption made in this literature is that each firm is searched either by a random fraction of consumers (for search with homogenous products, see Stigler (1961), Diamond (1971), Rothschild (1973, 1974), Salop and Stiglitz (1977), Reinganum (1979), Varian (1980), Burdett and Judd (1983), Stahl (1989), Janssen and Moraga-Gonzalez (2004), Baye, Morgan and Scholten (2006); for search with differentiated products, see Wolinsky 1984, 1986 and Anderson and Renault, 1999) or that firms are searched in a predetermined order (Arbatskaya, 2007; Armstrong, Vickers and Zhou, 2009; Zhou, 2011; Rhodes, 2011). Armstrong, Vickers and Zhou (2009) provide a differentiated products model in which one firm is exogenously prominent (all consumers search the prominent firm first), while non-prominent firms are searched randomly. They show that the prominent firm charges a lower price and earns a higher profit than non-prominent firms, and that the prominent firm charges a lower price than the random search price. The results from this exogenous search literature suggest that firms benefit from being searched early, influencing a series of more recent papers that deal with models in which firms can take costly actions to affect the order in which consumers search them. There are several ways in which firms can become prominent, such as persuasive advertising (Haan and Moraga-Gonzalez, 2011), commission payments, price-directed and history-directed search (Armstrong and Zhou, 2011; Haan, Moraga-Gonzalez and Petrikaite, 2014). Haan, Moraga-Gonzalez and Petrikaite (2014) show that when firms compete in prices for consumers' search, the resulting game has characteristics of a prisoner's dilemma: even though choosing to disclose prices leads to lower prices and profits, firms will still have an incentive to do so. On OTA's websites consumers observe all price before search and this result suggests that these prices are lower than they otherwise would be if firms could conceal their prices. In addition, Moraga-Gonzalez, Sandor and Wildenbeest (2014) identify a condition under which search costs can have pro-competitive effects. And finally, a third strand of the theoretical search literature focuses on the incentives of intermediaries to divert consumer search. Hagiu and Jullien (2011) identify two reasons for search diversion: intermediary's profit maximization and influencing the demand facing a firm

and therefore their pricing decisions. Berman and Katona (2013) show how firms might try to influence how intermediaries rank them. De Corniere and Taylor (2014) focus on a related question and ask how existing contracts between intermediaries and firms affect the quality of the product chosen by the firm. They show that the intermediary will promote the best firm, but because the firm is faced with a hold up problem (it has to choose its quality before the intermediary chooses how to rank it) it will choose to underinvest in quality. As search diversion has been shown to be such a pervasive phenomenon in the theoretical literature, I think of my counterfactual experiments as one way to quantify the extent of search diversion by comparing the average utility that consumers obtain from the current ranking with that from the best ranking for consumers.

In this section I reviewed both the empirical and the theoretical search literature in marketing and economics that is related to my paper. In the next section, I describe the data that I use for analysis.

3 Data

The Expedia data set that I use comes from a competition organized at the International Conference on Data Mining (ICDM) in December 2013 entitled "Learning to rank hotels to maximize purchases". This contest started in September 2013 and ended in November 2013 and was hosted by Kaggle.com.[2] The data is provided at the level of a search impression. A search impression is an ordered list of hotels and their characteristics (such as the number of stars, consumer reviews, and prices) seen by consumers in response to a search query describing the location and dates of their trip. The most important feature of this data set is the fact that only two thirds of the data set comes from search impressions under Expedia's proprietary ranking, while the rest of the data comes from search impressions where the ranking was randomly generated. A random ranking is a ranking where the position of the hotel does not depend on its characteristics or its past purchases, but rather is generated randomly. Search impressions with a random ranking are costly for Expedia since they generally lead to fewer purchases. Nevertheless, they are used by Expedia to train their ranking algorithm without the position bias of the existing algorithm. I will use this feature of the data to investigate the causal effect of the ranking on consumer search and purchase decisions. In this section, I describe the data, provide formal evidence of the experimental variation in my data set and discuss data limitation.

3.1 Description of the Data

The data set I use contains 7,986,074 observations on hotels from search impressions between November 1, 2012 and June 30, 2013.[3] The data comes from searches of 132,412 hotels located in

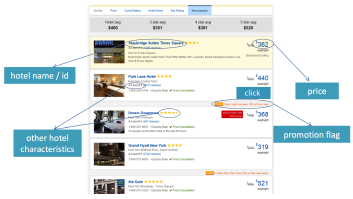
[2]See Appendix B 11.1 for details about learning to rank algorithms and about the winning algorithm in this competition using the algorithm LambdaMART.

[3]Appendix A contains details about data cleaning.

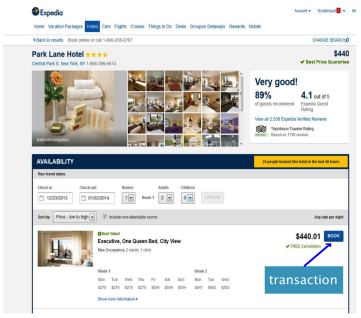
171 countries and 21,190 different destinations. In Figure 1 I summarize graphically the variables of interest present in the data. At the search query level, I have information on the date and time of the search, the destination ID (city, county or neighborhood), the length of stay (in days), the booking window (the number of days between the search and the first day of the trip), the number of adults and children traveling, the number of rooms searched, and an indicator for whether the trip includes a Saturday night. At the search impression level, I observe the first page of results that was displayed to consumers.[4] This contains the hotel ID and its characteristics (for example, the price and the number of stars) and position in the ranking.[5] I observe consumer choices in the form of their clicks and purchases at a particular hotel. Finally, less than 5% of observations also include information on the average star rating and average price of hotels previously purchased by a consumer, as well as the country in which the consumer lives. However, this is not enough information to link consumers who are making repeated searches over time.[6]

Figure 1: Information on the Data Observed

Search Query



Search Impression: Information Before the Click



Hotel's Page: Information After the Click

Table 1 provides summary statistics about search impressions. The data set contains 317,218 search impressions with an average number of 25 hotels shown in a search impression. Consumers on average search more than a month in advance of their trip for trips lasting approximately two days. About half of all impressions were for trips that included a Saturday night stay. The average search was for a trip for one hotel room and two adults traveling with no children. Search impressions contain a large fraction of hotels that are part of chain (64%) or that are on promotion (20%). One third of search impressions and 2,516,587 observations come from consumers who were shown a random ranking of hotels. There are a total of 352,523 clicks, with 118,149 clicks

[4] In a companion data set from Wharton Customer Analytics Initiative (WCAI) on consumers searching for hotels on a similar online travel agent in Manhattan, I find that in 67% of search impressions consumers only consider the first page of results.

[5] Hotel ID's are anonymized. As a result, the same brand located in two different parts of a city is given different hotel ID's. For example, "Hotel A City Center" and "Hotel A Airport" appear as different hotels in my data.

[6] In the same companion data set from WCAI I find that a significant fraction of consumers (more than 40%) only search once.

Table 1: Summary statistics: Search impressions

	Mean	Median	Std. Dev.	Min	Max
Number of Hotels Displayed	25.18	30.00	9.01	5	38
Trip Length (days)	2.36	2.00	2.08	1	59
Booking Window (days)	37.19	16.00	52.38	0	498
Saturday Night (percent)	0.51	1.00	0.50	0	1
Adults	1.99	2.00	0.87	1	9
Children	0.36	0.00	0.76	0	9
Rooms	1.11	1.00	0.43	1	8
Chain (percent)	0.64	0.71	0.29	0	1
Promotion (percent)	0.20	0.15	0.19	0	1
Random Ranking (percent)	0.33	0.00	0.47	0	1
Total Clicks	1.11	1.00	0.57	1	30
Two or More Clicks (percent)	0.06	0.00	0.25	0	1
Total Transactions	0.64	1.00	0.48	0	1
Observations	317,218				

under the random ranking. There is approximately one click per search impression, with 6% of search impressions including two or more clicks. Finally, two thirds of all search impressions end in a transaction for a total of 201,442 transactions. Only approximately 14,900 search impressions have historical information about the consumer. I find that consumers on average purchased in the past from hotels with 3.3 stars at a price of $170 per night, while hotels charged on average $170 in the last year.

The data is anonymized, so determining the exact country or city to which a consumer wishes to travel is not possible. However, there exists suggestive evidence that the largest country (labeled 219) is the U.S. The largest country has 5,236,418 observations and 203,858 search impressions. Out of those, 84% of searches are made by consumers also located in this country, suggesting that the country has a large territory with a large fraction of domestic travel. This is also consistent with information from Alexa which shows that in May 2015, 73% of Expedia's traffic come from visitors located in the U.S., while the second largest country in terms of traffic was South Korea with less than 2% of traffic. The prices charged are also consistent with the largest country being the U.S. According to the American Hotel and Lodging Association, the average price of a room in the U.S. in 2013 was $110.35.[7] In my data set, which contains only a subset of all properties in the U.S., the median price in 2013 was $118.

Table 2 shows how the characteristics of the hotels displayed vary by the type of ranking observed. I divide results by the type of the search impressions, Expedia or random, as well as by whether the search impression ended in a transaction. What is immediately clear is that Expedia's ranking displays more expensive hotels of higher quality, as measured by the number of stars and the reviews of the hotels. Also Expedia's ranking displays a larger proportion of chains and hotels with more promotions than the random ranking. Finally, search impressions that lead to a transaction, regardless of the ranking type, have cheaper hotels displayed. The last two

[7]See `http://www.ahla.com/content.aspx?id=36332`

columns in this table perform a t-test confirming that these differences are significant. Tables 13 and 14 in Appendix B 11.3 show that on average, clicked and purchased hotels are cheaper and of higher quality than those displayed. Also, Appendix B 11.4.3 shows how the characteristics of the hotels displayed (price, number of stars and reviews) vary by position and ranking type.

Table 2: Hotel characteristics displayed by search impression type

	No Tran.				Tran.				No Tran.	Tran.
	Random		Expedia		Random		Expedia			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Diff.	Diff.
Price	153.62	106.14	167.63	109.54	131.68	85.24	136.65	85.71	-14.01***	-4.97***
Stars										
Less than 3	0.20	0.40	0.14	0.35	0.26	0.44	0.21	0.41	0.06***	0.05***
3	0.42	0.49	0.37	0.48	0.45	0.50	0.43	0.49	0.05***	0.02***
4	0.30	0.46	0.38	0.49	0.24	0.43	0.29	0.46	-0.08***	-0.05***
5	0.08	0.27	0.11	0.31	0.05	0.21	0.06	0.24	-0.03***	-0.01***
Review Score										
Less than 2.5	0.09	0.28	0.05	0.21	0.07	0.25	0.05	0.22	0.04***	0.02***
Between 2.5 and 3	0.12	0.32	0.09	0.28	0.13	0.34	0.11	0.31	0.03***	0.02***
Between 3.5 and 4	0.46	0.50	0.48	0.50	0.46	0.50	0.48	0.50	-0.02***	-0.02***
Between 4.5 and 5	0.34	0.47	0.39	0.49	0.34	0.47	0.36	0.48	-0.05***	-0.02***
Chain	0.59	0.49	0.64	0.48	0.69	0.46	0.68	0.47	-0.05***	0.01***
Location Score	2.83	1.55	3.27	1.52	2.49	1.40	2.76	1.47	-0.44***	-0.26***
Promotion	0.19	0.39	0.30	0.46	0.15	0.36	0.22	0.41	-0.11***	-0.07***

Significance of differences obtained by means of a t-test.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.2 The Experiment

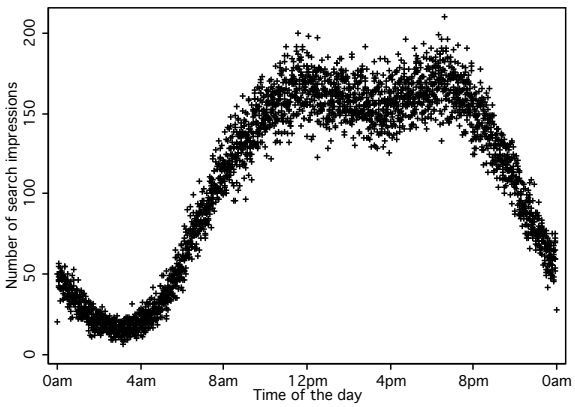
In this subsection, I test formally for the two types of randomness present in my data set: (i) consumers were randomly assigned to the two types of rankings and (ii) that in constructing the random ranking a hotel's quality did not affect its probability of being placed in any position, but rather the position of the hotel was randomly determined. These claims are also supported informally by discussions with the administrator of the competition.

3.2.1 Random assignment of consumers to the two types of rankings

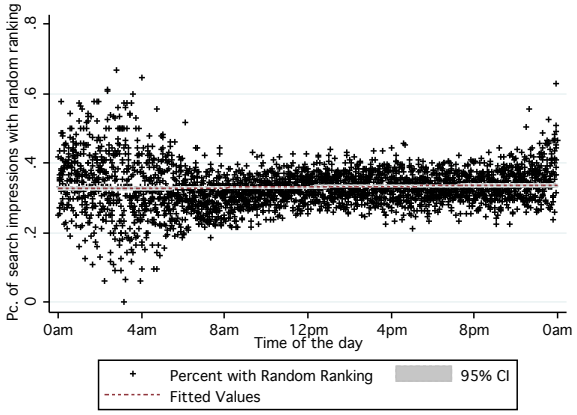
To show that consumers were randomly assigned to each type of ranking, I perform two tests. First, I test whether the time of arrival at the OTA's website is related to the type of ranking the consumer saw. One concern may be that different types of consumers visit the website at different times of the day and if the probability of observing one type of ranking is different at different times of the day, then this could bias the results. For example, if business travelers were known to search for hotels after 5pm and they are also more likely to purchase, then if after 5pm the probability of observing the random ranking is lower, there would be a correlation between higher purchases and Expedia's ranking in the data that is not due to the causal effect of the ranking observed, but rather to the way in which consumers were assigned to different rankings. However, Figure 2 shows that this is not a concern in the data. More precisely, the left panel plots the number of search impressions made by the time of the day and shows that more searches occur in the afternoon and evening. The right panel plots the fraction of search impressions seeing

the random ranking every 30 seconds during the course of one day, in the entire data set. Even though more consumers are searching in the second part of the day, the fraction seeing the random ranking is constant throughout the day. Thus, this figure suggests that consumers were randomly assigned to seeing either type of ranking.

Figure 2: Number of search impressions displayed and the fraction seeing the random ranking every 30 seconds in a day



Number of search impressions



Fraction of random search impressions

The second test I perform to check whether consumers were randomly assigned to see each ranking is to check whether consumer characteristics observed by the OTA prior to showing a ranking are different between the two rankings. When the consumer arrives at the OTA's website, she reveals details of her upcoming trip, such as her destination, the length of the trip, how long in advance she is searching for, the number of travelers and rooms requested, as well as whether her trip includes a Saturday night. For some consumers, the OTA also has historical information that is revealed when the consumer arrives at the website. One concern might be that the OTA takes all of this information into account when they decide which search impressions see the random ranking and which see Expedia's ranking. Table 3 shows that this also not a concern. Comparing search impressions with the same conversion across the two rankings by means of a t-test, I find that consumers seeing Expedia's ranking have very similar characteristics as those seeing the random ranking. Although in the full sample (first two columns) some of these difference are statistically significant, their magnitude is very small and the significance disappears when I condition on a particular destination (last two columns condition on the largest destination in the data set). Combined, these findings suggest that there are no systematic differences in consumer observables that lead the OTA to assign consumers differently to different types of rankings.

3.2.2 The random ranking

The second type of randomness in my data set comes from the construction of the random ranking. As stated in the competition description, Expedia's approach is a learning to rank approach.[8]

[8] `https://www.kaggle.com/c/expedia-personalized-sort/forums/t/5808/position-benchmark`.

Table 3: T-test: Search query characteristics revealed before the ranking

	Full Sample		Destination 4562	
Difference (Expedia-Random)	No Transaction	Transaction	No Transaction	Transaction
Trip Length (days)	0.1928***	0.0614***	0.1138	0.0686
	(0.0197)	(0.0155)	(0.1762)	(0.3523)
Booking Window (days)	5.0308***	1.0879*	-6.9376	-3.0726
	(0.4611)	(0.4372)	(4.0245)	(7.7610)
Adults	0.0455***	-0.0128	0.0197	-0.0124
	(0.0066)	(0.0084)	(0.0556)	(0.1250)
Children	0.0517***	0.0001	-0.0000	0.0316
	(0.0059)	(0.0073)	(0.0404)	(0.0919)
Rooms	0.0057	-0.0031	-0.0137	-0.0788
	(0.0034)	(0.0041)	(0.0274)	(0.0609)
Saturday Night	-0.0243***	-0.0063	-0.0178	-0.0126
	(0.0038)	(0.0049)	(0.0301)	(0.0809)
Consumer Hist. Stars	-0.0052	0.0511	-0.2330	NA
	(0.0355)	(0.0287)	(0.3028)	
Consumer Hist. Price	-5.4904	0.1174	-21.9471	NA
	(5.3397)	(4.4236)	(62.5292)	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Note: The entries in column (4) denoted by NA appear because there are no search impressions ending in a transaction in destination 4562 with historical consumer information under the random ranking. This occurs for several reasons: only 10 percent of search impressions end in a transaction under the random ranking and very few search impressions include historical information.

This means that the hotel's position at a point in time depends on its past performance. A machine learning algorithm computes a score of the hotel based on its past conversion and click through rates, its characteristics and its match with the consumer search query entries. A higher score is interpreted as a more desirable hotel that has a higher probability of purchase. Hotels are then ranked in decreasing order of these scores. This method for ranking hotels makes the position of the hotel endogenous. In contrast, under the random ranking, hotels are randomly assigned to positions, as stated by the competition administrator.[9] To show that this is indeed the case in the data, I run a rank ordered logit regression of position on hotel past conversion rate and its characteristics to mimic what a learning to rank algorithm does. My results can be found in Table 4 below. For this test, I restrict my attention to four of the largest destinations in my data set.[10] With very few exceptions, past performance of the hotel or its characteristics do not determine its position within the random ranking, while under Expedia's ranking there is a strong correlation between these characteristics and the hotel's position.[11] Moreover, this result provides important insights into how the Expedia's ranking is constructed. Expedia's ranking favors non-chains that were purchased more often in the past, that are cheaper, of higher quality and that are on promotion.

[9] `https://www.kaggle.com/c/expedia-personalized-sort/forums/t/5772/meaning-of-random-bool`.

[10] Destination 8192 has the largest number of observations (121,522), but has few observations with the random ranking, so I choose to focus on the next four largest destinations. See Appendix B 11.6 for summary statistics.

[11] In Appendix B 11.2, I show a symptom of displaying hotels based on past performance under Expedia's ranking, that of Expedia oversampling a small set of hotels to display at the top of the ranking.

Table 4: Effect of a hotel's past conversion rate and characteristics on position by ranking type

	Destination 4562		Destination 9402		Destination 8347		Destination 13870	
	Random Position	Expedia Position	Random Position	Expedia Position	Random Position	Expedia Position	Random Position	Expedia Position
Past CR	0.5893	2.7753***	1.1671	6.6953***	0.4661	3.1106***	1.9734	5.7512***
	(0.5558)	(0.2656)	(0.7776)	(0.3642)	(0.6967)	(0.2829)	(1.5981)	(0.5137)
Price	-0.0008***	-0.0061***	-0.0005	-0.0086***	-0.0002	-0.0021***	-0.0011	-0.0028***
	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0004)	(0.0003)	(0.0008)	(0.0003)
Stars	0.0253	0.4373***	0.0604	1.0209***	0.0254	0.3185***	0.0124	0.5431***
	(0.0292)	(0.0275)	(0.0381)	(0.0300)	(0.0441)	(0.0292)	(0.0701)	(0.0262)
Review Score	-0.0442*	0.0852***	-0.0809*	0.5471***	-0.0562	0.2600***	0.0273	0.3158***
	(0.0183)	(0.0190)	(0.0356)	(0.0391)	(0.0304)	(0.0327)	(0.0711)	(0.0377)
Chain	-0.1110**	-0.1188***	0.0201	0.4860***	0.0780	-0.3143***	0.0104	-0.3501***
	(0.0384)	(0.0282)	(0.0500)	(0.0358)	(0.0626)	(0.0324)	(0.0904)	(0.0330)
Location Score	0.0312*	0.4543***	-0.0231	0.4350***	0.0734**	0.0911***	-0.0100	0.1708***
	(0.0130)	(0.0178)	(0.0150)	(0.0182)	(0.0243)	(0.0184)	(0.0291)	(0.0155)
Promotion	0.1020*	0.6865***	0.1076*	1.1274***	0.1717**	0.8561***	-0.0061	0.8488***
	(0.0410)	(0.0280)	(0.0539)	(0.0307)	(0.0553)	(0.0355)	(0.0925)	(0.0310)
Observations	26,397	50,435	19,530	46,368	15,171	39,507	6,702	46,198
Log likelihood	-12,233	-21,077	-8,668	-17,619	-6,890	-16,718	-3,023	-18,435

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Rank ordered logit regression with dependent variable position. A positive coefficient means correlation with a top position. Positions greater than five are coded as incomplete. This is motivated by the observation that the learning to rank algorithm is engineered to correctly predict choices in top positions, with lower penalties for predicting a lower position wrong. As a result, to test whether the same algorithm is at play behind both rankings I focus on top positions.

3.3 Data Limitation

The data set is well suited to study the causal effect of rankings on choices. However, to avoid revealing Expedia's conversion rate and aid the machine learning algorithm, the data made available was chosen so that: (i) all search impressions have at least one click, and (ii) the fraction of searches leading to a transaction does not represent Expedia's or the random ranking's true conversion rate. Since the data was made available for a machine learning competition, including consumer choices (clicks and purchases) in the data is necessary to allow the ranking algorithm to learn consumers' preferences. Related to (i), observing at least one click per search is not typical of online click-stream data, where most search impressions receive no click.[12] However, other researchers, such as Chen and Yao (2014), focus on searches that end in a transaction, thereby also reducing their data set to one that has at least one click per search.

Related to (ii), from the WCAI companion data set for hotel searches in Manhattan, only 3% of search impressions with at least one click end in a transaction. However, in the Kaggle data set, 90% of search impressions under Expedia's ranking end in a transaction, compared to only 10% under the random ranking. The data made available was randomly sampled from searches ending and not ending in a transaction, with a larger weight placed on sampling searches ending in a transaction. Since this sampling was done randomly, the data set can be used to understand the causal effect of ranking on consumer choices.[13] However, this selection has two implications for my analysis. First, it means that I cannot compare conversion rates across the two ranking

[12]For example, De los Santos and Koulayev (2014) find that only 33% of consumers make a click.
[13]As assured by the administrator and as demonstrated in the previous section.

types or for the same hotel across two rankings, because this conversion rate is not representative of the performance of the two rankings. Second, if a large fraction of searches that contain one click lead to a transaction, one possible concern is that consumers discovered their ideal hotel on a previous visit and now return to purchase it. In this case, the ranking observed should have a minimal effect on consumer choices since the consumer will look for the previously identified hotel within those displayed and click on it regardless of its position in the current ranking. Figure 3 in Section 4.1 alleviates this concern by showing that higher ranked hotels receive more clicks under the random ranking. Since consumers were randomly assigned to the two ranking types and if the consumer had determined her ideal hotel before the current search, then under the random ranking there is no reason why this hotel would be displayed more often at the top of the ranking. As a result, observing more clicks at the top under the random ranking refutes this story.

Although this data set has its limitations, I conclude that the benefit of recovering the causal effect of rankings on choices and understanding the source of endogeneity and its magnitude are important questions that cannot be properly addressed without this data set that contains experimental variation in the position of hotel.

4 Reduced Form Evidence

In this section, I use the study the causal effect of rank on consumer choices and describe novel tests of two commonly made assumptions in the search literature.

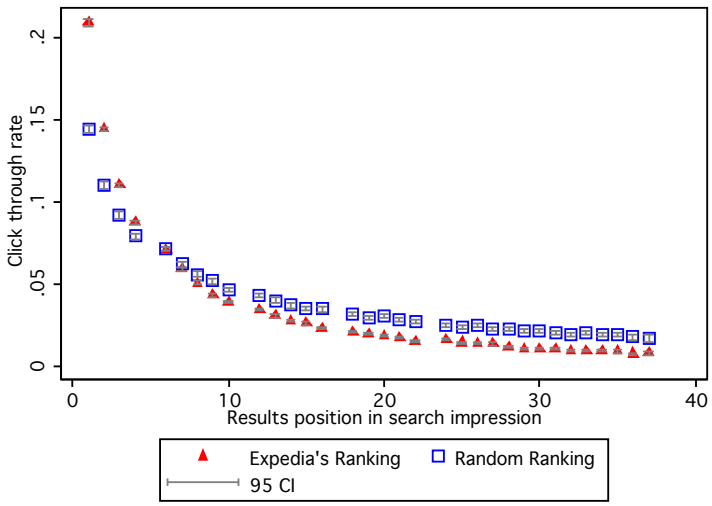
4.1 The Effect of Rank on Search and Choice

I start by considering the effect of the ranking observed on consumers' search. Consumers search by clicking on a hotel on the first page of results. In Figure 3, I illustrate the click through rate of a position. The click through rate of a position measures the fraction of times a position received a click out of all the times it was displayed. I restrict attention to search impressions that do not include a hotel in positions 5, 11, 17, 23. These positions are typically reserved for opaque promotions (deals where consumers get a large discount by booking a hotel they only learn about after they make a transaction). Only 13% of search impressions include a hotel in those positions, so that most search impressions include opaque offers. Two patterns are immediately obvious.[14] First, the click through rate of a position is surprisingly similar under the two rankings. Second, the click through rate is decreasing in position. The shape of the click through rate follows a power law pattern under both rankings. This power law pattern has been documented of other search intermediaries (e.g. Google) and can be expected of Expedia's ranking that ranks more relevant hotels at the top. However, the fact that a similar pattern holds for the click through rate of the random ranking is surprising. Hotels ranked at the top under the random ranking are not more likely to be of higher quality than those lower ranked, suggesting that the position

[14]The fact that the click through rate curves cross derives from the fact that all searches have at least one click.

of the hotel rather than its observable characteristics may play a larger role in determining the consumer's click.[15]

Figure 3: Click through rate (CTR) by results position and search impression type



The fact that the click through rate of a position is decreasing under the random ranking also alleviates a possible concern about the data selection. If most search impressions contain one click and lead to a transaction, one possible concern is that consumers made their search for the best hotel on previous visits, and thus the search observed in this data is one where the consumer has already identified her ideal hotel. However, this story is refuted by the fact that consumers were randomly assigned to the two ranking types and the fact that under the random ranking consumers are more likely to click on top positions. A hotel previously identified by the consumer is not more likely to be displayed at the top under the random ranking than at the bottom, so that higher click through rates in top positions refute this story.

After the consumer clicks on a hotel, she has the option of purchasing it. I study the effect of rankings on purchases by looking at the conversion rate (CR) of a position. The conversion rate measures the percent of clicks that end in a purchase. In Figure 4 I plot the conversion rate of a position for the two rankings separately.[16] I restrict attention to search impressions ending in a transaction to emphasize the difference in slope between the two figures that is separate from the click through rate pattern.[17] Two take-aways emerge from this figure. First, Expedia's conversion rate decreases with the position in the ranking, as higher ranked hotels lead to more purchases. Second, the conversion rate of the random ranking is constant across positions. In light of the similarity in click through rate patterns, the difference in slope in conversion rates by position under the two rankings is surprising and it suggests that consumers are able to identify which hotels they value after a click, but not before the click. It also implies that consumers' realized

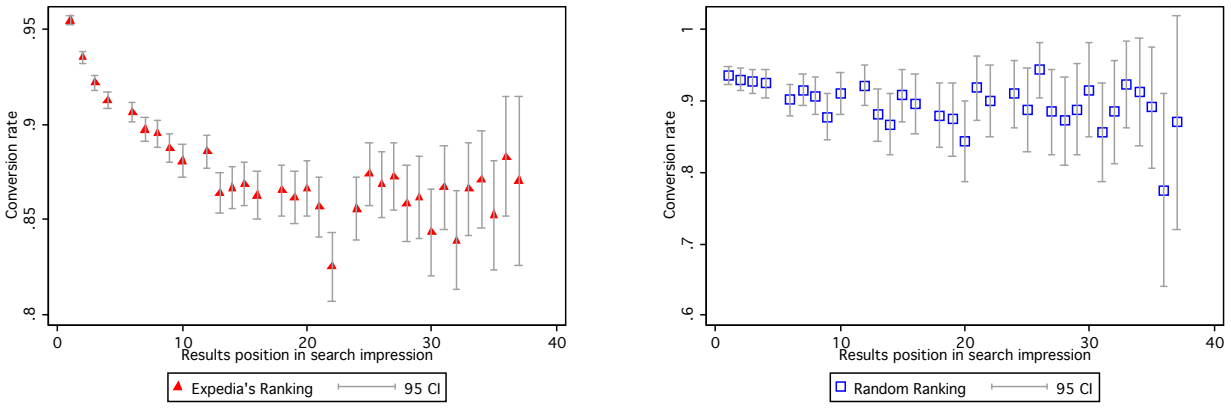
[15]See Appendix B 11.4.1 for various robustness checks. The same pattern as in Figure 3 holds.

[16]Note that the unconditional conversion rate of the two rankings is similar to the click through rate (see Figure 13 in Appendix B 11.4.2). Also, see Appendix B 11.4.2 for additional robustness checks.

[17]I thank Sergei Koulayev for this suggestion.

utility (their utility from hotel characteristics revealed both before and after the click) does not depend on the position of the hotel. These observations would not be possible without the random ranking. Finally, the stark difference in conversion rate under the two rankings suggests that Expedia's ranking is successful in identifying hotels that consumers want to purchase and displaying them at the top of the ranking, thereby potentially substantially reducing consumer's search costs.

Figure 4: Conversion rate conditional on a click by results position and search impression type



Expedia's Ranking Random Ranking

Note: Restrict attention to search impressions ending in a transaction.

The position of the hotel rather than its observable characteristics is important in determining the consumer's click. However, conditional on a click, higher ranked hotels sell more only under Expedia's ranking.[18] One possible interpretation of these patterns is the following. High quality hotels are not more likely to be displayed in the first position under the random ranking than in the last position. As a result, under the random ranking, the probability of a purchase conditional on a click is constant across positions. However, high quality hotels are more likely displayed at the top of Expedia's ranking, making its conversion rate higher in those positions. This suggests that Expedia's algorithm is able to identify those hotels that consumers want to purchase (those with large unobserved characteristics). This observations confirms the main benefit of search intermediaries: as information aggregators, they help consumers search more effectively by ranking first firms that they are more likely to find relevant. Even though theoretical concerns exist about intermediaries diverting consumer search (see Hagiu and Jullien, 2011), these figures show that they should be (at least partially) alleviated in this particular setting.

In this subsection I showed model free evidence of the effect of rank on consumer search and choice. In the next two subsections I describe novel tests of two common assumptions in the search literature.

[18]The same pattern can be shown in regression format. See Appendix B 11.7.

4.2 Informing the role of position in search

The last subsection established the influence of position on consumer choices. Figures 3 and 4 show that higher ranked hotels lead to more clicks under both rankings, but conditional on a click, top positions only lead to more transactions under Expedia's ranking. This suggests that the position of the hotel cannot affect the consumer's realized utility, since purchase decisions do not depend on the hotel's positions. However, the exact way in which positions affect consumer choices is unknown. The literature describes several mechanism through which the position may affect consumer choices. These mechanisms can be grouped into two main effects: rankings affect consumers' expected utility, either through signaling (Nelson, 1974; Kihlstrom and Riordan, 1984) or consumer learning about the relation between position and the relevance of an alternative (Varian, 2007; Athey and Ellison, 2011), or rankings affect consumer search costs (Ghose et al. 2012b; Chen and Yao, 2014).

I propose a novel test of whether the position of the hotel mainly affects search costs or expected utility. This test exploits the fact that I observe both searches where the nth displayed hotel is displayed in the nth position, and searches where it is displayed in the $n+1$th position. More precisely, in most search impressions, positions 5, 11, 17, 23 are reserved for opaque offers. In this case, no hotel is displayed in these positions, but rather an offer to purchase an unidentified hotel at a discount. As a result, the fifth displayed hotel will then be shown in position 6 instead of position 5. Approximately 13% of search impressions do not include such offers, and in this case, the fifth displayed hotel will be shown in position 5. I exploit this variation in the position of the fifth displayed hotel to test whether the position of the hotel mainly affects consumers' search costs or their expected utility (expectation of their match ϵ_{ij}). More formally, I look at search impressions that do not contain a click in the first four positions. In this case, consumers' expected utility at the fifth displayed hotel should be the same regardless of whether the fifth displayed hotel is in position 5 or 6. Any difference in the click through rate of position 5 and 6 will then be attributed to differences in search costs, not expected utility. In Table 5 I show my results. I run a linear probability model with dependent variable a click on an indicator for whether the fifth displayed hotel is shown in position 6, controlling for all observable characteristics of the hotels. I find that when the fifth hotel is displayed in position 6 it receives fewer clicks than when it is displayed in position 5. Although the coefficient is not significant, it points in the direction of the position mainly affecting consumers' search costs.

Table 5: Estimates of click on the position of the fifth displayed hotel

	Click
Position of fifth displayed hotel	-0.0099
	(0.0073)
Stars	0.0121***
	(0.0011)
Review Score	0.0070***
	(0.0009)
Chain	0.0039*
	(0.0018)
Location Score	-0.0029***
	(0.0006)
Price	-0.0003***
	(0.0000)
Promotion	-0.0057**
	(0.0019)
Observations	167,985
R^2	0.0069

Standard errors in parentheses
Note: Linear probability model with dependent variable a click of the probability of a click happening in the fifth displayed hotel, conditional on no click occurring in the first four displayed hotels. Search impressions with opaque offers will display the fifth hotel in position 6, while those without will display it in position 5.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3 Identifying Consumers' Search Method

If the two rankings display different types of hotels, then this difference can be used to determine whether consumers are searching simultaneous or sequentially.[19] Knowing that impressions from Expedia's ranking have better quality hotels displayed, if consumers were searching sequentially, then they are expected to terminate their search earlier when faced with Expedia's ranking than when faced with the random ranking. If consumers were searching simultaneously, then the number of clicks they make should not depend on the quality of hotels revealed by clicking, so they should make the same number of clicks under Expedia's ranking as under the random ranking.

Table 6: T-test: Number of clicks by search impression type

	Difference (Expedia-Random)
Total clicks in search impression	-0.0238***
	(0.0022)
Observations	317,218

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

These predictions and the differences in characteristics between the two rankings can be used to test for consumer's search method. To test for consumers' search method I perform a t-test of the difference in the number of clicks under the two rankings. In Table 6 I present my results. A negative sign means that consumers click less when faced with Expedia's ranking, consistent with consumers using a sequential search method. Since this result may in part be due to the fact that most search impressions have exactly one click, I perform the same t-test only on search impressions with at least two clicks. My results are in Table 7 and they also support the claim that consumers are searching sequentially.

Table 7: T-test: Number of clicks by search impression type in search impressions with at least two clicks

	Difference (Expedia-Random)
Total clicks in search impression	-0.0918***
	(0.0220)
Observations	20,592

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In this section I provided evidence of the effect of rankings on consumer search and choices, as well as described novel tests of commonly made assumptions in the literature. In the next section I present a sequential search model that can describe consumer's choices from an ordered list of alternatives.

[19] I thank Pradeep Chintagunta for suggesting this test.

5 Model

To understand the causal effect of rankings on consumer searches and choices, I can rely solely on the reduced form evidence and exploit the exogenous feature of the ranking observed. However, in order to quantify consumers' average utility under the two rankings and to perform counterfactuals constructing better rankings, I use a model of consumer search which I describe next.

5.1 Utility

When the consumer arrives at the OTA's website, she types in the destination, the exact dates of her trip, the number of guests traveling and the number of rooms she is looking to book. In response to this search query, she gets a search impression, i.e. an ordered list of hotels that match her search criteria. Such a search impression includes a lot of valuable information that the consumer observes without clicking on a particular hotel, i.e. without searching. For example, this list contains information about the name of the hotel, the number of stars it has, its review score and its the price. By clicking on a particular hotel, the consumer discovers more detailed information about it. More precisely, she can see more pictures of the hotel, can locate it on the map, read past consumer reviews, and learn about different amenities. I model this information as the consumer's match value with the hotel. The consumer readily observes this information after clicking on a hotel and can determine how much utility she derives from it, but from the econometrician's perspective, this information is unobserved. Therefore, I model the match value as a random error term. I follow Kim et al. (2010) and assume that the match value follows a standard normal distribution.

These considerations lead to the following model for consumer i's utility for hotel $j \in \{1, \ldots, J\}$

$$u_{ij} = v_{ij} + \epsilon_{ij} \tag{1}$$

where v_{ij} contains consumer i's valuation over hotel j's characteristics such as the number of stars, the review score and the price. This part of the utility function is known to the consumer even without searching. The match value ϵ_{ij} is only discovered by paying a search cost to click. The consumer also has an outside option denoted by $j = 0$, that of not booking a hotel, booking a hotel at a later time or choosing a different firm to book the trip. I do not have information about the exact outside option that the consumer chooses, so I model here the outside option as $u_{i0} = \epsilon_{i0}$. Thus the outside option is an i.i.d. random term that follows a standard normal distribution.

5.2 Search Cost

The consumer observes v_{ij} for all j's displayed on the first page of results for free. To learn about the match value ϵ_{ij} of a particular hotel, the consumer has to pay a search cost. I model consumer i's search cost for hotel $j \in \{1, \ldots, J\}$ as

$$c_{ij} = exp(l_{ij} + \eta_{ij}) \tag{2}$$

where l_{ij} contains the consumer's sensitivity to the booking window and the position of the hotel, consistent with my findings in the previous section and the literature (Ghose et al., 2013; Chen and Yao, 2014). Unlike most of the literature, I assume that the full search cost is observed by the consumer before searching, but not by the econometrician.[20] Adding an idiosyncratic shock to consumers' search costs is meant to capture the idea that, conditional on observables such as booking window, and characteristics and position of the hotel, consumers might click on different hotels in a different order. The order in which consumers click on hotels reveals important information about their preferences and search costs.[21] The random shock η_{ij} to search costs follows a standard normal distribution. The lognormal distribution of the search costs is consistent with prior literature (Kim et al. 2010, Ghose et al. 2013, Chen and Yao 2014) and it ensures that search costs are positive. The standard normal assumption on the random shock η_{ij} is chosen to simplify the estimation. To observe the outside option the consumer pays no search cost.

5.3 Optimal Search

To compute the optimal search strategy of consumers I rely on Weitzman (1979) who provides the solution to a general ordered search problem. His solution indicates that it is optimal for consumers to begin by ranking firms in order of their reservation utility. Reservation utilities are defined as the level of utility that the consumer would have to have in hand before searching a particular hotel to make her indifferent between searching that hotel or not. Weitzman (1979) shows that reservation utilities can be computed by equating the expected marginal gains from searching firm j with its marginal cost as in

$$c_{ij} = \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) f(u_{ij}) du_{ij} \tag{3}$$

where the z_{ij} that solves this equation is consumer i's reservation utility from searching j.

Kim et al. (2010) show that equation (3) can be rewritten by taking advantage of the distributional assumptions made. More precisely, if $\epsilon_{ij} \sim N(0,1)$, then $u_{ij}|v_j \sim N(v_j, 1)$. Using this and the expression for the expectation of the truncation of normally distributed random variables, equation (3) can be rewritten as

[20] Most papers estimating search models assume that search costs are completely deterministic. To the best of my knowledge, the only exception is Moraga-Gonzalez, Sandor and Wildenbeest (2015), which include a random component in their search cost specification, which has a T1EV distribution with linear search costs.

[21] Note that the Kaggle data set I use does not provide information on the order in which consumers clicked (also, most consumers only clicked once). As a result I infer the order of consumers' clicks from the position of the hotel. See Appendix B 11.8 for evidence supporting this assumption from the companion WCAI data set that includes click order.

$$\begin{aligned} c_{ij} &= (1-\Phi(m_{ij}))(\lambda(m_{ij}) - m_{ij}) \\ &= B(m_{ij}) \end{aligned} \tag{4}$$

where $\lambda(\cdot) = \frac{\phi(\cdot)}{1-\Phi(\cdot)}$ is the hazard function and where $m_{ij} = z_{ij} - v_j$. The result in equation (4) provides a straightforward way of computing the reservation utility z_{ij}. More precisely, it says that given any search cost c_{ij}, one can invert equation (4) and solve for m_{ij}.[22] Then, using the definition of m_{ij}, the reservation utility is given by $z_{ij} = m_{ij} + v_j$. Note that this specific function relating search costs and reservation utilities depends on the normality assumption of ϵ_{ij}. To speed up computation, I follow Kim et al. (2010) and construct a look-up table for $c_{ij} = B(m_{ij})$ outside the estimation loop. During estimation, for a particular value of search costs, I use the table to look up the value of m_{ij} and construct the reservation utility.

Once the consumer computes all reservation utilities z_{ij}, the following strategy due to Weitzman (1979) characterizes her optimal search

1. (Selection Rule): If a search is to be made, the firm with the highest reservation utility should be searched next.

2. (Stopping Rule): Search should terminate when the maximum utility observed exceeds the reservation utility of any unsearched firm.

3. (Choice Rule): Once the consumer stops searching, she will purchase from the firm with the highest realized utility of those searched.

These rules, demonstrated by Weitzman (1979) to characterize optimal search, inform the likelihood function I use in estimation.

5.4 Likelihood

Suppose there are J firms that consumer $i \in \{1, \ldots, I\}$ can search. Order these firms by consumer i's reservation utility. Denote by $R_i(n)$ the identity of the firm with the nth highest reservation utility. Suppose consumer i searched a number $h \leq J$ of these firms, so that $R_i = [R_i(1), \ldots, R_i(h)]$ gives the set of searched firms and the order in which they were searched. The outside option is always searched (denote it for simplicity as either $j = 0$ or $R_i(0)$). Observing the exact sequence of searches that the consumer performed and her purchase decision allows me to draw the following conclusions from Weitzman's optimal search strategy. First, if the consumer makes an nth search, then her reservation utility from that firm must exceed her reservation utility from all firms searched next and all those not searched. Formally, it must be that

$$z_{iR_i(n)} \geq \max_{k=n+1}^{J} z_{iR_i(k)}, \ \ \forall n \in \{1, \ldots, J-1\} \tag{5}$$

[22] Kim et al. (2010) show that the function $B(\cdot)$ is monotonic and decreasing in its argument and that a unique solution to $c_{ij} = B(m_{ij})$ exists. This is a specific application of the more general result proven by Weitzman (1979) on the existence and uniqueness of the reservation utility. Thus, this inversion is possible.

otherwise, using the selection rule, the consumer would have searched another firm next that had a higher reservation utility. Note that if search costs were fully observed by the econometrician, the reservation utilities z_{ij} would be exactly determined and thus the statement above would not a probability statement and it would not allow additional learning about consumer preferences and search costs. Second, if the consumer makes an nth search, then her reservation utility from that firm must exceed her utility from all firms searched so far, including the outside option. Otherwise, according to the stopping rule, the consumer would have stopped searching. Formally,

$$z_{iR_i(n)} \geq \max_{k=0}^{n-1} u_{iR_i(k)}, \ \ \forall n \in \{1, \ldots, J-1\} \tag{6}$$

Third, all unsearched firms must have a lower reservation utility than all searched alternatives, including the outside option,

$$z_{iR_i(m)} \leq \max_{k=0}^{h} u_{iR_i(k)}, \ \ \forall m \in \{h+1, \ldots, J\} \tag{7}$$

otherwise, according to the stopping rule, the consumer should have continued searching. Finally, if the consumer choose to purchase from firm j, including choosing the outside option, then her utility from this choice must exceed all utilities searched. Formally,

$$u_{ij} \geq \max_{k=0}^{h} u_{iR_i(k)}, \ \ \forall j \in R_i \cup \{0\} \tag{8}$$

If consumers search sequentially, then their search and purchase decisions are not separate. This means, that the probability of observing a certain outcome is characterized by a joint probability. Putting all of these conditions together, the probability P_{ijR_i} that i searches exactly in the order R_i and purchases from firm j (including the outside option) is given by

$$\begin{aligned} P_{ijR_i} &= Prob(z_{iR_i(n)} \geq \max_{k=n+1}^{J} z_{iR_i(k)} \cap z_{iR_i(n)} \geq \max_{k=0}^{n-1} u_{iR_i(k)} \cap z_{iR_i(m)} \leq \max_{k=0}^{h} u_{iR_i(k)} \cap \ u_{ij} \geq \max_{k=0}^{h} u_{iR_i(k)}, \\ &\quad \forall n \in \{1, \ldots, J-1\}, \forall m \in \{h+1, \ldots, J\}, \forall j \in R_i \cup \{0\}) \\ &= \int \int I(cond) \phi(\epsilon_i) d\epsilon_i \phi(\eta_i) d\eta_i \end{aligned} \tag{9}$$

where *cond* stands for the four conditions I derived from Weitzman's optimal search rule and where $I(\cdot)$ is an indicator for whether these conditions hold. The log-likelihood function is given by

$$LL = \sum_i \sum_{R_i} \sum_j d_{ijR_i} log P_{ijR_i} \tag{10}$$

where $d_{ijR_i} = 1$ if i chose search order R_i and purchased from j (including outside option). The integral in equation (9) does not have a closed form solution.[23] Thus, I replace the choice probability P_{ijR_i} with the simulated choice probabilities $\hat{P}_{ijR_i}$ which replace the integral in (9) with a summation over D draws of the two error terms (ϵ, η) from their respective distribution. This results in the following simulated log-likelihood

[23]Note that the integral in equation (9) has no closed form solution even if search costs are exactly determined.

$$SLL = \sum_i \sum_{R_i} \sum_j d_{ijR_i} log \hat{P}_{ijR_i} \tag{11}$$

The choice probability can be simulated in a number of ways. The most straightforward and widely used simulator is accept-reject (AR). It was originally proposed by Manski and Lerman (1981) for probits. This simulator approximates P_{ijR_i} by the proportion of draws from the appropriate distribution that satisfy the conditions (9). However, using the AR simulator in maximizing the SLL can be problematic for two reasons. First, any finite number of draws D can result in a reject, so that $\hat{P}_{ijR_i}$ is zero and the log of zero is undefined. This possibility is especially likely if the data contains very few choices, so that the true probability is low. This is the case with my data set. Each search impression contains on average 25 hotels, making searching in a particular order and buying from a particular hotel especially unlikely. The second difficulty comes from the fact that the choice probabilities are not twice differentiable, so the simulated probabilities will not be smooth. Thus, finding a maximum by optimizing the SLL using first and second derivatives will not be effective. Even though there is a way to circumvent this problem and use an approximation of the gradient to the SLL instead, Train (2009) concludes that in practice AR is difficult to use.

The GHK simulator after Geweke (1989, 1991), Hajivassiliou (Hajivassiliou and McFadden, 1998), and Keane (1990, 1994) would be another option, which is widely used as a probit simulator. The GHK operates on utility differences between the chosen alternative and those not chosen. As a result it requires knowledge of the distribution of the difference in utility and reservation utility of alternatives or the difference in reservation utilities of two alternatives. Utility error terms are normally distributed, while m in the reservation utility has a distribution given by $F_m(\mu) = 1 - F_c(B(\mu))$, where $F_c(\cdot)$ is the distribution of search costs. Thus, there is no closed form expression for the distribution of utility and reservation utility differences.[24] For these reasons, I choose not to use the GHK simulator, and instead replace the indicator function in the AR simulator with a smooth function. Any function that is increasing in the chosen alternative and that has defined first and second derivatives can be used. As suggested by McFadden (1989), I choose the logit function that satisfies these conditions and is convenient to use. This is known as the *logit-smoothed AR simulator*. It has also been successfully used by Honka (2014) and Honka and Chintagunta (2014) in the consumer search setting and by many others in simulating probit.

I now describe the steps I use to simulate $\hat{P}_{ijR_i}$ using the logit-smoothed AR simulator.

1. Draw $d = \{1, \ldots, D\}$ samples of $(\epsilon_{ij}^d, \eta_{ij}^d)$ for each consumer and each firm.

[24] I show here how to compute the distribution of m_{ij}. Drop subscripts and define $m \equiv z - v$, and $B(m) = [1 - \Phi(m)][\lambda(m) - m]$ and $c = B(m)$. Because B is a one-to-one function it has an inverse, denoted by B^{-1}. Then m equals $m = B^{-1}(c)$. Given a distribution for the search cost, F_c and using formulas for the distribution of a function of random variables, I can write the cdf of m as

$$F_m(\mu) = Pr(m < \mu) = Pr(B^{-1}(c) < \mu) = Pr(c > B(\mu)) = 1 - F_c(B(\mu)) \tag{12}$$

and the density of m is given by

$$f_m(\mu) = f_c(B(\mu)) |\frac{dB(\mu)}{d\mu}| \tag{13}$$

where $|\frac{dB(\mu)}{d\mu}| = 1 - \Phi(\mu)$.

2. Use $(\epsilon_{ij}^d, \eta_{ij}^d)$ to form utility u_{ij}^d and search cost c_{ij}^d.

3. Use the relation $c_{ij}^d = B(m_{ij}^d)$ to compute m_{ij}^d and form reservation utilities z_{ij}^d.

4. Define the following expressions for each draw d

   (a) $\nu_1^d = z_{iR_i(n)}^d - \max_{k=n+1}^{J} z_{iR_i(k)}^d$

   (b) $\nu_2^d = z_{iR_i(n)}^d - \max_{k=0}^{n-1} u_{iR_i(k)}^d$

   (c) $\nu_3^d = \max_{k=0}^{h} u_{iR_i(k)}^d - z_{iR_i(m)}^d$

   (d) $\nu_4^d = u_{ij}^d - \max_{k=0}^{h} u_{iR_i(k)}^d$

5. Put these expressions into the logit formula and compute S^d for each draw d

$$S^d = \frac{1}{1 + \sum_{n=1}^{4} e^{-\frac{\nu_n^d}{\lambda}}} \tag{14}$$

   where $\lambda > 0$ is a scaling parameter.

6. The simulated choice probability is the average over D draws of the error terms,

$$\hat{P}_{ijR_i} = \frac{1}{D} \sum_d S^d \tag{15}$$

There is little guidance in choosing the scaling parameter λ. As $\lambda \to 0$, the simulator is unbiased because it approaches the AR simulator. So, the researcher should use a small enough λ, but not too small to reintroduce the numerical problems one faces when optimizing with a non-smooth function.

5.5 Identification

The main difficulty in separately identifying preferences and search costs in differentiated products models, as pointed out by Sorensen (2001) and Hortaçsu and Syversion (2004), comes from understanding the consumer's stopping decision. More precisely, a consumer may stop searching because her search costs are very large or because she observed an alternative that provides her with a large utility gain. The key identification strategy relies on the idea that search decisions are determined both by utility and by search costs, while purchase decisions are only determined by utility differences. Thus, the set of covariates that enter search costs (search window and position), but do not enter utility can be used to identify preferences and search costs separately. In addition, all consumers see a different set of firms in their search impression, thus providing rich variation in terms of hotel characteristics observed, prices and positions of hotels. Using a similar identification strategy as in De los Santos and Koulayev (2014), fixing any set of hotel characteristics, I can find variation in other characteristics that identify the effect of the fixed characteristics. In addition, covariates that enter search costs, but not utility serve as exclusion restrictions (Chen and Yao, 2014). Finally, the nonlinearity in search costs also aids identification.

Both the price of the hotel and its position (in Expedia's ranking) may be endogenous. I will show evidence to alleviate concerns about price endogeneity and instead focus on the endogeneity of position. Price may be endogenous for two reasons. First, an unobserved quality shock may affect both the consumer's choices and hotel's prices. Second, consumer specific choice probabilities may affect what prices hotels set. The prices set by hotels are a function of their marginal costs and a markup term that depends on the quality of the hotel. If at least part of this quality is unobserved to the econometrician, but observed by consumers before making choices and by hotels, then the price set by the hotel may be correlated with the error term in the utility function of the consumer, and therefore be endogenous. A permanent or temporary common shock to the unobserved quality of the hotel can shift consumers' preferences and hotel's pricing decisions. Examples include the construction of a stadium next to the hotel, permanently decreasing the comfort of staying at the hotel or the organization of a conference near or at a hotel that temporarily increases demand for a particular hotel.

The most common method to alleviate price endogeneity concerns is instrumental variables. However, in the hotel industry, very few instruments (if any) are available. Insight provided by Coventure Analysis into the cost structure of the industry reveals that roughly 65% of the industry's costs in the period 2009-2014 came from two sources: labor and costs of goods sold (bedding and meals). As a reference, marketing accounts for only 2% in 2014 according to the same report. Labor costs can be treated as constant within a location, while the class of the hotel (its number of stars) may be a good approximation for the cost of goods sold. This insight suggests that a possible instrument for price may be the average price of the same hotel in a different location or the same star hotel in a different location. These Hausman style instruments are meant to capture the marginal costs of the hotel. The identifying assumption here is that the prices may be correlated across locations because of common marginal costs, but controlling for the hotel or the class of hotels, market specific valuations are uncorrelated across locations. Unfortunately, I cannot use the first instrument since I do not observe the same hotel in different locations. Even though the same hotel may be displayed in rankings in different destinations, these destinations are usually different neighborhoods of the same city or overlapping cities, making the identifying assumption difficult to satisfy. The second instrument is also problematic since destinations and countries are anonymous so taking the average price across very different destinations (for example cities on different continents) will make the assumption that the average price is capturing marginal costs hard to satisfy. Other possible instruments are lagged prices of the same hotel. However, if the unobserved quality of the hotel is correlated over time, the lagged prices will not be valid instruments as lagged prices would be correlated with the current period shock. Another option is using region dummies as proxies for marginal costs, but I do not observe regions and determining whether a destination is a neighborhood or an entire city will not provide an accurate enough approximation. Finally, as another instrument for price, one can use the average price of other hotels for the same trip, excluding the focal hotel, as well as the focal hotel's non-price characteristics. These instruments are similar to the one's used by Chen and Yao (2014) in the online hotel industry application, and by Hortacsu and Syverson (2004) and

BLP in different settings. These instruments capture the position in characteristics space of the focal hotel relative to all others, assuming that characteristics are predetermined or exogenous. However, this last assumption may not be tenable in my case.

Even though price instruments in the hotel industry are difficult to obtain, concerns about the endogeneity of price may be partially alleviated by the observation that prices are set by the hotel's revenue management system and thus not set in response to individual consumers' preferences.[25] Revenue management teams collect data on consumer demand, price and competition in a market that allows them to segment consumers in so called "micro-markets" based on their predicted willingness to pay. An optimization algorithm then finds prices that maximize the firm's revenue within each micro-market. Hotel revenue management systems work in a similar fashion, with segmentation based on willingness to pay, price-elasticity with respect to available substitutes, and group discounts (Cross et al. 2009; Mauri, 2013).[26] If hotels strive to offer targeted pricing, then observing the price that one consumer sees reveals important information about their underlying response parameters (this idea is similar to the one found in Manchanda et al. 2004 on physician detailing). However, I will show that even this form of endogeneity does not pose a significant concern. To show this I run a regression of price on observable characteristics to show that these capture most of the variation in prices. My results can be found in Table 8. In the first column, I only regress price on hotel and trip date fixed effects in destination 4562 and obtain an adjusted R^2 of 0.766. This provides suggestive evidence that segmentation at a particular hotel is mostly based on the trip date. In other words, specific dates command different prices, but all consumers searching for a hotel for the same trip date will see the same price for a particular hotel. In the next column I add additional trip characteristics requested, such as the length of the trip or the number of travelers, and show that these contribute to explaining the additional variation in price. Similarly, adding search characteristics, such as the number of days before the trip that the search is conducted or the time of the day or the day of the week of the search, also contribute to explaining the variation in price. Finally, including information about the average prices of similar hotels for the same trip and whether the hotel is running a promotion, provides the largest increase in variation explained and leads to an adjusted R^2 of 0.813. The last three columns of Table 8 show that a similar pattern happens across all four destinations.[27]

This analysis suggest that observable characteristics explain most of the variation in price of a hotel, with the trip date explaining the majority of it. From discussions with an employee at an large hotel chain and from previous literature, the remaining price variation may be due either to (i) different suppliers selling the particular hotel, or (ii) experimental price variation (Einav et

[25]Koulayev (2014) makes a similar observation.

[26]Note that Expedia does not set hotel prices, it merely ranks hotels with the characteristics set by hotel managers.

[27]In Appendix B 11.5 I also report the adjusted R^2 from running similar separate regressions as in Table 8 on each hotel in a destination. For these histograms I include all hotels in a destination that have more than one observation. In Figure 19 I find that the regression recovers most of the variation in price for most hotels. In Appendix B 11.5, I investigate which hotels have a larger unexplained portion of the variation in price (see Figure 20). I find that such hotels have a smaller number of displays, are chains with fewer than 3 stars and lower hotel location score.

Table 8: Predicting price with observable characteristics

Destination	4562	4562	4562	4562	9402	8347	13870
Trip characteristics							
Trip Length (days)		1.3723***	1.1195***	1.6314***	2.4010***	0.1274	0.3313***
		(0.1232)	(0.1266)	(0.1168)	(0.1403)	(0.1144)	(0.0945)
Adults		9.8311***	9.5846***	8.1004***	3.2964***	4.4646***	3.4776***
		(0.3201)	(0.3215)	(0.2943)	(0.3018)	(0.2219)	(0.1853)
Children		13.9633***	13.8446***	11.3240***	6.2819***	3.4076***	3.5209***
		(0.4252)	(0.4255)	(0.3887)	(0.3050)	(0.2049)	(0.1211)
Rooms		-4.7943***	-4.9863***	-3.9110***	-5.2620***	-3.6941***	-8.1948***
		(0.7051)	(0.7058)	(0.6439)	(0.7167)	(0.6114)	(0.4839)
Saturday Night		0.0200	0.3161	-0.4788	0.0522	1.6732	4.0309***
		(1.0484)	(1.0485)	(0.9573)	(1.1569)	(0.9064)	(0.6819)
Search characteristics							
Booking Window (days)			0.0729***	0.0482***	-0.0323***	-0.0243***	-0.0904***
			(0.0071)	(0.0065)	(0.0060)	(0.0065)	(0.0043)
Time of day							
9am-6pm			-0.9025	-0.6583	-2.5246***	-1.6776**	-0.5700
			(0.6637)	(0.6045)	(0.6086)	(0.5673)	(0.4203)
6pm-midnight			-2.8066***	-2.0094**	-2.8446***	-2.3766***	-1.3685**
			(0.7235)	(0.6595)	(0.6395)	(0.6447)	(0.4328)
Weekend			-0.7446	-0.0903	-1.9055***	0.2331	0.8280**
			(0.5574)	(0.5085)	(0.4848)	(0.5049)	(0.3115)
Competition							
Avg. prices of similar hotels				-1.7492***	-2.0091***	-1.8734***	-2.0930***
				(0.0165)	(0.0199)	(0.0256)	(0.0226)
Promotion				-21.2012***	-26.1707***	-17.7625***	-14.1015***
				(0.5656)	(0.5952)	(0.6917)	(0.5572)
Hotel and trip date fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	60,968	60,968	60,968	58,486	58,855	38,598	51,458
Adjusted R^2	0.766	0.774	0.775	0.813	0.801	0.842	0.842

Standard errors in parentheses
OLS regression with dependent variable price. Time of day of the search is with respect to the left out variable, searches performed between midnight and 9am (local time).
The average price of similar hotels is computed as the average price of hotels with the same number of stars and reviews and same type (chain or independent) as the focal hotel for the same trip date (excluding the focal hotel).
I restrict attention to hotels that are displayed at least 100 times to be able to include hotel fixed effects in all specifications above.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

al., 2015; Koulayev, 2014). Since both of these explanations are not demand related (supply or experimental), I conclude that the price variation observed conditional on parameters of request, is unlikely to be correlated with the utility error term and thus does not need an instrument. Thus, the price observed by the consumer does not reveal any additional information than what is contained in her search query characteristics. Since all estimation is done conditional on the search query, the price observed by a consumer does not reveal additional valuable information, alleviating price endogeneity concerns. Finally, a hotel's price did not vary by ranking type. As a result, any remaining concerns about price endogeneity should not translate into concerns about the price coefficient estimates being biased across the two rankings.

Having addressed concerns about price endogeneity, I will focus on position endogeneity in this paper. The position that a hotel is displayed in is endogenous because it is chosen by the OTA to maximize the probability that a consumer makes a purchase. OTA's use learning to rank algorithms to decide where to display a particular hotel and these algorithms are a function of the hotel's characteristics, their match with the consumer's search query and their past performance (past click through rate and conversion rate). See Appendix B 11.1 for details about learning to rank algorithms. This non-random nature of the ranking creates a simultaneity problem in estimating the search model: the current position of the hotel affects its performance (clicks/purchases) and its performance affects its future position. This means that consumers' choices depend on the hotel's current position and the hotel's position depends partially on consumer preferences. I can describe the problem as follows.[28] The probability of an outcome (purchase or click) of consumer i at hotel j conditional on the preferences of the consumer, can be decomposed into the probability of that outcome, conditional on the position of j and consumer's preferences and the marginal probability of observing j in a given position

$$Pr(\text{outcome}_{ij}|\theta_{ij}) = Pr(\text{outcome}_{ij}|X_{ij}, \text{position}_j, \theta_{ij})Pr(\text{position}_j|X_{ij}, \theta_{ij}) \tag{16}$$

where θ_{ij} denotes consumer i's preference parameters for hotel j and X_{ij} denotes the hotel's characteristics revealed at the time of i's search. The marginal probability $Pr(\text{position}_j|X_{ij}, \theta_{ij})$ depends on hotel's characteristics and on consumer's preferences because they are chosen to maximize the probability that the consumer will buy from j. Thus, in maximizing $Pr(\text{outcome}_{ij}|\theta_{ij})$ to estimate θ_{ij}, one cannot omit the marginal probability from the likelihood, because it would lead to biased estimates of θ_{ij} (similar to Manchanda, Rossi and Chintagunta, 2004).

The approach taken in this paper to eliminate the endogeneity bias in the hotel's position is twofold. First, I use the random ranking subsample to estimate the model. When the ranking is random, the marginal probability is independent of θ_{ij} making maximization of the conditional likelihood sufficient. Second, I propose a method to eliminate the endogeneity bias in Expedia's ranking by modeling the marginal probability $Pr(\text{position}_j|X_{ij}, \theta_{ij})$ and I validate it through comparison with the random ranking results. This method is in the spirit of Manchanda, Rossi and Chintagunta (2004) and addresses simultaneity.

[28] I adapted this explanation from Nair et al. (2014).

In this subsection, I discussed identification and un the next subsection, I present simulation results that confirm that my model is identified.

5.6 Simulation

In this section, I describe simulation results on generated data that show that Simulated Maximum Likelihood using the logit-smoothed AR simulator recovers parameters well.[29] To this end, I generate a data set of 1000 consumers, each searching among five firms, which leads to a total of 5,000 observations. To construct the utility of the consumer, I draw the characteristics/quality of the hotel to represent the number of stars of the hotel from a normal distribution with mean 3 and variance 1 and I draw prices for each hotel from a normal distribution with mean 3 and variance 1/4. This ensures that at least some consumers want to search or make a transaction. I determine the position of each hotel in a consumer's ranking by drawing without replacement from the set $1, \ldots, 5$. The booking window is a random draw from a lognormal distribution with mean zero and variance 1. Given search costs, I use the relationship $c = B(m)$ to compute m and form reservation utilities for each consumer and product combination. By ordering hotels for each consumer by their reservation utility, I can determine what clicks the consumer will make, in what order, and what transactions she will make.

Using this generated data set, I estimate the parameters using the logit-smoothed AR simulator. I use 50 draws from the distribution of the utility and the search cost error terms for each consumer and hotel combination and a scaling factor $\lambda = 1/5$. The results of this simulation (which was repeated 50 times) are given in Table 9. On the left hand side in parenthesis are the true parameters and on the right hand side I show the estimated parameters. I find that my method works well in recovering the parameters of interest. In the second column, I contrast my estimates with those from a model that does not account for the order of clicks. As described previously, this is a model where search costs are deterministic so that the order of clicks does not inform estimates of preferences and search costs, i.e. it lacks the additional inequalities relating the order of reservation utilities. I find that such a model cannot recover parameter estimates as well, especially for the number of stars and the search cost estimates. I supplement this claim with evidence on the number of observations estimated correctly in Table 10. Here the data set is ordered by the estimated reservation utility of each consumer. In the first panel, I count how many hotels' reservation utilities match the order of those hotels in the generated data set. The second and third panel perform a similar calculation for observations with at least one click and with a transaction. I find that in all three cases, the model that accounts for order improves the performance of the estimates, especially for observations that contain consumer choices, i.e. clicks and transactions.

Having introduced the model and discussed its identification, in the next section I apply the model to data on consumer online searches for hotels.

[29] I thank Elisabeth Honka for the hints she gave me on running this simulation.

Table 9: Simulation results

	Accounting for Order	Not Accounting for Order
Preferences		
Constant (1)	0.9599***	0.7282***
	(0.2666)	(0.2026)
Stars (0.5)	0.5232***	0.2003***
	(0.0260)	(0.0197)
Price (-1)	-1.1478***	-0.5638***
	(0.0714)	(0.0557)
Search cost		
Constant (-3)	-2.9126***	-0.7391***
	(0.0020)	(0.0026)
Search Window (-1)	-0.9052***	-0.1249***
	(0.0008)	(0.0023)
Position (1)	1.1247***	0.1968***
	(0.0001)	(0.0028)
Observations	5,000	5,000
Log-likelihood	-3,307	-2,862

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Simulation results after 50 repetitions, 50 draws from the distribution of utility and search cost error terms for each consumer and hotel combination and scaling factor $\lambda = 1/5$.

Table 10: Number and percentage of observations estimated correctly under the two models

	Number	Percentage
All observations (5,000)		
Not accounting for order	1,893	36.78
Accounting for order	3,434	68.68
Observations with at least one click (1,224)		
Not accounting for order	597	48.77
Accounting for order	1,080	88.24
Observations with a transaction (407)		
Not accounting for order	196	48.16
Accounting for order	362	88.94

Note: An observation is a hotel ordered correctly by its reservation utility for a consumer.

6 Application of the Model

6.1 Utility and Search Costs

In this section, I describe the application of the sequential search model to the data on consumers searching for hotels on Expedia. I model consumer i's utility for hotel j as

$$u_{ij} = x_j\beta - \alpha p_{ij} + \epsilon_{ij} \tag{17}$$

where x_j gives the characteristics of the hotel and p_{ij} is the price that j charges at the time that i makes the search query. More precisely, x_j contains a constant (a value for the baseline utility), the number of stars of the hotel, its review score, a location score, an indicator for whether the hotel is part of a chain, and an indicator for whether the hotel is running a promotion. Because the data set is anonymized I cannot tell the exact brand or name of the hotel so these characteristics are meant to best capture the relevant information about the brand of the hotel. The random shock to utility ϵ_{ij} follows a standard normal distribution. I model search costs as follows

$$c_{ij} = exp(L_{ij}\gamma + \eta_{ij}) \tag{18}$$

where L_{ij} contains a constant (the baseline search cost of the consumer), the booking window (the number of days before the trip that the consumer searches) and the position of the hotel in the ranking that the consumer observes. The random shock to search costs η_{ij} follows a standard normal distribution.

Since my data is at the search impression level, I do not observe the consumer making the search. I thus consider a search in the data as being performed by a unique consumer. As a result, I do not observe any within consumer variation in tastes for hotel characteristics or search costs, which motivates my assumption of constant parameters $[\beta, \alpha, \gamma]$. In addition, the hotel characteristics observed on the first page of results, such as the number of stars or the reviews of the hotel, are more likely to be considered by consumers as vertical characteristics. In this case, I expect there to be limited across consumer variation in their valuation for these characteristics. Consumers are more likely to treat characteristics observed on the hotel's page as horizontal (for example, interpreting the text in a review or seeing detailed pictures of the hotel). Thus, consumer heterogeneity is captured by the idiosyncratic shock to utility that models the characteristics that consumers search for and that are revealed on the hotel's page. The random component in the consumer's search cost models the idea that the consumer observes her exact search costs, whereas the econometrician has only partial information about this cost that is revealed from the booking window and the position of the hotel. I allow this random shock to be heterogeneous across consumers and hotels.

The purpose of estimation is recovering $\theta = [\beta, \alpha, \gamma]$. I estimate parameters using Simulated Maximum Likelihood where I simulate the choice probabilities using the logit-smoothed AR simulator with 50 draws for each consumer-hotel combination of utility and search cost error terms and a scaling factor of $\lambda = 5$. The next subsection presents my results.

6.2 Estimation Results (preliminary)

For the present analysis I restrict my attention to one of the four largest destinations in my data set, destination 13870.[30] I further restrict attention to search impressions of the same length. This allows me estimate consistent position effects. More precisely, if across ranking types some search impressions were longer than others, then the estimated position effect may be in part driven by the fact that one ranking type has shorter search impressions potentially leading to a larger position effect. I thus focus on search impressions longer than the average search impression length (25).

Table 11 shows the main estimation results. In Panel A I present the coefficient estimates, while in Panel B I derive results based on these estimates that facilitate interpretation. There are two columns in Table 11 for each ranking type. In general the estimates of preference and search costs are economically meaningful and significant. For example, consumers derive higher utility from additional number of stars, better review scores, better location, more promotions and lower prices. The first column, where I estimate the search model on the random ranking subsample, should be interpreted as the unbiased estimates. Expedia's ranking is constructed on the premise of ranking the best hotels at the top. This is realized by learning which hotels perform well and which characteristics consumers value the most from past hotel performance. Past conversion and click through rates, together with hotel characteristics and their match with consumer search query observables, contribute to the determination of the position of a hotel in a search impression. As a result, the position of the hotel under Expedia's ranking is endogenous and not controlling for this endogeneity may lead to biased results. This is exactly what I find in Table 11. Not controlling for endogeneity leads in many cases to an underestimation of the importance of preference parameters, for example for the price sensitivity of the consumer, her preference for reviews, chains or promotions. In contrast, search costs parameters are greatly overestimated when not controlling for endogeneity.[31] For example, baseline search costs are equal to \$12.56 ($exp(-3.0674)/(0.0037)$), but by not controlling for endogeneity (column 2), the baseline search costs are estimates to be equal to \$19.81. Moreover, an increase in position by one is associated with an increase in search costs of 53 cents, but the effect of position is inflated to 80 cents when not controlling for endogeneity.

Incorrectly estimating the importance that consumers attribute to the alternatives that they are displayed hinders evaluation of the performance and the improvement of the current ranking. Having access to a data set where the ranking was randomly generated allows me to circumvent this endogeneity bias and correctly recover the causal effect of rankings. In the next section I use the estimation results from this section to measure the average utility gain of consumers under the current Expedia ranking and compare it to the welfare from an improved ranking.

[30] See Appendix B 11.6 for summary statistics for this destination.

[31] Recall that search costs are modeled as $c_{ij} = exp(L_{ij}\gamma + \eta_{ij})$, while the table presents the coefficient γ.

Table 11: Main Estimation Results: Destination 13870

Ranking	Random	Expedia
Panel A: Coefficients		
Preferences		
Price (100 $)	-0.3704***	-0.2453***
	(0.1005)	(0.0304)
Stars	0.1061*	0.2116***
	(0.0508)	(0.0205)
Review Score	0.2055**	0.0274
	(0.0625)	(0.0295)
Location Score	0.0466*	0.07133***
	(0.0234)	(0.0123)
Chain	0.0306	0.0175
	(0.0730)	(0.0292)
Promotion	0.0692	-0.0350
	(0.0674)	(0.0272)
Constant	0.6735	0.9883***
	(0.7114)	(0.2157)
Search Cost		
Position	0.0408***	0.0394***
	(0.0009)	(0.0000)
Booking Window (100 days)	-1.4359***	-1.3543***
	(0.0077)	(0.0031)
Constant	-3.0674***	-3.0240***
	(0.0076)	(0.0042)
Observations	2,475	14,525
Log-likelihood	-334	-2,095
Panel B: Equivalent Change in $		
Position	**0.53**	**0.80**
Stars	28.64	86.23
Baseline search costs	12.56	19.81
Booking Window	-0.19	-0.28

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6.3 Counterfactuals

6.3.1 Average Utility Gain of a Ranking

The preference and search cost estimates from the random ranking allow me to compare rankings, assuming the distribution of preferences does not change. To avoid the potential bias resulting from not knowing consumers' outside option, I focus here on the average utility of consumers who purchased under Expedia's ranking. For these consumers, I propose to measure the average utility of the ranking displayed as follows. Consider search impressions ending in a transaction and seeing the ranking r. Define the *average net utility* of a ranking as a function of parameters θ as

$$U_r(\theta) = \frac{1}{N} \sum_i \left[u_{ij*}(\theta) - \sum_j c_{ij}(\theta) \right] \tag{19}$$

for all consumers i who purchase a hotel $j*$ and searched hotels j (including the purchased hotel), where N gives the number of search impressions ending a transaction.

I use this metric for two purposes. First, I use it to measure the average net utility of the purchasing consumers under Expedia's ranking. Second, I use it construct a counterfactual ranking. One option for constructing this ranking is taking all hotels in a certain destination and ranking them according to the average utility (average net utility plus total search costs) with the parameters I estimated from the ranking ranking model. I consider this counterfactual to be unrealistic, because there might be good reasons (e.g. availability) why the OTA chose a particular set of hotels to display to consumers. I thus consider a different counterfactual ranking that is constructed as follows: beginning with the hotels that were displayed to the consumer under Expedia's ranking, the Best ranking rearranges hotels in descending order of their average utility (as estimated from the random ranking). In other words, this ranking measures the welfare gain from reordering the hotels that were displayed to consumers under Expedia's ranking, without changing their characteristics. I call this ranking the "Best Ranking" for consumers.

In destination 13870, 374 out of 581 consumers made a purchase. Out of these 374 purchases, 49 consumers continue purchasing from the same hotels under the Best ranking as they did under Expedia's ranking. In Table 12 I compare the characteristics of the hotels purchased when consumers purchase from different hotels under the two rankings. I find that the average consumer gains approximately \$57 from an improved ranking, partially from searching less and partially from purchasing from a better hotel. However, the average transaction price decreases, suggesting that Expedia's revenues would decrease by moving toward the improved ranking for consumers. More precisely, hotels purchased under the Best ranking are on average \$14 cheaper and have half a star and half a review point more than those purchased under Expedia's ranking. This means that compared to the ideal ranking for consumers, Expedia's current ranking displays hotels that are too expensive, with too few stars and review scores, not enough chains and location scores that are too high. Finally, in Figure 5 I show a histogram of the average net utility under the two rankings. This figure clearly shows that the best ranking shifts the distribution of average

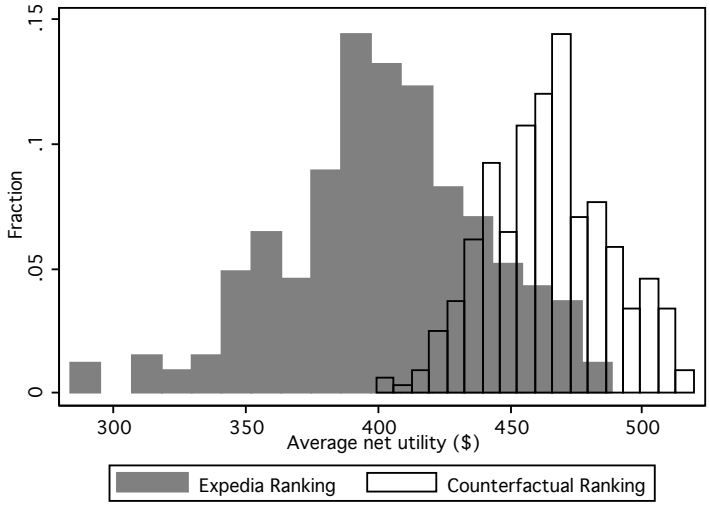
Table 12: T-test: Hotel characteristics of purchased hotels: Destination 13870

	Difference (Best-Expedia)
Average net utility	57.18***
	(24.18)
Average search cost	-5.812***
	(-9.59)
Price	-14.33***
	(-5.58)
Stars	0.545***
	(10.14)
Review Score	0.494***
	(16.06)
Chain	0.271***
	(9.16)
Location Score	-0.267***
	(-3.77)
Promotion	0.0523
	(1.38)
Position	0.437
	(0.78)
Observations	650

t statistics in parentheses
Note: Comparing average characteristics of hotels purchased under Expedia's ranking and in the counterfactual ranking, excluding consumers who purchased the same hotel under both rankings and consumers who did not purchase under Expedia's ranking.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

net utility to the right, increasing consumers' welfare.

Figure 5: Histogram of Average Net Utility Gain from an Improved Ranking



Note: Restrict attention to consumers who purchase under both rankings. Average net utility is defined as the average utility of the consumer minus her total search costs.

In this section, I estimated consumer preferences and search costs and shown both the direction and the magnitude of the endogeneity bias by comparing the estimated parameters under the random ranking and Expedia's ranking. I then used the estimated parameters from the random ranking to measure the average net utility of consumers from Expedia's current ranking and compared it to the welfare gain from an improved ranking. The next section concludes.

7 Limitations and Future Research

One area of research I plan to focus on next is proposing a method to eliminate the endogeneity bias inherent in the ranking. This method can be validated by comparing estimates using it with those from the random ranking. I plant use this method to estimate preferences and search costs in the companion data set which does not suffer from this data limitation, but which also does not have the luxury of a random ranking.

8 Conclusion

In this paper, I quantify the role of search intermediaries rankings in affecting consumer search and purchase decisions. Search intermediaries aggregate information and rank alternatives for consumers in order of relevance. As a result, the ranking observed in the data is endogenous, making it difficult to separate the role of the position of the alternative in the ranking from all other characteristics of the alternatives. Instrumental to my research has been a unique data set from a popular online travel agent, Expedia, that contains observations on search impressions both from Expedia's curated ranking and from a random ranking. Having access to data from the random ranking facilitates the study of the causality of the ranking. In addition, it allows me to show both the direction and the magnitude of the endogeneity bias of position under Expedia's ranking by comparing it to the random ranking. Finally, using a model of sequential search, I estimate consumers' preference and search costs and use these estimates to construct several counterfactual experiments of interest comparing the welfare of Expedia's ranking with that of improved rankings.

9 References

1. Anderson, S. P., and R. Renault (1999): "Consumer Information and Firm Pricing: Negative Externalities from Improved Information," International Economics Review, 41, 721-742.

2. Ansari, A., and C. Mela (2003): "E-customization," Journal of Marketing Research, 40, 131-145.

3. Ansari, A., S. Essegaier, and R. Kohli (2000): "Internet Recommendation Systems," Journal of Marketing Research, 37, 363-375.

4. Armstrong, M., J. Vickers, and J. Zhou (2009): "Prominence and Consumer Search," RAND Journal of Economics, 40, 209-233.

5. Armstrong, M., J. Vickers, and J. Zhou (2011): "Paying for Prominence," Economic Journal, 121, F368-395.

6. Arbatskaya, M. (2007): "Ordered Search," RAND Journal of Economics, 38, 119-126.

7. Athey, S., and G. Ellison (2011): "Position Auctions with Consumer Search," Quarterly Journal of Economics, 126, 1213-1270.

8. Baye, M., J. Morgan, and P. Scholten (2006): "Information, Search, and Price Dispersion," T. Hendershott (ed.) Handbook of Economics and Information Systems, Elsevier Press, Amsterdam.

9. Baye, M, B. De los Santos, and M. Wildenbeest (2014): "What's in a Name? Measuring Prominence, and Its Impact on Organic Traffic from Search Engines," Working paper.

10. Berman, R., and Z. Katona (2013): "The Role of Search Engine Optimization in Search Marketing," Marketing Science, 32, 644-651.

11. Blake, T., C. Nosko, and S. Tadelis (2014): "Consumer Heterogeneity and Paid Search Effectiveness: A Large Scale Field Experiment," Working paper.

12. Breiman, L., J. Friedman, C. Stone, and R. Olshen (1984): "Classification and Regression Trees," Wadsworth Statistics/Probability.

13. Burdett, K., and K. L. Judd (1983): "Equilibrium Price Dispersion," Econometrica, 51, 955-69.

14. Burges, C., T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender (2005): "Learning to Rank Using Gradient Descent," In Proceedings of the 22nd International Conference on Machine Learning, 89-96. ACM.

15. Burges, C., R. Ragno, and Q.V. Le (2006): "Learning to Rank with Non-Smooth Cost Functions," Advances in Neural Information Processing Systems.

16. Burges, C. (2010): "From RankNet to LambdaRank to LambdaMART: An Overview," Technical report, Microsoft Research Technical Report MSR-TR-2010-82.

17. Chan, T.Y., and Y.H. Park (2015): "Consumer Search Activities and the Value of Ad Positions in Sponsored Search Advertising," Marketing Science, Articles in Advance, 1-18.

18. Chapelle, O., and Y. Chang (2011): "Yahoo! Learning to Rank Challenge Overview," Journal of Machine Learning Research-Proceedings Track, 14, 1-24.

19. Chen, Y., and S. Yao (2014): "Sequential Search with Refinement: Model and Application with Click-stream Data," Working paper.

20. Cross, R., J. Higbie, and D. Cross (2009): "Revenue managementÕs renaissance: a rebirth of the art and science of profitable revenue generation," Cornell Hospitality Quarterly 50: 56-81.

21. Diamond, P. A. (1971): "A Model of Price Adjustment," Journal of Economic Theory, 3, 156-168.

22. De los Santos, B., and S. Koulayev (2014): "Optimizing Click-through in Online Rankings for Partially Anonymous Consumers, " Working paper.

23. De los Santos, B., A. Hortaçsu, and M. Wildenbeest (2012): "Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior," American Economic Review, 102, 2455-2480.

24. De Corniere, A., and G. Taylor (2014): "Quality Provision in the Presence of a Biased Intermediary," Working paper.

25. Einav, L., T. Kuchler, J. Levin, and N. Sundaresan (2015): "Assessing Sale Strategies in Online Markets Using Matched Listings," American Economic Journal: Microeconomics, 7(2), 215-247.

26. Geweke, J. (1989): "Bayesian Inference in Econometric Models using Monte Carlo Integration," Econometrica, 57, 1317-1339.

27. Geweke, J. (1991): "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints," in E.M. Kermidas, ed., Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface, Interface Foundation of North America, Inc., Fairfax, 571-578.

28. Ghose, A., P. Ipeirotis, and B. Li (2012a): "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowd-Sourced Content," Marketing Science, 31, 492-520.

29. Ghose, A., P. Ipeirotis, and B. Li (2012b): "Surviving Social Media Overload: Predicting Consumer Footprints on Product Search Engines, " Working paper.

30. Ghose, A., P. Ipeirotis, and B. Li (2013): "Examining the Impact of Ranking on Consumer Behavior and Search Engine Revenue," forthcoming at Management Science.

31. Ghose, A., and S. Yang (2009): "An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets," Management Science, 55, 1605-1622.

32. Haan, M.A., and J.L. Moraga-Gonzalez (2011): "Advertising for Attention in a Consumer Search Model," The Economic Journal, 111, 552-579.

33. Haan, M.A., J.L. Moraga-Gonzalez and V. Petrikaite (2014): "Advertising, Consumer Search and Product Differentiation," Working paper.

34. Hagiu, A., and B. Jullien (2011): "Why do Intermediaries Divert Search?," RAND Journal of Economics, 42, 337-362.

35. Hajivassiliou, V., and D. McFadden (1998): "The Method of Simulated Scores for Estimation of LDV Models," Econometrica, 66, 863-896.

36. Hausman, J. (1996): "Valuation of New Goods Under Perfect and Imperfect Competition," in The Economics of New Goods, Studies in Income and Wealth, ed. by T. Bresnahan, and R. Gordon, 207-248. National Bureau of Economic Research.

37. Yoganarasimhan, H. (2014): "Search Personalization," Working paper.

38. Hong, H., and M. Shum (2006): "Using Price Distributions to Estimate Search Cost," RAND Journal of Economics, 37, 257-275.

39. Honka, E. (2014): "Quantifying search and switching costs in the U.S. auto insurance industry," forthcoming in RAND Journal of Economics.

40. Honka, E., and P. Chintagunta (2014): "Simultaneous or Sequential? Search Strategies in the U.S. Auto Insurance Industry," Working paper.

41. Hortaçsu, A., and C. Syverson (2004): "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds," Quarterly Journal of Economics, 119, 403-456 .

42. Janssen, M. C., and J. L. Moraga-Gonzalez (2004): "Strategic Pricing, Consumer search and the Number of Firms," Review of Economics Studies, 71, 1089-1118.

43. Jerath, K., L. Ma, Y. Park, and K. Srinivasan (2011): "A Position Paradox in Sponsored Search Auctions," Marketing Science, 30, 612-627.

44. Jeziorski, P., and S. Moorthy (2014): "Advertiser Prominence Effects in Search Advertising," Working paper.

45. Jeziorski, P., and I. Segal (2012): "What Makes them Click: Empirical Analysis of Consumer Demand for Search Advertising," Working paper.

46. Keane, M. (1990): "Four Essays in Empirical Macro and Labor Economics," PhD Thesis, Brown University.

47. Keane, M. (1994): "A Computationally Practical Simulation Estimator for Panel Data," Econometrica, 62, 95-116.

48. Kihlstrom, R. E., and M. H. Riordan (1984): "Advertising as a Signal," Journal of Political Economy, 92, 427-450.

49. Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2010): "Online Demand under Limited Consumer Search," Marketing Science, 29, 1001-1023.

50. Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2014): "The Probit Choice Model under Sequential Search with an Application to Online Retailing," Working paper.

51. Koulayev, S. (2014): "Search for Differentiated Products: Identification and Estimation," RAND Journal of Economics, 45, 553-575.

52. Manchanda, P., P.E. Rossi, and P.K. Chintagunta (2004): "Response Modeling with Non-random Marketing-Mix Variables," Journal of Marketing Research, 41, 467-478.

53. Manski, C., and S. Lerman (1981): "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambrdige, MA, 305-319.

54. Mauri, A.G. (2013): "Hotel Revenue Management: Principles and Practices," Pearson.

55. McFadden, D. (1989): "A Method of Simulating Moments for Estimation of Discrete Response Models Without Numerical Integration," Econometrica, 57, 995-1026.

56. Mehta, N., S. Rajiv, and K. Srinivasan (2003): "Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation," Marketing Science, 22, 58-84.

57. Moraga-Gonzalez, J., and M. Wildenbeest (2008): "Maximum Likelihood Estimation of Search Cost," European Economic Review, 52, 820-848.

58. Moraga-Gonzalez, J.L., Z. Sandor, and M. Wildenbeest (2015): "Consumer Search and Prices in the Automobile Market," Working paper.

59. Moraga-Gonzalez, J.L, Z. Sandor, and M. Wildenbeest (2014): "Do Higher Search Costs Make Markets Less Competitive?," Working paper.

60. Nair, H. S., S. Misra, W. J. Hornbuckle IV., R. Mishra, and A. Acharya (2014): "Big Data and Marketing Analytics in Gaming: Combining Empirical Models and Field Experimentation," Working paper.

61. Narayanan, S., and K. Kalyanam (2014): "Position Effects in Search Advertising: A Regression Discontinuity Approach," Working paper.

62. Nelson, P. (1974): "Advertising as Information," Journal of Political Economy, 82, 729-754.

63. Reinganum, J. F. (1979): "A Simple Model of Equilibrium Price Dispersion," Journal of Political Economy, 87, 851-858.

64. Rothschild, M. (1973): "Models of Market Organization with Imperfect Information: A Survey," Journal of Political Economy, 81, 1283-1308.

65. Rothschild, M. (1974): "Searching for the Lowest Price When the Distribution of Prices Is Unknown," Journal of Political Economy, 82, 689-711.

66. Rhodes, A. (2011): "Can Prominence Matter Even in an Almost Frictionless Market?," The Economic Journal, 121, F297-F308.

67. Salop, S., and E. Stiglitz (1977): "Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersion," The Review of Economic Studies, 44, 493-510.

68. Seiler, S. (2013): "The Impact of Search Costs on Consumer Behavior: A Dynamic Approach," Quantitative Marketing and Economics, 11, 155-203.

69. Stigler, G.J. (1961): "The Economics of Information," Journal of Political Economy, 69, 213-225.

70. Sorensen, A. T. (2001): "Price Dispersion and Heterogeneous Consumer Search for Retail Prescription Drugs," NBER working paper 8548.

71. Stahl, D. O. (1989): "Oligopolistic Pricing with Sequential Consumer Search," American Economic Review, 79, 700-712.

72. Train, K. (2009): "Discrete Choice Methods with Simulation," Cambridge University Press.

73. Varian, H. R. (1980): "A Model of Sales," American Economic Review, 70, 651-659.

74. Varian, H. R. (2007): "Position Auctions," International Journal of Industrial Organization, 25, 1163-1178.

75. Weitzman, M. L. (1979): "Optimal search for the best alternative," Econometrica, 47, 641-654.

76. Wilson, C. M. (2010): "Ordered Search and Equilibrium Obfuscation," International Journal of Industrial Organization, 28, 496-506.

77. Wolinsky, A. (1984): "Product Differentiation with Imperfect Information," Review of Economic Studies, 51, 53-61.

78. Wolinsky, A. (1986): "True Monopolistic Competition as a Result of Imperfect Information," Quarterly Journal of Economics, 101, 493Ð511.

79. Yang, S., and A. Ghose (2010): "Analyzing the Relationship Between Organic and Sponsored Search Advertising: Positive, Negative, or Zero Interdependence?," Marketing Science, 29, 602-623.

80. Yao, S., and C. F. Mela (2011): "A Dynamic Model of Sponsored Search Advertising," Marketing Science, 30, 447-468.

81. Zhou, J. (2011): "Ordered Search in Differentiated Markets," International Journal of Industrial Organization, 29, 253-262.

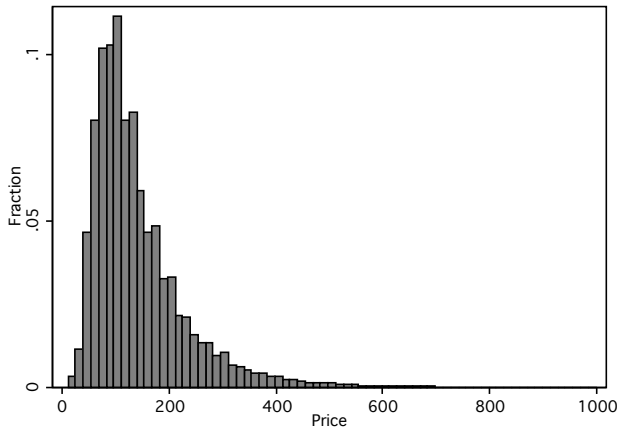
10 Appendix A: Data Cleaning

The training data set contains 9,917,530 observations. I make two necessary changes to the raw data set and use 7,986,074 observations for my analysis.

1. First, that data set contains some errors in the way price information was stored. Some hotel prices are either very high, more than $19 million per night, or very low, $0.01 per night. I used the following method to remove searches that include such outliers. In the data set I observe not only the average price displayed for a hotel, but also the total spent by the consumer (i.e. price multiplied by the number of nights and number of hotel rooms booked, plus taxes and fees). I used these two numbers to correct for outliers. More precisely, I removed searches that contain at least one observation where the total amount spent exceeds the price paid multiplied by the length of the trip and the number of rooms book plus taxes (not exceeding 30% of the price). This meant removing 1,618,626 observations.

2. Second, I choose to focus on "typical" searches and remove searches that include prices lower than $10 or higher than $1000 per night. This further reduces the data set by 312,830 observations. I focus only on these searches for two reasons. First, not having very high or very low prices helps mitigate the first problem above for searches not ending in a transaction, but which are likely to suffer from the same problems. Second, there are very few searches that include such extreme prices. The histogram in Figure 6 below shows that hotels with prices close to $1000 are very rare.

Figure 6: Histogram of Prices Displayed by Hotels



Note: Very few hotels have prices close to $1000.

11 Appendix B: Further Evidence

11.1 Further Evidence for Section 3.1: Learning to Rank Algorithm

In this subsection I describe the basic learning to rank problem, its evaluation and describe the winning algorithm, LambdaMART, for the Expedia challenge on Kaggle.[32] This algorithm has also won the 2010 Yahoo! Learning to Rank Challenge (Chapelle and Chang, 2011) and had been used by Yoganarasimhan (2014) to improve on the winning algorithm of the Yandex competition using personalization. Yoganarasimhan (2014) provides a more in depth discussion of similar concepts. I will first describe how rankings are evaluated and then present the formal learning problem.

11.1.1 Evaluation

The most commonly used evaluation method in the learning to rank literature is NDCG (Normalized Discounted Cumulative Gain). It is also used by Kaggle to evaluate the quality of the rankings proposed in the Expedia challenge, as well as in many other settings (for example the Yandex competition studied by Yoganarasimhan, 2014). I will thus also use this metric and describe it briefly here.

Denote by k the number of possible hotels to be ranked on one results page (in my application, this was 38). Define the Discounted Cumulative Gain (DCG) as the following

$$DCG_k = \sum_{p}^{k} \frac{2^{rel_p} - 1}{log_2(p+1)}$$

where rel_p gives the relevance assigned by the consumer to the hotel in position p. For example, for the Expedia competition if a consumer purchased the hotel in position p, then rel_p was assigned a value of 5, if the consumer clicked in position p then $rel_p = 1$ and $rel_p = 0$ if the consumer did not consider the hotel in position p. The NDCG is formed by dividing the DCG by the ideal DCG (IDCG) which is a ranking ordered by the revealed relevance scores of the consumer. As a result, the $NDCG \in \{0, 1\}$.

As an illustration of this metric, consider the following example. Suppose there are three hotels A, B, C and a consumer is shown these hotel in this order. Further suppose that the consumer clicks on A, purchases from B and does not consider C. In this case,

$$DCG_3 = \frac{2^1 - 1}{log_2(2)} + \frac{2^5 - 1}{log_2(3)} + 0 \approx 20.56$$

The ideal ranking would have displayed B before A, in which case

$$IDCG_3 = \frac{2^5 - 1}{log_2(2)} + \frac{2^5 - 1}{log_2(3)} + 0 \approx 32.63$$

[32] Winning algorithms by Owen Z. and Jun Wang are available at `https://www.kaggle.com/c/expedia-personalized-sort/details/winners`

It follows that the ranking proposed A, B, C achieved a score of $NDCG_3 = \frac{DCG_3}{IDCG_3} = 0.63$. As a comparison, the winning algorithm for Expedia's competition on Kaggle earned a $NDCG_{38} = 0.54$ improving on Expedia's default ranking which scored $NDCG_{38} = 0.50$.

11.1.2 Learning problem

Machine learning is a subfield of computer science that studies algorithms to learn patterns and make predictions from data. Learning to rank algorithms are an example of such algorithms with the goal of ranking documents based on relevance. In the current application, a document is a hotel and its characteristics. The system, Expedia, maintains a collections of these hotels and when a consumer makes a search query, it proceeds to rank the hotels. The ranking task can thus be summarized by a ranking model $f(x_{ih}, \delta)$ that takes the characteristics x_{ih} of hotel h in response to a query by consumer i and computes a score $\hat{s}_{ih} = f(x_{ih}, \delta)$ such that hotels with a higher score are ranked closer to the top. Note that a ranking model $f(x_{ih}, \delta)$ can be constructed even without learning by only taking into account the characteristics of the hotel, such as price and number of stars. In contrast, a learning ranking model exploits the availability of data on so called relevance scores of the users. What this means is that in many instances (including in the present application) data on consumer clicks or bookings are available. This data gives an indication of how relevant the ranking proposed was for the consumer. Learning to rank algorithms thus use consumer observed choices to construct $f(x_{ih}, \delta)$ in addition to the characteristics of the hotel.

The purpose of the model is then to learn a function that will rank most relevant hotels at the top. More concretely, this means that the goal is to find a function that will rank the hotel that will be purchased by the consumer at the top (the most relevant), followed by hotels that the consumer will click and finally by those that will not be considered by the consumer. Denote by s_{ih} the score given by consumer i to hotel h, where s_{ih} is highest if i purchases h (for example, for the competition, $s_{ih} = 5$ if h was purchased by i, $s_{ih} = 1$ if h was clicked and zero if the consumer did not consider h). The purpose of the learning to rank algorithm is then to take the data on hotel characteristics x_{ih} for a query performed by i and i's clicks/purchases s_{ih} for each h and learn a function

$$f(x_{ih}, \delta) = \hat{s}_{ih}$$

so that the ranking order of predicted scores $\hat{s}_{ih}$ are exactly equal to the ranking order of observed s_{ih}.

Consider two hotels h and j. The observed relevance score for a consumer can be interpreted as

$$\begin{cases} s_{ih} > s_{ij}, & \text{if } h \text{ is prefered to } j \\ s_{ih} = s_{ij}, & \text{if } h \text{ is equal to } j \\ s_{ih} < s_{ij}, & \text{if } j \text{ is prefered to } h \end{cases} \tag{20}$$

The researcher than proposes a model to predict how relevant a consumer will find two hotels. A ranking algorithm that is at the basis of LambdaMART is called RankNet (see Burges et al, 2005, 2010). This method models the probability that a hotel h is more relevant than j using a sigmoid function as follows

$$P = \frac{1}{1 + e^{-\sigma(\hat{s}_{ih} - \hat{s}_{ij})}}$$

where σ determines the shape of the sigmoid function. The log-likelihood of observing the data is then given by

$$LL = -\bar{P}log(P) - (1 - \bar{P})log(1 - P)$$

where $\bar{P}s$ are the actual probabilities observed in the data.

Define

$$y_{ihj} \begin{cases} 1, & \text{if } s_{ih} > s_{ij} \\ 0, & \text{if } s_{ih} = s_{ij} \\ -1, & \text{if } s_{ih} < s_{ij} \end{cases} \tag{21}$$

Using this new notation, the log-likelihood becomes

$$LL = \frac{1}{2}(1 - y_{ihj})\sigma(\hat{s}_{ih} - \hat{s}_{ij}) + log(1 + e^{-\sigma(\hat{s}_{ih} - \hat{s}_{ij})})$$

When the function f is known, then estimation proceeds using maximum likelihood to recover the parameters δ of the function. However, when the function f is not known, both it and its parameters must be recovered through estimation. LambdaMART and other learning to rank algorithms provide a solution to this problem. The solution starts by useing stochastic gradient descent algorithm to determine δ iteratively by updating from

$$\delta \to \delta + \eta \frac{\partial LL}{\partial \delta}$$

where $\frac{\partial LL}{\partial \delta} = \lambda(\frac{\partial \hat{s}_{ih}}{\partial \delta} - \frac{\partial \hat{s}_{ij}}{\partial \delta})$ and $\lambda = \sigma\left(\frac{1}{2}(1 - y_{ihj}) - \frac{1}{1+e^{-\sigma(\hat{s}_{ih} - \hat{s}_{ij})}}\right)$ and where η is the rate at which the researcher wants the algorithm to learn. These results follow from differentiation. However, Burges et al. (2006) show that modifying the expression for λ so that it is weighted by change in NDCG from changing two hotels' positions performs better. More precisely, this means defining λ as

$$\lambda = \frac{-\sigma}{1 + e^{-\sigma(\hat{s}_{ih} - \hat{s}_{ij})}} |\Delta NDCG|$$

where $|\Delta NDCG| = |NDCG(\hat{s}_i) - NDCG(\hat{s}_i^{h,j})|$ and $NDCG(\hat{s}_i^{h,j})$ is the NDCG score of $\hat{s}_i$ with the entries for h and j switched. The model is then trained using gradient boosted regression trees (MART-Multiple Additive Regression Tree). A regression tree is a method to determine

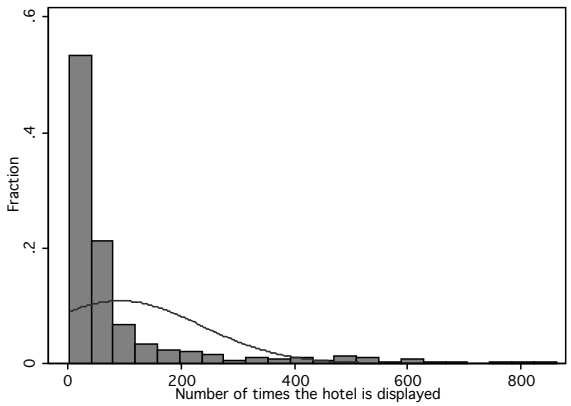
the effect of one variable on another that classifies the predictor variable into sets and tries to determine the threshold for classification based on these. In addition, gradient boosted methods perform classification based on residuals (for more details, see Breiman et al. (1984) for an introduction).

Learning to rank is a supervised learning task and thus uses training, validation and testing phases. What this means is that the researcher usually has access to three types of data sets. The training data set typically consists of search queries performed by the consumer (the details of the trip that the consumer requested) and the ranking she was displayed (the ordered list of hotels she observes in response to her query). It also includes how relevant the consumer perceived the ranking to be. This relevance is manifested in clicks and/or bookings of the hotels displayed. Hotels booked are interpreted to be the most relevant, followed by those clicked and lastly by those not clicked. In technical terms, the training data set is considered labeled. The validation data set contains the same features as the training data set, while the testing data set contains the displayed hotels, but does not reveal consumers' clicks and purchases. The researcher estimates several models on the training data set, and tests which model performs best in terms of out of sample prediction on the validation data set. The chosen model is then used to predict choices made on the test data set that does not contain relevance scores. In technical terms, the test data is not labeled.

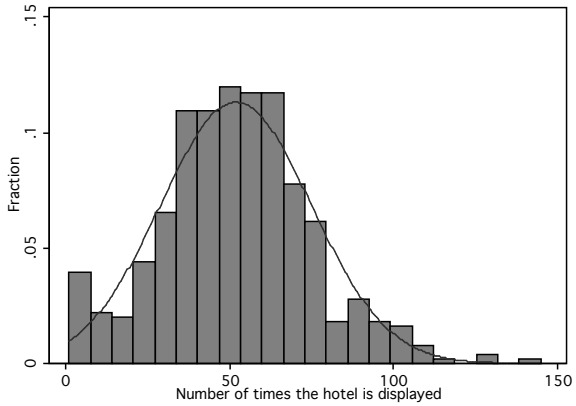
11.2 Further Evidence for Section 3.1: Two Types of Randomness

The fact that the random ranking is constructed by ignoring past performance of the hotel or its match with the consumer search query makes individual hotels have a more even probability of being displayed. This symptom can be illustrated with a simple plot. In Figure 7 I plot the histogram (with a normal density) of the number of times a hotel in destination 4562 is displayed on the first page of results under each type of ranking. I find that under Expedia's ranking a few hotels are displayed 800 times (out of approximately 1600 possible search impressions), while the median hotel is only displayed 35 times. This observation is consistent with the idea of Expedia's ranking oversampling some hotels as the ranking algorithm displays more favorable hotels more frequently. In contrast, the right panel shows the histogram of the number of times a hotel in the same destination is displayed under the random ranking. Here the distribution looks closer to a normal distribution, with the mean and median number of displays around 52 (out of 980 possible search impressions). This finding is consistent with the idea that under the random ranking, hotels have a more even probability of being displayed. The same pattern that appears in these graphs holds even if I only look at hotels in top positions instead of all displayed or if I look across different destinations.

Figure 7: Number of times a hotel is displayed by search impression type: Destination 4562



Expedia's Ranking



Random Ranking

11.3 Further Evidence for Section 4: Hotel characteristics clicked and purchased

Table 13: Hotel characteristics clicked by search impression type

	No Tran.				Tran.				No Tran.	Tran.
	Random		Expedia		Random		Expedia			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Diff.	Diff.
Price	147.81	94.81	156.13	95.52	117.00	56.76	118.77	58.58	-8.32***	-1.77**
Stars										
Less than 3	0.15	0.36	0.10	0.30	0.20	0.40	0.17	0.38	0.05***	0.03***
3	0.40	0.49	0.34	0.47	0.46	0.50	0.44	0.50	0.05***	0.02***
4	0.36	0.48	0.44	0.50	0.30	0.46	0.34	0.48	-0.08***	-0.05***
5	0.09	0.29	0.11	0.32	0.04	0.20	0.05	0.22	-0.02***	-0.01***
Review Score										
Less than 2.5	0.07	0.25	0.04	0.18	0.03	0.17	0.03	0.16	0.03***	0.01***
Between 2.5 and 3	0.09	0.29	0.07	0.26	0.10	0.30	0.08	0.28	0.02***	0.02***
Between 3.5 and 4	0.48	0.50	0.52	0.50	0.52	0.50	0.53	0.50	-0.03***	-0.01**
Between 4.5 and 5	0.35	0.48	0.38	0.49	0.35	0.48	0.36	0.48	-0.02***	-0.01
Chain	0.59	0.49	0.64	0.48	0.69	0.46	0.67	0.47	-0.05***	0.02***
Location Score	2.85	1.53	3.22	1.51	2.57	1.38	2.78	1.44	-0.37***	-0.21***
Promotion	0.24	0.43	0.36	0.48	0.24	0.43	0.31	0.46	-0.11***	-0.07***

Significance of differences obtained by means of a t-test.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 14: Hotel characteristics of purchased hotels by search impression type

	Random Ranking		Expedia's Ranking		
	Mean	SD	Mean	SD	Diff.
Price	116.55	55.50	118.07	57.42	-1.52**
Stars					
Less than 3	0.20	0.40	0.17	0.38	0.03***
3	0.46	0.50	0.44	0.50	0.02***
4	0.30	0.46	0.34	0.47	-0.04***
5	0.04	0.20	0.05	0.22	-0.01**
Review Score					
Less than 2.5	0.03	0.17	0.02	0.15	0.01***
Between 2.5 and 3	0.10	0.30	0.08	0.27	0.02***
Between 3.5 and 4	0.52	0.50	0.53	0.50	-0.02**
Between 4.5 and 5	0.35	0.48	0.36	0.48	-0.01
Chain	0.70	0.46	0.67	0.47	0.03***
Location Score	2.54	1.37	2.74	1.43	-0.20***
Promotion	0.24	0.43	0.31	0.46	-0.07***

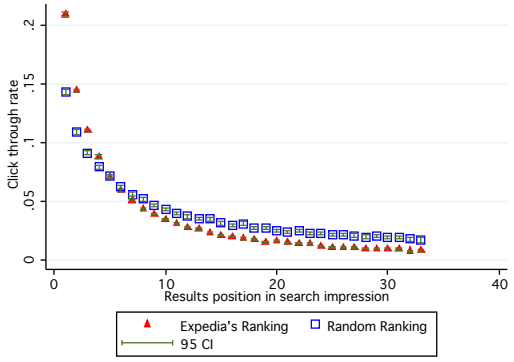
Significance of differences obtained by means of a t-test.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

11.4 Further Evidence for Section 4.1: The Effect of Rankings on Search and Choice

11.4.1 Click through rate

Figure 8: Click through rate by results position and search impression type: Search impressions without sorting



Note: Even though I do not observe whether the consumer sorted results by a criterion such as price or distance, I can check whether hotels seem to be ordered by this criterion in the data. For this figure I eliminate approximately 50,000 of the almost 8 million observations that seem to be ordered by price. From the WCAI companion data set, I find that only 34% of search impressions contain filtered results. Out of those search impressions that are filtered, most filtering happens by distance (60%) and by price (34%).

Figure 9: Click through rate by results position and search impression type: Search impressions 30 (median) or longer

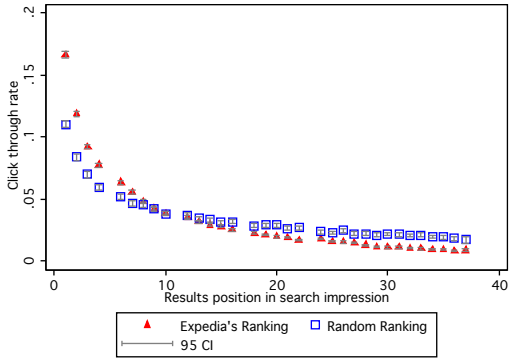


Figure 10: Click through rate by results position and search impression type: Search impressions ending in a transaction

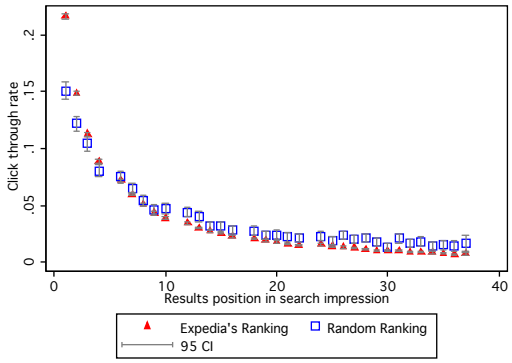


Figure 11: Click through rate by results position and search impression type: Search impressions not ending in a transaction

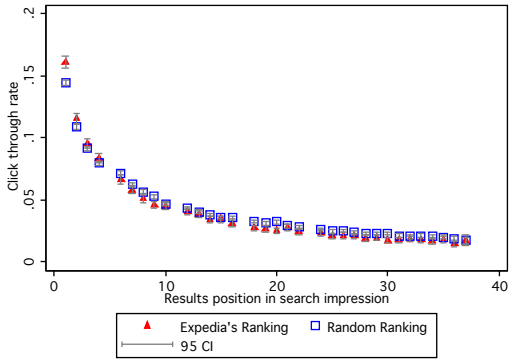
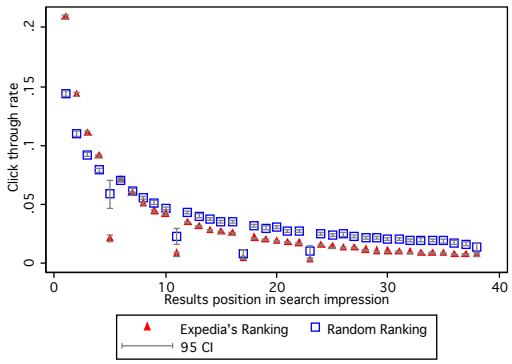
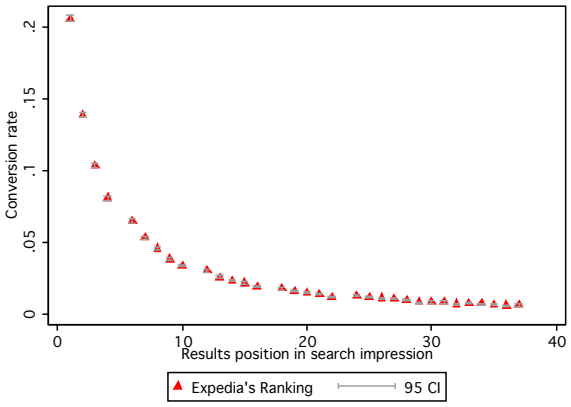


Figure 12: Click through rate by results position and search impression type: Search impressions with and without positions reserved for opaque offers

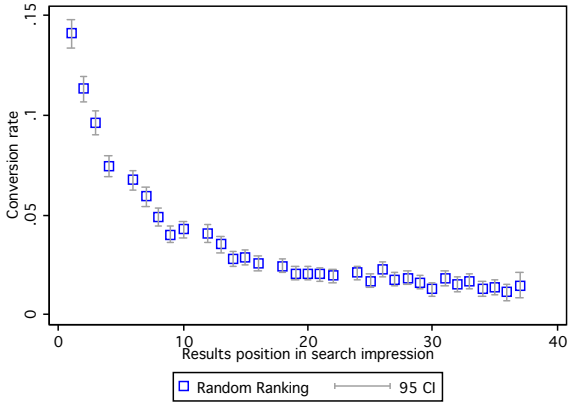


11.4.2 Conversion Rate

Figure 13: Conversion rate unconditional on click by results position and search impression type



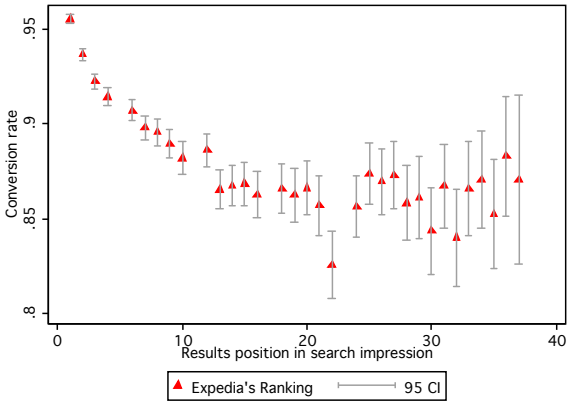
Expedia's Ranking



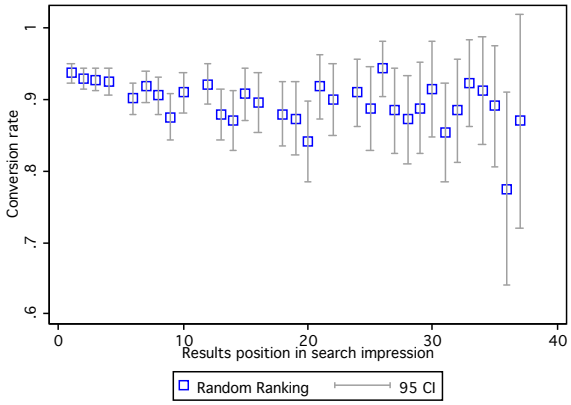
Random Ranking

Note: Restrict attention to search impressions ending in a transaction.

Figure 14: Conversion rate by results position and search impression type: Search impressions without sorting

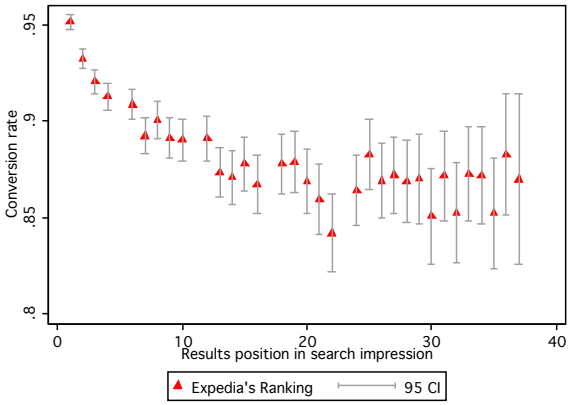


Expedia Ranking

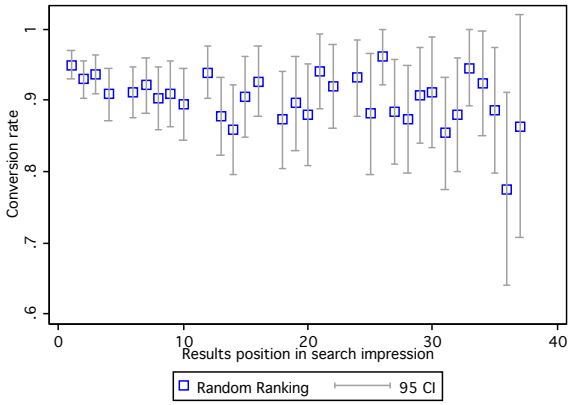


Random Ranking

Figure 15: Conversion rate by results position and search impression type: Search impressions without sorting that are longer than 30 (median)



Expedia Ranking

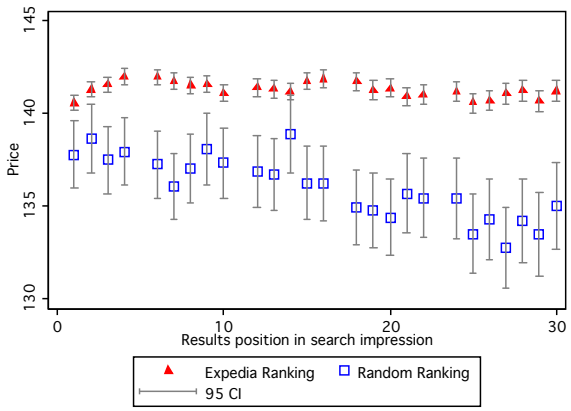


Random Ranking

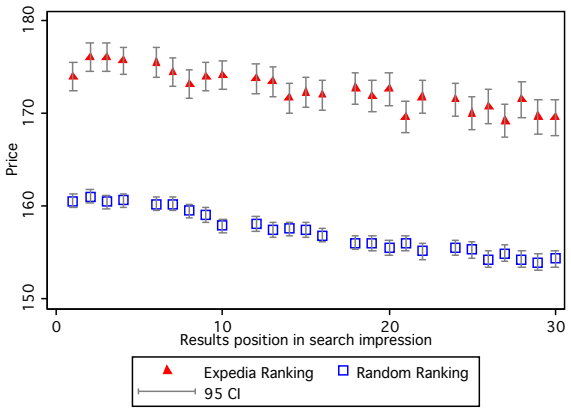
11.4.3 Characteristics Displayed by Position

In this subsection I plot the average characteristics of the hotels displayed by position under each type of ranking. I consider the average price, the number of stars and the review score of the hotels displayed.[33] I restrict attention to the first 30 positions in the ranking, because of the high volatility in the number of observations for position higher than that. This allows me to plot approximately 90% of the data set in these figures.

Figure 16: Characteristics Displayed by Position: Price

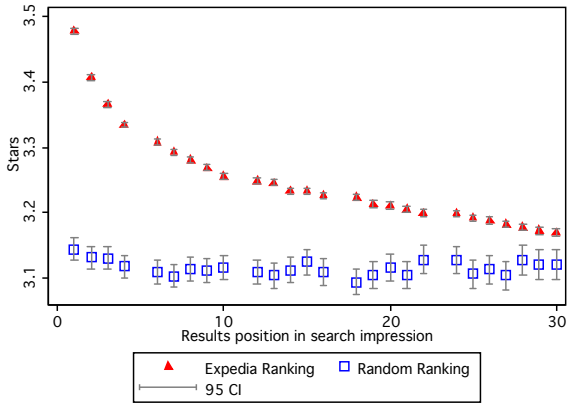


Ending in transaction

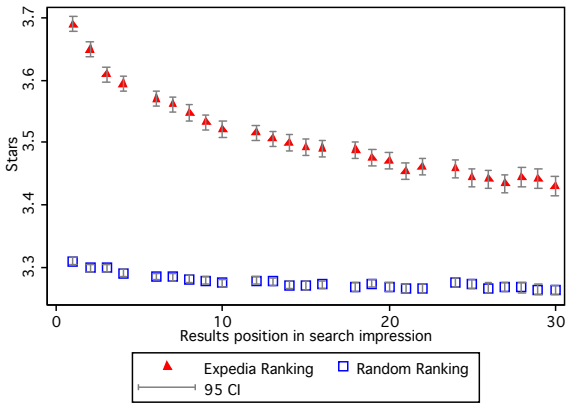


Not ending in transaction

Figure 17: Characteristics Displayed by Position: Stars



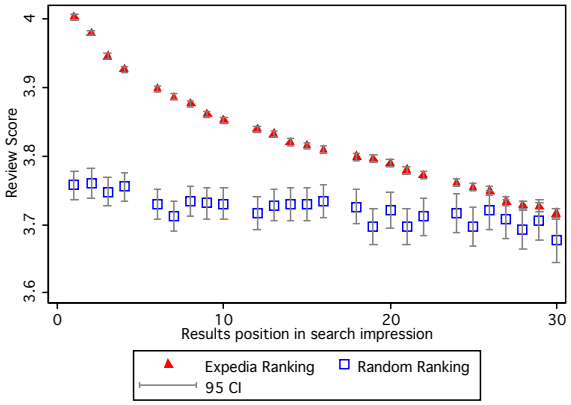
Ending in transaction



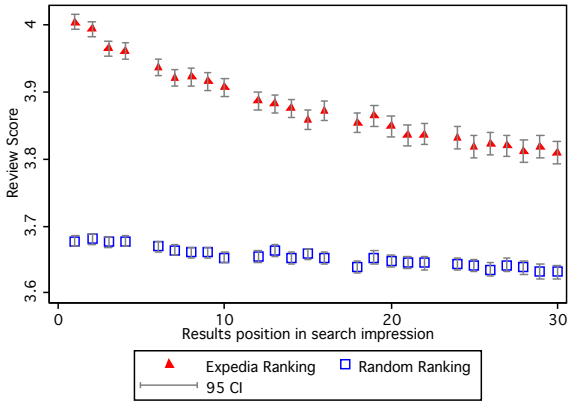
Not ending in transaction

[33]The figures for the other characteristics (e.g. fraction of chains displayed) are available upon request.

Figure 18: Characteristics Displayed by Position: Review Score



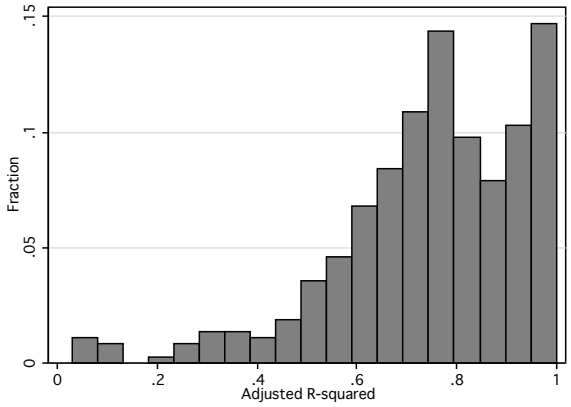
Ending in transaction



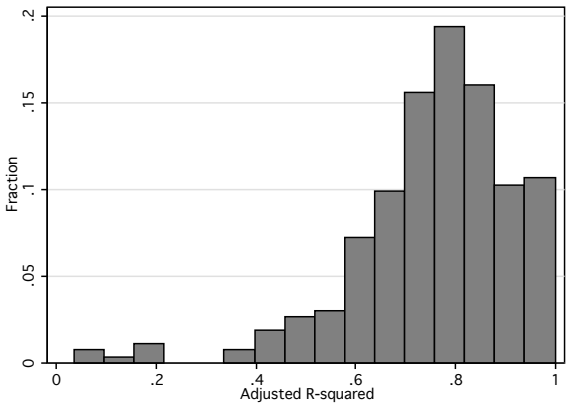
Not ending in transaction

11.5 Further Evidence for Section 5: Price Endogeneity Concerns

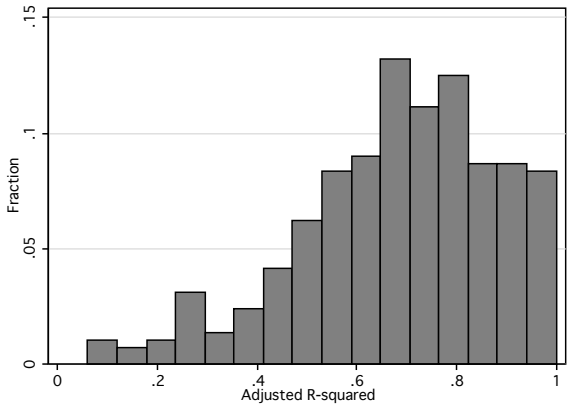
Figure 19: Histogram of the adjusted R^2 from running separate regressions of price on observable characteristics for each hotel in a destination



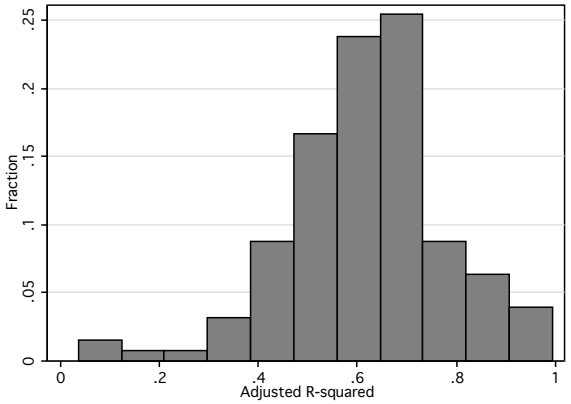
Destination 4562



Destination 9402

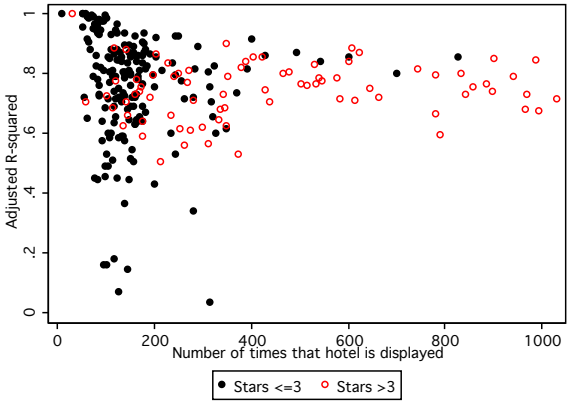


Destination 8347

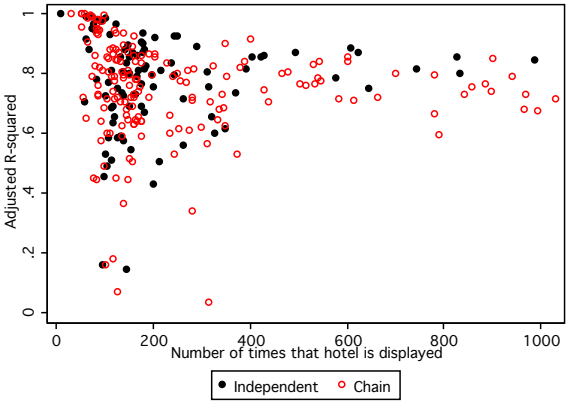


Destination 13870

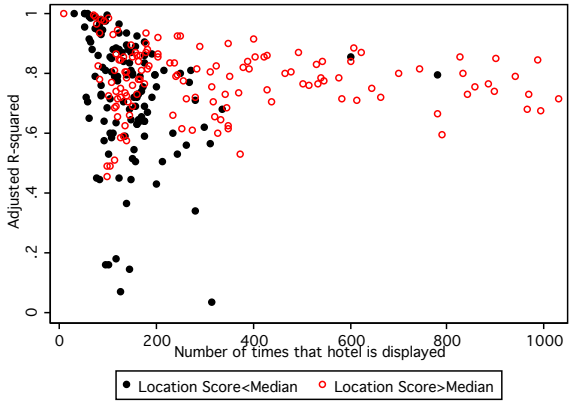
Figure 20: Graph of the adjusted R^2 from running separate regressions of price on observable characteristics for each hotel in Destination 9402



Hotel Stars



Chain



Hotel Location Score

11.6 Further Evidence for Sections 3.1 and 6.2: Summary Statistics for the Four Destinations used in Estimation

For the structural estimation, I choose to focus on four of the largest destinations in my data set: 4562, 9402, 8347, and 13870.[34] This allows me to control for differences across destinations and not confound results. All four of these destinations are in the largest country, 219, which I have shown earlier is likely the U.S. I summarize their characteristics in Tables 15, 16 and 17. These four destinations have a total of 250,825 observations with 27% of those from the random ranking. This is comparable to the 33% for the full data set (only destination 13870 has fewer observations with the random ranking). There are more than eight thousand search impressions and more than 1,600 hotels, the majority of which appear in both rankings.[35] Each search impression has at least one click and there are 4,752 transactions.

Table 15: Summary statistics by destination

	Number				
Destinations	4562	9402	8347	13870	Total
Observations	76,920	65,981	54,747	53,177	250,825
Observations with Random Ranking	26,446	19,551	15,202	6,758	67,957
Search Impressions	2,652	2,170	1,892	1,715	8,429
Hotels	637	329	450	223	1,639
Hotels Displayed in Both Rankings	502	305	433	140	1,380
Clicks	3,058	2,446	2,179	1,900	9,583
Transactions	1,331	1,389	950	1,082	4,752

Table 16 reveals that these four destinations have similar search query characteristics as the full sample. The number of hotels displayed is approximately 30, which is the median of the full sample. Search queries are for trips that are on average one day longer than in the full sample, and are searched approximately 10 days earlier than in the full sample. All other characteristics align closely with the ones in the full sample.

Table 17 is the equivalent of Table 2 for these four destinations. I again split the data into search impressions ending in a transaction and those that do not. I find that Expedia's ranking displays more expensive hotels of higher quality, regardless of whether the search impression ends in a transaction. As in the full data set, search impressions ending in a transaction are generally those that are cheaper.

[34] Destination 8192 has the largest number of observations (121,522), but has few observations with the random ranking, so I choose to focus on the next four largest observations that have a fraction of random rankings that is closer to the one in the full data set.

[35] Recall that I only observe the first page of results, so even though a particular destination contains a fixed number of hotels and both rankings display the same hotels, they need not list the same hotels on the first page, which is why I do not expect that all hotels will appear on the first page of results under both rankings.

Table 16: Summary statistics: Search impressions by destination

	Mean			
Destination	4562	9402	8347	13870
Number of Hotels Displayed	29.00	30.41	28.94	31.01
Trip Length (days)	3.24	2.58	3.75	2.89
Booking Window (days)	48.76	42.69	45.07	47.12
Saturday Night (percent)	0.44	0.48	0.38	0.41
Adults	1.88	1.96	2.21	2.21
Children	0.30	0.36	0.92	1.18
Rooms	1.12	1.11	1.15	1.10
Chain (percent)	0.62	0.69	0.74	0.74
Promotion (percent)	0.36	0.29	0.42	0.24
Random Ranking (percent)	0.37	0.31	0.30	0.14
Total Clicks	1.15	1.13	1.15	1.11
Two or More Clicks (percent)	0.08	0.08	0.07	0.06
Total Transactions	0.50	0.64	0.50	0.63
Observations	2,652	2,170	1,892	1,715

Table 17: Hotel characteristics displayed by search impression type, conversion and destination

	Random Ranking				Expedia Ranking			
	No transaction		Transaction		No transaction		Transaction	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Destination 4562								
Price	241.99	141.49	215.89	114.77	261.44	116.09	237.20	105.41
Stars	3.37	0.91	3.33	0.90	3.76	0.74	3.72	0.77
Review Score	3.84	0.95	3.80	0.97	4.05	0.75	4.02	0.80
Chain	0.70	0.46	0.71	0.45	0.59	0.49	0.55	0.50
Location Score	3.80	1.83	3.67	1.93	4.90	1.40	5.02	1.34
Promotion	0.25	0.43	0.22	0.42	0.42	0.49	0.44	0.50
Destination 9402								
Price	193.18	109.78	175.83	104.63	205.20	101.65	187.64	92.89
Stars	3.19	0.91	3.19	0.90	3.53	0.83	3.49	0.86
Review Score	3.92	0.71	3.92	0.69	4.11	0.56	4.07	0.59
Chain	0.73	0.44	0.73	0.44	0.65	0.48	0.65	0.48
Location Score	3.40	1.58	3.47	1.57	4.41	1.25	4.39	1.25
Promotion	0.18	0.38	0.23	0.42	0.36	0.48	0.36	0.48
Destination 8347								
Price	115.49	82.38	107.07	80.95	162.47	118.79	128.39	80.81
Stars	3.02	0.71	2.97	0.68	3.47	0.77	3.39	0.73
Review Score	3.79	0.84	3.74	0.78	4.06	0.66	3.97	0.63
Chain	0.78	0.41	0.77	0.42	0.76	0.43	0.68	0.47
Location Score	2.83	0.96	2.88	0.99	2.68	0.93	2.86	0.87
Promotion	0.31	0.46	0.32	0.47	0.36	0.48	0.53	0.50
Destination 13870								
Price	118.34	63.43	116.43	58.97	132.02	70.55	133.79	70.14
Stars	2.68	0.71	2.66	0.70	2.97	0.75	3.07	0.75
Review Score	3.62	0.70	3.59	0.72	3.82	0.63	3.89	0.55
Chain	0.76	0.43	0.76	0.43	0.75	0.44	0.72	0.45
Location Score	3.00	1.37	2.77	1.40	3.29	1.26	3.63	1.03
Promotion	0.16	0.36	0.18	0.38	0.21	0.41	0.27	0.44

11.7 Further Evidence for Section 4.1: Effect of position on consumer clicks and purchases

Table 18: Differential effect of position on clicks and purchases

	Click Random	Click Expedia	Transaction Random	Transaction Expedia
Position	-0.0018***	-0.0025***	-0.0002	-0.0043***
	(0.0000)	(0.0000)	(0.0001)	(0.0002)
Price	-0.0001***	-0.0002***	-0.0002***	-0.0010***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Stars	0.0157***	0.0074***	0.0022	0.0091**
	(0.0004)	(0.0003)	(0.0024)	(0.0029)
Review Score	0.0008**	0.0013***	0.0054**	0.0164***
	(0.0003)	(0.0003)	(0.0020)	(0.0032)
Chain	0.0014*	0.0032***	0.0056	0.0069
	(0.0006)	(0.0004)	(0.0035)	(0.0039)
Location Score	0.0051***	0.0028***	0.0086***	0.0280***
	(0.0002)	(0.0002)	(0.0013)	(0.0019)
Promotion	0.0102***	0.0039***	0.0155***	0.0179***
	(0.0006)	(0.0003)	(0.0034)	(0.0036)
Destination fixed effects	Yes	Yes	Yes	Yes
Observations	683,799	1,646,147	28,653	63,832
Adjusted R^2	0.014	0.026	0.031	0.101

Standard errors in parentheses

Note: Linear probability model of clicks/transactions as a function of hotel characteristics and position. The last two columns report regression coefficients for purchases conditional on a click. I restricted attention to destinations with at least 10,000 observations. This allows me to include destination fixed effects in the regression.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

11.8 Further Evidence for Section 5: Evidence that the position of the hotel is a good proxy for the order in which consumers click

In this subsection I show that the position of a hotel is a strong predictor of the order in which consumers click. To do so I use the companion data set from WCAI that contains information on consumers' click order in the form of time stamps associated to each click. I then ask what fraction of searches with at least two clicks had a click ordered that matched the position of the hotels clicked. I also compare this fraction with the fraction ordered by price. Table 19 shows my results. I find that in 35% of all searches with at least two clicks and in the majority of searches (65%) with exactly two clicks the position of the hotel exactly matches the click order of the consumer. In contrast, the price of the hotels clicked only matches the order of 20% of the clicks. This finding allows me to model consumers' click order even in the absence of information on the order of clicks.

Table 19: Percentage of clicks ordered by price, position or either one: Evidence from Manhattan (WCAI)

	Percentage		
	Price	Position	Price or Position
Searches with at least two clicks	20	35	40
Searches with exactly two clicks	49	65	77