Experiments as Instruments: Heterogeneous Position Effects in Sponsored Search Auctions*

Matthew Goldman† UC San Diego

Justin M. Rao Microsoft Research

October 29, 2014

Abstract

Google and Bing employ generalized second price (GSP) auctions to allocate billions of dollars of sponsored search advertising. The theoretical work establishing the VCG equivalence of the GSP hinges on strong, and largely untested, assumptions about the causal impact of ad position on user click probability and the value of resulting clicks. We re-purpose internal business experimentation to test these assumptions using a broad cross-section of advertisers. We find a minimal, homogenous impact of position on average click quality, providing a valuable assurance in that domain. In contrast, we find substantial heterogeneity of the impact on click through rates. For brand queries, off-brand competitors show much steeper click curves than the advertiser matching the query. For generic queries, higher quality and less popular websites benefit more from position. These findings strongly reject the conventional multiplicatively-separable model and the associated VCG equivalence, raising serious question about the efficiency properties of the GSP in sponsored search.

*We would like to thank the Microsoft Corporation for allowing us to publish this work and Jim Andreoni, Susan Athey, Dean Foster, Sebastién Lahaie, R. Preston McAfee, Andrey Simonov and Vasilis Syrgkanis for helpful discussion and comments.

†Much of this work was done while Goldman was an intern at Microsoft Research.

1 Introduction

Sponsored search links—the paid advertisements on a search engine results page—generate over forty billion dollars annually for search engines. Google and Bing/Yahoo![1] use a generalized second price auction (GSP) to allocate and price ad slots using a "pay per click" bidding model. Since payment is contingent on a user clicking, ranking is not based simply on the contingent bids, but also an advertisement's "clickability" and relevance to the query. These parameters directly impact efficiency and must be estimated by the platform. Despite these substantial departures from classical auctions, the seminal works of Edelman et al. (2007) and Varian (2007) show that under certain assumptions the GSP is payoff equivalent to the dominant strategy Vickrey-Clarke-Groves (VCG) mechanism. It is hard to understate the importance of this result in both academia, where the papers have been extensively cited, and industry, where the authors hold, or held, prominent positions at search engines.[2] By linking payoffs in the potentially problematic GSP, which evolved out of a simple rank-by-bid first price auction designed in the late 90's, to the VCG, a virtual gold standard in mechanism design, the authors provided valuable reassurance about a mechanism that was in place largely due to historical accident.

Yet this result did not come without some fairly strong assumptions such as complete information, which is challenged in Athey and Nekipelov (2010) and the "envy-free" equilibrium refinement, which is reexamined in Gomes and Sweeney (2012) and Lucier et al. (2012). Proponents of the GSP have countered that these assumptions usefully approximate long-run play and the VCG-equivalent equilibrium has remained popular. In this paper, we empirically assess another assumption required for these results to hold, namely the *separability of the click curve*: the impact of moving up the page (towards the top slot) is assumed to scale click-through-rate (CTR) by the same multiplicative factor for *all advertisers*.[3] In other words, the search engine estimates the differential value of each slot via a single, multiplicative *click curve*. This curve is used to convert observed CTRs into "baseline clickability," which are, in turn, used to predict the CTR of ad-position pairs in any proposed allocation. It is clear that the accuracy of the estimated click curve is essential for the GSP to achieve an efficient allocation. If a single click curve does not apply to all advertisers: 1) the notion of a scalar-valued "baseline clickability" loses meaning and efficient allocation of multiple ad positions by a GSP-style mechanism is impossible, 2) for any chosen definition, "baseline clickability" will be miss-estimated, and 3) there are no known equilibrium properties of the GSP and certainly no VCG equivalence.

However, estimating *advertiser-specific* position effects is difficult because position is an en-

[1] Microsoft's AdCenter operates the auction for both Yahoo! and Bing.

[2] Varian is the chief economist at Google. Schwarz was a principle economist at Yahoo at the time of writing and is now at Google.

[3] This assumption is frequently maintained in subsequent work and used in published auction guidance by the major search engines. The only paper we are aware of that explicitly removes the assumption is a theory paper, Hummel and McAfee (2014), examining two click curves from an algorithmic perspective.

dogenous outcome of the auction. Ads are ranked higher when they are expected to get more clicks by the search engine (because of pay-per-click pricing) or advertiser (due to events like sales or geographic targeting). Our approach uses experiments conducted as a part of normal business practice as instrumental variables. These experiments randomize users into conditions to test new algorithms, such as those applying to click prediction or relevancy score estimation. Since the experiments were not designed to induce variation in ranking *per se*, we have to address a variant of the weak instruments problem. Intuitively, an experiment is only "turned on" when it induces a different ranking of ads from the "control" for a given query. This shuffling only occurs when an advertiser's bid and clickability metric drift into a region where the employed ranking algorithms "disagree." A variety of regularization approaches exist to select a set of instruments based on first stage fit, but most are designed for a large, fixed class of candidate instruments (Belloni et al., 2012; Okui, 2011). In our case, the strength of each instrument is moderated by a continuous covariate, time, so a natural solution is the regularization approach presented in Carrasco (2012), which non-parametrically constructs optimal instruments that apply more weight to time periods when an experimental condition has higher relevance. The approach is computationally efficient and allows us to estimate position effect curves for many of the top 20,000 (by revenue) advertiser-query pairs on Bing. In contrast, nearly all empirical work on sponsored search uses a handful of queries or works with one focal advertiser.

Our results are unequivocal: there is substantial heterogeneity in position effects across advertisers. The heterogeneity is not only highly statistically significant, but the magnitudes of the differences are large. The impact of position varies by more than 100% across three key mainline types we identify: 1) on-brand ads appearing on brand queries (ex. a Samsung ad on "Samsung smart phone") 2) off-brand ads on brand queries (ex. a Nokia ad on "Samsung smart phone") 3) ads on non-brand, product queries (ex. any ad on "smart phone"). Within each segment, position effects vary as well. For product queries, high quality (based on independently gathered user engagement metrics) and less well-known websites benefit more from a higher position. In other words, quality is a compliment to position whereas popularity is a substitute. For brand queries, the features of the on-brand advertiser are not systematically related to position effects. For the off-brand advertisers, less-well known firms benefit more from position, but the magnitude is smaller than for product queries.

We also investigate the *homogeneity of click value* assumption that is maintained in the foundational papers and subsequent work, namely: the value of a click is assumed not to depend on an ad's position. Three recent papers examining the impact of position on the probability of a sale ("conversion rate"), all focusing on a single advertiser, reach different conclusions. One finds a positive impact of moving up the page (Ghose and Yang, 2009), one negative (Agarwal et al., 2011) and one zero (Narayanan and Kalyanam, 2014). Further complicating matters, theories of search and the psychology of choice can produce arguments for all three patterns (Lana, 1963; Brunel

and Nelson, 2003; Joachims et al., 2005; Kempe and Mahdian, 2008) as does work on equilibrium allocation under different models of consumer search (Jerath et al., 2011; Athey and Ellison, 2011). Using the same methodology as described above, we estimate a statistically significant, positive impact of moving up the page on conversion rate in aggregate, but the size of the effect is less than two percent of the baseline mean. This may be interpreted as a (less than) 2% change to the value of a click. Since our standard errors are smaller than one percent and we do not find significant heterogeneity we conclude that the homogeneous click value assumption is not problematic for GSP efficiency. Moreover, the scale of our analysis allows us to resolve an outstanding inconsistency in the literature.

So while the homogeneity of click value assumption holds approximately, the separability of the click curve fails decisively. This calls into serious question the efficiency of one of the most celebrated economic mechanisms of recent times. Billions of dollars are likely being lost every year as compared to VCG, which is perfectly capable of incorporating heterogeneous position effects. Making such a serious claim requires the scale and precision our study was designed specifically to deliver.

Stepping back a bit, our analysis is relevant to a broader class of contingent-bid mechanisms used online. Content providers like Facebook, Twitter, and LinkedIn and search platforms like Kayak and Yelp all sell advertising on a per-action basis (usually the action is a click). The popularity of this bidding format over a per-impression framework where each advertiser must determine how impressions map to actions implies that platforms have a better understanding of this mapping than advertisers would in a decentralized setting. This could be because the platform can use data from all advertisers to inform estimates for any given advertiser or due to simple economies of scale in estimation (or both). Both factors improve the value proposition for advertisers compared traditional media, such as television, where action-contingent payments are not currently possible. But the efficiency of the "centralized solution" relies on the platform choosing the right mechanism and accurately estimating the relevant quantities. And while the GSP remains a popular choice, Facebook famously opted for VCG indicating the situation is certainly fluid. New providers face an important mechanism design choice and should carefully consider empirical evidence, such as estimates contained in this paper, in making this choice.[4]

Shifting to a methodological perspective, we show that OLS, even with advertiser-specific controls for time-related confounds, is biased. Both of our IV methodologies estimate click curves that are significantly flatter on average, as would be expected based on the endogeneity of position. Additionally, Dummy TSLS produces standard errors about three-fold larger than our preferred Nonparametric IV estimator, and the first stage regression was not viable most of the time. The regularization approach was thus crucial for our ability to make useful generalizations. More broadly,

[4] Amazon Marketplace and Ebay currently take a fixed cut of sales, which simplifies their ranking problem. If they moved to a model in which sellers could place bids in terms of the cut the provider gets, then they would have to put some auction mechanism for slots in place.

"experiments as instruments" has many potential applications since active business experimentation is becoming increasingly common.[5] A company that experiments can easily tally thousands of randomized control trials per year. We show how this pool can be usefully re-purposed for causal inference on applications where direct experimentation is infeasible, or as in this case, financially prohibitive. Since these experiments only impact ad ranking for some values of time, we relied on a recent advance to this problem that was ideally suited to regularizing over this continuous underlying dimension moderating instrument strength. Related approaches, such as Belloni et al. (2012) and Okui (2011), would be natural to apply in discrete settings.

The remainder of this paper proceeds as follows: in Section 2 we explain the rules of sponsored search auction in detail, in Section 3 we describe the data and platform experimentation, in Section 4 we discuss our methods of estimation, results are presented in Sections 5 and 6, and a discussion and conclusion follows in Section 7.

2 The Generalized Second Price Auction

A sample search engine results page for a popular commercial query is displayed in Figure 1. It is divided into three key areas: "north" or "mainline" sponsored listings (ads), algorithmic results (the information retrieved for the query by the search engine) and "east" or "sidebar" ads which are less prominent and displayed further down the page. Mainline ads, if shown, are located directly above the algorithmic results and constitute the vast majority of revenue for the search engine. During the time period of study, the major U.S. search engines display 4 mainline ads (and never more than 4) on most popular commercial queries.

These four ad slots can be thought of as consecutively diminishing prizes. In this spirit, they may be referred to as (in descending order) ML1, ML2, ML3, and ML4. These are allocated according to a generalized second price auction (GSP) as follows. For a given query, each ad i is assigned a "rank score" (s_i) according to

$$s_i \equiv S(b_i, q_i; \alpha), \tag{1}$$

where b_i is bid, q_i is an index of ad quality – generally taken to be an advertiser's "baseline clickability" as estimated by the search engine and discussed in the previous section – and α gives ranking parameters set by the search engine[6]. A common formulation is given by $s_i = b_i \cdot q_i^{\alpha}$. Here, if α is set to 0, the auction is rank-by-bid. If α is set to 1, the ranking is by value per search. The

[5] Firms use field experiments to measure advertising effectiveness (Lewis and Reiley, 2013), improve ranking and recommendation (Radlinski et al., 2008; Craswell et al., 2008) and optimize user interfaces Kohavi et al. (2007, 2012). It has evolved beyond simply a measurement tool to become an integral component of market design (Ostrovsky and Schwarz, 2011; Manzi, 2012).

[6] In practice, the rank score is also effected by additional indices penalizing irrelevant ads, but these are left out for clarity.

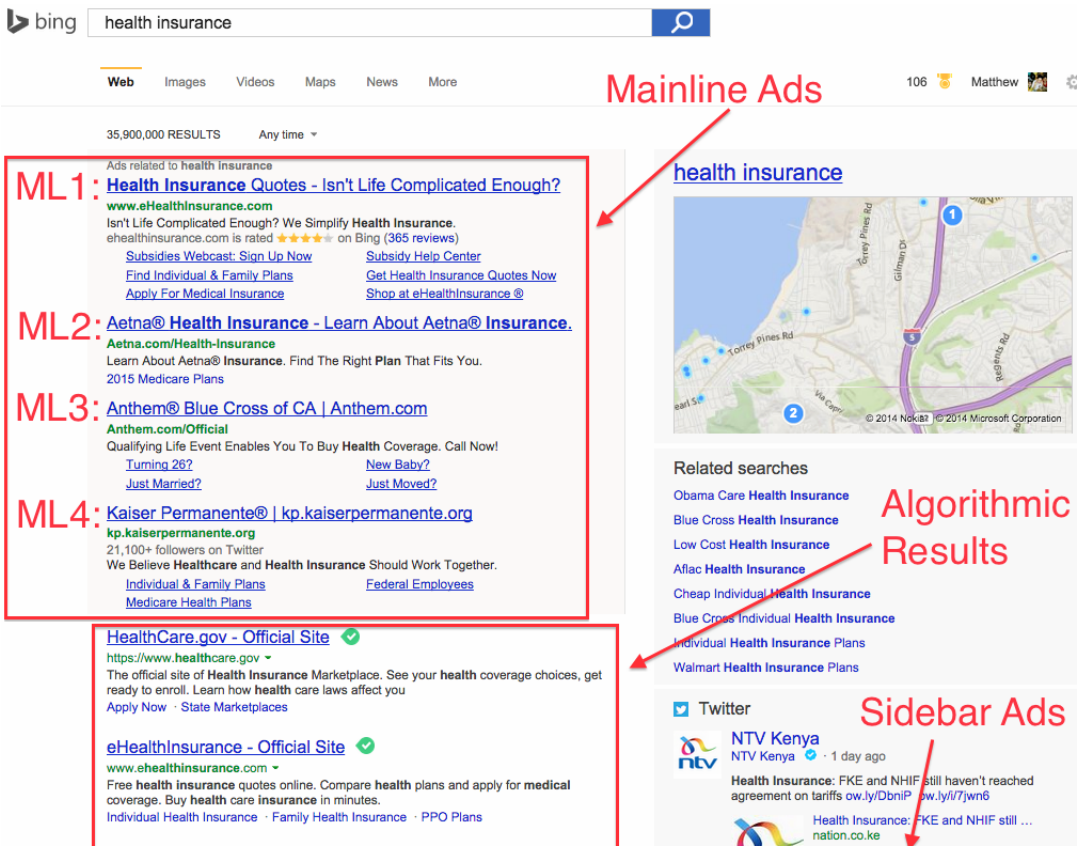


Figure 1: An example results page for a popular commercial query. The top 4 listings are mainline advertisements. Below these are the unpaid (algorithmic) results generated by the Bing search engine. Sidebar ads are much smaller and displayed further down the right most column of the page.

revenue optimal α will depend on the joint distribution of q and b, but usually lies somewhere in between these extremes (Lahaie, 2006; Athey and Nekipelov, 2010).

Ads are ranked by s_i and the top four[7] are shown. The advertiser in slot k is assigned the smallest per click payment that maintains her superior rank score: $s_k > s_{k+1}$. That is, her cost-per-click (CPC) is given by

$$c_k = S^{-1}(s_{k+1}; b_k, \alpha)$$
$$c_4 = S^{-1}(r^*; b_k, \alpha),$$

where S^{-1} gives the bid necessary for an advertiser of given quality to reach a certain rank score and r^* is a reserve value set to anchor the auction and insure that the advertiser in the final mainline

[7]More generally, the number of ads may range from 0-4 depending on reserve prices, auctioneer policies, and the number of qualifying bidders.

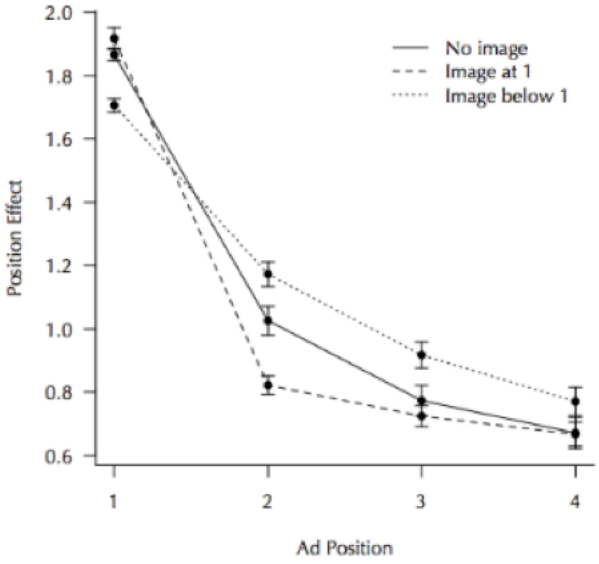


Figure 2: A multiplicatively seperable click curve estimated in Metrikov et al. (2014).

slot has a non-zero payment. Letting $p_{i,k}$ give the probability of click for advertiser i in slot k, expected revenue per impression is

$$\sum_{k=1}^{4} c_k * p_{i,k}. \tag{2}$$

Typically $p_{i,k}$ is assumed to be a function only of the quality index (q_i) and a multiplicatively separable slot effect (μ_k). That is, for any advertiser i,

$$p_{i,k} = q_i \cdot \mu_k$$

and the values of μ are said to give the shape of a global click curve, an example of which is shown in Figure 2.

This formulation is maintained both in the theoretical literature and is part of standard industrial practice for estimating q_i from observed data. It is referred to herein as the assumed *separability of the click curve*. The theoretical literature also typically assumes bidders have a linear preference for clicks given by v_i. Notable here is that the value of a click is not allowed to depend on the position k in which the click occurs. We refer to this assumption as the *homogeneity of click value*. Both the *separability of the click curve* and the *homogeneity of click value* are core theoretical properties of these auctions that underlie their VCG equivalence (Edelman et al., 2007; Varian, 2007).

3 Data

Our data covers the outcomes of billions of commercial searches on `bing.com` by English-speaking Americans from April through July of 2013. From this, we draw the 20,000 unique mainline combinations from the top of the revenue distribution for that period.[8] For each search, we have the participating bidders, quality and relevance parameters attached to each bidder, the ranking and pricing parameters chosen by the auctioneer, serving logs, click logs and conversion tracking (for a subset of advertisers). Importantly, we do not have access to any geographic or individual-level information about where the query was served. These latent variables may impact both ad position and click probability and present a potentially problematic source of endogeneity. However, each user is randomly assigned to an experimental traffic cell in which the method of ad ranking and display may be altered. As discussed in Section 3.2, many of these experiments will meet the criteria of being valid instrumental variables.

Each ad that is displayed on a results page is referred to as an *impression*. An *ad click* occurs when a user selects one of these advertisements (as opposed to one of the unpaid search results) and an ad's *click-through-rate* (CTR) refers to its probability of getting a click per impression. After a consumer makes a click, *dwell time* measures the amount of time she spends on the resulting page before hitting the back button. If the consumer never returns to the results page, then dwell time is not recorded and will be said to be "infinite." Additionally, a subset of advertisers choose to report *conversions* which indicate whether or not consumer's met some advertiser-defined standard of participation at the advertiser's website. The definition of a conversion and the resulting *conversion rate* will vary based on this definition, so our main analysis will use a normalized version of this metric. Pooled averages of these metrics for ads in each position are displayed in Table 1.[9]

Table 1: Average user behavior by ad position

	Mainline 1	Mainline 2	Mainline 3	Mainline 4	Sidebar
Ads shown per Impression	1.00	1.00	1.00	1.00	3.89
Click Through Rate	0.321	0.087	0.054	0.039	0.014
Conversion Rate	0.033	0.039	0.039	0.040	0.051
Average Dwell Time† *(seconds)*	357.72	252.31	237.39	225.30	243.98
Probability of Infinite Dwell Time	0.387	0.285	0.263	0.251	0.271
Avg. Share of Revenue (%)	56.257	21.617	12.039	7.888	2.199

† Dwell times are averaged conditional on being reported.

[8] To ensure good coverage, we classified each ad by its primary position (ML1, ML2, ML3, ML4, or sidebar) and selected the 4,000 highest revenue ads from each categorization.

[9] Standard errors are omitted from Table 1 because they are too small to represent with appropriate significant digits. Table 1 obscures substantial heterogeneity in CTRs at the advertiser level. This is demonstrated in the dispersion of the histograms presented in Figures 11 and 12 in the appendix.

Ads displayed in the first mainline spot are clicked on much more often and tend to have much more engaged clicks (higher dwell time). Conversion rate is not higher in the top slot, but dwell time is, indicating that differing conversion standards are likely playing a role here. The first mainline slot generates a majority of revenue in our sample. These advertisers are very popular and get roughly two-thirds of all ad clicks. However, given the pricing rule used in the GSP, a higher clickability not only improves ranking but also reduces cost-per-click. The net effect, which may seem surprising at first, is that clicks in the first position generate the least revenue per click for the queries in our sample.

3.1 Measures of User Engagement

We use two measures as proxies for the value of a click to an advertiser: conversions and dwell time. Conversions are only tracked for a subset of advertisers and provide binary information about user engagement. For some advertisers, a conversion may indicate only an expensive purchase, whereas for others it may indicate merely that a user completed a free registration or signed up for an email list. As such, baseline conversion rates can vary dramatically across advertisers. However, for a given advertiser, changes in conversion rate are meaningful and we will interpret a $x\%$ increase in conversion rate as representing a corresponding x% increase in average click value.

Dwell times are available for all advertisers and are a continuous measure of the time between a user's initial click on a sponsored link and her eventual return to the results page. If a user never returns, the dwell time is labeled as "infinite." Intuitively, long and infinite dwell times are desirable as they indicate better matches and more engaged users, but they are not directly interpretable. We thus calibrate them using conversion data by estimating the relationship between conversions and dwell time according to

$$P(\text{Conversion}_{i,t} = 1 | \text{Dwell}_{i,t} = d) = \alpha_i \cdot \phi(d),$$

for a click that has occurred on ad i on impression t. α_i is an ad-specific effect allowing for heterogeneity in baseline conversion rates and the nonparametric function ϕ gives a multiplicative shock to conversion probability based on an observed dwell time. This model lacks some generality, but it allows the ϕ function to take the interpretation of a percentage change in conversion probability and thus click value. The estimated value of ϕ is presented by the thick red line in the right panel of Figure 3 alongside a histogram of the distribution of pooled dwell times (left panel) for scale.

We define *mapped dwell time* as the additional regressor, $md \equiv \phi(\text{d})$. This can be thought of as a normalized conversion probability, that is a $x\%$ increase in the average value of md as representing an $x\%$ increase in any given advertiser's conversion probability which we, in turn, interpret as an $x\%$ increase in click value.

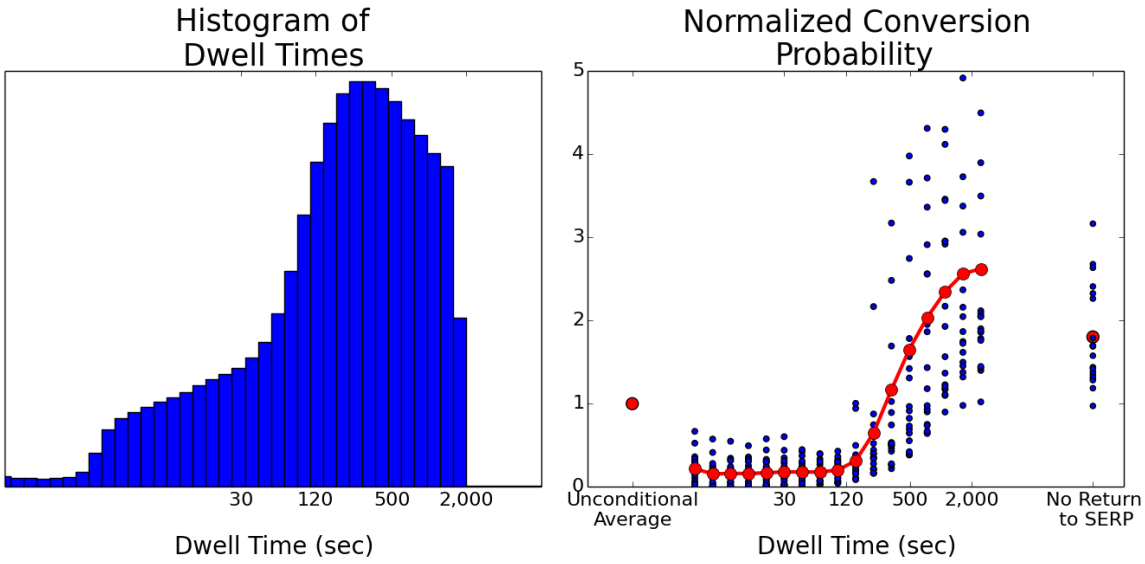


Figure 3: The unconditional distribution of dwell time (left panel) and an estimate of the multiplicative mapping from dwell time to click value (right panel). The blue dots in the right panel represent normalized conversion probability for the twenty ads in our data that convert the most often.

3.2 Experimentation

As is evident from equation (2), search engine revenue is directly effected by the parameters of the rank score function (Lahaie, 2006; Athey and Ellison, 2011). This motivates constant experimentation with new click prediction, relevance, and ranking algorithms. At a given time, Microsoft's Ad Center is operating many experiments on the Bing platform, each of which alters one or more parameters that govern ad ranking or display. Search results pages are randomized into these experiments at either the user or search level. Randomization is based upon a mathematical operation on a user's browser cookie, which assures that searchers are randomized identically, regardless of any geographic or individual-level factors. Additionally, advertisers have no control over which experiments they are involved in and cannot bid differently by experimental condition. The probabilistic weights that determine how many users go into each experiment are altered over time as older experiments expire or are expanded and new ones are introduced. Thus, experimental assignment is exogenous to a host of potential advertiser related confounds, but only after conditioning on time.

These experiments have a wide range of goals, such as improving user experience through more relevant ads, improving ad quality estimation, or changing the ranking of ads to maximize revenue. Although the experiments were designed for many proximate reasons, many have the effect of shuffling the ranking of ads. This is intuitive; the ranking of ads is the primary impact

the auctioneer has on the marketplace. We label the largest experimental cell as our *control* and use the other experiments as instrumental variables for causal inference. A challenge we have to address is that in addition to experimenting with alogrithmic components of ad serving that impact ranking, the search engine also tries out different visual displays of ads, which can directly effect user behavior. These experiments produce endogenous IVs in our causal model and must be carefully pruned. Prior to any estimation, these experiments can be detected if they produce a statistically distinct click probability (relative to the control) while an ad is held in a fixed position.

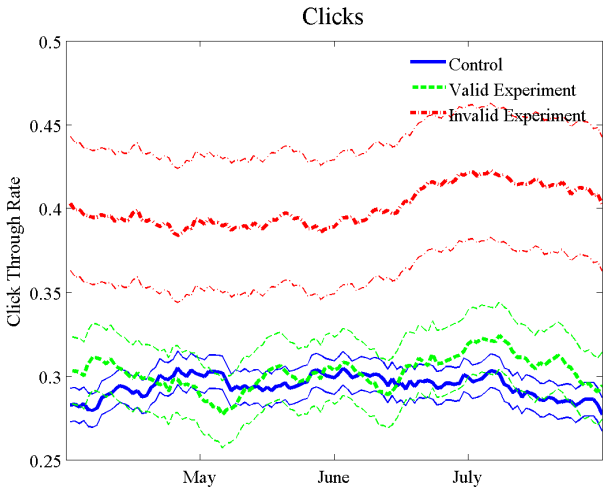


Figure 4: Local constant estimation of the average click probability over time for one particular ad in the control (green) and two other experimental conditions.

Figure 4 gives such an example where an invalid instrument is clearly identified. This advertisement was exclusively listed in the first mainline position in all three experimental conditions. The green (labeled as valid) experiment gives statistically identical click probabilities to the control indicating that it may be a valid instrument. However, the red experiment gives significantly larger click probabilities indicating that the ad was displayed differently in that condition. This pattern of results, would leave us to conclude that data from the red experiment should be removed from our analysis. If an experiment is shown to form an invalid instrument for a statistically significant subset of our ad-query pairs, it demonstrates the invalidity of that experiment globally. Scaling the analysis presented in Figure 4, gives us much greater power to prune invalid experiments.[10]

After pruning invalid experiments and aggregating experiments into relevance clusters, each of

[10]More generally, identifying valid instruments in sample is typically a difficult statistical problem that can only be attacked with weak, often underpowered, over-identification tests (Hall, 2000) and can complicate the properties of the resulting estimation(Leeb and Pötscher, 2005). However, we are able to leverage the combined information about these experiments in all 20,000 ad-query combinations in order to address this problem. Since our estimates will be performed on the level of the ad-query combination (utilizing only 1/20,000 of the data relevant to the model selection step), the critique from (Leeb and Pötscher, 2005) does not have the same bite.

which produces identical distributions of ad position across all 20,000 ad-query combinations, we are left with 40 unique experiments. A histogram of this distribution can be found in the Appendix.

3.3 Auxiliary Regressors

In order to understand heterogeneity in position effects, we collect a number of additional regressors for our 20,000 mainline combinations. Most importantly, we label each query as either "brand" or "product." Brand queries are specific to one particular company or manufacturer, such as "Samsung smart phone," while product queries are more generic, such as simply "smart phone." Within the class of brand queries, we further differentiate between a "on-brand" advertiser that represents the brand in the query and "off-brand" competitors. On-brand advertisers are identified by matching the brand term in the search query to the web address an ad points to.[11] Figure 5(a) shows the pooled distribution of each type of mainline pair across position. On-brand advertisers occupy the top slot with a very high frequency. Figure 5(b) plots the corresponding pooled CTRs by position for each type. These estimates have no causal interpretation of position, rather they show that brand queries in general, and on-brand advertisers in particular, tend to get significantly more ad clicks than ads displayed on product queries.

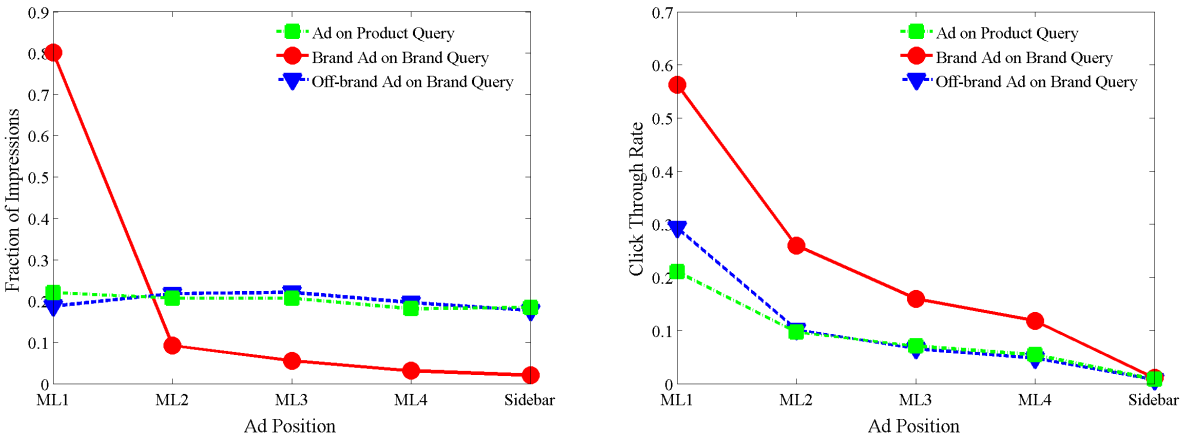


Figure 5: Pooled distribution of ad impressions (left) and observed click through rates (right) for ads on product queries, on-brand ads on brand queries, and off-brand ads on brand queries.

We also use the destination URL that each ad points to in order to obtain additional regressors through Alexa.com.[12] These additional regressors are intended to characterize the overall popularity

[11] Note this procedure will tend to miss brand terms that are not the company's name or in the web address. This may include, instance, colloquial terms or marketing slogans.

[12] Alexa.com is a subsidiary of Amazon.com that audits and makes public the frequency of visits on various Web sites. It's statistics are based on the observed browsing behavior of a subset users who have the Alexa toolbar installed.

and quality of these websites. Of particular interest, are the regressors *US Rank*, which gives a rank in overall popularity in terms of the number of unique visitors per day and *Bounce Rate* which gives an idea of how frequently visitors to a given website are dissatisfied and leave very quickly. Table 2 provides a summary of our auxiliary regressors and their correlation matrix is given in the appendix.

Table 2: Summary statistics

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**N**
On-brand Ad	0.03	0.171	0	1	20000
Off-brand Ad	0.721	0.449	0	1	20000
Ad on Product Query	0.249	0.433	0	1	20000
Global rank	776927.14	2290149.06	1	20421230	16623
US rank	39198.864	77175.539	0.286	2044261	14534
Bounce rate	0.359	0.142	0	2.7	14499
Pages per day	4.269	5.768	1	583	14499
Sec per day	250.906	168.103	0	5784	14499

4 Estimation

Our estimation procedure is designed to produce estimates at mainline pair level. For a given query, Each ad could be displayed in any of four mainline position or the in the sidebar. Our parameters of interest are the impact on click probability of moving an advertisement up one slot to position j, given by β_j, and the baseline click through rate for an ad in position $j+1$, given by α_{j+1}. In the next section, we also estimate these measures for click value. In order to allow for an unrestricted shape of the click curve, we will present separate estimates for the impact of each of our four positional goods. For example, for $j = 4$, we estimate the impact of moving an ad from the sidebar to ML4.

We will consider four estimation strategies, starting with OLS, which is likely biased. For OLS, the estimating equation is

$$\text{OLS:} \quad P(\text{Click} = 1|\text{Pos} \in \{j, j+1\}) = \alpha_{j+1} + \beta_j \cdot 1\{\text{Pos} = j\} + U, \qquad U \perp \text{Pos}. \tag{3}$$

This estimator simply pools all the data for a particular mainline combination. It improves on the summary statistics of the previous section by removing confounds that vary across advertisers, regardless of time, but it is still vulnerable to any time-varying shocks to click probability (U) that also correlate with ad position. In particular, one may expect that advertisers are placed in better ad positions during periods of high clickabillity. This could happen mechanically through the search engines estimation of quality or if U is positively correlated with changes to an advertiser's bid. In both case we expect overestimate β_j, which we call a *steepness bias* of the click curve.

To help control for these confounds, we consider a second specification that includes fixed effects at the four-hour level.[13] The corresponding equation is

$$\text{Time FE:} \quad P(\text{Click} = 1|\text{Pos} \in \{j, j+1\}) = \alpha_{j+1,t} + \beta_j \cdot 1\{\text{Pos} = j\} + U, \qquad U \perp \text{Pos}|t. \quad (4)$$

This specification should effectively control for any shocks to time-varying clickability and to an extent mitigate the steepness bias in (3). However, we may still be confounded by endogenous shocks at the geographic level (location of the user is inferred from the IP address). This could come in two forms: (1) competitors geo-target their ads to some locales but not others in an endogenous way or (2) geo-specific estimation of ad quality leads ads to be ranked higher where they are most popular. Intuitively, both of these factors would lead to steepness bias. Thus, we may expect a smaller, but still positive, steepness bias in estimates of equation (4).

While, we cannot directly observe these potential geographic confounds, we can still take advantage of our experiments to do robust causal estimation. The simplest approach is to create a vector of dummy variables for the presence of a given impression in each of our 40 relevance clusters[14] in our data. This vector $\mathbf{Z}$ is assigned randomly and guaranteed to be orthogonal to all possible confounds at a given point in time. Thus, we consider a *Dummy TSLS* approach given by a standard TSLS procedure with the estimating equations in (5).

$$\text{Dummy TSLS:} \quad P(\text{Click} = 1|\text{Pos} \in \{j, j+1\}) = \alpha_{j+1,t} + \beta_j \cdot 1\{\text{Pos} = j\} + U, \qquad U \perp \mathbf{Z}|t \quad (5)$$
$$P(1\{\text{Pos} = j\}) = \gamma_{j+1,t} + \mathbf{Z} \cdot \pi + V.$$

Crucial to the implementation of this approach is that experimental conditions can only be used during time periods for which they place every single impression of a given ad in either position j or $j+1$. That is to say, the population of users is held constant—the ad location varies by one slot at most. In equations (3) and (4) we assumed that the contemporaneous distribution of U would not vary with ad position and it was sufficient to simply eliminate those impressions in which the ad was not placed in position j or $j+1$. Doing so here, could compromise the validity of our instruments if the selected sample differs in some systematic way.[15]

Given these concerns, we eliminate experiments that do not respect this inclusion restriction. Cutting down our data like this combined with the weak instrument problem made it very difficult

[13] Due to computation issues related to the volume of data, we were not able to perform a fixed effect analysis at any more narrow intervals.

[14] Recall that each relevance cluster, as defined in Section 3.2, is a collection of experiments that produce identical ad rankings.

[15] An alternative approach involves using our instruments to estimate a multidimensional IV model so that experiments that place ads in more than two positions can be used. However, relevance in such settings is a much steeper challenge for our instruments and can complicate identification. See the Appendix for a brief discussion. Various bounds on parameter values and the distribution of unobservables can provide sharper inference, efficiently incorporating these may be a useful direction for future work.

to obtain sufficient first stage relevance using the dummy TSLS approach.[16] To increase power, it is necessary to allow more flexibility in the first stage fit, thereby increasing the weight on instruments when they are "turned on." Given the time-varying strength of our instruments, a logical way to accomplish this is to consider interactions between experimental condition and time as additional first stage regressors.

A variety of regularization approaches exist to select a set of instruments for a good first stage fit, but most are designed for a large, but fixed, class of candidate instruments (Belloni et al., 2012; Okui, 2011) and become computationally intensive because the discrete interaction terms make the data set very wide. An ideally suited solution for our setting is to instead generate instruments via a nonparametric smoothing of the conditional distribution of ad position for each experimental condition. These nonparametric IVs effectively "turn on" only for time periods where experiments have a contemporaneous impact on the distribution of ad position. This procedure is developed by Carrasco (2012), who shows it to be optimal for econometric models based on conditional moment restrictions with continuous covariates.

An example of how the first stage fit of a procedure based on these nonparametric instruments compares to our Dummy TSLS specification is shown in Figure 6. Nonparametric IV improves first stage fit by allowing one to sharpen the focus on natural experiments when they actually occur. Dummy TSLS flattens the experiment across the entire time period under study. This is equivalent to using an infinite bandwidth in the smoothing procedure. For the example shown, this adds noise to the estimates from time periods during early May and July when our experiments did not impact ranking.

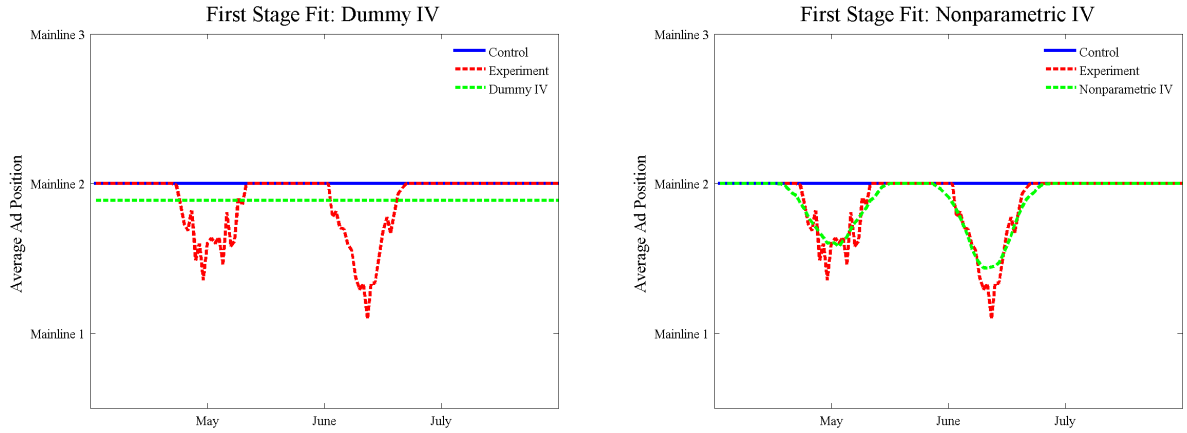


Figure 6: First stage fits generated by our Dummy TSLS and Nonparametric IV specifications for a hypothetical ad.

[16] We follow the standard convention of throwing out any estimates obtained with a first stage F-statistic less than 10.

Formally, the estimating equations for this model are presented in (6).

$$\text{Nonparametric IV:} \quad P(\text{Click} = 1|\text{Pos} \in \{j, j+1\}) = \alpha_{j+1,t} + \beta_j \cdot 1\{\text{Pos} = j\} + U, \qquad U \perp \mathbf{Z}|t$$
$$P(1\{\text{Pos} = j\}) = \phi_h(\mathbf{Z}, t) + V \tag{6}$$

The key difference is that the first stage is allowed to be a nonparametric function of both $\mathbf{Z}$ and t. As mentioned, Dummy TSLS is nested in this approach as it is equivalent to an infinite bandwidth. Bandwidth selection is clearly the key knob to for this method. Sending bandwidth to zero results in a "perfect" first stage fit and essentially returns an OLS estimate since the predicted X for the second stage regression is arbitrarily close to the observed X. Accordingly, very small bandwidths should be avoided as they generate exactly the kind of endogeneity bias we seek to avoid.[17] In choosing bandwidth, we thus took a conservative approach and opted for 1 week for all estimations. This was shown to induce a bias of less than 1% of the treatment effect in our worst case simulations allowing us to be confident in the inference of casual effects.[18]

5 Results: Click Curves

We estimate the four specifications of Section 4 on each of the 20,000 ad-query combinations in our data. Many advertisements did not spend much time in certain positions. For these, our effects of interest are not identified and estimates will not be presented. For many other ads, treatment effects could be estimated in our OLS or Time FE models, but lacked sufficient relevance for identification in the Dummy TSLS or sometimes even the Nonparametric IV model.

5.1 Endogeneity of OLS methods

In order to get a baseline sense of the extent to which endogeneity corrupts our more naive estimators, we summarize the estimation results for those queries that could estimated by all four techniques in Table 3. Treatment effects are summarized for each position by their simple and weighted average.[19]

Focusing on the first two rows, we see that, for all positions, the average estimated treatment effects is larger in OLS than in our Time FE method. This is statistically significant and especially

[17] Another way to eliminate this endogeneity bias, is to remove the data from a particular impression when computing the corresponding first stage fitted value. Then bandwidth selection is driven only by a desire to to maximize power via a strong first stage fit. Existing cross-validation techniques might be ideally suited to this purpose. Unfortunately, this type of "leave one out" approach was not computationally feasible in our setting.

[18] We believe that a MSE-optimal bandwidth could be derived – at least for the case of homoscedastic errors – by following the expansion presented in Okui (2011). This may be a profitable direction for more theoretically inclined work.

[19] The weighting is chosen to be consistent across estimation techniques. Since the Dummy TSLS method tended to have the largest standard errors, we chose our weights to be inverse to the standard error of those estimates.

Table 3: Results summary on mainline pairs that could be estimated by all four methods.

	ML 1 Bonus		ML 2 Bonus		ML 3 Bonus		ML 4 Bonus	
	Simple Mean	Weighted Mean	Simple Mean	Weighted Mean	Simple Mean	Weighted Mean	Simple Mean	Weighted Mean
OLS	7.65	6.73	2.15	1.60	0.99	0.62	2.84	1.68
	(0.03)	(0.04)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Time FE	7.48	6.60	2.08	1.56	0.90	0.57	2.81	1.66
	(0.06)	(0.05)	(0.03)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)
Dummy TSLS	7.57	6.50	1.84	1.42	0.68	0.47	2.68	1.53
	(0.21)	(0.15)	(0.15)	(0.09)	(0.14)	(0.07)	(0.10)	(0.04)
Nonparametric IV	7.69	6.58	1.94	1.44	0.67	0.45	2.73	1.58
	(0.15)	(0.11)	(0.17)	(0.09)	(0.10)	(0.05)	(0.07)	(0.03)
Ad-Query Pairs	269	269	394	394	380	380	634	634

† Aggregates for the sample of ads in which the Dummy TSLS model has standard errors $< 5\%$.
††Weights chosen identically across estimations and to be most friendly to IV.

true for ads near the top of the page, but is only marginal for ads moving from the sidebar to the last mainline ad. Going from the second to the third and fourth rows, we again see a general decline in average treatment effects, but this time the effect in concentrated on the less popular ads at the bottom of the page. In combination, these results strongly suggest that naive OLS estimation of click curves is significantly biased toward steepness, which is exactly the type of endogeneity one would expect given GSP mechanics promote ads during periods of high popularity. Furthermore, it is perhaps intuitive that the endogeneity in the more popular ads at the top of the page seems to be driven by time-varying factors—perhaps related to the overall popularity of the brand/site—while endogeneity for ads further down the page is driven by geographic factors.

In order to expand the analysis to more ads, we dropped the Dummy TSLS method, the most restrictive method in terms of coverage, from the comparison. This allows us average over many additional ad-query pairs that did not meet the first stage relevancy requirement, or had very large standard errors, in the Dummy TSLS analysis. The pattern of results is displayed in Table 4 and is consistent with those found above. The fact that the number of ads estimated triples is a testament to the superior efficiency of the Nonparametric IV approach.

5.2 Heterogeneous Click Curves

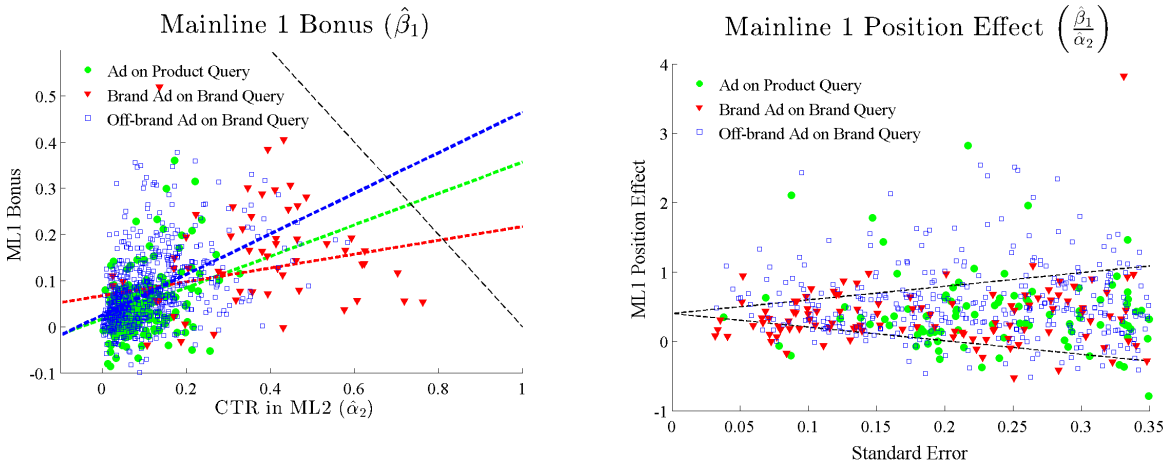
Given the bias in OLS and noisiness of the Dummy TSLS, we focus attention on the estimates derived from our Nonparametric IV method in order to explore heterogeneity in click curves. First, we look at the causal impact of moving an ad from the second to the first mainline position. Since most of the revenue in this auction comes from the first slot, it forms a natural starting point. Figure 7a plots the impact on CTR of moving to slot 1 on the y-axis against CTR in the second

Table 4: Results Summary on mainline pairs that could be estimated by our OLS and Nonparametric IV method.

	ML 1 Bonus		ML 2 Bonus		ML 3 Bonus		ML 4 Bonus	
	Simple	Weighted	Simple	Weighted	Simple	Weighted	Simple	Weighted
OLS	7.94	7.22	1.98	1.62	0.91	0.66	2.89	1.82
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Time FE	7.72	7.03	1.90	1.57	0.86	0.61	2.83	1.80
	(0.04)	(0.03)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
Nonparametric IV	7.93	7.10	1.79	1.51	0.53	0.43	2.41	1.50
	(0.12)	(0.09)	(0.10)	(0.05)	(0.10)	(0.05)	(0.06)	(0.03)
Ad-Query Pairs	783	783	858	858	828	828	1,880	1,880

† Aggregates for the sample of ads in which the Nonparametric IV model has standard errors $< 5\%$.
††Weights chosen identically across estimations and to be most friendly to Nonparametric IV.

mainline position on the x-axis. We differentiate between our three types of ads (ads on product queries, on-brand ads, and off-brand ads) and compute inverse variance-weighted GLS lines of best fit for each group. The graphical presentation is motivated by our interest in the hypothesis of *separability of the click curve*. Separability has the strong implication that all the dots on this graph should be along a ray coming out of the origin, the slope of which defines the (constant) ML1 position effect. Corresponding GLS regressions of these estimated treatment effects onto baseline CTR and our auxiliary regressors are presented in Table 5.



(a) (b)

Figure 7: Heterogeneous position effects for ML1: The left panel shows treatment effect of moving from the second slot (ML2) to the first slot (ML1) for each advertiser is plotted against baseline clickability in ML2. The right panel scatters the implied multiplicative effect for each ad against the corresponding standard error. Dots outside the black lines are those that are statistically different from the group level average at the 5% level.

Table 5: Auxillary regressions on $\hat{\beta}_1$.

	Brand Query On-Brand Ad			Brand Query Off-Brand Ad			Product Query		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Constant		6.69	4.83		2.49	.49		1.62	-2.88
		(2.23)***	(5.30)		(.21)***	(.57)		(.54)***	(1.30)**
CTR in ML2 ($\hat{\alpha}_2$)	.30	.15	.16	.62	.44	.37	.45	.34	.33
	(.03)***	(.06)***	(.06)***	(.02)***	(.02)***	(.02)***	(.03)***	(.05)***	(.05)***
Log US Rank			-.25			.36			1.03
			(.67)			(.08)***			(.17)***
Bounce Rate			11.78			.66			-10.38
			(10.66)			(1.29)			(2.82)***
Number of Ad-Query Pairs	91	91	90	1,242	1,242	931	412	412	305

Figure 7a immediately reveals three substantial rejections of the separability assumption. First, our three types of ads have clearly distinct slopes, with off-brand ads having the steepest position effect and on-brand ads having the flattest effect. The overwhelming statistical significance of this difference is confirmed the different estimates on $\hat{\alpha}_2$ obtained in specification (1) of Table 5. Second, the GLS line for on-brand ads does not appear to go through the origin, but rather has a positive intercept of around 7%. This means that on-brand ads with low baseline clickability experience a greater multiplicative return to position than those with higher baselines. The statistical significance here can be checked by the non-zero constant estimated in specification (2) of Table 5 for on-brand ads.

A final rejection of separability is that even within ad type, there is substantial, statistically significant heterogeneity in the slope of click curves. This is evident from Figure 7b, which presents a scatterplot of the estimated multiplicative position effect for each ad against the corresponding standard error. The dashed black lines start from the global average and have a slope of ± 1.96, so they represent a 95% confidence interval around the group mean. Multiplicative separability implies that no more than 5% of dots should be observed outside these lines.

Additional clues as to what is causing the excess heterogeneity can be gleaned from specification 3 in Table 5. All else equal, the ML1 treatment effect is smaller for product ads and off-brand ads that have a low website rank (meaning they are more popular). ML1 position and the popularity of a website thus work as substitutes for the purposes of attracting more clicks for these types of ads. Additionally, the position effect for ads on product queries was shown to be larger for ads with a low Bounce Rate. A lower Bounce Rate indicates fewer dissatisfied visitors and is generally regarded as an indicator of higher website quality. Interestingly, this means that for product queries the ads that benefit the most from the ML1 position are those that are not very popular but tend to generate high levels of user interest based on engagement metrics from Alexa.com.

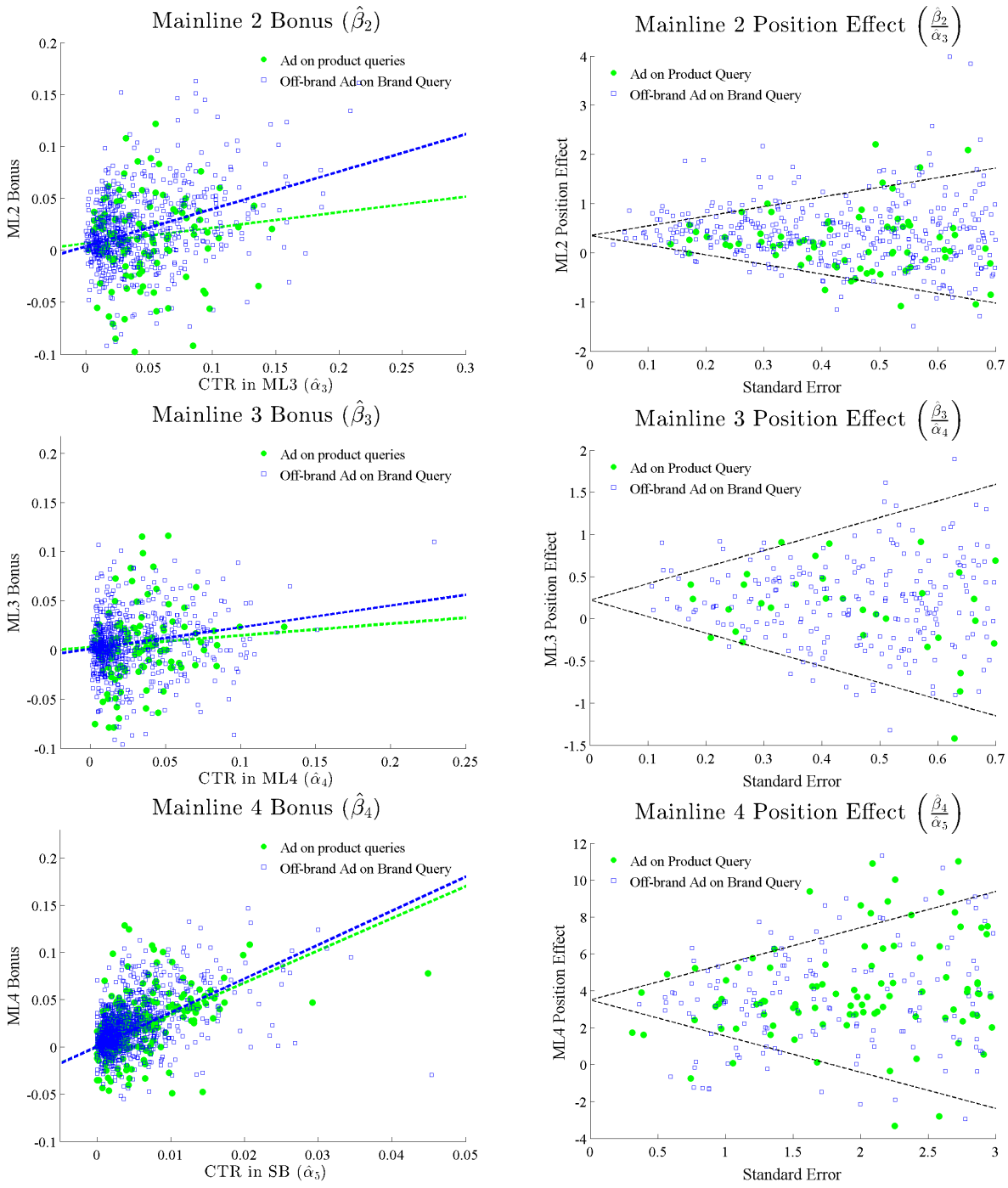


Figure 8: Heterogeneous positions effects for ML2-ML4. This was only close in the case of the Mainline 3 position effect. Here $238/264 = .9015$ dots were within the black lines. This fraction is distinct from .95 with $p = .0017$.

We now replicate the above analysis for positions 2–4. As previously noted, on-brand ads were rarely observed in lower positions on the page and thus we could not identify the rest of the click curve for these type of ads. As such, our analysis for the remaining positions will focus only on ads on product queries and off-brand ads on brand queries. Figure 8 collects scatterplots of these treatment effects for each of these 3 positions in analog to the results of Figure 7 and auxiliary GLS regressions are in the Appendix. The results generally adhere more closely to separability, but still have significant violations. Most obviously, for all three positions, off-brand ads again have significantly steeper position effects than ads on product queries. Also within each category, we still see statistically significant heterogeneity for each treatment effect as represented by the excess dispersion of dots in the right panel of each row. The "most separable" treatment effect is for ads moving from ML4 to ML3. Here the average ad had a 24% multiplicative bonus and 238 out of the 264 estimates with standard errors < 0.7 were statistically indistinct from the group level average. However, this is still only 90.15% and allows us to reject a null of no heterogeneity with $p = .0017$ (although we note the observed differences are much less economically meaningful than for the other positions).

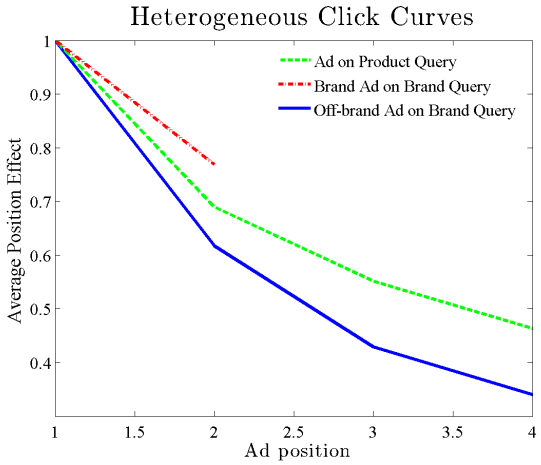


Figure 9: Average position effect by ad type. On-brand ads were very rarely observed outside of the first two positions. Reliable estimates of their full click curve were difficult to obtain and are not presented.

Finally, to visualize the differences we aggregate the results to graphically show the *average* click curve for our three types of ads. These are displayed in Figure 9. This summarizes the data ignoring within-class heterogeneity and could be a useful approximation for a search engine to employ. Placed on the same plot, the large degree of heterogeneity in click curves is stark.

6 Results: Click Value

We now turn our attention to estimating the impact of ad position on the value of a click. We'll use the measure of *mapped dwell time*, defined in Section 3.1, as our primary outcome metric. Recall that this metric is defined in order to be representative of proportional changes to conversion rate and thus click value. This gives it an easy interpretation that is directly relevant to an advertiser's decision to change their bid in order to seek (or avoid) a given position, namely the percentage change in the average value of a click. Further the definition of a conversion varies arbitrarily across advertisers (they are free to define it any way they wish). By contrast, this measure facilitates fair cross-advertiser comparisons.

The results for positions 1 and 2 are summarized for all four methods in the first 4 columns of Table 6. Since we only observe value metrics conditional on a click, we cannot estimate the IV methods for positions 3 and 4 (hence the "n/a's" in the table) since CTR tends to be low in these positions. Focusing on the top position, the non-parametric IV weighted mean indicates that the impact of moving from slot 2 to slot 1 is a 1.7% increase in conversion rate over an advertiser's baseline in slot 2. The standard error is 0.50, meaning we can tightly bound the mean effect to be $< 3\%$. So while we do document a positive impact on click value of moving up the page, the magnitude is very small. For ML2, the results are not statistically significant but again can be bounded near zero.

We examine the impact of moving from the sidebar to the mainline ads in the final 4 columns Table 6, but have to rely on OLS estimates, which are expected to have positive bias. We find a significantly negative impact of moving to the mainline. Comparing the ML4 bonus to the other three positions, we see that it is the only one to show an economically meaningful departure from zero. Given that sidebar are simply those ranked 5th or higher, we have no reason to expect a dramatically different endogeneity bias for these estimates and feel safe concluding this is representative of a causal effect.[20] A simple explanation is that this result reflects the changing population of clickers when an ad moves to the mainline. Combined with the very low CTR on sidebar ads, this may indicate that users who have gone to the the trouble of finding a sidebar ad, simply have a higher purchasing intent. However, since this result relies on OLS, there is naturally some uncertainty remaining.

7 Discussion and Conclusion

There has been extensive research on the revenue properties of the GSP in sponsored search. The seminal papers of Edelman et al. (2007) and Varian (2007) show clearly that under certain assump-

[20]Note it should not be interpreted as a causal effect on any population of clickers because ad position clearly changes the composition of users. Instead it is the causal effect facing an advertiser's average click value, which is the relevant quantity for their bidding decisions.

Table 6: ML1-4 Normalized Conversion Bonus ($\hat{\beta}_1$)

	ML 1 Bonus		ML 2 Bonus		ML 3 Bonus		ML 4 Bonus	
	Simple	Weighted	Simple	Weighted	Simple	Weighted	Simple	Weighted
OLS	-0.40	0.74	-0.56	-0.10	-0.68	-0.07	-8.96	-8.42
	(0.06)	(0.02)	(0.05)	(0.03)	(0.06)	(0.03)	(0.12)	(0.07)
Time FE	0.43	0.37	0.14	-0.02	0.08	0.15	-6.18	-6.09
	(0.15)	(0.12)	(0.19)	(0.16)	(0.25)	(0.21)	(0.86)	(0.81)
Dummy TSLS	3.30	3.24	2.80	2.26	n/a	n/a	n/a	n/a
	(1.47)	(1.05)	(3.18)	(2.62)				
Nonparametric IV	0.92	1.70	-3.29	-1.47	n/a	n/a	n/a	n/a
	(0.79)	(0.50)	(1.64)	(1.15)				

tions, all envy-free equilibria of the GSP obtain at least the revenue of the VCG mechanism.[21] Chief among these assumptions are the *separability of the click curve* and the *homogeneity of click value*. These have been tested empirically in recent papers (Agarwal et al., 2011; Ghose and Yang, 2009; Narayanan and Kalyanam, 2014), each focusing on a single advertiser. These papers reach contrasting conclusions on the issue of click value and, due to their limited sample, could not provide strong evidence on separability. Our paper provides robust causal inference on a wide cross-section of advertisers to test both assumptions at scale. Our results provide valuable reassurance for the homogeneity of click value, but strongly reject separability of the click curve, and with it, many desirable theoretical properties of the current GSP approach to position auctions.

It is immediately clear that this false presumption of separability can induce significant costs. As a brief exercise, consider two otherwise-identical advertisers (B for brand and O for off-brand) competing for the first mainline slot and let us normalize their click curves to be one in the second position, $\mu_2^O = \mu_2^B = 1$. Recall that in Specification (2) of Table 5, we estimated that the average off-brand advertisers has a multiplicative bonus of 62% when shifting to the first mainline position ($\mu_1^O = 1.62$), while the average on-brand advertiser only gets a 30% bonus ($\mu^B = 1.3$) from being in the top slot. Supposing a knife edge case, in which the GSP barely decides to rank the second ad on top, we see that combined surplus from the top two slots could be as much as $1 - \frac{1+\mu_1^B}{1+\mu_1^O} \approx 13\%$ less than would be achieved in an efficient allocation. The exact loss of revenue would depend on the rank score of the third place bidder, but could range anywhere from the estimated 13% to as much as 19%, if the third ranked bidder has a rank score of 0. So even if we ignore all click curve heterogeneity beyond just brand vs. off-brand averages, as much as 19% of revenue could be lost on some queries by improperly placing on-brand ads in the first mainline position over better suited, off-brand competitors.[22] Expanding the analysis to include all four slots, account for additional

[21] Relaxing this assumption, Lucier et al. (2012) shows that all Bayes-Nash equilibria still obtain at least half the revenue of the VCG.

[22] Since off-brand ads do not see any decline in click value when they are moved up the page, one could perhaps argue that such a reallocation would be efficient.

elements of heterogeneity, or allow for strategic bidding based on these misallocations could lead to a greater calculated revenue impact.

Eliminating these losses requires the incorporation of heterogeneous click curves into the mechanism. For this, there are two hurdles. The first is an empirical challenge. This is precisely what the methodology developed in this paper was designed to address. Our analysis has identified ad-query type, site popularity and quality as important dimensions along which the shape of click curves may vary, but also has uncovered substantial unexplained heterogeneity. As such, efficient, real-time estimation of an individual advertiser's click curve should be cast in the framework of a hierarchical Bayesian model. This could efficiently combine information on the population distribution of click curves with estimates for individual advertisers that can be gleaned from ongoing experimentation.

The second challenge is a theoretical one. All current equilibrium analysis–of which we are aware–presumes separability of the click curve. However, even without further theoretical inquiry, we see two options for improvement. One approach is to modify the GSP by simply incorporating estimated heterogeneity in the click curves into the calculation of an advertiser's quality. Current practice defines each advertiser's quality as their estimated probability of receiving a click in the first mainline position. This is calculated by deflating each advertisers observed click through rates by a common set of global position effects (like those presented in Figure 2) and results in exactly the type of misallocation discussed above. But, if each advertiser's click curve was known, then quality could be estimated without bias and we could be confident that the first mainline position – at least – was being allocated correctly. This approach is straightforward, would not require a fundamental change in the structure of the auction and could build directly off the estimates presented herein.

However, this approach still cannot eliminate the possibility of misallocation (and a corresponding misalignment of advertiser incentives) of slots further down the page. In fact, since any type of GSP auction must be organized around a scalar-valued index of bidder quality, it can never be certain to efficiently allocate more than one prize. Only a shift to a VCG mechanism could fully incorporate heterogeneously estimated click curves in a way that insured efficient allocation of (and incentive compatible bidding for) all four mainline slots.

More broadly, variants of the GSP are employed to allocate advertising space by many of the most popular search engines and online content providers. Their adoption, within a contingent bid framework in which advertisers pay for user action (usually clicks), makes the auctioneer the residual claimant on these actions. This could be because the auctioneer presumes a greater familiarity with the impression to clicks mapping on their platform and creates a competitive advantage over offline advertising where such action-based payment schemes are not possible. Yet we must conclude by sounding a note of caution. Such a centralized system (as opposed to simply selling impressions and letting advertisers work out their own payoff relevant factors) requires a careful understanding of the empirical realities of the auction. As demonstrated here, heterogeneity in how individual advertisers

interact with the platform may need to be factored into the mechanism if efficient allocation and incentive compatibility of bidding are to be maintained. Additionally, even if all parameters are perfectly estimated, a GSP-style auction cannot be sure to achieve efficient allocation, or VCG equivalence in revenue, if multiple goods are sold simultaneously to advertiser's with non-scalar heterogeneity.

References

Agarwal, A., Hosanagar, K., and Smith, M. (2011). Location, location, location: An analysis of profitability of position in online advertising markets. *Journal of Marketing Research*.

Athey, S. and Ellison, G. (2011). Position auctions with consumer search. *The Quarterly Journal of Economics*, 126(3):1213–1270.

Athey, S. and Nekipelov, D. (2010). A structural model of sponsored search advertising auctions. In *Sixth Ad Auctions Workshop*.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Brunel, F. F. and Nelson, M. R. (2003). Message order effects and gender differences in advertising persuasion. *Journal of Advertising Research*, 43(03):330–341.

Brynjolfsson, E., Dick, A. A., and Smith, M. D. (2010). A nearly perfect market?: Differentiation vs. price in consumer choice. *Quantitative Marketing and Economics*, 8(1):1–33.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.

Chakravarti, A., Janiszewski, C., and Ülkümen, G. (2006). The neglect of prescreening information. *Journal of Marketing Research*, pages 642–653.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 87–94. ACM.

Edelman, B., Ostrovsky, M., and Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259.

Ghose, A. and Yang, S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10):1605–1622.

Gomes, R. and Sweeney, K. (2012). Bayes-nash equilibria of the generalized second-price auction. *Games and Economic Behavior*.

Hall, A. R. (2000). Covariance matrix estimation and the power of the overidentifying restrictions test. *Econometrica*, 68(6):1517–1527.

Hummel, P. and McAfee, R. P. (2014). Position auctions with externalities and brand effects. *arXiv preprint arXiv:1409.4687*.

Jerath, K., Ma, L., Park, Y.-H., and Srinivasan, K. (2011). A position paradox in sponsored search auctions. *Marketing Science*, 30(4):612–627.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. ACM.

Kempe, D. and Mahdian, M. (2008). A cascade model for externalities in sponsored search. In *Internet and Network Economics*, pages 585–596. Springer.

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., and Xu, Y. (2012). Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 786–794. ACM.

Kohavi, R., Henne, R. M., and Sommerfield, D. (2007). Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 959–967. ACM.

Lahaie, S. (2006). An analysis of alternative slot auction designs for sponsored search. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 218–227. ACM.

Lana, R. E. (1963). Interest, media, and order effects in persuasive communications. *The Journal of Psychology*, 56(1):9–13.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01):21–59.

Lewis, R. and Reiley, D. (2013). Online ads and offline sales: Measuring the effects of retail advertising via a controlled experiment on yahoo! In *Working paper*.

Lucier, B., Paes Leme, R., and Tardos, E. (2012). On revenue in the generalized second price auction. In *Proceedings of the 21st international conference on World Wide Web*, pages 361–370. ACM.

Manzi, J. (2012). Uncontrolled: The suprising power of trial and error for business, politics and society.

Metrikov, P., Diaz, F., Lahaie, S., and Rao, J. (2014). Whole page optimization: how page elements interact with the position auction. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 583–600. ACM.

Narayanan, S. and Kalyanam, K. (2014). Position effects in search advertising: A regression discontinuity approach. *Working Paper*.

Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics*, 165(1):70–86.

Ostrovsky, M. and Schwarz, M. (2011). Reserve prices in internet advertising auctions: A field experiment. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 59–60. ACM.

Radlinski, F., Kurup, M., and Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 43–52. ACM.

Varian, H. R. (2007). Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178.

8 Additional summary tables

8.1 Auxiliary regressions for positions 2–4

Table 7: Auxillary regressions on $\hat{\beta}_{2-4}$.

		Brand Query			Product Query		
		(1)	(2)	(3)	(1)	(2)	(3)
Mainline 2 Bonus ($\hat{\beta}_2$)	Constant		.38 (.06)***	.05 (.17)		.66 (.38)*	-.41 (1.03)
	CTR in ML3 ($\hat{\alpha}_3$)	.44 (.02)***	.36 (.02)***	.34 (.02)***	.25 (.05)***	.15 (.08)*	.09 (.11)
	Log US Rank			.02 (.02)			.03 (.12)
	Bounce rate			.76 (.38)**			2.12 (2.09)
	Number of Ads	1,148	1,148	999	221	221	153
Mainline 3 Bonus ($\hat{\beta}_3$)	Constant		.10 (.04)***	.18 (.11)		.27 (.27)	.18 (.11)
	CTR in ML4 ($\hat{\alpha}_4$)	.26 (.02)***	.22 (.03)***	.23 (.03)***	.19 (.06)***	.12 (.08)	.23 (.03)***
	Log US Rank			-.006 (.01)			-.006 (.01)
	Bounce Rate			-.04 (.33)			-.04 (.33)
	Number of Ads	1,039	1,039	870	207	207	870
Mainline 4 Bonus ($\hat{\beta}_4$)	Constant		.05 (.02)**	.12 (.09)		.008 (.08)	1.06 (.38)***
	CTR in SB ($\hat{\alpha}_5$)	3.68 (.09)***	3.59 (.10)***	3.46 (.11)***	3.41 (.15)***	3.40 (.18)***	3.64 (.22)***
	Log US Rank			-.01 (.01)			-.10 (.05)**
	Bounce Rate			.52 (.23)**			-.43 (.87)
	Number of Ads	2,066	2,066	1,592	598	598	392

8.2 Primary and Secondary Clicks

The first click a user makes on an advertisement on a given search page is referred to as *primary*. Often users return to the search page after clicking on an advertisement, thus affording them the opportunity to make *secondary* clicks, each of which is charged to the appropriate advertiser. However, as demonstrated in Figure 10 this is relatively rare.

However, it is interesting to note that these *secondary* clicks appear to be considerably less

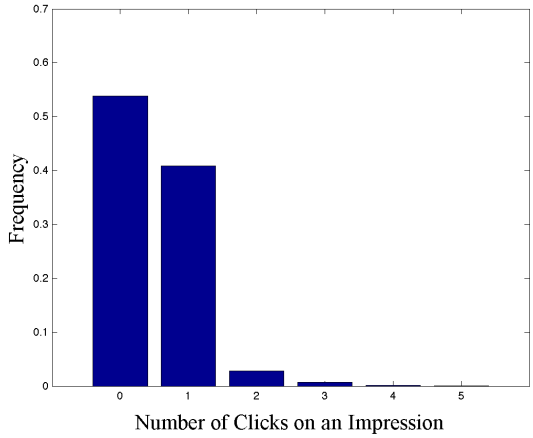


Figure 10: Distribution of clicks per impression.

valuable as shown in Table 8.

Table 8: User engagement on primary and secondary clicks

Metric	Click	Mainline 1	Mainline 2	Mainline 3	Mainline 4	Sidebar
Probability of Conversion	Primary	0.030	0.040	0.041	0.041	0.050
	Secondary	0.027	0.029	0.030	0.031	0.035
Average Dwell Time	Primary	321.188	234.135	224.334	215.307	228.467
	Secondary	199.261	164.150	148.236	147.539	176.131
Probability of Infinite Dwell Time	Primary	0.400	0.293	0.273	0.262	0.280
	Secondary	0.267	0.209	0.177	0.176	0.211
Mapped Dwell Time	Primary	1.197	0.956	0.915	0.886	0.923
	Secondary	0.851	0.720	0.655	0.647	0.725

8.3 Heterogeneous Click Through Rates

The histograms in Figure 11 and 12 demonstrate the heterogeneity of click through rates for ads shown commonly in each mainline position. This distributions are purely observational and have no causal interpretation.

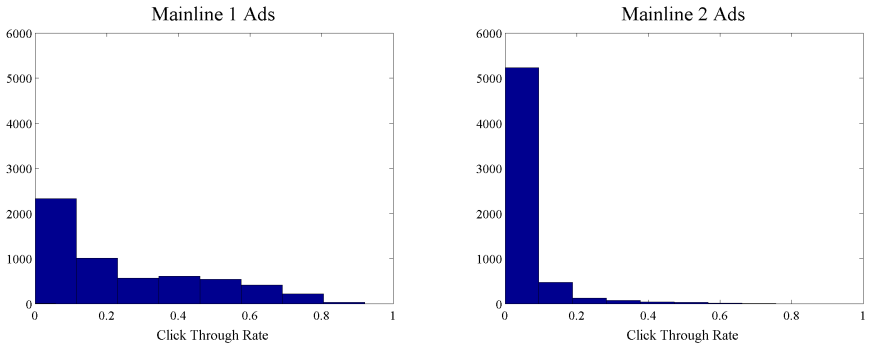


Figure 11: Histogram of average CTR at the advertisement level for positions ML1 and ML2. Ads are only included if they have at least 3,00 impressions in that position.

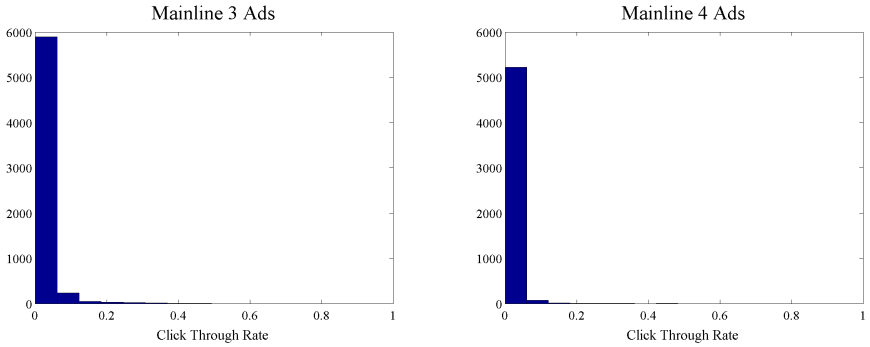


Figure 12: Histogram of average CTRs at the advertisement level for positions ML3 and ML3. Ads are only included if they have at least 3,00 impressions in that position.

8.4 Alexa Stats

Table 9 demonstrates some correlations between our different website level metrics from Alexa.com.

Table 9: Correlations

	Log Global rank	Log US rank	Bounce rate	Pages per day	Seconds per day
Log Global rank	1				
Log US rank	0.980***	1			
Bounce rate	0.464***	0.474***	1		
Pages per day	-0.209***	-0.271***	-0.0978***	1	
Sec per day	-0.453***	-0.445***	-0.528***	0.373***	1

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

8.5 Heterogeneous Click Curves with Dummy TSLS Estimates

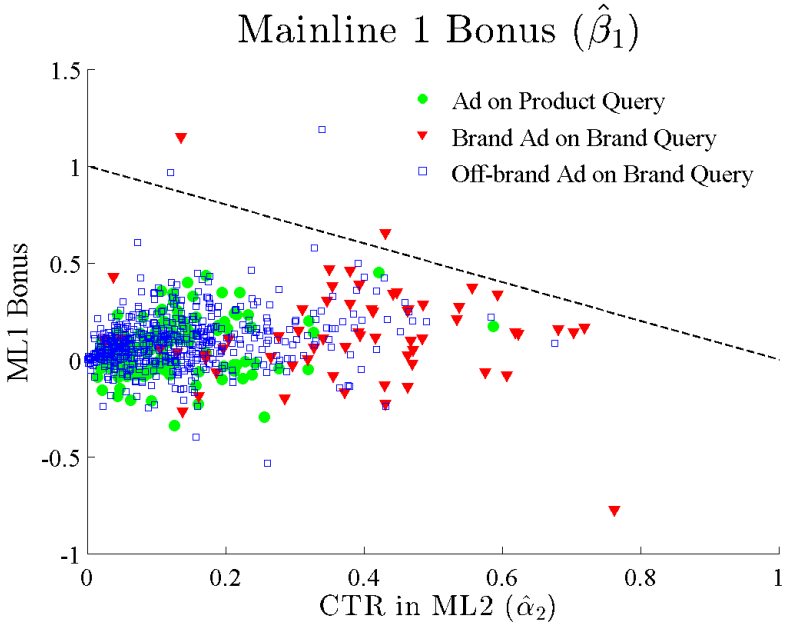


Figure 13: Position effect of moving from the second slot (ML2) to the first slot (ML1) by advertiser. x-axis gives the average CTR over ML1 and ML2. In Panel 1 we show OLS estimates, which are biased. Panel 2 gives lower IV estimates which uncovers a diminishing return of position for dominant ads.

Table 10: Auxillary regressions on $\hat{\beta}_1$ with Dummy TSLS estimates.

	Brand Query On-Brand Ad			Brand Query Off-Brand Ad			Product Query		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Constant		3.39	10.92		1.67	.22		2.02	-4.98
		(5.62)	(12.56)		(.31)***	(.90)		(1.29)	(3.76)
CTR in ML2 ($\hat{\alpha}_2$)	.37	.30	.32	.65	.53	.47	.49	.38	.40
	(.07)***	(.14)**	(.14)**	(.03)***	(.03)***	(.04)***	(.07)***	(.10)***	(.11)***
Log US Rank			-2.45			.39			1.65
			(1.84)			(.13)***			(.53)***
Bounce rate			34.96			-1.41			-18.88
			(23.97)			(1.95)			(7.36)**
Ad-Query Pairs	40	40	40	380	455	345	112	112	85

Table 11: Auxillary regressions on $\hat{\beta}_2 - 4$ with Dummy TSLS estimates.

		Brand Query			Product Query		
		(1)	(2)	(3)	(1)	(2)	(3)
Mainline 2	Constant		.34 (.10)***	.24 (.34)		.57 (.96)	-.21 (2.38)
	CTR in ML3 ($\hat{\alpha}_3$)	.38 (.03)***	.32 (.03)***	.30 (.04)***	.33 (.09)***	.24 (.17)	.21 (.21)
	Log US Rank			-.02 (.05)			-.07 (.26)
	Bounce rate			.86 (.66)			2.78 (4.18)
	Ad-Query Pairs	459	459	404	78	78	61
Mainline 3	Constant		-.02 (.05)	-.01 (.21)		.89 (.42)**	-1.29 (2.30)
	CTR in ML4 ($\hat{\alpha}_4$)	.28 (.04)***	.29 (.05)***	.26 (.05)***	.26 (.12)**	.06 (.15)	.001 (.23)
	Log US Rank			.04 (.03)			.16 (.28)
	Bounce rate			-.75 (.62)			1.64 (3.03)
	Ad-Query Pairs	405	405	351	64	64	37
Mainline 4	Constant		.23 (.05)***	.12 (.15)		.85 (.23)***	1.25 (1.12)
	CTR in SB ($\hat{\alpha}_5$)	3.65 (.17)***	3.18 (.20)***	3.33 (.20)***	3.57 (.31)***	2.87 (.35)***	2.75 (.48)***
	Log US Rank			-.04 (.02)**			-.10 (.13)
	Bounce rate			1.15 (.39)***			2.04 (1.84)
	Ad-Query Pairs	569	569	448	115	115	83

8.6 Heterogeneous Click Curves with OLS Estimates

We now replicate the analysis of Section 4.2 with our OLS estimates derived from equation (3). Results are presented in Tables 12 and 13.

8.7 Experimental Frequencies

As discussed in Section 2.2, our data is divided up into 40 distinct valid relevance clusters that differentially impact ranking. Each of these relevance clusters is composed of multiple Bing experiments. A histogram of their frequencies is presented in Figure 15.

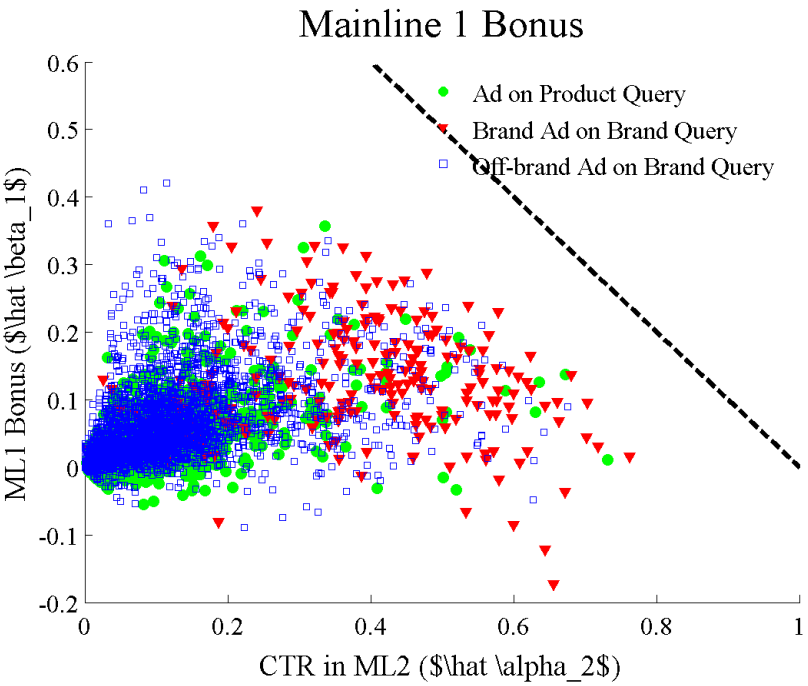


Figure 14: Position effect of moving from the second slot (ML2) to the first slot (ML1) by advertiser. x-axis gives the average CTR over ML1 and ML2. In Panel 1 we show OLS estimates, which are biased. Panel 2 gives lower IV estimates which uncovers a diminishing return of position for dominant ads.

Table 12: Auxillary regressions on $\hat{\beta}_1$ with OLS estimates.

	Brand Query On-Brand Ad			Brand Query Off-Brand Ad			Product Query		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Constant		24.51	33.41		2.90	1.03		1.00	.33
		(1.34)***	(2.59)***		(.08)***	(.22)***		(.09)***	(.26)
CTR in ML2 ($\hat{\alpha}_2$)	.35	-.19	-.21	.47	.32	.29	.37	.31	.31
	(.02)***	(.03)***	(.03)***	(.004)***	(.005)***	(.006)***	(.005)***	(.008)***	(.009)***
Log US Rank			-1.33			.19			.12
			(.32)***			(.03)***			(.03)***
Bounce Rate			5.34			2.58			-.58
			(5.88)			(.54)***			(.58)
N	290	290	276	7,681	8,603	6,455	3,493	3,493	2,462

Table 13: Auxillary regressions on $\hat{\beta}_2 - 4$ with OLS estimates.

		Brand Query			Product Query		
		(1)	(2)	(3)	(1)	(2)	(3)
Mainline 2	Const.		.57 (.01)***	.18 (.04)***		.67 (.03)***	.36 (.09)***
	CTR in ML3 ($\hat{\alpha}_3$)	.31 (.003)***	.24 (.003)***	.22 (.003)***	.25 (.004)***	.16 (.005)***	.14 (.006)***
	Log US Rank			.04 (.006)***			.03 (.01)**
	Bounce Rate			.25 (.11)**			.66 (.19)***
	Ad-Query Pairs	10,707	10,707	8,580	4,232	4,232	2,941
Mainline 3	Const.		.21 (.008)***	.14 (.03)***		.33 (.02)***	.14 (.03)***
	CTR in ML4 ($\hat{\alpha}_4$)	.23 (.003)***	.18 (.003)***	.19 (.003)***	.18 (.004)***	.12 (.005)***	.19 (.003)***
	Log US Rank			.007 (.004)**			.007 (.004)**
	Bounce Rate			.05 (.07)			.05 (.07)
	Ad-Query Pairs	10,523	10,523	8,303	4,117	4,117	8,303
Mainline 4	Constant		.59 (.02)***	.31 (.04)***		1.27 (.04)***	2.17 (.14)***
	CTR in SB ($\hat{\alpha}_5$)	2.74 (.03)***	2.17 (.03)***	3.15 (.04)***	3.87 (.05)***	2.71 (.06)***	2.63 (.07)***
	Log US Rank			.009 (.006)			-.12 (.02)***
	Bounce Rate			.15 (.12)			.62 (.27)**
	Ad-Query Pairs	9,107	9,107	7,111	3,618	3,618	2,493

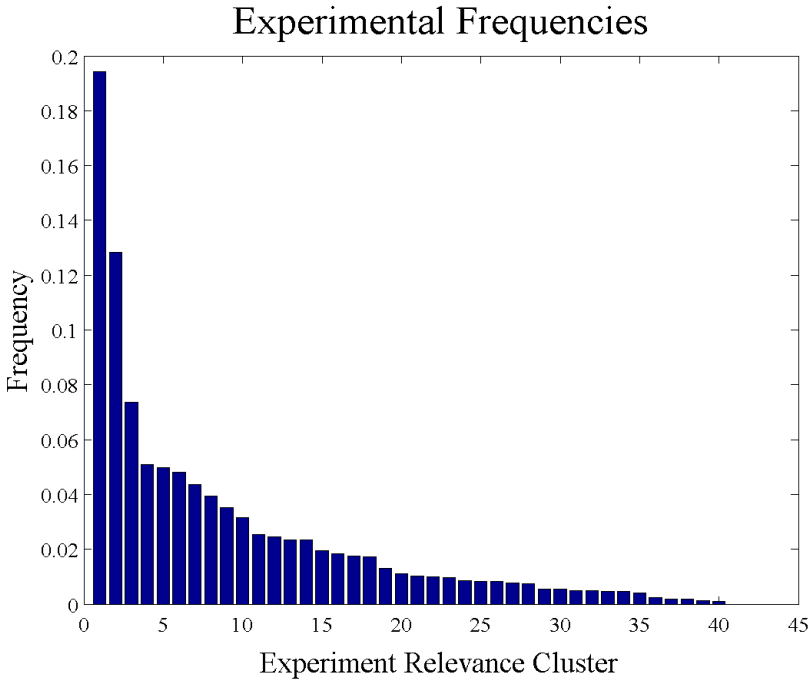


Figure 15: Histogram of experimental frequencies