

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Essays on Economic Behavior in Field Settings

Permalink

<https://escholarship.org/uc/item/2hb1942q>

Author

Goldman, Matthew

Publication Date

2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Essays on Economic Behavior in Field Settings

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Economics

by

Matthew Goldman

Committee in charge:

Professor James Andreoni, Chair
Professor Karsten Hansen
Professor Paul Niehaus
Professor Yixiao Sun
Professor Kenneth Wilbur

2015

Copyright
Matthew Goldman, 2015
All rights reserved.

The dissertation of Matthew Goldman is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2015

DEDICATION

To Roberta Sue and Ronald Lyle Goldman, for sitting through my
baseball games.

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Table of Contents	v
List of Figures	viii
List of Tables	xi
Acknowledgements	xii
Vita	xiii
Abstract of the Dissertation	xiv
Chapter 1 Experiments as Instruments:	
Heterogeneous Position Effects in Sponsored Search Auctions	1
1.1 Introduction	2
1.2 The Generalized Second Price Auction	8
1.2.1 Evolution of Sponsored Search Mechanisms	9
1.2.2 Current Sponsored Search Practice	11
1.2.3 Role of the Click Curve and VCG Equivalence	14
1.3 Data	19
1.3.1 Measures of User Engagement	21
1.3.2 Experimentation	23
1.3.3 Auxiliary Regressors	25
1.4 Estimation	27
1.4.1 Formal Data Generating Process (DGP)	28
1.4.2 Identification	30
1.4.3 Constructing optimal instruments	32
1.4.4 Summarizing “Experiments as Instruments”	38
1.4.5 Alternative Estimators	40
1.5 Results: Click Curves	41
1.5.1 Endogeneity of OLS methods	42
1.5.2 Heterogeneous Click Curves	43
1.5.3 Within Query Heterogeneity	50
1.6 Results: Click Value	52
1.7 Discussion	54
1.7.1 The GSP and Efficiency of Sponsored Search	54

1.7.2	Mechanism Design Choice	56
1.7.3	Experiments as Instruments as a Method	57
1.8	Conclusion	58
1.9	Additional summary tables	59
1.9.1	Auxiliary regressions for positions 2–4	59
1.9.2	Primary and Secondary Clicks	59
1.9.3	Heterogeneous Click Through Rates	61
1.9.4	Alexa Stats	62
1.9.5	Heterogeneous Click Curves with Dummy TSLS Es- timates	62
1.9.6	Heterogeneous Click Curves with OLS Estimates . .	63
1.9.7	Experimental Frequencies	65
	Acknowledgments	68
Chapter 2	Optimal Stopping in the NBA:	
	An Empirical Model of the Miami Heat	69
2.1	Introduction	70
2.1.1	Related Work	74
2.2	Strategic Elements and Data Overview	76
2.2.1	Basketball basics and time/risk neutrality	76
2.2.2	Game log data	77
2.2.3	Structure of a possession	78
2.2.4	Basketball as a stopping problem	80
2.2.5	Sample restriction to the “half-court offense”	81
2.2.6	Lineups vs. individuals	82
2.3	Formal Model	84
2.4	Identification	86
2.4.1	A simple test using average efficiency	87
2.4.2	The shot clock as an instrument to infer opportunity distributions	87
2.4.3	Parametric structural model	94
2.5	Results	96
2.5.1	Sample Restrictions	96
2.5.2	Main results	97
2.5.3	Value of time	104
2.6	Discussion and conclusion	106
2.7	Appendix	108
2.7.1	Parametric Model Specification Tests	108
2.7.2	Computing the win value of time	109
2.7.3	Additional plots and tables	111
	Acknowledgments	113

Chapter 3	Loss Aversion around a Fixed Reference Point in Highly Experi- enced Agents	114
3.1	Introduction	114
3.2	Background and Preliminaries	119
3.3	Theoretical Framework	120
3.3.1	“Standard” Model of a Win-Maximizing Team . . .	121
3.3.2	Reference-Dependent Preferences over Outcomes . .	125
3.3.3	Reference dependent preferences over game state . .	126
3.3.4	Summary of Predictions	127
3.4	Main Results	128
3.4.1	Quantifying loss aversion	129
3.4.2	Fixed vs. quickly adjusting reference point	134
3.5	Additional Results and Robustness	138
3.5.1	Breaking down loss aversion: Effort intensive tasks .	139
3.5.2	Breaking down loss aversion: focus intensive tasks .	141
3.5.3	Do Officials Favor The Losing Team	142
3.6	Conclusion	144
	Acknowledgments	148
Bibliography		149

LIST OF FIGURES

Figure 1.1:	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 side. . . .	9
Figure 1.2:	A multiplicatively separable click curve estimated in Metrikov et al. (2014). Note that the slot effect for the top position is not normalized to 1 as it is common practice.	18
Figure 1.3:	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 represent normalized conversion probability for the twenty ads in our data that convert the most often.	22
Figure 1.4:	Local constant estimation of the average click probability over time for one particular ad in the control (green) and two other experimental conditions.	25
Figure 1.5:	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. The CTRs in the right panel do not have causal interpretation.	26
Figure 1.6:	First stage fits generated by our Dummy TSLS and Optimal TSLS specifications for a hypothetical ad.	37
Figure 1.7:	Heterogeneous position effects for ML1: In (a), the treatment effect of moving from the second slot (ML2) to the first slot (ML1) for each advertiser is plotted against baseline clickability in ML2. . . .	45
Figure 1.8:	A scatterplot of 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.	46
Figure 1.9:	Heterogeneous position 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 $p = .0017$	48
Figure 1.10:	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. . . .	50
Figure 1.11:	Within query analysis	51
Figure 1.12:	Distribution of clicks per impression.	61
Figure 1.13:	Histogram of average CTR at the advertisement level for positions ML1 and ML2. Ads are only included if they have at least 3,000 impressions in that position.	62

Figure 1.14:	Histogram of average CTRs at the advertisement level for positions ML3 and ML3. Ads are only included if they have at least 3,000 impressions in that position.	62
Figure 1.15:	Position effect of moving from ML2 to 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. .	63
Figure 1.16:	OLS estimates of the 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.	64
Figure 1.17:	Histogram of experimental frequencies	66
Figure 2.1:	Shooting hazard by how the possession originates as a function of time remaining on the shot clock.	82
Figure 2.2:	Histograms of t-statistics, negative values indicate overshooting. The thinner dashed line ($t=-1.645$) indicates significance for one lineup. The thicker line ($t=-2.93$) provides a multiple testing correction. . .	88
Figure 2.3:	Illustrating the identification of cut-thresholds under the exogeneity of the shot clock.	90
Figure 2.4:	The left panel computes a simple average of estimated reservation thresholds and continuation values across our thirty lineups. The right panel zooms in on periods 4-12. These pictures reveal no aggregate violations of dynamic efficiency.	98
Figure 2.5:	Chi-square statistics for the joint adherence to dynamic efficiency for each of our thirty lineups (left) and measures of DWL (center) as calculated according to equation (2.7). The right panel gives DWL as a percentage of total value gained by dynamic decision making. .	100
Figure 2.6:	Player-by-player t-statistic for deviations from dynamic optimality. Positive values indicate “under-shooting.”	103
Figure 2.7:	Chi-square statistics for optimal steepness” (left) and a t-statistic for excess flatness (right).	104
Figure 2.8:	Shooting hazard broken down by the valuation a team should place on time.	105
Figure 2.9:	Efficiency residuals from our parametric specification.	108
Figure 2.10:	Scatterplot of the combinations of efficiency and usage observed aggregated across all of our core lineups. The blue brackets give the corresponding 95% CI for efficiency in each shot clock period. The dashed green line shows the linear fit from a GLS regression. . . .	109
Figure 2.11:	Lineup t-statistic for deviations from dynamic optimality scattered against lineup characteristics.	112

Figure 3.1:	The marginal impact of a point on win probability, as a function of score margin and time remaining in the game for the first three quarters (left) and only the final quarter (right).	120
Figure 3.2:	Prediction of the win-maximizing model with symmetric performance function, $\rho = 1 + .2(e_o^{1/2} - e_d^{1/2})$, specified identically for both the home and away team.	123
Figure 3.3:	Predictions of the win-maximizing model with symmetric performance function, $\rho = 1 + .2(e_o^{1/2} - e_d^{1/2})$ and a stronger home preference for victory $U_{win,h} = 1.5 \cdot U_{win,a}$	123
Figure 3.4:	Predicted offensive performance for a win-maximizing team with performance functions for the home and away team (respectively) given by $\rho_H = 1.2 + .3e_o^{1/2} - .2e_d^{1/2}$ and $\rho_A = 1 + .2e_o^{1/2} - .3e_d^{1/2}$	124
Figure 3.5:	Predicted performance premium of a ref-dependent model over the score margin. $\gamma_A=.8, \gamma_B=.9, \lambda = 1, \beta = .8$. $\rho = 1 + .2(e_o^{1/2} - e_d^{1/2})$ is specified identically for each team.	127
Figure 3.6:	Offensive efficiency achieved by “replacement players” across game states as estimated according to (3.7). Holding player quality constant, both home and away teams improve dramatically as they fall behind and the effect is only slightly smaller in the playoffs.	135
Figure 3.7:	Nonparametric estimates of the impact of game state over the last 36 minutes broken up by performance over the first 12.	137
Figure 3.8:	Nonparametric derivative of the estimates in Figure 3.6. Both Home and Away teams appear to show the sharpest changes in relative performance around a score margin of zero.	138
Figure 3.9:	Nonparametric estimates of the impact of game state broken up by the pre-game spread.	139
Figure 3.10:	Parametric projections of win probability conditional on score margin and time remaining for the home team in an even match-up; Panel 2: non-parametric estimates of the same function. Non-parametric estimator is censored where its point-wise standard error exceeds .03 .	146

LIST OF TABLES

Table 1.1:	An example auction illustrating rank score and payment calculation .	14
Table 1.2:	Average user behavior by ad position	20
Table 1.3:	Summary statistics	27
Table 1.4:	Results summary on mainline pairs that could be estimated by all four methods. (S) indicates a simple averaging while, (W) indicates a weighted average with weights inverse to the variance of each estimate.	42
Table 1.5:	Results Summary of mainline pairs that could be estimated by our OLS and Optimal TSLS methods.	44
Table 1.6:	Auxillary regressions on $\hat{\beta}_1$	47
Table 1.7:	ML1-4 Normalized Conversion Bonus ($\hat{\beta}_1$)	53
Table 1.8:	Auxillary regressions on $\hat{\beta}_{2-4}$	60
Table 1.9:	User engagement on primary and secondary clicks	61
Table 1.10:	Correlations	63
Table 1.11:	Auxillary regressions on $\hat{\beta}_1$ with Dummy TSLS estimates.	64
Table 1.12:	Auxillary regressions on $\hat{\beta}_{2-4}$ with Dummy TSLS estimates.	65
Table 1.13:	Auxillary regressions on $\hat{\beta}_1$ with OLS estimates.	66
Table 1.14:	Auxillary regressions on $\hat{\beta}_{2-4}$ with OLS estimates.	67
Table 2.1:	Data Overview	78
Table 2.2:	Possession Usage and Action Frequencies	79
Table 2.3:	Correlation Coefficients	112
Table 3.1:	Total impact of lead on offensive performance (Points/100 possessions)	130
Table 3.2:	Player (only) impact of lead on offensive performance (Points/100 possessions)	133
Table 3.3:	Impact of lagged score margins on offensive performance (Points/100 possessions)	136
Table 3.4:	Impact of game state on Offensive Rebounding Rate (Offensive Rebounds /100 missed shots)	140
Table 3.5:	Impact of game state on Turnover Rate (Turnovers /100 possessions)	140
Table 3.6:	Impact of Game State on Free Throw Percentage (%)	141
Table 3.7:	Impact of Game Type on Free Throw Percentage (%)	142
Table 3.8:	Lineup fixed effect estimates of the impact of lead on offensive foul rate.	143

ACKNOWLEDGEMENTS

All three chapters of this dissertation are coauthored with Justin Rao from Microsoft Research. The first chapter could not have been completed without the extensive support and data provided by Microsoft Corporation. A particular thanks is extended to James Andreoni, David Kaplan and Justin Rao for their wisdom as mentors, goodness as friends and patience as coauthors, without which my development as an economist would have been impossible.

VITA

2008	B. S. in Economics & Mathematics, College of William and Mary
2009-2014	Teaching Assistant –UCSD Department of Economics Undergraduate Program
2011-2012	Teaching Assistant –UCSD Department of Economics PhD Program (<i>Mathematics for Economists</i>)
2012-2014	Research Assistant for Professor James Andreoni
2013	Research Intern at Microsoft Research in New York
2015	Ph. D. in Economics, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Essays on Economic Behavior in Field Settings

by

Matthew Goldman

Doctor of Philosophy in Economics

University of California, San Diego, 2015

Professor James Andreoni, Chair

This dissertation includes three empirical investigations of economic behavior in field settings. The first chapter studies the generalized second price (GSP) auction used to allocate billions of dollars of advertising on web-search platforms. The GSP has a tight link with the favorable properties of the truthful Vickrey-Clarke-Groves mechanism, which hinges on a critical assumption: higher slots must increase click-through-rate by the same scaling factor for all ads. Since position is endogenous, this assumption is largely untested. We develop a novel method to re-purpose internal business experimentation to estimate the causal impact of position for 20,000 search ads. We strongly reject

the multiplicatively-separable model, position effects differ by 100% across ads. This heterogeneity is partially explained by advertiser attributes.

The remaining chapters study the prevalence of standard behavioral biases in the expert population of professional basketball players. In particular, the second chapter considers their adherence to optimal stopping rules. By rule, teams must shoot within 24 seconds of the start of a possession. At each second of the “shot clock,” optimal play requires that a lineup’s reservation shot value equals the continuation value of the possession. Using a structural stopping model, we find that most lineups adopt a reservation threshold that matches the continuation value function very closely. Mistakes we do observe come in the form too low a threshold and excess steepness. Overall, the lineups we study capture 84% of the gains of a dynamic threshold vs. an optimal fixed threshold.

Finally, the third chapter studies how reference dependence and loss aversion motivate effort in this same population. We find a very large “losing motivates” effect, an average team scores like a league leader when trailing by ten points and a bottom dweller leading by ten. Detailed data on players’ actions shows this effect comes through differential exertion of effort. Using betting spreads and lagged score margin, we test if expectations influence effort as would be predicted by any theory with a reference point updating over the course of the game; they do not. The reference point appears remarkably stable, far less malleable than previously found in experimental work studying less experienced agents.

Chapter 1

Experiments as Instruments:

Heterogeneous Position Effects in

Sponsored Search Auctions

1.1 Introduction

Sponsored search links—the paid advertisements on a search engine results page—generate over forty billion dollars annually for search engines. Google, Bing and Yahoo! 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 bids, but also an advertisement’s “clickability” and relevance to the query. In other words, bids are weighted by parameters that must be estimated by the platform. Despite these substantial departures from classical auctions, the seminal works of Edelman, Ostrovsky and Schwarz (2007; henceforth EOS) and Varian (2007) show that under certain assumptions the GSP is payoff equivalent to the dominant strategy Vickrey-Clarke-Groves (VCG) mechanism.¹ 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. By linking payoffs in the potentially problematic GSP, which evolved out of a simple rank-by-bid first price auction designed in the late nineties, to the truthful VCG, a virtual gold standard in mechanism design,² 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

¹The VCG is an auction that induces each bidder to truthfully reveal their private valuation and, as such, has many desirable theoretical properties (Groves, 1973). Its implementation in the context of sponsored search was discussed in EOS (2007) and is described for completeness in Section 2.3 of this paper.

²In package goods auctions, such electromagnetic spectrum licenses, the VCG can have poor revenue properties. However, this can only occur if units in the package are compliments (Ausubel and Milgrom, 2006). In the search context, users are substitutes for each other and bids are not for a package of users, so these concerns do not apply.

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 a given ad in a given position for any candidate 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. These concerns were noted in EOS (2007):⁴

[Our]analysis would have to change considerably if there are specific advertiser-position effects. The magnitude of these specific advertiser-position effects is ultimately an empirical question, and we do not have the kind of data that would allow us to answer it, but judging from the fact that the two major search engines effectively ignore it in their mechanisms (Yahoo! ignores CTRs altogether;⁵ Google computes an advertiser’s estimated CTR conditional on the advertiser attaining the first position), we believe it to be small.

Indeed, estimating *advertiser-specific* position effects was (and continues to be) difficult, even for search engines with access to voluminous data, because an ad’s position is an endogenous outcome of the auction. Ads are ranked higher when they are expected

³This 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.

⁴We note that Hal Varian’s paper had similar caveats and he is Google’s chief economist.

⁵This is no longer true. We discuss the evolution of the GSP mechanism in detail in the next section.

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 a relevant instrument when it induces a different ranking of ads from the “control” for a given query. This only occurs when competing advertisers have bids and clickability metrics that 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 ad types: 1) “on-brand” ads appearing on brand queries (e.g. a Samsung ad on “Samsung smart phone”) 2) “off-brand” ads on brand queries (e.g. a Nokia ad on

“Samsung smart phone”) 3) ads on unbranded product queries (e.g. any ad on “smart phone”). Within each type, 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 (Brunel and Nelson, 2003; Joachims et al., 2005; Kempe and Mahdian, 2008; Lana, 1963) as does work on equilibrium allocation under different models of consumer search (Athey and Ellison, 2011; Jerath et al., 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 on this average are small (less than one percent) and we do not find significant evidence for heterogeneity across advertisers, we conclude that there are unlikely to be any economically significant deviations from the homogeneity of click value assumption. Moreover, the scale of

our analysis allows us to resolve an outstanding inconsistency in the literature, broadly endorsing the finding of Narayanan and Kalyanam (2014).

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 of revenue and advertiser value are likely being lost every year as compared to the 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.

More generally, 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). When bids and payments are instead based on impressions, each advertiser has to determine how impressions map to the relevant intermediate actions, such as clicks, and then finally to sales. By paying for the intermediate action, the contingent-bid model puts the advertiser one step closer to its end goal and can thus provide a competitive edge over traditional media, such as television, where payments are technologically restricted to impressions. However, the efficiency of this model depends on platforms having a better understanding of the impressions-to-clicks mapping than advertisers would in a decentralized setting. This could arise 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). In addition, the efficiency of the “centralized solution” relies on the platform choosing the right mechanism and accurately estimating the relevant quantities. The GSP remains a popular choice to accomplish this goal but 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.⁶

Shifting to a methodological perspective, we show that OLS estimations of the impact of ad position on click probability are biased, even when using advertiser-specific controls for time-related confounds. Both standard Two Stage Least Squares (herein “Dummy TSLS” because experiments are dummy variables in the first stage) and our optimal TSLS procedure estimate click curves that are significantly flatter than OLS, as we expect based on the endogeneity of position. Dummy TSLS produces standard errors about three-fold larger than our preferred Optimal TSLS estimator, and the first stage regression was not viable most of the time. Our smoothing approach to instrument construction was thus crucial for our ability to make useful generalizations. More broadly, “experiments as instruments” has many potential applications since active business experimentation is becoming increasingly common—many firms tally thousands of randomized control trials per year.⁷ We show how this pool can be usefully repurposed for causal inference on applications where direct experimentation is infeasible, or as in this case, financially prohibitive.⁸ In our application, experiment relevance is modulated by time so the ideal solution is to regularize over this continuous determinant of instrument strength. Related approaches, such as Belloni et al. (2012) and Okui (2011), are natural to apply in discrete settings.

The remainder of this paper proceeds as follows: in Section 2 we explain the

⁶Amazon Marketplace and eBay currently take a fixed cut of sales, which simplifies their ranking problem. If they moved to a more general 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.

⁷Firms use field experiments to measure advertising effectiveness (Lewis and Reiley, 2013), improve ranking and recommendation (Craswell et al., 2008; Radlinski et al., 2008) and optimize user interfaces (Kohavi et al. (2012, 2007)). It has evolved beyond simply a measurement tool to become an integral component of market design (Manzi, 2012; Ostrovsky and Schwarz, 2011).

⁸Direct experimentation for our example is done at small scale, but since it results in shuffling the ranking in ways that fundamentally depart from the allocation selected by the auction, it is only done for a small subset of traffic to estimate a global click curve.

rules of the 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, a discussion follows in Section 7 and we briefly conclude in Section 8.

1.2 The Generalized Second Price Auction

A sample search engine results page for a popular commercial query is displayed in Figure 1.1. The page can be 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, all major U.S. search engines showed four mainline ads (and never more than four) on popular commercial queries.

These four ad slots can be thought of as prizes of consecutively diminishing value. In this spirit, they are traditionally referred to by their slot numbers ML1, ML2, ML3, and ML4. Any sponsored search allocation mechanism takes a given query (associated with a given user/time/location) and allocates advertisers to the available slots. The number of slots available is typically determined by the combination of a reserve price and a hard cap (no more than X ads no matter what bids are, $X = 4$ for the period of study). To set context, we’ll now present a short history of the evolution of the sponsored search auction as practiced by the largest search engines.

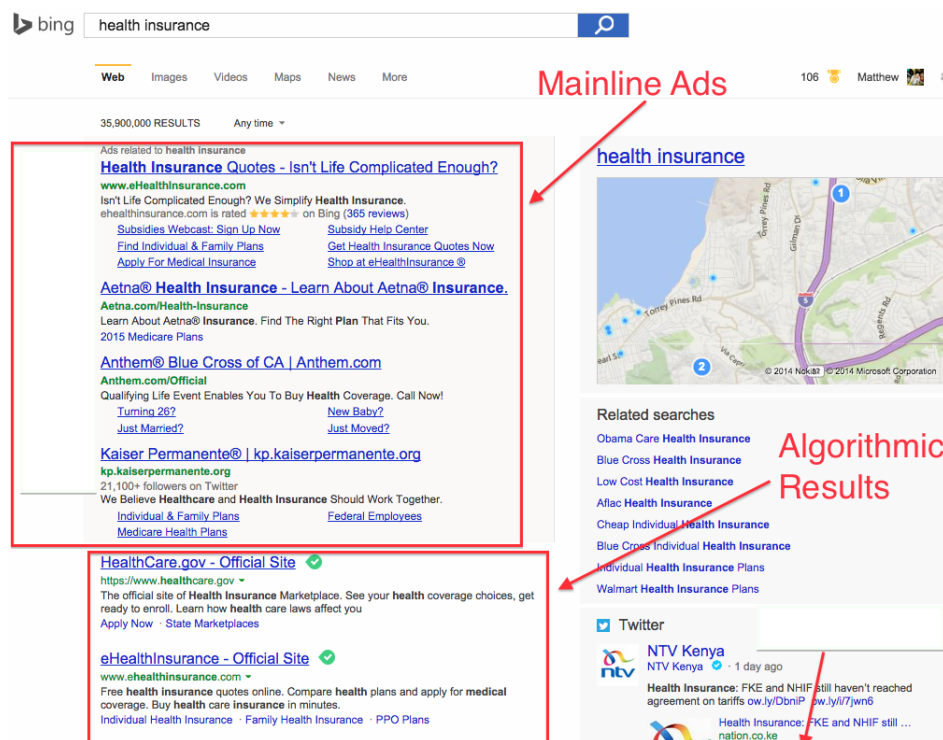


Figure 1.1: An example results page for a popular commercial query. The top 4 listings are mainline advertisements. Below these are the unpaid (algorithmic) results generated by the Bing search engine. Sidebar ads are much smaller and displayed further down the right side.

1.2.1 Evolution of Sponsored Search Mechanisms

The first advertising on the internet came in the form of banner or “display” ads. Following traditional advertising media, these ads were sold by impression (loading the ad on a webpage) and prices were quoted in cost-per-thousand impressions (*cost-per-mille*, CPM). Advertisers specified the web location of the ad, for instance on `aol.com`, and the number of impressions it was interested in purchasing. Purchases were made through negotiated “hand shake deals.”⁹ In 1997, a search engine named GoTo (later known as Overture) introduced pay-per-click pricing and auction-based allocation. Instead of

⁹This model is still used today and has evolved to incorporate targeting dimensions like user demographics, although display ad inventory is now more commonly sold through real-time auctions.

paying each time an ad is shown, payments are only made if a user clicks an advertiser's link. Advertiser's specified willingness to pay per click in the form of bids on specific *keywords*. If one of these keywords appeared in a user's search query, the advertiser's bid was entered into a real-time auction for slots on the page. GoTo used a simple first price auction: advertisers were ranked in descending order of click-contingent bids and if an advertiser received a click, it simply paid its bid. Since auctions for a given keyword were run many times each day, the first price rule naturally led to bid cycling. For example, upon winning the top slot, a bidder could lower its bid in attempt to discover the lowest payment it could make to retain the top slot. Since lower-placed bidders had the same incentive, cycles would spin incessantly (Edelman and Ostrovsky, 2007).¹⁰

Despite these shortcomings, the pay-per-click model proved very popular among advertisers. Overture served the ads for Yahoo!, Microsoft and AOL and was later acquired by Yahoo!. At the turn of the century, the first price, rank-by-bid auction was the dominant, if not only, mechanism in sponsored search. Google entered the sponsored search business in 2002 and licensed patents associated with the Overture model as well, but made two key modifications. First, Google opted for a second price rule to guard against bid cycling: each advertiser's per-click payment is a function of the bid of the advertiser ranked below.¹¹ Second, Google began weighting each bid by the baseline "clickability" of the associated ad. Since the platform is only paid when clicks occur, weighting bids can lead to substantial gains to revenue and advertiser surplus (Lahaie and Pennock, 2007). With these two modifications, the modern *generalized second price auction* (GSP) was born.

¹⁰McAdams and Schwarz (2007) discuss how the socially inefficient investment in bidding agents impacts the auction beyond just instability.

¹¹Yahoo! switched to the second price rule around this time as well, Overture remained on the first price rule. Edelman and Ostrovsky (2007) use this difference to estimate the loss in efficiency of bid cycling. Another difference was that Overture maintained open book of bids (which anyone could examine), whereas Google introduced sealed bids, which Yahoo! quickly adopted as well.

Further validating these innovations, a pair of seminal papers (Varian, 2007; Edelman, Ostovsky and Schwarz, 2007) showed that under certain assumptions and an intuitive equilibrium refinement, all of which are discussed in section 2.3, the GSP efficiently allocates ad slots and achieves at least as much revenue as would be obtained in the dominant strategy equilibrium of the VCG mechanism. These papers were hugely influential within industry and academia. Indeed it is hard to imagine papers having more visibility within industry—Hal Varian was (and is) Google’s chief economist, Michael Schwarz was Yahoo’s lead economist working in marketplace design, and Ben Edelman and Michael Ostrovsky are leading academic economists that have worked closely with major technology companies. Beyond providing valuable assurance with regards to revenue, these papers specified an equilibrium that could be easily computed in simulations, allowing, for instance, calculation of optimal rank score parameters and reserve prices. Shortly after publication, the three major search providers had converged on the GSP auction as the dominant mechanism to allocate and price advertising slots.¹² In 2010, Yahoo! entered into an agreement with Microsoft in which Microsoft would operate their sponsored search marketplace. At the time of this writing, Bing/Yahoo! and Google cover nearly the entire North American search market and employ a version of the GSP that is still closely aligned with the works of Edelman et al. (2007); Varian (2007). In the next subsection, we detail this mechanism.

1.2.2 Current Sponsored Search Practice

At the time of each user-submitted query, the advertisers with bids on matching keywords are entered into a real-time GSP auction.¹³ Each ad i is assigned a “rank score”

¹²Yahoo! made the switch before the actual publication of the papers.

¹³A keyword bid can match a query in a few ways. The advertiser can bid on specific words in a query, known as “exact match.” Alternatively it can bid on specified keywords and allow the search engine to

(s_i) according to

$$s_i \equiv S(b_i, q_i; \alpha), \quad (1.1)$$

where b_i is bid and q_i is an index of ad quality that is generally taken to be an advertiser's "baseline clickability," as estimated by the search engine, but in practice also includes factors such as semantic relevance to the query (EOS, 2007)¹⁴. Finally, α gives ranking parameters set by the search engine. A commonly used rank score formula that captures the most important elements is:

$$s_i = b_i \cdot q_i^\alpha. \quad (1.2)$$

Here, if α is set to 0, the auction is rank-by-bid. If α is set to 1, the ranking is by expected CTR times advertiser bid (thus sometimes referred to as "rank by advertiser value"). The revenue optimal α will depend on the joint distribution of q and b and generally lies somewhere between these extremes (Athey and Nekipelov, 2010; Lahaie and Pennock, 2007).

Ads are ranked in descending order by s_i and up to (but never more than) four ads are shown if their rank score exceeds a reserve value (r^*) set by the auctioneer. High commercial intent queries (like those in our empirical analysis) will always have four advertisers clearing the reserve value. The advertiser with the j^{th} highest rank score is placed in slot j and is assigned a per-click payment (c_j) defined to be the lowest bid possible to maintain a higher rank score than the advertiser in position $j + 1$. That is, we

match the bid to a larger set of queries via a machine learning algorithm, this is known as "broad match." In the case that an advertiser has both an exact and broad match bid, only the exact match bid is used.

¹⁴These additional factors penalize irrelevant ads because they lead to confusion on the behalf of the user, even if they are "clickable." This helps eliminate deceptive ads, because the relevancy includes an analysis of the web address the ad points to.

solve for c_j in:

$$S(c_j, q_j; \alpha) = S(b_{j+1}, q_{j+1}; \alpha).$$

For a specific example, suppose we use the rank score function as given by (1.2), then by construction

$$s_j \equiv b_j \cdot q_j^\alpha > b_{j+1} \cdot q_{j+1}^\alpha \equiv s_{j+1}. \quad (1.3)$$

Since the per-click price is defined as the lowest bid necessary to maintain position, we replace b_j with c_j and set $c_j \cdot q_j^\alpha = b_{j+1} \cdot q_{j+1}^\alpha$. This solves to

$$c_j = b_{j+1} \cdot \left(\frac{q_{j+1}}{q_j} \right)^\alpha \leq b_j,$$

where the inequality follows by rearranging (1.3). If the advertisers have the same baseline clickability, then the price for slot j is simply the bid of the advertiser in slot $j + 1$. Suppose instead that $\alpha = 1$ and the ad in slot j is twice as clickable ($q_j = 2q_{j+1}$), then it will pay half the bid of the advertiser in slot $j + 1$. Finally, we note that for the advertiser in slot 4, cost is set to equalize rank score to the reserve value: $c_4 = \frac{r^*}{q_4^\alpha} \leq b_4$.

More generally, cost-per-click (CPC) is given by the bid necessary to exceed the rank score of the advertisement placed directly below:

$$c_k = S^{-1}(s_{k+1}; q_k, \alpha)$$

$$c_4 = S^{-1}(r^*; q_4, \alpha),$$

where S^{-1} inverts the rank score function for the bid argument. In other words, the payment of the advertiser in slot j is given by the bid of the advertiser in slot $j + 1$, but adjusted for their relative quality, where in practice this adjustment depends on parameters set by the search engine. A complete ranking and pricing example for an auction with $\alpha = 1$ and $r^* = .1$ is shown in Table 1.1. Notice the ranking of ads may

diverge from a simple rank-by-bid and that ads in higher positions often pay less per click; both features are due to the quality adjustment and are commonly observed.

Table 1.1: An example auction illustrating rank score and payment calculation

Advertiser	Bid (b_i)	Quality (q_i)	α	Rank Score (s_i)	Position	CPC (c_i)
Health Insurance Quotes	\$2.25	0.6	1.00	1.35	ML1	\$2.125
Aetna Health Insurance	\$4.25	0.3	1.00	1.275	ML2	\$0.50
Anthem Blue Cross	\$1.00	0.15	1.00	0.15	ML3	\$0.933
Kaiser Permanente	\$1.40	0.1	1.00	0.14	ML4	\$1.00

1.2.3 Role of the Click Curve and VCG Equivalence

The degree to which slots differ from each other is naturally a key determinant of bidding incentives. Let $p_{i,j}$ represent the probability that some fixed population of users will click on advertisement i when it's displayed in slot j . Suppose advertisers A and B have click curves given by

$$\mathbf{p}_A = [0.14, 0.12, 0.10, 0.08] \quad (1.4)$$

$$\mathbf{p}_B = [0.15, 0.14, 0.13, 0.12].$$

Ad position causally influences CTR for both advertisers, but the impact is smaller for advertiser B . In this case, we say that advertiser B has a “flatter click curve.” The slope of an advertiser's click curve determines how differentiated the slots are and thus can be

directly linked to how much it would like to shade its bid in equilibrium.

Let's consider an extreme case to build intuition. Suppose an advertiser has the same CTR regardless of slot (a “flat” click curve). In this case, if she wins the top slot she should immediately lower her bid because slots 2–4 are cheaper and equally valuable. Indeed, this advertiser is best off in either fourth place or off the page entirely. Consistent with this intuition, Gomes and Sweeney (2012) shows that the GSP will not have a pure strategy equilibrium if positions are not differentiated “enough.” In this case, a uniform price auction (5th price auction) is a more natural mechanism since all winning bidders pay the same price.

Since payments to the auctioneer and value to the advertiser only accrue when a click is made, the click curves also determine the revenue and efficiency associated with a given allocation. Specifically,

$$\text{Revenue} = \sum_{j=1}^4 c_j * p_{i,j} \quad (1.5)$$

$$\text{Surplus} = \sum_{j=1}^4 v_{i,j} * p_{i,j}, \quad (1.6)$$

where $v_{i,j}$ denotes the value for advertiser i of receiving a click when his ad is listed in slot j and the vector $\mathbf{v}_i$ may be referred to as the click value curve for advertiser i . Suppose both the click rate curve and click value curve are known. A combinatorial auction like the Vickrey-Clarke-Groves (VCG) mechanism could ensure efficient allocation and incentive compatible behavior by bidders. To see how this would work in practice, we first note that since the mechanism has truthful reporting of values as a (weakly) dominant strategy, we can safely assume access to the click value curves. Advertiser specific click curves would have to be estimated, but suppose this is possible with reasonable precision. We can then maximize per-impression surplus in (1.6) subject to the constraint that each

advertiser occupies at most one slot. The real genius of the VCG is in the pricing rule, which ensures truthful reporting. Each advertiser pays the *externality* imposed on the other advertisers by occupying their allocated slot. This can be computed as follows: 1) take the surplus maximizing ranking with all advertisers and then remove advertiser i and re-optimize the ranking 2) compute the differences in expected per impression payoffs for the other bidders as compared to the original ranking 3) advertiser i pays the sum of these differences.¹⁵

The VCG represents a dramatic departure from the current marketplace because it does not rely on a uni-dimensional ranking. It will be instructive to walk through an alternative, non-combinatorial scheme that ranks by per-search value ($v_{i,j} \cdot p_{i,j}$) slot by slot. That is, compute $v_{i,1} \cdot p_{i,1}$ for each ad i and fill the top slot with the highest ranked advertiser, remove this advertiser from the pool and continue down the line. However, it is easy to come up with examples to show this would not work generally. For instance, recall the example click curves given for advertisers A and B in (1.4) and assume that each advertiser values a click in any slot at \$1.00 (that is, $v_{A,j} = v_{B,j} = 1$ for all j). Advertiser B has the highest per-search value from slot 1, but because of A 's steeper click curve, he has twice the marginal benefit of occupying the first (instead of the second) position and surplus would be maximized by putting A in the top slot and B in the second slot.

More generally, if either the click or click value curve is allowed to vary arbitrarily across advertisers, the ranking problem necessarily possesses a *combinatorial* structure, meaning that any mechanism that defines advertisers with a scalar (e.g. rank score as defined in the previous section) can never be sure to produce an efficient ranking. However, the calculation of (and ranking by) a scalar-valued “rank score” is a core feature

¹⁵Note that payments do not need to be computed at the time the ads are served, which must be done very quickly. Computational techniques can be, and are, used to compute the ranking with a low enough latency to be used in practice.

of the GSP as outlined in the previous subsection. This is why GSP auctions can only have the desirable properties found in Varian (2007) and EOS (2007) under two important assumptions restricting advertiser heterogeneity. The first of these assumptions, the *homogeneity of click value* imposes that for a given advertiser, all slots are valued equally (for any advertiser i and position j , $v_{i,j} = v_i$). The second assumption, the *separability of the click curve* imposes that $p_{i,j}$ is the product of a quality index (q_i) and a multiplicatively separable slot effect (μ_j). That is, for any advertiser i ,

$$p_{i,j} = q_i \cdot \mu_j.$$

The vector $(\mu_1, \mu_2, \mu_3, \mu_4)$ defines a global click curve, an example of which is shown in Figure 1.2. The top slot is typically normalized to 1 ($\mu_1 = 1$). Under this formulation, q_i captures the expected CTR of an ad *if it occupied the top slot*. Of course many ads are not observed in the top slot, so the click curve is used to convert the observed CTRs into this measure of quality.

The separability of the click curve and homogeneity of click value assumptions are maintained both in the theoretical literature and in industrial practice. It is easy to see how they greatly simplify the allocation problem. Advertisers are completely characterized by the scalar $v_i \cdot q_i$ (the product of their value and baseline clickability). An efficient ranking simply assembles advertisers in descending order according to this quantity. Any mechanism that induces advertisers with higher valuations to bid more can produce the efficient ranking. However, candidate mechanisms may have multiple equilibria, suffer from instability or have poor revenue guarantees, so not just any mechanism will be an attractive option to use in practice. The work of EOS (2007) and Varian (2007) is not seminal because it showed the GSP can achieve an efficient ranking—under these assumptions, any sensible auction can achieve this—but rather because they introduced

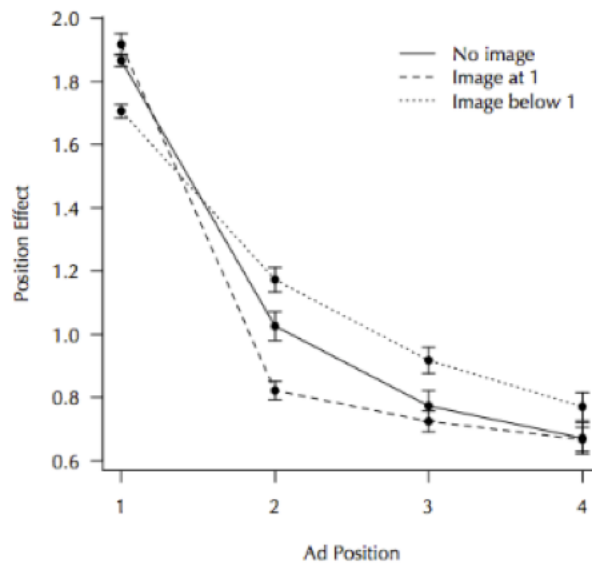


Figure 1.2: A multiplicatively separable click curve estimated in Metrikov et al. (2014). Note that the slot effect for the top position is not normalized to 1 as it is common practice.

an intuitive equilibrium refinement known as the *envy free condition*.¹⁶ Under this refinement, the GSP has unique equilibrium that is payoff equivalent to the dominant strategy equilibrium of the VCG mechanism where advertisers truthfully bid their value. This equivalence provided precisely the revenue and stability properties sought by search engines.

Although our treatment has been brief, hopefully it is clear how the separability of the click curve and homogeneity of click value simplify the allocation and pricing problem. If these assumptions hold, advertisers can be usefully ranked by the product of their valuation on clicks v_i and baseline clickability q_i . Given this single dimension,

¹⁶Interested readers are directed to the original papers for a complete description. Briefly, the envy-free condition requires that an advertiser in slot k will not want to switch to any slot k' and pay the price that the advertiser in k' is currently paying. The standard individual rationality constraint ensures this will be true for slots below k . However for the slot above k , it need not hold. For instance, suppose the advertiser bid the minimum amount to win slot k . In this case, the advertiser in $k - 1$ is paying, all else equal, the same price as slot k , which creates “envy.” This envy is eliminated by bidding higher to the point that the cost of the advertiser above is raised enough so a switch of positions is no longer preferred. Colloquially speaking, advertisers bids aggressively enough to keep the competitor above “honest.”

ranking is relatively straightforward and VCG equivalence in payoffs follows from an intuitive equilibrium refinement. Moreover, baseline clickability can be easily computed by simply adjusting observed CTRs for the multiplicative effect of the occupied position. If these assumptions fail, the GSP (or any other ranking based mechanism) is fundamentally incapable of achieving efficient allocation, there may not exist a single index of baseline clickability and current practice for estimating clickability from observed CTRs is biased.

1.3 Data

Our data covers the outcomes of billions of commercial searches on `bing.com` by American users searching in English from April through July of 2013. We restrict ourselves exclusively to queries with “high commercial intent,” for which four mainline ads were displayed on every query. From this, we draw 20,000 unique ad-query combinations from the top of the revenue distribution for that period.¹⁷ 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

¹⁷An ad-query combination includes all data on which a particular ad was displayed on a particular query. To ensure a representative sample, we classified each ad-query combination by its primary position (ML1, ML2, ML3, ML4, or sidebar) and selected the 4,000 highest revenue ads from each bin.

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 consumers 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 1.2.¹⁸

Table 1.2: Average user behavior by ad position

	Pos 1	Pos 2	Pos 3	Pos 4	Sidebar
<i>Ads shown per Impression</i>	1.00	1.00	1.00	1.00	3.89
<i>Click Through Rate (Normalized)</i>	1.00	0.271	0.168	0.121	0.044
<i>Conversion Rate</i>	0.033	0.039	0.039	0.040	0.051
<i>Average Dwell Time[†] (seconds)</i>	357.72	252.31	237.39	225.30	243.98
<i>Prob. “Infinite” Dwell Time</i>	0.387	0.285	0.263	0.251	0.271
<i>Avg. Share of Revenue (%)</i>	56.257	21.617	12.039	7.888	2.199

[†] Dwell times are averaged conditional on being reported.

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

¹⁸Standard errors are omitted from Table 1.2 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 1.13 and 1.14 in the appendix.

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.

1.3.1 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 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 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.” Dwell time is a noisy measure as there are many possible reasons a user might not return quickly (or at all) to the search page. But on average, 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 1.3 alongside a histogram of the distribution of pooled dwell times (left panel) for scale.

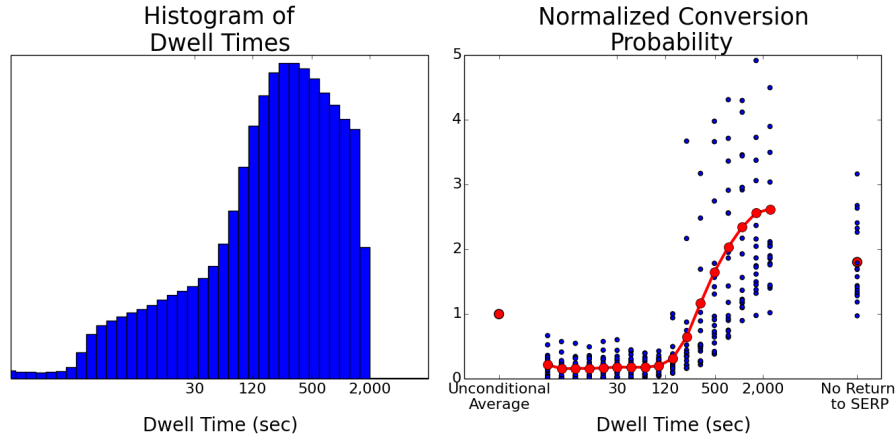


Figure 1.3: The unconditional distribution of dwell time (left panel) and an estimate of the multiplicative mapping from dwell time to click value (right panel). The blue dots represent normalized conversion probability for the twenty ads in our data that convert the most often.

We define *mapped dwell time* as the additional regressor, $md \equiv \phi(d)$. 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.

1.3.2 Experimentation

As is evident from equation (1.5), search engine revenue is directly affected by the parameters of the rank score function (Athey and Ellison, 2011; Lahaie and Pennock, 2007). This motivates constant experimentation with new click prediction, relevance, and ranking algorithms. At a given time, Microsoft’s AdCenter is operating many experiments on the Bing/Yahoo 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 differing proximate reasons, they often 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 chose to 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 algorithmic components of ranking, the search engine also tries out different visual displays of ads, which can directly affect user behavior. Any experiment with different display

parameters cannot be usefully compared to our “control” and can be said to produce endogenous IVs in our causal model. 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 1.4 shows an example where this occurs. The advertisement depicted was exclusively shown in the first position in all three experimental conditions. The experiment represented by the green line gives statistically identical click probabilities to the control, indicating that it may be a valid instrument. However, the experiment displayed in red gives significantly larger click probabilities despite the ad holding the same position, indicating that the ad had different visual properties.¹⁹ This pattern of results indicates that data from the red experiment should be removed from our analysis.

It the last example we used a case where the ad was always ranked in the same position for all experimental conditions. Obviously even the valid experiment would not make an attractive instrument *for this ad* because it has zero first stage relevance by construction. However, it very well could have relevance for other ads. Similarly, the invalid flight might not induce as dramatic a CTR shift for other ads, but is likely invalid globally. Our pruning method relies on the huge number of ad-query pairs available to make these sorts of comparisons. Ads that have stable positions across experiments constitute the test cases for validity. 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 1.4 across all 20,000 of our ad-query pairs, gives us much greater power to prune invalid experiments as compared to standard over-identification tests that must operate within sample.²⁰

¹⁹In this example, the confound in question was additional “deep links” present on the ad.

²⁰More generally, identifying valid instruments in sample is typically a difficult statistical problem

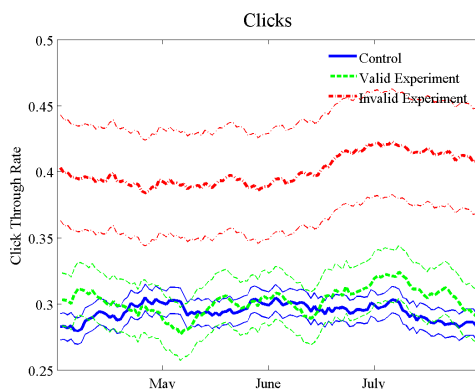


Figure 1.4: Local constant estimation of the average click probability over time for one particular ad in the control (green) and two other experimental conditions.

We prune invalid experiments by identifying excess violations of the requirement that if position remains constant for a given ad on a given query, CTR must remain constant as well. After pruning, we cluster experiments, as denominated in the platform’s system, into a single entity if they produce an identical distributions of ad position across all 20,000 ad-query combinations. In the end we are left with 40 unique experiments. A histogram of this distribution can be found in the Appendix in Figure 1.17.

1.3.3 Auxiliary Regressors

In order to understand heterogeneity in position effects 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 unbranded, such as simply “smart phone.” Within the class of brand queries, we further differentiate

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.

between an “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.²¹ Figure 1.5(a) shows the pooled distribution of each ad type across the four positions. On-brand advertisers occupy the top slot with a very high frequency. Figure 1.5(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. They also show that empirically the higher slots get more clicks for all query types.

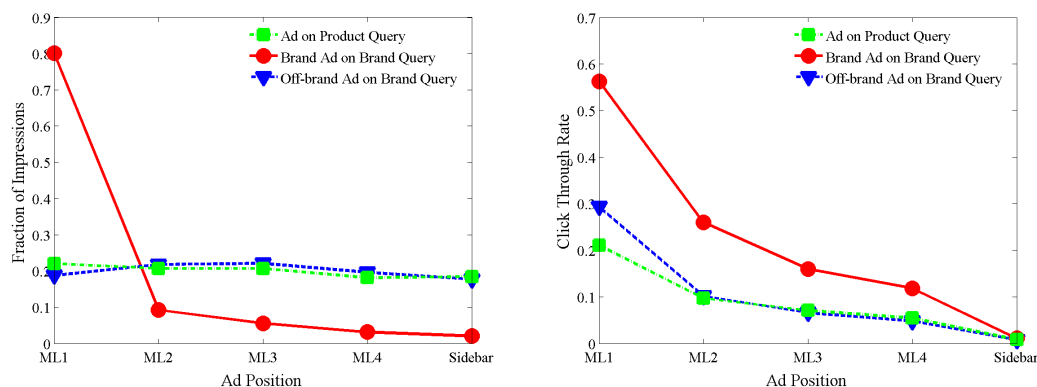


Figure 1.5: Pooled distribution of ad impressions (left) and observed click through rates (right) for ads on product queries, on-brand ads on brand queries, and off-brand ads on brand queries. The CTRs in the right panel do not have causal interpretation.

We also use the destination URL that each ad points to in order to obtain additional features through Alexa.com.²² These additional regressors are intended to characterize the overall popularity and quality of these websites. Of particular interest, are *US Rank*,

²¹Note this procedure will tend to miss brand terms that are not the company’s name or in the web address. This may include, for instance, colloquial terms or marketing slogans.

²²Alexa.com is a subsidiary of Amazon.com that audits and makes public the frequency of visits on various Web sites. Its statistics are based on the observed browsing behavior of a subset of users who have the Alexa toolbar installed.

which gives a rank in overall popularity in terms of the number of unique visitors per day and *Bounce Rate* which gives the fraction of visitors to a given website who are dissatisfied and leave very quickly. Table 1.3 provides a summary of our auxiliary regressors and their correlation matrix is given in the appendix in Table 1.10.

Table 1.3: Summary statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
On-brand Ad	0.03	0.171	0	1	20,000
Off-brand Ad	0.721	0.449	0	1	20,000
Ad on Product Query	0.249	0.433	0	1	20,000
Global rank	776,927.14	2,290,149.06	1	20,421,230	16,623
US rank	39,198.864	77,175.539	1	2,044,261	14,534
Bounce rate	0.359	0.142	0	2.7	14,499
Pages per day	4.269	5.768	1	583	14,499
Sec per day	250.906	168.103	0	5,784	14,499

1.4 Estimation

In this section, we describe our method for identifying and estimating causal parameters of the click curve. Throughout we will do this by focusing on one advertiser (i) and one pair of adjacent ad positions ($j, j + 1$) at a time. Specifically, we will seek to obtain causal estimates using the following linear probability model

$$P(\text{Click}_{i,s} = 1 | \text{Pos}_{i,s} \in \{j, j + 1\}) = \alpha_{i,j+1} + \beta_{i,j} \cdot 1\{\text{Pos}_{i,s} = j\}, \quad (1.7)$$

where s denotes a particular impression, i a particular advertiser, and j gives ad position. $\alpha_{i,j+1}$ is the mean “baseline clickability” for advertisement i when displayed in position $j + 1$ (one slot below) and $\beta_{i,j}$ can be interpreted as the causal impact on click probability of moving an ad up from position $j + 1$ to j .

Importantly, (1.7) is not a structural equation, but rather gives the reduced form

parameters we will seek to estimate. The model is motivated by our interest in separability, which makes the strong prediction that for all advertisers i , the additive causal impact of moving up one slot, β_j , is equal to the baseline CTR times a fixed ratio: $\beta_{i,j} = \frac{\mu_j}{\mu_{j+1}} * \alpha_{i,j+1}$, where the vector μ is a global click curve, like the one presented in Figure 1.2. For example, suppose $\frac{\mu_j}{\mu_{j+1}} = 1.2$, separability requires that position j raises CTR by 20% for all advertisers.

In subsection 4.1 we will formally describe the structure of sponsored search and the resulting data generating process. In subsection 4.2, we demonstrate how this structure and associated experimentation on the platform facilitate the identification of (1.7) in an IV framework. In subsection 4.3 we discuss how to optimally construct instruments. Subsection 4.4 summarizes the entire ‘experiments as instruments’ approach and subsection 4.5 contrasts our IV estimators with more naïve specifications.

1.4.1 Formal Data Generating Process (DGP)

For any particular query, there may be an arbitrarily large number of advertisers who submitted a matching bid. Each advertiser i has the option to specify a unique bid in each geographic region g and to change her bid at any point in time t . Each advertiser also has a quality score that is internally estimated by the search engine. This may be separately estimated in each geographic region and is updated continuously. Neither bid nor quality may be conditioned on any finer level of personalization than geographic region. As such, we conceptualize bid and quality as being given by fully arbitrary functions of g and t .²³

²³Unlike bid, quality score can update continuously over time. For example, suppose, as is commonly observed, bids remained fixed for a given interval of time. If quality scores change in this time period the ranking can get shuffled, which allows one to infer causal effects through a regression discontinuity approach as discussed and applied in Narayanan and Kalyanam (2014).

Additionally, each search s is randomized into a single experimental condition z_s . Let's call the value $z_s = 0$ as corresponding to an impression s being in the “control” or baseline method for ranking ads. z_s may take other discrete values to indicate that impression s is in a different experimental condition. For example, $z_s = 1$ could correspond to impression s being in an experiment which uses a new quality score estimation algorithm. In practice, there are some additional parameters that go into the rank score formula whose value could depend on the experimental condition, z , but for ease of exposition we will allow only the value of α in the rank score formula, the rule for estimating quality (q) and the mainline reserve value (r^*) to depend on z .²⁴ The value of z_s is drawn unconditionally from Φ_{t_s} , where Φ_t is a (potentially) time-varying distribution giving the frequency with which impressions are randomized into different experiments. Importantly, Φ_t cannot vary with the geographic region.

Formally, for a particular search s occurring in geographic region g_s and at time t_s , we have

$$\begin{aligned} z_s &\sim \Phi_{t_s} & z_s &\perp g_s | t_s, \\ b_{i,s} &= \eta_i(g_s, t_s) & q_{i,s} &= \zeta_i(g_s, t_s, z_s), \\ r_{i,s} &= b_{i,s} \cdot q_{i,s}^{\alpha(z_s)}. \end{aligned} \tag{1.8}$$

Ad i will be displayed on search s with $Pos_{i,s} = j$, if and only if $r_{i,s}$ is the j^{th} largest rank score on impression s and is greater than the reserve value $r^*(z_s)$. Since the set of competitor ads is fixed conditional on g and t , we can define Π_i as a deterministic mapping onto the position of ad i on impression s such that

$$\Pi_i(z_s, g_s, t_s) = Pos_{i,s}.$$

²⁴These parameters cover the primary drivers of ranking and thus revenue and advertiser surplus.

Finally, once an ad is placed in position j , we write the probability of receiving a click on search s as

$$\gamma_i(j, g_s, t_s) \equiv P(\text{Click}_{i,s} = 1 | \text{Pos}_{i,s} = j). \quad (1.9)$$

In Section 3, we discussed the fact that certain experiments alter the display of advertisements relative to the control. For these experiments, the value of z_s would enter γ directly, compromising identification in our causal model. For this reason, these experiments were carefully pruned from our data set using the procedure outlined in Section 3.2.

For each advertisement i , we will be interested in identifying baseline click probabilities (in position $j + 1$) and average treatment effects of moving up to position j , given by

$$\begin{aligned} \alpha_{i,j+1} &\equiv \int \gamma_i(j+1, g, t) d\mathcal{F}_{g,t} \\ \beta_{i,j} &\equiv \int [\gamma_i(j, g, t) - \gamma_i(j+1, g, t)] d\mathcal{F}_{g,t}, \end{aligned}$$

where these values are each calculated over some common distribution $\mathcal{F}$ of time periods and geographic locations.

1.4.2 Identification

Given the structure of the previous subsection, we now outline a simple instrumental variable approach to identify values of $\alpha_{i,j+1}$ and $\beta_{i,j}$. For each ad i and impression s , we observe the exact time of the search (t_s), the experiment it is randomized into (z_s), the position of advertisement i on the page ($\text{Pos}_{i,s}$) and whether or not a click occurs ($\text{Click}_{i,s}$). We do *not* observe the geographical information represented by g_s . This means

that identification must be based off of the reduced form objects,

$$\mathcal{X}_{i,j}(z,t) \equiv P(Pos_{i,s} = j | z_s = z, t_s = t) \quad (1.10)$$

$$\mathcal{Y}_i(z,t) \equiv P(Click_{i,s} = 1 | z_s = z, t_s = t),$$

$$\mathcal{Y}_i(z,t,j) \equiv P(Click_{i,s} = j | z_s = z, t_s = t, Pos_{i,s} = j), \quad (1.11)$$

which define distributions of ad position and click probability conditional on observed covariates. For identification, it suffices to suppose the existence of two experiments, z^0 and z^1 , and some interval of time $\mathcal{T}$ that satisfy the following two conditions:

Assumption 1. For all values of g and for all t in some time interval $\mathcal{T}$, impressions in both experiments are mapped either to position j or $j + 1$. That is,

$$\exists \mathcal{T} \text{ s.t. } \forall g \text{ and } \forall t \in \mathcal{T}, \quad \Pi_i(z^0, g, t), \Pi_i(z^1, g, t) \in \{j, j + 1\}.$$

Assumption 2. There must exist at least one value of g such that our two experiments will yield different ad rankings during some sub-interval of $\mathcal{T}$. That is,

$$\exists g^*, \mathcal{T}^* \subset \mathcal{T} \text{ s.t. } \forall t \in \mathcal{T}^*, \quad \Pi_i(z^1, g^*, t) = j \text{ and } \Pi_i(z^0, g^*, t) = j + 1.$$

The first assumption grants us an interval of time during which we can restrict the problem to one with only two levels of treatment without censoring any geographic regions. This ensures that the set of users is identical across experiments. To see why this is important, suppose instead that experiment A put the ad in slot 1 for all queries and experiment B put in slot 2 or off the page entirely. This could occur because the change in quality score meant two closely ranked competitors leap-frogged the ad. In this case, experiment B results in the ad shown in different underlying geographies, which

could be more (or less) interested in clicking on the ad. The second assumption ensures us that we have at least one experiment that can be re-purposed as an instrument that is relevant in at least one geographic region g^* during some subinterval $\mathcal{T}^*$. If assumption 1 was weakened to allow for a multivariate distribution of ad position, identification would still be possible, but would require multiple experiments that had linearly independent effects on the distribution of ad position. We will not explore estimators based on such arguments in this paper.

Under assumptions 1 and 2, the standard logic of local average treatment effects is sufficient to identify the parameters of (1.7). Specifically, we can express our objects of interest as,

$$\begin{aligned}\beta_{i,j}(g^*, \mathcal{T}^*) &\equiv \mu_i(j, g^*, \mathcal{T}^*) - \mu_i(j+1, g^*, \mathcal{T}^*) \\ &= \frac{\mathcal{Y}_i(z^1, \mathcal{T}) - \mathcal{Y}_i(z^0, \mathcal{T})}{\mathcal{X}_{i,j}(z^1, \mathcal{T}) - \mathcal{X}_{i,j}(z^0, \mathcal{T})}, \\ \alpha_{i,j+1}(g^*, \mathcal{T}^*) &\equiv \mu_i(j+1, g^*, \mathcal{T}^*) \\ &= \frac{\mathcal{Y}_{i,j+1}(z^1, \mathcal{T}) \cdot \mathcal{X}_{i,j+1}(z^1, \mathcal{T}) - \mathcal{Y}_{i,j+1}(z^0, \mathcal{T}) \cdot \mathcal{X}_{i,j+1}(z^0, \mathcal{T})}{\mathcal{X}_{i,j+1}(z^1, \mathcal{T}) - \mathcal{X}_{i,j+1}(z^0, \mathcal{T})},\end{aligned}$$

both of which consist exclusively of observed quantities defined in (1.10).

1.4.3 Constructing optimal instruments

The discussion in the previous subsection suggests that accurate inference on the parameters of (1.7) can be achieved in two steps. First, restrict the sample to an interval of time ($\mathcal{T}$) and a series of K experimental conditions ($z^0, z^1, \dots, z^K$) during which ad i was displayed exclusively in positions j and $j+1$. Second, use experimental classification as an instrumental variable for ad position. The first step is straightforward, but the second allows considerable leeway in the construction of instruments. The simplest procedure is

to simply construct dummy variables for experimental classification. This leaves us with K instruments given by

$$Z_{i,s}^k = 1\{z_s = z^k\}, \text{ for } k \in \{1, \dots, K\}, \quad (1.12)$$

or the corresponding K -vector $\mathbf{Z}_{i,s}$. A simple and valid approach is to use these K instruments in a standard two stage least squares (TSLS) framework, while using narrowly bracketed fixed effects to control for the direct impact of time. We will refer to this as *Dummy TSLS*, because it is based off of the dummy instrumental variables defined in (1.12). Formal estimation equations for this are given below.

$$\text{Dummy TSLS: } P(\text{Click}_{i,s} = 1) = \alpha_{i,j+1,t} + \beta_{i,j} \cdot 1\{\text{Pos}_{i,s} = j\}, \quad (1.13)$$

$$\hat{P}(1\{\text{Pos} = j\}) = \gamma_{i,j+1,t} + \mathbf{Z}_{i,s} \cdot \boldsymbol{\pi}_{i,j},$$

where $1\{\text{Pos}_{i,s} = j\}$ is an indicator of the ad being placed in position j (treatment) and α and γ give time-varying fixed effects that are bracketed every four hours, and $\boldsymbol{\pi}_{i,j}$ is the estimated K -vector of first stage coefficients and is not permitted to vary with time. The second equation in (1.13) gives the fitted values of a first-stage regression, which are then plugged into the first equation to produce TSLS estimates.

However, this procedure may (and often does) leave us with a weak instruments problem. As we discussed in the introduction, this occurs because the experiments were not designed to directly change ad ranking and in practice the impact on ranking is not uniform across our time interval $\mathcal{T}$ or geographies, as is essentially implied by (1.13). The ideal instrument is one that always induces a change in position and holds all other factors equal. This effectively replicates a randomized control trial specifically tailored for this purpose of changing an ads rank. Specifically, we would like to have access to the scalar:

$$Z_{i,s}^* = E[1\{Pos_{i,s} = j\} | g_s, t_s, z_s]. \quad (1.14)$$

This instrument would then take a value of 1 only in those combinations of experiment, time period and geographic location that induced the ad to be ranked in slot j . By assumption there are no additional factors on which ad ranking is based, which implies these instruments would perfectly replicate observed ad position. The resulting estimation is thus reduced to a simple comparison of mean click through rates within the regions of geography (g^*) and time (T^*) where experiments were relevant. There would be no need for a first stage regression in this case.

This theoretical optimum reconstructs the ideal randomized control trial. We do not have access to geographic information in this study, as such we can only construct instruments by interacting our experimental classifications with time.²⁵ Again following Carrasco (2012), our optimal instruments are now given by the scalar

$$Z_{i,s}^* = E[1\{Pos_{i,s} = j\} | t_s, z_s]. \quad (1.15)$$

These instruments effectively “turn on” only for time periods where experiments have a contemporaneous impact on the distribution of ad position. Generally speaking, experiments induce a distribution in observed position. Even conditional on time, the distribution is typically not deterministic because it often contemporaneously places ad i in different positions when the user is in different geographic locations—this is reflected in instrument values between 0 and 1. In contrast to the fully optimal formulation, these instruments do not perfectly predict variation in position. However, by conditioning on time they will often be much stronger than the dummy IVs defined in (1.12), which

²⁵In practice there are other factors in addition to geography that we do not have access to.

effectively take the unconditional distribution an experiment induces in position.

We have access to all the information necessary to estimate the instruments in (1.15), but need to ensure that we do so without over-fitting our first stage regression. A variety of regularization approaches exist to select a set of instruments for a good first stage fit, while meeting exactly this challenge. However, most are designed for a large, but fixed, class of candidate instruments (Belloni et al., 2012; Okui, 2011). These approaches are computationally intensive because the discrete interaction terms under consideration (in our case these would be interactions between experimental condition and blocks of time) make the data set very wide.²⁶ An ideally suited solution, which was first suggested in Carrasco (2012), is to instead generate instruments via a nonparametric smoothing of the conditional distribution of ad position for each experimental condition. Similar to the ideal case, these instruments are formulated so that a first stage regression is redundant.

Formally, we estimate our optimal instruments by conditioning on the impression's experimental classification and using a Gaussian kernel to smooth over the time dimension. We will refer to the results of this smoothing as our *estimated optimal instruments* and denote them by $\hat{Z}_{i,s}^*$. The construction requires the selection of a bandwidth for our Gaussian kernel. Choice of bandwidth reflects an important trade off: smaller bandwidths allow more flexibility in the construction of our instruments and may allow us to achieve a stronger first stage fit, but if bandwidths are too small then the estimated values of our instruments may be pulled toward the realized ad position on *a particular impression*, resulting in some endogeneity creeping into our instruments. Intuition can be grasped by considering the two extremes. A bandwidth arbitrarily close to zero estimates our instruments $\hat{Z}_{i,s}$ as exactly equal ad position on that impression ($Pos_{i,s}$)

²⁶Indeed they were not computationally feasible in our setting. The raw data exceeded 3 terabytes and the interaction terms would have added greatly to this.

and thus incorporates the endogeneity we are trying to remove; the resulting second stage estimates will exactly match a naïve OLS procedure. In contrast, a bandwidth of infinity results in very inflexible instruments and weak IV problem analogous to the one encountered in the Dummy TSLS procedure of (1.13).

The best solution to this problem is to use a jack-knife procedure when constructing the optimal instruments. Specifically, we recommend “leaving out” data from impression s when calculating the value of $\hat{Z}_{i,s}^*$. Jack-knife approaches have a long history of improving the properties of TSLS estimators (Angrist et al., 1995) and in our case they obviate any concern of creeping endogeneity and allow us to choose the bandwidth purely in order to maximize the strength of fit in the first stage regression. Operationally, this could be accomplished with a cross-validation approach. However, in our application the jackknife is not computationally feasible. We took what we believe is a conservative approach and opted for one week bandwidth for all estimations. This can be shown to induce a *worst case* bias of less than 5% of the treatment effect for all of the estimators presented herein, which allows us to be confident in the inference of causal effects. Calculation of a worst case bound for this bias is straightforward as it is proportional to the variance of our first stage nonparametric estimator.²⁷

Given the estimated values of $\hat{Z}_{i,s}^*$, we present formal estimation for what we term the *Optimal TSLS* approach:

$$\begin{aligned} \text{Optimal TSLS: } P(\text{Click}_{i,s} = 1) &= \alpha_{i,j+1,t} + \beta_{i,j} \cdot 1\{\text{Pos}_{i,s} = j\}, \\ \hat{P}(1\{\text{Pos} = j\}) &= \hat{Z}_{i,s}^*, \end{aligned} \tag{1.16}$$

In our framework, first stage fit is determined by the construction of the instrument

²⁷Future versions of the paper will include an appropriate expansion of our estimator and proof of this result.

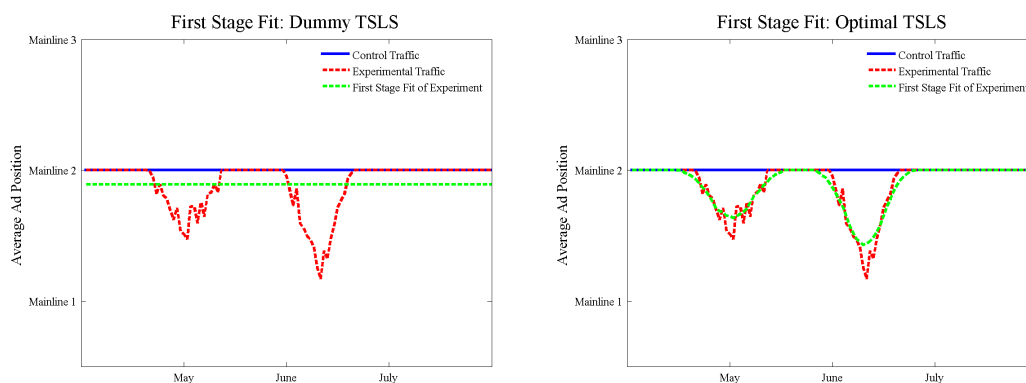


Figure 1.6: First stage fits generated by our Dummy TSLS and Optimal TSLS specifications for a hypothetical ad.

itself. An example of how the first stage of our optimal instruments compares to our Dummy TSLS specification is shown in Figure 1.6. Two experimental conditions are depicted, the control, in blue, and an experiment, in red. The experiment tends to promote the ad to the first position, whereas the control keeps it fixed in position two. The green lines give the fitted instrument values. In the left panel, it is clear that Dummy TSLS effectively “flattens” the experiment across the entire time period under study. It is easy to see that this is equivalent to using an infinite bandwidth in the smoothing procedure. For the example shown, this effectively adds noise to the estimates from time periods during early May and July when our experiments did not impact ranking. In the right panel, we smooth over time as described above. Our optimal instruments improve the first stage fit by allowing us to sharpen the focus on ranking experiments when they actually occur.

1.4.4 Summarizing “Experiments as Instruments”

In this subsection, we briefly summarize the steps discussed in Sections 3 and 4 that were necessary for us to effectively re-purpose the search engine’s internal experimentation in order to achieve identification of our desired causal effects.

1. *Identify the causal model.*

We are interested in the impact of a change in ad position on click through rates. We chose to estimate the linear probability model in (1.7) in order to obtain estimates of both baseline clickability (in a lower slot, $j + 1$) and a treatment effect of moving up (to the higher slot, j). This model is without any loss of generality and is ideal for testing our main hypothesis of interest: separability of the click curve.

2. *Prune invalid instruments from the data set.*

Some experiments on the search platform directly impact the visual properties of advertisements. These experiments may impact user click decisions even without changing the ranking of ads and thus would result in endogenous IVs in our causal model. Generally, it may be possible to prune such experiments if all of their parameters are clearly and carefully documented. In our case, we were able to effectively pretest our data for invalid experiments by searching across all 20,000 ad-query pairs in our data and checking to see if a given experiment was able to shift that ad’s click probability (relative to the control) even when it did not induce a shift in ranking. This was discussed formally in Section 3.2.

3. *Focus the data on natural experiments.*

Consistent with assumption 1 in our identification argument, we pruned data from each advertisement to focus only on time periods where the “control” experiment

($z_s = 0$) ranked a given ad i exclusively in positions j and $j + 1$. Then we restricted our sample to the control and to other experiments which also had this same bivariate ranking of the ad. Thus our restricted sample was one in which only two levels of treatment need to be modeled and this could be done without concern that we were selectively censoring impressions from certain geographic regions.

4. *Estimate optimal instruments.*

The best possible instruments for causal inference are ones that match ad position as closely as possible without being endogenous to the ad's ranking on that particular impression. Our approach is to follow the advice of Carrasco (2012) by non-parametrically estimating these optimal instruments as detailed in Section 4.3. This is not a necessary step for causal inference, but allows for much greater power in the parameter estimates for each advertiser. Generally, we suggest a jackknife style procedure for estimating optimal instruments. This obviates any concern of creeping endogeneity and allows one to tailor the nonparametric estimation exclusively to maximize first stage fit and, in turn, power. When a jackknife procedure is not computationally feasible (as in our setting), one must be careful not to choose a very small bandwidth or risk some degree of endogeneity bias in the resulting estimates.

5. *Run Two Stage Least Squares.*

This part of the approach is completely standard. We note only that it is critical to include fixed effects for time in the second state regression.

1.4.5 Alternative Estimators

In addition to the TSLS estimators discussed above, we will also present naïve estimators of the click curve based on OLS methods. We will do this to illuminate the degree of endogeneity in the ranking of ads and motivate our methods as a necessary alternative.

First we consider a straightforward OLS methodology. Here, the estimating equation is

$$OLS: \quad P(\text{Click}_{i,s} = 1) = \alpha_{i,j+1} + \beta_{i,j} \cdot 1\{\text{Pos}_{i,s} = j\}. \quad (1.17)$$

This estimator simply pools all the data for a particular mainline combination. It improves on the summary statistics of Table 1.3 by removing confounds that vary across advertisers, but it is still vulnerable to time-varying shocks to click probability that are correlated with ad position. Recalling the structural functions defined in subsection 4.1, we could say that an OLS methodology would be valid if either of the following were true: (1) neither geography (g) nor time (t) has any impact on the click probability function μ or (2) neither geography (g) nor time (t) has an impact on ad position (Π). Neither one of these assumptions is particularly plausible. In fact, one may expect that advertisers are placed in better ad positions during time periods (or when searched in geographic regions) where they are especially clickable. This could happen mechanically through the search engines estimation of quality (q) or through an advertiser's changing bid. In both case we expect values of β_j estimated from (1.17) to be overestimated, which we call a *steepness bias* of the click curve.

We also consider a second specification that includes time fixed effects at the four-hour level. These are identical to the fixed effects used in the TSLS models of the

previous section. The formal estimation equation is

$$\text{Time FE: } P(\text{Click}_{i,s} = 1) = \alpha_{i,j+1,t} + \beta_{i,j} \cdot 1\{\text{Pos}_{i,s} = j\}, \quad (1.18)$$

where the addition of the fixed effects (the t subscript on α) is the only modification. This specification should effectively control for any shocks to time-varying clickability and to an extent mitigate the steepness bias in (1.17). However, we may still be confounded if geography is impacting both clickability (μ) and ranking (Π). As discussed above, this type of endogeneity is still likely to lead to a steepness bias. Thus, we may expect a smaller, but still positive, steepness bias in estimates of equation (1.18).

1.5 Results: Click Curves

As explained in the last section, our goal is to identify the causal impact of moving up a slot to position j . Our estimating equations produce the additive CTR bonus from moving up one slot for a given advertiser on a given query. For example, $\beta_{i,1}$ captures the additive impact on CTR of moving from slot 2 to slot 1 for query-ad pair i . Taking the estimated $\hat{\alpha}_{i,2}$ as the baseline, we can also convert this additive factor into the more familiar multiplicative factor. Ideally we would like to estimate all four position factors for each of the 20,000 ad-query pairs in our data. However, many advertisements do not spend much time in certain positions—when the effects of interest are not properly identified we do not report them. Additionally, our optimal instruments framework can only be used when there is sufficient first stage relevance. The same applies for Dummy TSLS and we’ll see the coverage for this estimator is about one-third that of optimal instruments.

1.5.1 Endogeneity of OLS methods

In order to get a baseline sense of the extent to which endogeneity confounds our naïve estimators, we summarize the estimation results for those queries that could be estimated by all four techniques in Table 1.4. Treatment effects are summarized for each position by their simple and weighted average and are presented as percentage points.²⁸

Table 1.4: Results summary on mainline pairs that could be estimated by all four methods. (S) indicates a simple averaging while, (W) indicates a weighted average with weights inverse to the variance of each estimate.

	ML 1 Bonus		ML 2 Bonus		ML 3 Bonus		ML 4 Bonus	
	(S)	(W)	(S)	(W)	(S)	(W)	(S)	(W)
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
<i>OLS</i>	7.65 (0.03)	6.73 (0.04)	2.15 (0.01)	1.60 (0.01)	0.99 (0.01)	0.62 (0.01)	2.84 (0.01)	1.68 (0.01)
<i>Time FE</i>	7.48 (0.06)	6.60 (0.05)	2.08 (0.03)	1.56 (0.02)	0.90 (0.02)	0.57 (0.01)	2.81 (0.02)	1.66 (0.01)
<i>Dummy</i>	7.57	6.50	1.84	1.42	0.68	0.47	2.68	1.53
<i>TSLs</i>	(0.21)	(0.15)	(0.15)	(0.09)	(0.14)	(0.07)	(0.10)	(0.04)
<i>Optimal</i>	7.69	6.58	1.94	1.44	0.67	0.45	2.73	1.58
<i>TSLs</i>	(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 those ads in which the Dummy TSLs model has standard errors < 5%.

†† Weights chosen identically across estimations and to be most friendly to IV.

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

²⁸The 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.

and fourth rows, we again see a general decline in average treatment effects, but this time the effect is concentrated on the less popular ads at the bottom of the page. In combination, these results strongly suggest that naïve 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 when and where they are most popular. 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 drop the Dummy TSLS method, the most restrictive method in terms of coverage, from the comparison. This allows us to 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 1.5 and is consistent with those found above. The fact that the number of estimates triples is a testament to the superior efficiency of the Optimal TSLS approach.

1.5.2 Heterogeneous Click Curves

Given the bias in OLS and noisiness of the Dummy TSLS, we focus attention on the estimates derived from our Optimal TSLS method 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. This is a natural starting point since most of the revenue in this auction comes from the first slot. Figure 1.7 plots the additive impact on CTR of moving to slot 1 on the y-axis against CTR in the second mainline position on the x-axis. The graphical presentation is motivated by our interest in the hypothesis of *separability of the*

Table 1.5: Results Summary of mainline pairs that could be estimated by our OLS and Optimal TSLS methods.

	ML 1 Bonus		ML 2 Bonus		ML 3 Bonus		ML 4 Bonus	
	(S)	(W)	(S)	(W)	(S)	(W)	(S)	(W)
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
<i>OLS</i>	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)
<i>Time FE</i>	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)
<i>Optimal TSLS</i>	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 those ads in which the Optimal TSLS model has standard errors < 5%.

††Weights chosen identically across estimations and to be most friendly to Optimal TSLS.

click curve. Separability implies that all these points should lie on a single ray from the origin, with noise due to sampling variation alone. An initial glance at the figure reveals this is not the case. In Panel B we compute the implied multiplicative factors to show that there is substantial heterogeneity that cannot be explained by sampling variation.

As previously discussed, we further explore heterogeneity by identifying three key types of ads. On-brand ads on brand queries (e.g. Apple’s ad on “Apple laptop”), off-brand ads on brand queries (e.g. Sony’s ad on “Apple laptop”) and unbranded product queries (e.g. all ads on “laptop”). For each type we compute inverse variance-weighted GLS lines of best fit for each group. Specifically, we regress $\hat{\beta}_{i,j}$ onto $\hat{\alpha}_{i,j+1}$ with a weight inverse to the estimate standard error of $\hat{\beta}_{i,j}$.²⁹ Separability has the strong implication that the causal impact of position is the same for all three types. We can strongly reject this is the case. First, the types of ads have clearly distinct slopes, with off-brand ads having the

²⁹We note that $\hat{\alpha}_{i,j}$ is a generated regressor in the sense that it is estimated with some variability from a prior model. This tends to bias down the multiplicative click curves estimated in Table 1.6. However, this bias can be calculated and is proportional to the sample variance of $\hat{\alpha}_{i,j}$, which tends to be quite small. Future versions of this paper will include a correction (or at least bound) for this bias.

steepest position effect and on-brand ads having the flattest effect. The overwhelming statistical significance of this difference is confirmed by the differences in the coefficient estimates for $\hat{\alpha}_2$ obtained in specification (1) of Table 1.6. 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 1.6 for on-brand ads.

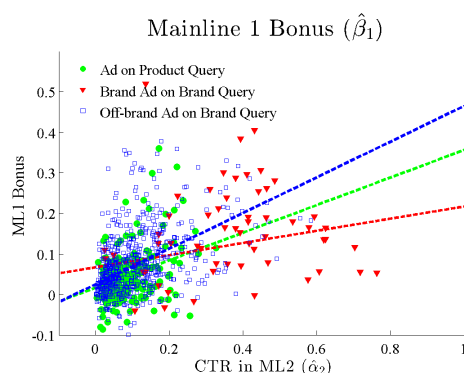


Figure 1.7: Heterogeneous position effects for ML1: In (a), the treatment effect of moving from the second slot (ML2) to the first slot (ML1) for each advertiser is plotted against baseline clickability in ML2.

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 1.8, 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.

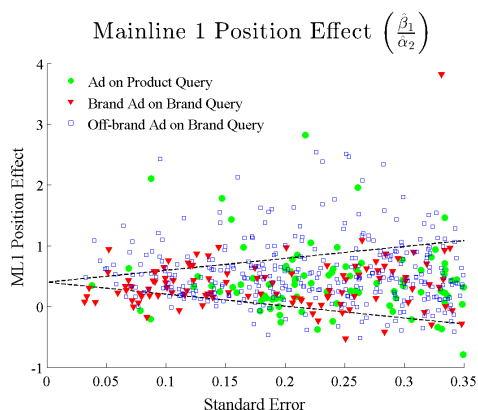


Figure 1.8: A scatterplot of the implied multiplicative effect for each ad against the corresponding standard error. Dots outside the black lines are those that are statistically different from the group level average at the 5% level.

Three important observations are worth noting about Figure 1.8. First, a group of on-brand advertisers in the bottom left have a tightly estimated small or zero effect of position. Other advertisers along the bottom of the graph also do not have a substantial position effect. These bidders prefer a lower slot, since they can play less per click and get roughly the same amount of clicks. This provides an incentive to dramatically shade bid from true valuation and if two (or more) of these advertisers are present, cycling will occur. Both create problems for the GSP. Bid shading removes price support for advertisers ranked above and makes it difficult to recover valuations from bids. Cycling creates instability and inefficient investment. We note that the focal advertiser in Narayanan and Kalyanam (2014) has estimated position effects of roughly zero and appears to be in this minority. Second, advertisers towards the top of the figure benefit tremendously from position. Uniform position normalization will incorrectly estimate these advertisers' quality depending on which position they are most commonly observed in. An accurate inference of quality is essential for efficient allocation. Additionally, based on GSP pricing, overestimating quality results in a lower CPC than that based on

Table 1.6: 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 (2.23)***	4.83 (5.30)		2.49 (.21)***	.49 (.57)		1.62 (.54)***	-2.88 (1.30)**
CTR in ML2 ($\hat{\alpha}_2$)	.30 (.03)***	.15 (.06)***	.16 (.06)***	.62 (.02)***	.44 (.02)***	.37 (.02)***	.45 (.03)***	.34 (.05)***	.33 (.05)***
Log US Rank			-.25 (.67)			.36 (.08)***			1.03 (.17)***
Bounce Rate			11.78 (10.66)			.66 (1.29)			-10.38 (2.82)***
Ad-Query Pairs	91	91	90	1,242	1,242	931	412	412	305

auction logic. Finally, it is clear that bidding incentives can vary dramatically across advertisers, meaning the simple treatment of the GSP fundamentally fails to capture its complexity in practice.

Additional clues as to what is causing the excess heterogeneity can be gleaned from specification 3 in Table 1.6. All else equal, the ML1 treatment effect is smaller for product ads and off-brand ads that have a rank closer to the top of the distribution (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 within the industry 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

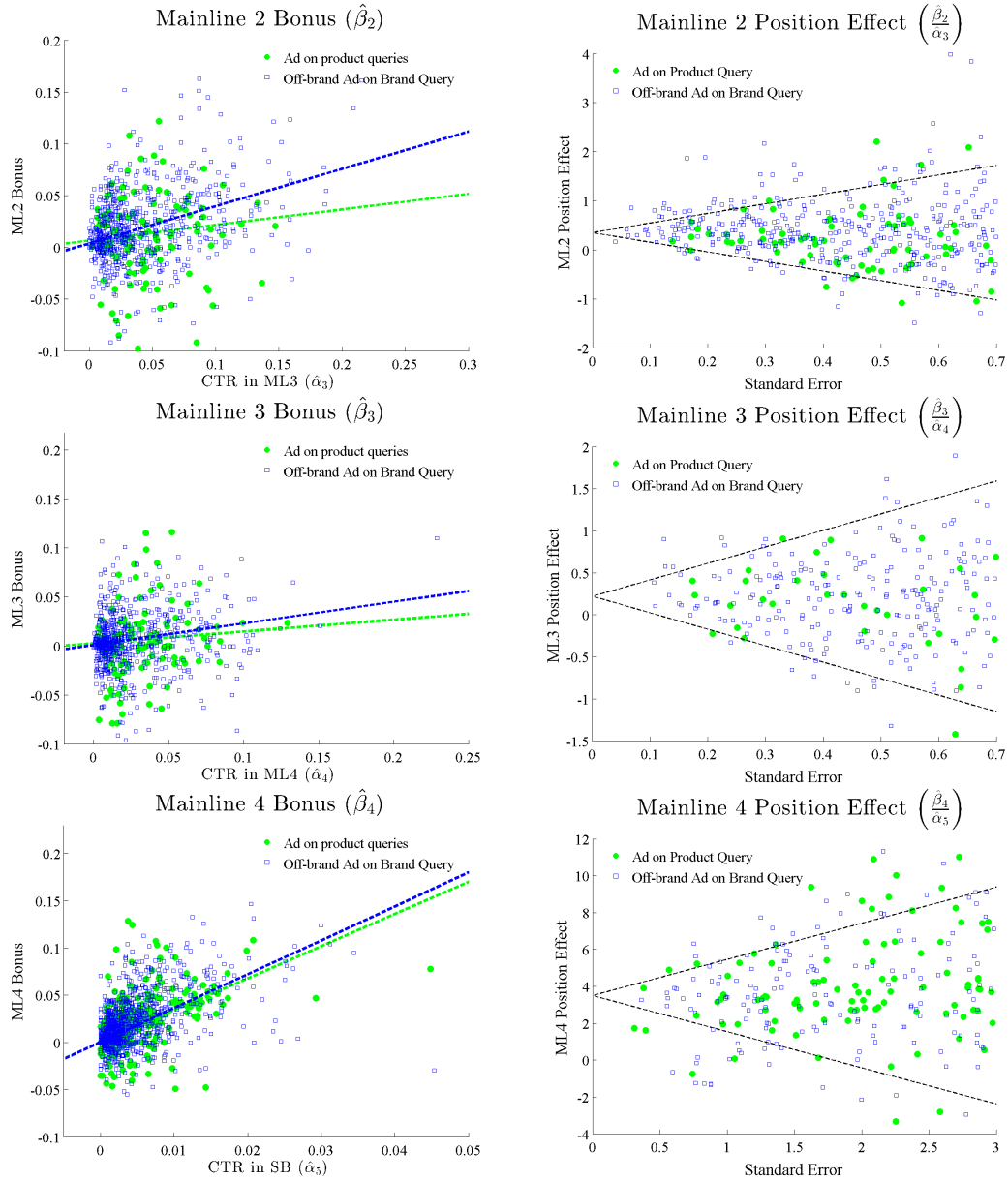


Figure 1.9: Heterogeneous position effects for ML2-ML4. This was only close in the case of the Mainline 3 position effect. Here $238/264 = .9015$ dots were within the black lines. This fraction is distinct from .95 with $p = .0017$.

interest based on engagement metrics from Alexa.com.

We now report the same estimates for positions 2–4. As was shown in Figure 1.5,

on-brand ads were rarely observed in lower positions on the page and thus we are unable to identify the rest of the click curve for these type of ads. Our analysis for the remaining positions will focus only on ads on product queries and off-brand ads on brand queries. Figure 1.9 collects scatterplots of these treatment effects for each of these 3 positions in analog to the results of Figures 1.7 and 1.8 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).

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 1.10. They are an analog to the global click curve presented in Figure 1.2, but allow a separately estimated curve for each ad type. 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.

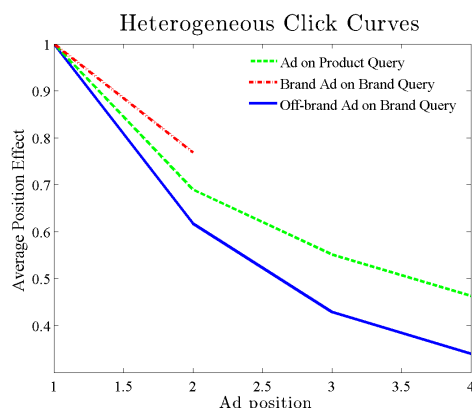


Figure 1.10: Average position effect by ad type. On-brand ads were very rarely observed outside of the first two positions. Reliable estimates of their full click curve were difficult to obtain and are not presented.

1.5.3 Within Query Heterogeneity

We now turn our attention to a formal demonstration of click curve heterogeneity *within particular queries*. One could imagine that a multiplicatively separable click curve might approximately hold for all advertisers participating in the auction on a particular query string, but that differences in the shape of this object *across queries* have driven our results on nonseparability thus far. If this were the case, then the GSP could easily be salvaged by estimating a unique click curve for the advertisers on each query. Unfortunately, this does not appear to be so.

We test this on 32 branded query strings for which we were able to separately estimate (with a standard error of less than 100%) average multiplicative position effects for participating on- and off-brand advertisers as they move from the second mainline position to the first. The availability of estimates for multiple advertisers on these 32 queries is a fortuitous byproduct of our original analysis and we suspect that a more targeted approach could yield an expanded population of suitable queries.

Regardless, these 32 show a strong and significant pattern of greater position

effects for off-brand advertisers. Specifically, results are displayed in Figure 1.11. The scatter-plot on the left shows the weighted average position effect for off-brand advertisements (x-axis) and on-brand advertisements (y-axis) on these 32 branded queries. The presence of a separable click curve for advertisers on each query would predict that all of these dots lie close to the 45° line. Instead, we find that the average position effect in this sample for off-brand advertisers 121% compared to only 47% for on-brand advertisers. The paired t-statistic for this difference in means is $t = 3.97$ ($p < .001$). Additionally, for 25 of our 32 queries, the corresponding dot lies below the 45° line indicating that the ML1 position effect is larger. As such, a simple sign test also allows us to reject that the median difference between on-brand and off-brand position effects is zero ($p < .001$).

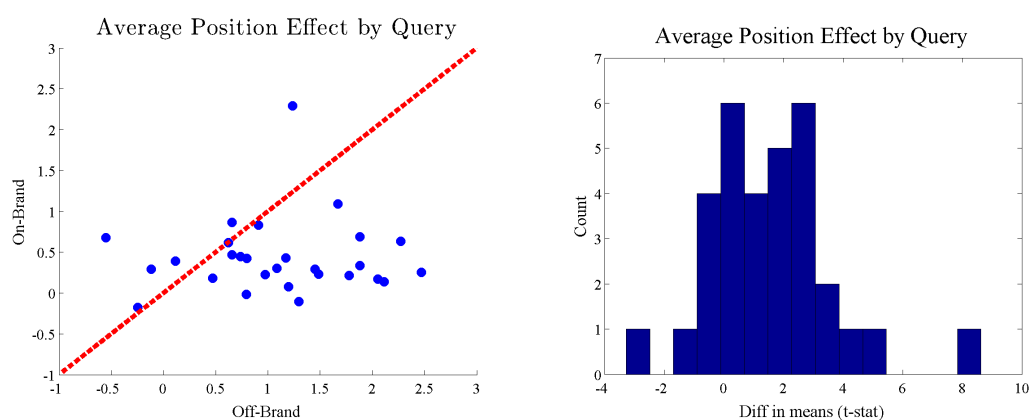


Figure 1.11: Within query analysis

Finally, we use the standard errors for each of our estimated position effects to perform a separate t-test for the equivalence of on- and off-brand position effects on that query. The right panel of Figure 1.11 shows a histogram of the 32 resulting t-statistics, where a positive t-statistic indicates a larger position effect for off-brand advertisers. We find that 12 of these 32 queries have position effects that are significantly ($p < .05$) larger for off-brand advertisers and only one query has a position effect that is significantly

larger for on-brand advertisers.

Taken together, these results show a strong pattern of advertiser heterogeneity in position effects within brand queries.

1.6 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 to represent a proportional change in an advertiser's 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 1.7. 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.

Table 1.7: ML1-4 Normalized Conversion Bonus ($\hat{\beta}_1$)

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

We examine the impact of moving from the sidebar to the mainline ads in the final 4 columns Table 1.7, 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.³⁰ 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 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.

³⁰Note 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.

1.7 Discussion

We'll now discuss the important implications of our findings and potential extensions of our methods.

1.7.1 The GSP and Efficiency of Sponsored Search

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 1.6, 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 close to 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.³¹ Expanding the analysis to: include all four slots, account for additional elements of heterogeneity, or allow for strategic bidding based on this misallocation could lead to a greater calculated revenue impact.

³¹Since 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.

Eliminating these losses requires the incorporation of heterogeneous click curves into the mechanism. There are two hurdles to do so. The first is an empirical challenge—advertiser specific position effects must be estimated. 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. Additionally, we suggest that platforms carefully track all data relevant to their ranking algorithm in order to get as much power as possible out of experimentation. In Section 4, we showed how the missing data on the geographic location of a particular impression limited the power of our estimates for any given advertiser.

The second challenge is a theoretical one. All current equilibrium analysis that we are aware of 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 1.2) 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. A transition from a mechanism like the GSP where bid shading is optimal to one where values should be truthfully reported, like the VCG, can be daunting for a platform. The risk is that bidders won't immediately understand the change and there will be a large revenue drop in the transition period.

1.7.2 Mechanism Design Choice

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 impressions-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. These “centralized solutions” (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

carefully estimated and modeled 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. While the GSP may be simpler to explain to bidders who are used to strategic shading, the flexibility of the VCG has clear advantages as well. Based on these considerations, selection of an optimal mechanism is an empirical question and should also hinge on a behavioral analysis of bidder sophistication.

1.7.3 Experiments as Instruments as a Method

Finally, this paper is the first of its kind to address a new methodological problem: How can large-scale business experimentation be re-purposed for causal inference on treatments beyond the original intent of the experiment? No part of the econometric theory applied here is brand new, but usefully applying existing ideas to this setting required a number of non-standard steps. We detailed these steps in Section 4.4 and described in detail throughout the paper. The two most significant pieces are: (1) the pruning of experiments that would procure invalid instruments via large scale over-identification testing; (2) the use of nonparametric smoothing to convert experimental designation into “optimal” instruments that allow us to focus on natural experiments on our treatment of interest (ad positions) whenever they are induced by the firm's intentional experimentation. The existing econometric literature on both of these topics was not designed with applications like ours in mind and that additional work on the properties and practical implementation of these techniques in analogous big data settings, especially with respect to computational constraints, could prove very useful.

We believe the experiments as instruments toolkit could be valuable in a range

of alternative scenarios, especially given how widespread business experimentation has become. For example, suppose the object of interest was some aspect of preferences for consumers in an online marketplace such as Amazon or eBay. The platform would presumably like to measure price sensitivity, but direct experimentation on pricing is often difficult for a variety of reasons such as public relation risks from consumers who may view them as “unfair,” or the risk of being “gamed” by strategic consumers arbitraging in secondary markets. However, ongoing experimentation with an algorithm that selects the listed products could provide useful variation in the “price choices” faced by a consumer. For example, Amazon often makes multiple versions of the same product available for purchase through different retailers. However, only one of these retailers is assigned the “Buy Box.” The listed price of the overall product is then given by this retailer’s price and a consumer selecting this product will see the retailer with the “Buy Box” displayed most prominently. Presumably the “Buy Box” is assigned based on a host of considerations such as the price offered, the rating of the retailer, Amazon’s own inventory holdings of the product, estimated shipping times and so forth. Thus a number of parameters could be introduced to the “Buy Box” algorithm and experimentation on these parameters could produce exogenous variation in the relative prices shown to the consumer.

1.8 Conclusion

This paper approaches the novel methodological problem of re-purposing large scale online experimentation for causal inference on alternative parameters. In doing so, we are able to demonstrate a broad and substantial deviation from typically maintained assumptions within the extensive theoretical research on sponsored search auctions.

The seminal papers of EOS (2007) and Varian (2007) show clearly that under these assumptions, all “envy-free” equilibria of the GSP obtain at least the revenue of the VCG mechanism.³² 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 sample, could provide limited evidence on separability.³³ 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.

1.9 Additional summary tables

1.9.1 Auxiliary regressions for positions 2–4

1.9.2 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

³²Relaxing this assumption, Lucier et al. (2012) shows that all Bayes-Nash equilibria still obtain at least half the revenue of the VCG.

³³Narayanan and Kalyanam (2014) is also able to reject separability within the three competing advertisers in their study. Specifically, they find that the most popular advertiser faces a statistically insignificant return to ad position, while less popular advertisers on the same query received significant increases in click probability from moving up the page. This is consistent with the broader pattern of results presented here.

Table 1.8: Auxillary regressions on $\hat{\beta}_{2-4}$.

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

to the appropriate advertiser. However, as demonstrated in Figure 1.12 this is relatively rare.

However, it is interesting to note that these *secondary* clicks appear to be considerably less valuable as shown in Table 1.9.

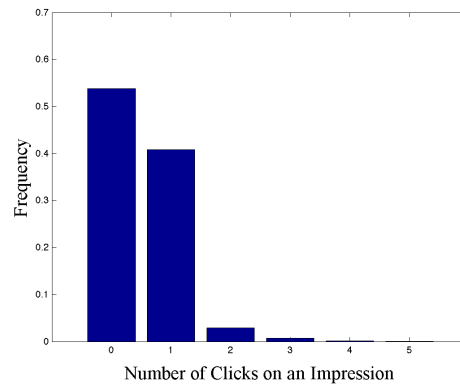


Figure 1.12: Distribution of clicks per impression.

Table 1.9: User engagement on primary and secondary clicks

Metric	Click	ML1	ML2	ML3	ML4	Sidebar
<i>Probability of Conversion</i>	Primary	0.030	0.040	0.041	0.041	0.050
	Secondary	0.027	0.029	0.030	0.031	0.035
<i>Average Dwell Time</i>	Primary	321.188	234.135	224.334	215.307	228.467
	Secondary	199.261	164.150	148.236	147.539	176.131
<i>Probability of Inf. Dwell Time</i>	Primary	0.400	0.293	0.273	0.262	0.280
	Secondary	0.267	0.209	0.177	0.176	0.211
<i>Mapped Dwell Time</i>	Primary	1.197	0.956	0.915	0.886	0.923
	Secondary	0.851	0.720	0.655	0.647	0.725

1.9.3 Heterogeneous Click Through Rates

The histograms in Figure 1.13 and 1.14 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.

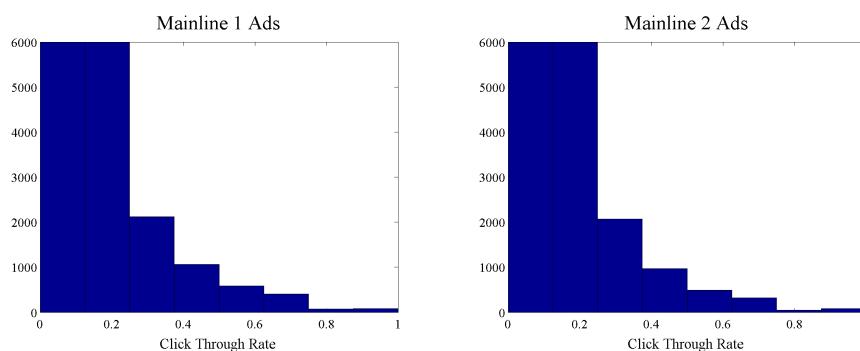


Figure 1.13: Histogram of average CTR at the advertisement level for positions ML1 and ML2. Ads are only included if they have at least 3,000 impressions in that position.

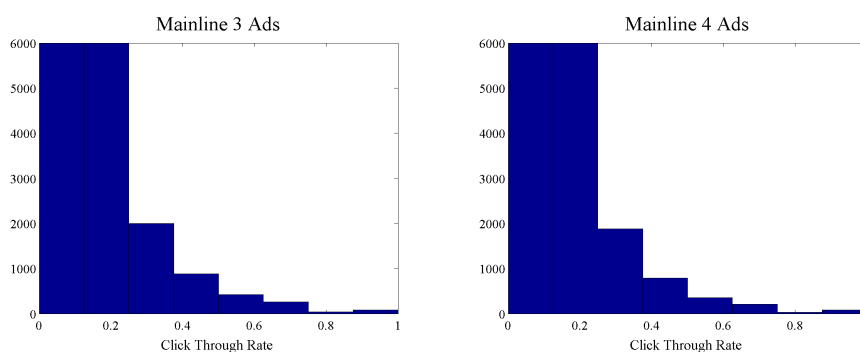


Figure 1.14: Histogram of average CTRs at the advertisement level for positions ML3 and ML3. Ads are only included if they have at least 3,000 impressions in that position.

1.9.4 Alexa Stats

Table 1.10 demonstrates some correlations between our different website level metrics from Alexa.com.

1.9.5 Heterogeneous Click Curves with Dummy TSLS Estimates

We now replicate the analysis of Section 4.2 with our OLS estimates derived from equation (1.17). Results are presented in Figure 1.15— and Tables 1.11 and 1.12.

Table 1.10: 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

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

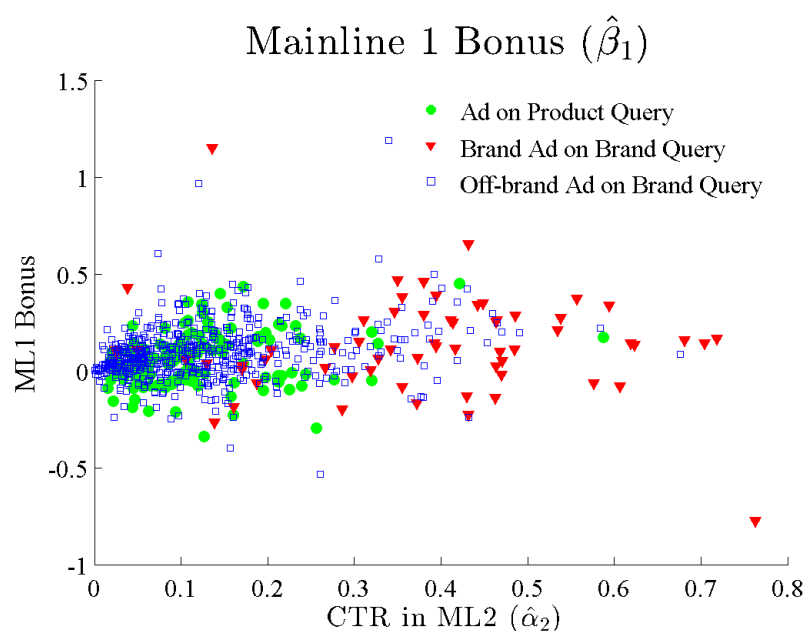


Figure 1.15: Position effect of moving from ML2 to ML1 by advertiser. x-axis gives the average CTR over ML1 and ML2. In Panel 1 we show OLS estimates, which are biased. Panel 2 gives lower IV estimates which uncovers a diminishing return of position for dominant ads.

1.9.6 Heterogeneous Click Curves with OLS Estimates

We now replicate the analysis of Section 4.2 with our OLS estimates derived from equation (1.17). Results are presented in Figure 1.16— and Tables 1.13 and 1.14.

Table 1.11: 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 (5.62)	10.92 (12.56)		1.67 (.31)***	.22 (.90)		2.02 (1.29)	-4.98 (3.76)
CTR in ML2 ($\hat{\alpha}_2$)	.37 (.07)***	.30 (.14)**	.32 (.14)**	.65 (.03)***	.53 (.03)***	.47 (.04)***	.49 (.07)***	.38 (.10)***	.40 (.11)***
Log US Rank			-2.45 (1.84)			.39 (.13)***			1.65 (.53)***
Bounce rate			34.96 (23.97)			-1.41 (1.95)			-18.88 (7.36)**
Ad-Query Pairs	40	40	40	380	455	345	112	112	85

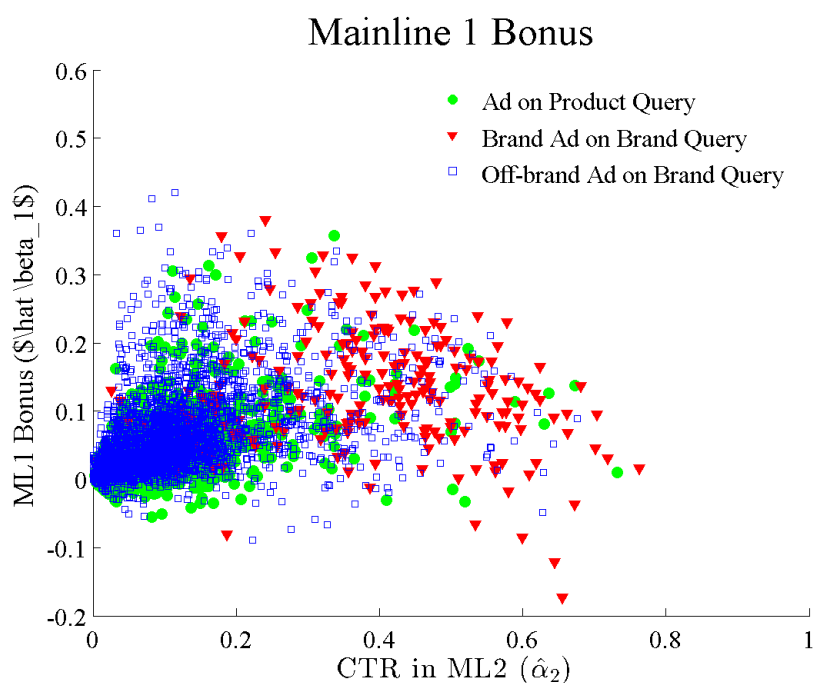
**Figure 1.16:** OLS estimates of the position effect of moving from the second slot (ML2) to the first slot (ML1) by advertiser. x-axis gives the average CTR over ML1 and ML2.

Table 1.12: 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

1.9.7 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 1.17.

Table 1.13: 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 (1.34)***	33.41 (2.59)***		2.90 (.08)***	1.03 (.22)***		1.00 (.09)***	.33 (.26)
CTR in ML2 ($\hat{\alpha}_2$)	.35 (.02)***	-.19 (.03)***	-.21 (.03)***	.47 (.004)***	.32 (.005)***	.29 (.006)***	.37 (.005)***	.31 (.008)***	.31 (.009)***
Log US Rank			-1.33 (.32)***			.19 (.03)***			.12 (.03)***
Bounce Rate			5.34 (5.88)			2.58 (.54)***			-.58 (.58)
N	290	290	276	7,681	8,603	6,455	3,493	3,493	2,462

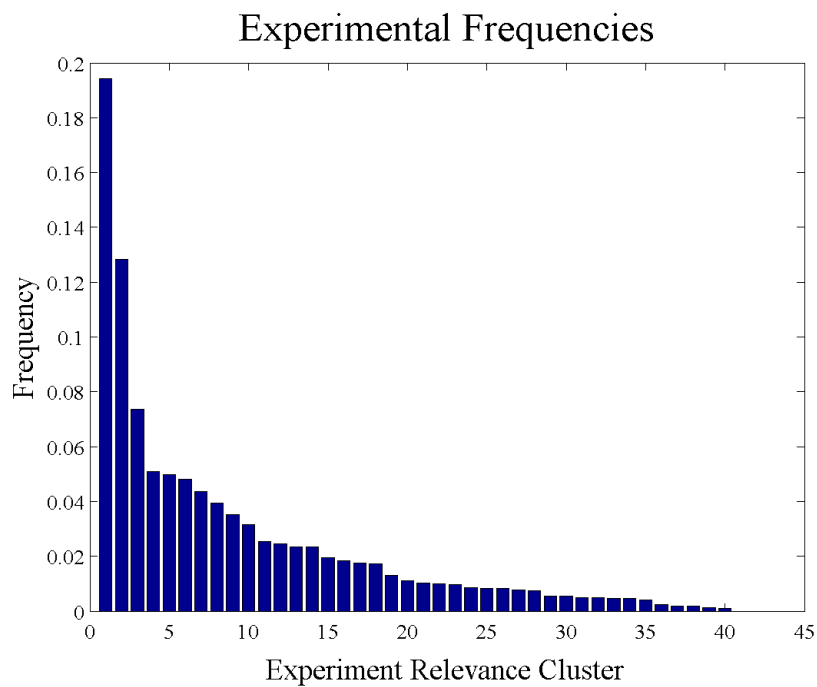
**Figure 1.17:** Histogram of experimental frequencies

Table 1.14: 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

ACKNOWLEDGMENTS

The preceding chapter of this dissertation is coauthored with Justin Rao from Microsoft Research and could not have been completed without the extensive support and data provided by Microsoft Corporation. This paper is being prepared for publication.

Chapter 2

Optimal Stopping in the NBA:

An Empirical Model of the Miami Heat

2.1 Introduction

In the world's foremost professional basketball league, the National Basketball Association (NBA), each time a team possesses the ball they are allocated 24 seconds to take a shot. Failing to do so results in zero points and forfeiture of the ball to the other team. This rule makes optimal stopping a key strategic element of basketball. The offense should use a reservation threshold, based on the time remaining on the shot clock, that determines if a scoring opportunity should be taken or passed up in favor of continuing the possession. The logic of this choice, comparing a current opportunity to the option value of continuing a dynamic interaction, underlies the theoretical requirements of rational decision making in many contexts, such as repeated games and bargaining, as well as real world business decisions, like whether to terminate workers (Arlotto et al., 2013). But it also happens to be a component of decision making that laboratory subjects generally struggle with (see for instance, Binmore et al. 2002; Lee 2006; Ochs and Roth 1989). This naturally raises the question of whether people can adhere to the predictions of economic theory for decision problems involving these types of trade-offs. Since NBA players are paid millions of dollars a year and face the shot clock stopping problem thousands of times per season, they form an attractive population on which to investigate this question.

We do so using a rich data set of 1.3 million possessions in the NBA from 2006–2010. The data includes the players on the court, the timing and location of shots and other key actions/outcomes. We model the behavior of five offensive teammates (a “lineup”) facing the decision to “use” the current “scoring opportunity”¹ at a given point in the shot clock versus continuing the possession in hopes of observing a better

¹“Using a scoring opportunity” can mean taking a shot or it can mean making an aggressive play that may lead to either a turnover or a shot.

opportunity before the shot clock expires. A lineup's *reservation threshold* specifies, for each point of the shot clock, the worst shot players are willing to take as compared to continuing the possession. This threshold can also be thought of as a team's marginal (or worst) shot at a given point of the shot clock. Optimal play requires that the reservation threshold is set equal to the *continuation value* of the possession. That is, a shot should only be taken if its expected point value exceeds the expected, forward-looking value of continuing the possession. We refer to this standard as *dynamic efficiency*.

Like all stopping problems in the field—see Rust's (1984) seminal paper for a discussion—the key challenge is to compute the “marginal counterfactual.” In our setting, we have to estimate a lineup's shot *opportunity* distribution based on the shots we *actually observe*. Knowledge of the opportunity distribution tells us what would happen if a team shot more or less by adjusting the reservation threshold, or in other words, allows us to model the marginal shot. Our identification scheme, which we detail in Section 4, exploits variation in shooting frequencies across the shot clock to simultaneously illuminate each lineup's opportunity distribution and the chosen reservation threshold for using opportunities at each point of the shot clock. Nonparametric identification of these reservation values is characterized in Theorem 2, but our main estimation relies on a parametric model of the opportunity distribution.

To ensure adequate sample sizes, we focus on the most commonly played lineup for each of the thirty NBA franchises in our data set, noting that this selection leave us with a team's better players who have considerably more experience playing together than a randomly selected lineup. We find that these lineups are quite adept at solving the stopping problem they face each possession. A typical lineup adopts a reservation threshold, throughout the entire range of the shot clock, very close to the possession's continuation value. Impressively, the chosen reservation values match both the level and

shape of the non-linear continuation value function.

We do, however, observe statistically significant deviations from optimality. To calibrate the size of these mistakes we simulate decision making using a fixed threshold. Specifically, we restrict each lineup to set a single reservation threshold for all periods of the shot clock, except the last period where we allow them to always shoot (this avoids an unrealistically high number of shot clock violations, giving us a meaningful baseline). In comparison to this optimal “fixed” threshold, the lineups we study capture 84% of the gains from dynamic efficiency and all but a single lineup capture at least 66%. Moreover, we estimate that switching from an optimal dynamic threshold to an optimal fixed threshold would cost a team three wins per season—an impact that is certainly valued at millions of dollars and far more for teams on the cusp of qualifying for the post-season or those in title contention.

Possible mistakes come in two flavors. The first is a level shift away from optimality, indicating a fixed propensity to be too eager or too reluctant to shoot for the entire range of the shot clock. The second is an incorrect slope, indicating under- or over-responsiveness to time remaining. We find statistically significant evidence of both classes of mistakes, but always in the same direction. When the level is off, the threshold is too low. When the slope is off, the threshold is too steep. Both mistakes push a lineup towards shooting too frequently (“impatience”). We correlate “excess shots” (the result of a mistakenly downward shifted reservation threshold curve) with the number of possessions a lineup played together and find a strong relationship. The lineups with the most playing experience (with each other) show the strongest adherence to optimality, while lineups with less shared experience tend to be impatient. Since coaches allocate playing timing based on a lineup’s performance, it is unclear whether this finding represents learning or selection. In either case, it is interesting that lineups

that make better decisions tend to play more together.

An important caveat is worth mentioning. While we model shot selection as a group (lineup) decision problem, in reality the player holding the ball is the proximate decision maker. A natural concern is that players may have an incentive shoot more frequently than is optimal for the team (by using too low a reservation threshold) in order to increase individual point output. If a player's points-per-game average is rewarded in salaries without an eye to scoring efficiency, then we'd expect this to be a major concern. NBA teams, however, typically employ many statisticians to evaluate player quality and shooting percentage and related efficiency metrics get substantial attention in the media. Nonetheless, since any failure of group incentives likely pushes players to selfish play—overshooting—there is some uncertainty as to whether the overshooting we observe (in some lineups) reflects a failure of dynamic optimization or a failure of incentives.

Our final piece of analysis yields an additional test of optimality. Toward the end of the game, the trailing team should place a positive value on time because they need more possessions to catch-up (the leading team has the opposite incentive). In our main analysis we are careful to remove these cases where time has intrinsic value. When time-invariance fails, an obvious implication is that the shooting hazard rate should shift up for the trailing and down for the leading team as compared to the time-invariant baseline. A more subtle requirement is that as the shot clock nears zero, the hazards should converge to the time-neutral baseline—with less and less time remaining, the amount of “time saved” by taking an early shot converges to zero, while the cost of sacrificing quality on marginal shots is constant throughout a possession. Indeed, we observe that teams respect both of these additional requirements of optimal play.

2.1.1 Related Work

There is a deep experimental literature investigating subjects' ability to effectively use continuation values. The earliest papers in this literature studied the “classical secretary problem” in which a subject observes a series of “applicants” and wins a prize if she “hires” the best one. A robust finding is that subjects “under-search” by hiring a candidate too quickly (Kahan et al., 1967; Rapoport and Tversky, 1970). Recent work that has relaxed many assumptions present in the classical formulation has also found early stopping (Bearden et al., 2005, 2006; Seale and Rapoport, 2000; Zwick et al., 2003). Most relevant, when payments are simply the chosen draw, like realizing a scoring opportunity in basketball, the solution requires a reservation threshold that declines monotonically with draws remaining; yet the majority of subjects use flat thresholds (Lee et al., 2004), even when the time horizon is short (Lee, 2006).²

Continuation values are important in a wider class of repeated games and subjects typically struggle to respond to them correctly. In alternating offer bargaining games, offers do not respect equilibrium continuation values (Ochs and Roth, 1989), a pattern that has been linked to a “limited lookahead” strategy (Johnson et al., 2002). In the repeated prisoner's dilemma, subjects tend to be less responsive to continuation probability than theory predicts, typically leaving surplus from cooperation on the table (Murnighan and Roth, 1983; Palfrey and Rosenthal, 1994).

Perhaps the most closely related paper, Romer (2006), looks at football coaches' decisions to “play” versus “punt” on 4th down in the NFL. Romer finds that coaches punt too often, implicitly undervaluing the continuation value of the possession. The key difference between Romer's study and ours is that his subjects are the coaches making

²With a very long time horizon, a high constant threshold does quite well since one will get a good draw with high likelihood. The short horizon games are a better analog to basketball.

deliberative decisions that occur infrequently (only 10–20 times per season per team).³ In contrast, we study quick decisions made by the players over the course of thousands of possessions played each year. Indeed, all the past empirical work (which we are aware of) on stopping problems outside of the laboratory has focused on deliberative decisions—those that are typically made with the help of computer software—such as harvesting trees (Provencher, 1997), renewing patents (Pakes, 1986) and replacing bus engines (Rust, 1987). These studies have generally found adherence to theoretical predictions. To the best of our knowledge, our study is the first evidence that individuals can reliably solve stopping problems of this level of complexity “on the fly.”

Finally there is a literature that generally uses the observability of actions in professional sports as a lens into predictive power of economic theory. Mixed-strategy Nash Equilibrium (MSNE) has been tested in the field using 2×2 simultaneous move games analogous to matching pennies. In soccer penalty kicks (Chiappori et al., 2002; Palacios-Huerta, 2003) and tennis serves (Hsu et al., 2007; Walker and Wooders, 2001), players typically randomize across actions consistent with minimax play. Our paper shares the same broad motivation for studying professional athletes as these papers, but focuses on a different decision problem.⁴

The remainder of the paper proceeds as follows. Section 2 provides background and describes the data, Section 3 presents the model, Section 4 describes identification, Section 5 gives the results and Section 6 concludes.

³A team faces many more 4th downs than ten per year but for the vast majority there is no real tension between punting and playing.

⁴Authors have also used professional sports to study behavioral theories such as loss aversion Berger and Pope (2011); Pope and Schweitzer (2011).

2.2 Strategic Elements and Data Overview

In this section, we give an overview of the strategic elements in the game of basketball, as it is played in the NBA, and describe the data used in the study.

2.2.1 Basketball basics and time/risk neutrality

In a NBA game, two teams face each other in a 48-minute contest. The team scoring the most points wins, with ties broken by a 5-minute overtime period. Five players from each team, referred to as the “lineup,” share the court at any given time. Points are scored by shooting the ball in the basket. Long-range shots earn three points, normal shots earn two points and shots taken after a foul, unguarded 15-foot attempts known as “free throws,” earn one point each (a player is awarded the number of free throws corresponding to the value of the shot he was fouled on). In our sample period, teams scored an average of 92.5 points per game and the median margin of victory was 9 points.

A game is effectively broken up into a series of non-overlapping “offensive possessions.” Each possession starts when one team secures the ball (this team is then referred to as the offense), at which point the 24-second shot clock begins counting down. If the offense fails to attempt a shot within the 24-second limit, then the ball is automatically turned over to the defense, resulting in zero points. The shot clock is prominently displayed above each basket. Since the offense generally faces the near basket and the defense generally faces the offense, meaning their backs are to the basket, the offense can more easily observe the shot clock than the defense.

In practice, each team gets around ninety possessions per game. The large number of possessions and high frequency of scoring (roughly half of all possessions

result in some points for the offense) are crucial features of basketball that make structural modeling and estimation feasible. We formally show in the Appendix that for the majority of possessions teams are best off behaving like a risk-neutral point maximizer. Inducing a policy that scores fewer points in expectation in order to save/burn time, or to avoid/seek risk, allows the offense to sacrifice mean output for a favorable shift in variance. However, with many possessions (N) remaining, the shift in mean can be shown to have an $O(N)$ impact on win probability as compared to the $O(\sqrt{N})$ impact of a variance shift. For large N , this is an unattractive tradeoff. Toward the end of games, or when the point disparity is large, this logic does not go through and we are careful to remove such cases from our main analysis.

For the remainder of our data, we assert that the offense should be neutral to both time and risk considerations. This allows us to tractably model the optimal stopping problem faced each possession because it implies: 1) a team only cares about expected points 2) for two shots that have the same ex-post value, the team is indifferent as to when they were taken over the course of the shot clock.

2.2.2 Game log data

Our data consists of play-by-play “game logs” from four NBA seasons (2006-2010).⁵ The game logs, taken from a feed provided by the NBA, capture the on court play through the measures presented in Table 2.1. In total, we have 1,376,893 possessions in our sample.

Because the quality of scoring opportunities depends critically on the personnel in the game, the bulk of our estimation is conducted separately by offensive “lineup.”

⁵Approximately 100 games (out of the 4,920 played during this time period) are missing from this data set. We have no reason to believe their omission is anything other than random.

Table 2.1: Data Overview

Event/Action	Description and Level of Granularity
Offensive/defensive lineup	Players on the court for each team
Game time	Minutes and second of each event
Shot clock value	Calculated from game time and action logs
Shooter	Player/time of the action
Rebound, assist, block	Player/time of the action
Foul	Shooting, non-shooting, flagrant, illegal defense
x,y coordinate of shot	Physical location of a shot attempt
Turnover	Bad pass, dribbling error, charge, lost ball

To ensure adequate power, we restrict the analysis to the lineup that played the most possessions for each of the thirty NBA franchises. Each of these is observed for at least 892 (and as many as 7,784) total possessions.

While we do look at heterogeneity in adherence to optimal play across different lineups, it is important to note that our sample restriction selects better players who have more experience playing together than a randomly selected NBA lineup. Lineups consisting of many inexperienced players who share the court infrequently typically only occur when the outcome of the game is effectively determined (a “blowout”) and thus suffer from not only small sample problems but also a lack of meaningful incentives. We are unable to say anything about these lineups, or ones which the coach tries for a short period of time before switching to a preferred combination of players.

2.2.3 Structure of a possession

We define the offensive output of a possession as the total points scored before the opponent starts its next possession. For instance, if a shot is taken, but missed, and the offense rebounds the ball, this does not count as a new possession. This means the points

assigned to a specific shot are not just the outcome of that shot, but rather a forward looking metric that captures total offensive output until the ball changes hands.

Table 2.2: Possession Usage and Action Frequencies

Proximate Outcome	<i>N</i>	Average per-game	Points Per Possession
Made 2-pointer	390,763	26.3	2.06
Made 3-pointer	88,009	5.9	3.02
Turnover	190,443	12.8	0
Fouled in the Act of Shooting	141,737	9.6	1.57
Missed 2 pointer	438,231	29.6	0.35
Missed 3 pointer	134,777	9.1	0.28
All (total possessions)	1,376,893	92.8	1.06

A player is said to “use the possession” if he is the first player from his team to shoot or turn the ball over. A turnover occurs when no shot is taken, but the defense secures the ball (for example, by intercepting a pass). We model a turnover as a choice to use the possession and model it as a missed shot that returns zero points. For example, and ignoring the possibility of rebounds for expositional purposes, a player who turns the ball over 5% of the time he has the ball in a shooting position and shoots with 55% accuracy is roughly equivalent to a player who shoots with 50% accuracy but does not turn the ball over.

Table 2.2 shows the empirical frequencies of the different ways in which a possession can be used. The second column gives the raw count (*N*), the third gives the per-game averages and the fourth gives the average point return. For instance, a missed 2-pointer averages 0.35 points for the team because there is some chance the team will secure an offensive rebound and the possession will continue. Whereas a made 2-pointer averages 2.06 points for the team, because there is some chance the player was fouled and received an additional free throw. A turnover always nets zero points, because it ends

the possession.

2.2.4 Basketball as a stopping problem

The rule⁶ requiring the offense to shoot before the 24-second shot clock expires puts a stopping problem at the heart of NBA basketball: which opportunities are more valuable than continuing the possession and thus should be used? We formally present the model in the following section, but give the core intuition here.

In each second-long period of the shot clock, we assume one player is holding the ball and observes a scoring opportunity. Many latent factors such as the player's shooting ability, his location on the court and how well he is being defended impact the expected returns to attempting to score. This player must, acting as a decision maker for the entire lineup, decide whether to use the possession now or to continue. While he also faces many other choices related to game-play (which direction to dribble, instructions to tell to teammates, where to look, to pass or keep the ball, etc.), we focus only on the decision to shoot vs. continue. These other decisions are well beyond the scope of our study and will be rolled up into the parameters that govern the quality of shots the lineup selects from.⁷

We solve for equilibrium using the appropriate tools from dynamic programming. Equilibrium requires that offensive players adopt a *reservation threshold* that monotonically declines with time remaining on the shot clock. This threshold determines which opportunities should be used and which should be passed up—if an opportunity is observed to be above the threshold, the shot is taken. The threshold should be set equal to

⁶Rule 7 of the NBA rule book. The rule was introduced in 1954 to increase the pace of play and make the game more entertaining.

⁷Perhaps most notably, this excludes the study of whether a lineup optimally shares possession of the ball.

the continuation value of the possession, a property we call *dynamic efficiency*. Dynamic efficiency ensures that expected points are maximized given the time constraint of the shot clock.

With this setup in mind, it is clear our primary challenge is modeling scoring opportunities observed by the player, but not the econometrician. We describe our approach in Section 4 after presenting our formal model in the next section.

2.2.5 Sample restriction to the “half-court offense”

In order to properly evaluate player decision making, we must take care that the possessions in our data are comparable to one another at each point in the shot clock. Figure 2.1 shows the hazard rate with which teams use possessions at different values of the shot clock, broken down by how the possession originated. Possessions originating with a “steal” tend to have a high number of shots taken early in the shot clock. These are what are known as “fast-breaks,” in which the team rushes down the court to take an unguarded shot—scoring opportunities used in this interval are clearly drawn from a different distribution. Possessions that originate when the game clock is stopped, known as a “dead ball” situation, effectively “start” later than those originating with steals or defensive rebounds—the hazard is very low until about seventeen seconds remaining, indicating the ball had not been brought into the offensive range yet. This occurs to a greater degree for possessions following made baskets.

All the hazards except for possessions following made basket’s converge at the twelve second mark. By this point, we can be confident that the offense and defense are both set in the appropriate “half-court” and that the method of possession origin no longer matters. Are possessions following made baskets fundamentally different in terms of game play? Almost surely not. The discrepancy occurs because the game clock,

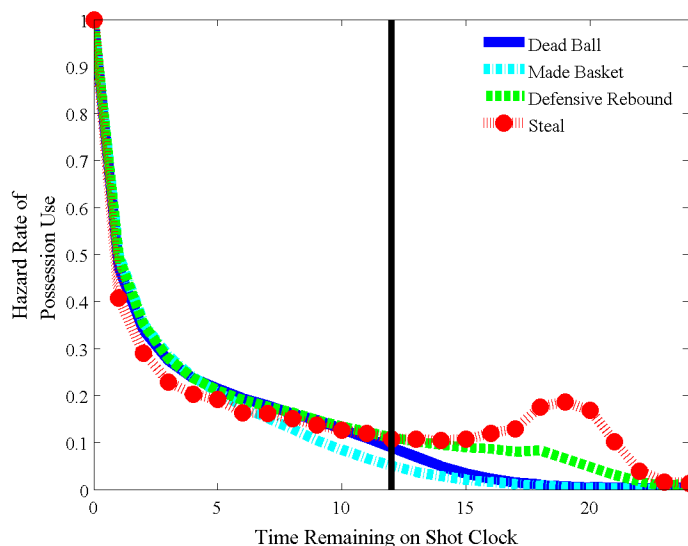


Figure 2.1: Shooting hazard by how the possession originates as a function of time remaining on the shot clock.

which we use to infer shot clock value, continues to run after a made basket but the shot clock does not start until the ball is inbounded, rendering our inference of the shot clock measurement unreliable in this case. For this reason, we drop all possessions from our data set that originate after made baskets and those that are used before the 12-second mark on the shot clock, to eliminate quick, opportunistic “breaks.” These restrictions ensure that once the shot clock hits twelve, both the offense and defense are engaged in five-on-five strategic interaction that is comparable across the possessions we retain for analysis.

2.2.6 Lineups vs. individuals

Our model is one of group decision making, but the actual choice to use the possession or continue is made in the flow of game play by whichever player happens to possess the ball at that point in the shot clock. Since possession of the ball (except when

certain actions occur) is not recorded in our data, we have two options. The first is to place assumptions on how the ball is distributed between teammates and to incorporate this into a model of the decision making of individual players. This approach was taken in some of our preliminary work (Goldman and Rao, 2011). However, the necessary assumptions are restrictive and impossible to test. The second option, and the one taken in this paper, is to instead model shooting decisions at the lineup level. While more straightforward, this has the disadvantage that individual players may be subject to additional incentives.

Specifically, it is easy to imagine that players may prefer to score lots of points in order to increase their status on the team and bolster their scoring statistics for future salary negotiations. If the team is a sufficiently weak organization, it may be unable to stop this behavior. This could lead to the appearance of “overshooting” by the team, when in reality it just reflects the selfish preferences of individual players. However, this behavior would only benefit individuals if suboptimal play was not reflected in salaries. By construction, overshooting means a player is using too low a reservation threshold, meaning they are selecting shots that offer lower expected points than their teammates. Thus while overshooting would increase a player’s points-per-game average, it would tend to lower his shooting percentage (two of the most commonly reported metrics of player performance) and lead to a reputation as an inefficient scorer. Nonetheless, since it is unclear how salaries reflect this tradeoff, we must be cautious in interpreting observed overshooting as it could be selfish play resulting from a failure incentives. Given the focal nature of points per game, it is hard to tell a convincing story in the other direction.

2.3 Formal Model

On each possession, the offensive lineup seeks to maximize its expected number of points by dynamically selecting from available scoring opportunities. We discretize the duration of the shot clock into one-second intervals (matching the granularity of our data). At each interval, the offense has the opportunity to either *use* the possession or to *continue* until the next period of the shot clock. With t second remaining on the shot clock, the offense draws an unbiased⁸ measure of the expected number of points the team would get from immediate use of the possession. This signal represents not just expected points on an immediate shot, but also whatever value the team can expect to get before ending the possession. This includes additional points scored after offensive rebounds or from foul shots, but also the risk that the team should turn the ball over and get zero points. Formally, the expected value of the offense's scoring opportunity with t seconds remaining is

$$\eta_t \sim F_t. \quad (2.1)$$

The team observes η_t and decides if it is a sufficiently valuable opportunity to *use* the possession. If so, the game is terminated and the offense receives expected value η_t . Otherwise, the game proceeds to period $t - 1$ and the offense receives a new draw. If the offense fails to *use* the possession in period $t = 0$, a shot clock violation occurs and the game is terminated with zero points for the offense.

For purposes of tractability, we assume that each F_t is continuous and has strictly positive density on its support; this insures the existence of a unique inverse CDF. F_t may also depend on the quality of the offensive and defensive lineups, but this notation is

⁸Note, persistent pessimism or optimism in the observation of a scoring opportunity is observationally equivalent to choice mistakes. As such, if this assumption does not hold it impacts the interpretation of the results, not the workings of the model.

suppressed.

If the game starts in period T , the offense's strategy is summarized by a $T + 1$ -vector of *reservation values*, $\mathbf{c}$. These correspond to the lowest quality (or “marginal”) shooting opportunity that the lineup is willing *use* in period $t \in \{0, 1, \dots, T\}$. That is, the offense will use the possession in period t if and only if $\eta_t \geq c_t$.

A risk neutral offense should chose each c_t to maximize the objective function,

$$V_t = V_{t-1} + \int_{c_t}^{\infty} (x - V_{t-1}) dF_t(x) \quad (2.2)$$

$$V_{-1} = 0,$$

where V_t denotes the expected point value of an unused possession with t seconds remaining on the shot clock. This framework makes clear that the selection of optimal reservation thresholds can be solved with dynamic programming and that they must satisfy the concept of dynamic efficiency, formally defined below.

Dynamic Efficiency: A lineup of offensive players is said to satisfy dynamic efficiency if for all t , it sets $c_t = V_{t-1}$.

We now state Theorem 1, which formalizes the conditions under which dynamic efficiency characterizes the selection of optimal reservation values in this game.

Theorem 1. *Suppose that:*

- (a) *The offense has **risk and time neutral** preferences over their point output,*
- (b) *The offense observes η_t as an **unbiased** measure of point output if player t uses the possession in period t .*

(c) η_t is drawn from a **continuous** distribution with pdf f_t having **positive density** on its support.

Then optimal offensive strategy is given by a unique vector of reservation values **(c)** satisfying **dynamic efficiency**.

Proof. Since the offense is risk-neutral and observes the true expectation of each scoring opportunity, it should maximize the objective function in (2.2) for each t . Suppose, in period t the offense selected $c_t^* \neq V_{t-1}$. Denote their period t value as $V_t(c_t^*)$. Referring to (2.2), the difference between this value and the value attained under dynamic efficiency is given by

$$V_t(c_t^*) - V_t(V_{t-1}) = \int_{c_t^*}^{V_{t-1}} (x - V_{t-1}) f_t(x) dx < 0,$$

where the strict inequality follows so long as f_t takes positive values in some neighborhood of V_{t-1} , which is insured by (c). If f_t has no density in some neighborhood of V_{t-1} , then the offense's optimal behavior may not be able to be uniquely expressed as a cut-threshold. Given knowledge of the opportunity distribution, these cut-thresholds are easily solved for dynamically using equation (2.2). $\square$

2.4 Identification

In this section, we develop the framework to take our model to the data. We begin by only considering the direct implications of the model as presented. Then we introduce an additional assumption and characterize the resulting identification of marginal efficiency. Finally, we make a distributional assumption on shot opportunities to estimate a structural parametric model that allows for point identified reservation thresholds for individual lineups.

2.4.1 A simple test using average efficiency

At any point in the shot clock, the chosen reservation threshold is equal to the expected value of the marginal shot a team must take in order to slightly increase its shooting frequency. As is the case for nearly any stopping problem, we do not observe marginal quantities and must infer them. Average efficiency—simply the average point value of possession usage for that time period—is directly observable. Since average efficiency exceeds marginal efficiency, optimality implies it must be greater than the continuation value of the possession. Otherwise, the team would be better off if they did not shoot at all in period t (this is very unlikely to be optimal, but it would be preferred to the team's actual behavior).

Efficiency averages are calculated over all possessions for which usage was initiated at shot clock value t and continuation values are estimated by averaging over those possessions used at $s < t$. Conducting this test for each of our thirty core lineups at $t = 6, 9$, and 12 yields only one lineup that is a statistically significant violator. However, this result does not hold up when an appropriate multiple testing correction is applied as shown in Figure 2.2. It is thus clear that lineups which flagrantly violate dynamic efficiency by shooting so much that their average shot value is lower than the continuation value of the possession are not tolerated in the NBA.

2.4.2 The shot clock as an instrument to infer opportunity distributions

Since there could be a large difference between the marginal shot a lineup is willing to take (c_t) and average shot value (e_t), the preceding test only rules out shooting far more frequently than is optimal. In order to improve upon this baseline, we need

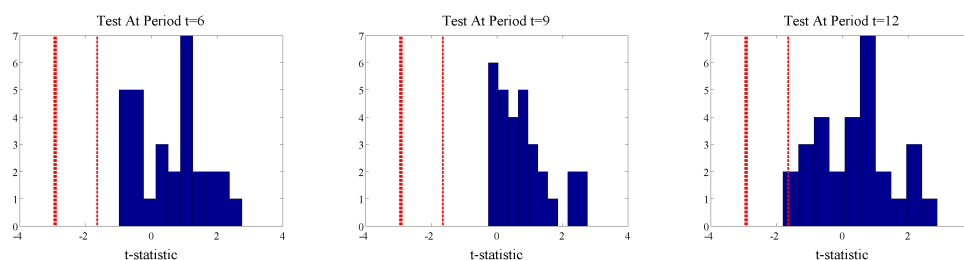


Figure 2.2: Histograms of t-statistics, negative values indicate overshooting. The thinner dashed line ($t=-1.645$) indicates significance for one lineup. The thicker line ($t=-2.93$) provides a multiple testing correction.

to estimate the appropriate counterfactual of a lineup's observed choice: what would happen if they shot a little bit more or less in a given period of the shot clock?

Our strategy exploits variation in shooting hazard rates over the course of the shot clock. Figure 2.1 showed the average lineup's hazard broken down by how the possession starts. Recall that we have restricted our sample to the final twelve seconds. As time winds down, shots are taken at an increasing frequency (conditional on reaching that point). This change in behavior represents the relevancy of our instrument and will facilitate identification.⁹ The validity of this instrument depends crucially on the assumption that the offense's opportunity distribution is the same in all periods of the shot clock ($F_t = F_s \equiv F$) that we retain for empirical analysis. To be clear, this does not require that the actual scoring opportunities used in each period of the shot clock are of the same average quality. As can plainly be seen in Figure 2.10 in the Appendix, this is not the case—shots taken in the final periods of the shot clock offer a substantially lower average point return. Instead, we only require that shots are drawn from the same distribution of scoring opportunities and that the differences in average efficiency are only driven by the offense's willingness to use lower quality opportunities in these later

⁹If player's used a fixed threshold this property would not hold. Our instrument relevancy depends at least some response to time, but the response could be greater or less than optimal.

periods.

Since the shot clock value does not directly constrain an offense's scoring options, independence appears to be a reasonable assumption. The main concern is that the defense could use the shot clock as an ally. In the final seconds of a possession they could devote disproportionately more attention to the ball handler, shifting the offense's opportunity distribution down, without fear that the offense could create a better shot a few seconds later. There are a few reasons we believe this is not a crippling concern. First, the shot clock is situated behind the defense, hindering the success of strategy that requires precise timing to be effective. Second, professionals can pass the ball so quickly that it would be risky to leave someone open in a good scoring position. Third, such actions are not part of typical basketball parlance or strategy and rarely mentioned by announcers or discussed by coaches.

A related concern is that teams may often run intricate plays to generate scoring opportunities. Several seconds of opportunities may effectively be sacrificed in order to generate a high-value opportunity at (for example) $t = 11$. If the team passes on this opportunity the draws observed in the next few seconds may be worse in expectation, perhaps because the offense chooses to "reset." However, this only differentially impacts the opportunity distribution (F_{11}) to the extent that these plays end at $t = 11$ in a way that they do not end in other periods. While not impossible (there could be clustering due to the time it takes to bring the ball up the court and run a "standard play"), it seems unlikely given that such predictable timing is unlikely to be optimal.

We investigate both potential violations of the i.i.d. assumption using over identification tests of our structural model, examining model residuals across periods of the shot clock. We discuss these tests in greater detail further on but note here that they not reveal any significant departures from our identifying assumptions or parametric

specifications.

We'll now consider an illustrative example and then provide formal results to characterize identification. Let u_t denote the hazard rate of possession use (henceforth referred to as a *usage rate*) for a given lineup with t seconds remaining on the shot clock. Consider two periods of the shot clock, e the earlier period and l the later period and suppose we directly observe the following reduced form statistics in these periods:

$$u_e = 0.08 \quad e_e = 1.1$$

$$u_l = 0.12 \quad e_l = 1.$$

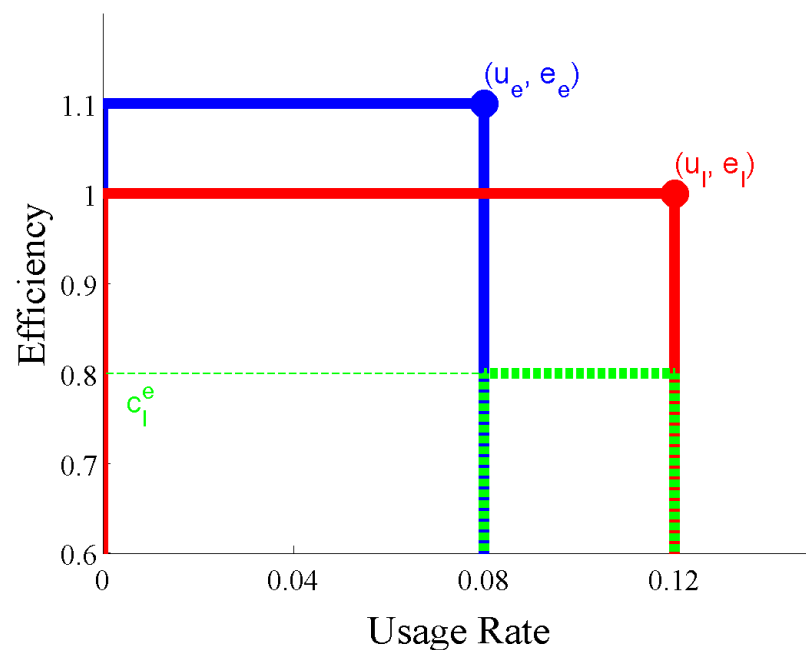


Figure 2.3: Illustrating the identification of cut-thresholds under the exogeneity of the shot clock.

This is as illustrated in Figure 2.3. In period e , this lineup selected the best 8% of its shooting opportunities and was able to average 1.1 points. Effectively, they are

integrating over the top 8% of their opportunity distribution and getting $.08 \times 1.1 = .088$ points. In period l , they were more aggressive, trading in the best 12% of their scoring opportunities for $.12 \times 1 = .12$ points. These point totals are given by the areas of the blue and red rectangles (respectively) in Figure 2.3. We have assumed that the opportunity distribution is the same in periods e and l . This means that the best 8% of scoring opportunities available in period l will also return .088 points. The remaining (or marginal) .04 possessions used in period l must return the remaining $.12 - .088 = .032$ points. This area is represented by the dashed green rectangle, whose area is exactly the difference between the red (period l) and blue (period e) rectangles. Then the value of these marginal shots when going from period e to l is given by,

$$c_l^e \equiv \frac{e_l u_l - e_e u_e}{u_l - u_e} = 0.80.$$

This value is not precisely the value of the marginal shot in period e or l , but rather the average value of those shots taken in period l , but not in period e . It is a lower bound for the marginal shot in the earlier period and an upper bound for the marginal shot in later period. With this intuition in hand, we can now characterize parametric and non-parametric identification of optimal behavior. Formally:

Theorem 2. *Suppose a lineup is observed to play the game described in Section 3 and that*

- (a) *In every period of the shot clock, average efficiencies (e_t) and hazard rates of possession use (u_t) are observed,*
- (b) *The distribution of shooting opportunities faced by the offensive lineups is invariant to the period of the shot clock. That is, for all shot clock periods s, t ,*

$$F_t = F_s \equiv F.$$

Then,

- (i) For any shot clock periods e, l in which $u_l > u_e$, we have the following bounds on cut-thresholds,

$$c_l \leq \frac{e_l u_l - e_e u_e}{u_l - u_e} \quad (2.3)$$

$$c_e \geq \frac{e_l u_l - e_e u_e}{u_l - u_e}. \quad (2.4)$$

- (ii) Suppose $\mathcal{T}$ is a K -vector of time periods in which a lineup is observed to select distinct hazard rates of possession use ($\mathbf{u}_{\mathcal{T}}$). Consider a parameterization of the opportunity distribution as $F(x) = G(x; \theta)$ with θ in the parameter space $\Theta \subset \mathbb{R}^L$. Now define

$$E(u; \theta) = u^{-1} \int_{G(1-u; \theta)}^{G(1; \theta)} x \cdot dG(x; \theta),$$

as the implied mapping between chosen usage rates and resulting efficiencies for a lineup with parameter θ . Then we may point identify θ and $\mathbf{c}$ for this lineup if the mapping $\Theta \rightarrow E(\mathbf{u}_{\mathcal{T}}; \Theta)$ is one-to-one.

The proof is given below. Part (i) gives nonparametric bounds by formalizing the logic accompanying Figure 2.3. Part (ii) provides conditions for parametric identification. The requirement of distinct usage rates is intuitive. If a lineup did the same thing in every period of the shot clock, then the shot clock is not a relevant instrument and we have no hope of learning about the tradeoff between usage and efficiency. An intuitive order condition is immediately implied by (ii). Identification of a L dimensional parametric model requires the observation of unique hazard rates in at least $K \geq L$ shot clock periods. That is we must have at least L dimensional relevance of our instrument.

Proof. (i) The proof is straightforward and mirrors the logic at the beginning of Section

4.2. Formally, we note the convenient facts that for any t ,

$$u_t = P(\eta_t > c_t) = 1 - F(c_t),$$

$$e_t = E(\eta_t | \eta_t > c_t) = \frac{\int_{c_t}^{\infty} x dF(x)}{u_t}$$

Now observe that $u_l > u_e \implies c_e > c_l$. Now we manipulate the right hand side of (2.3) as follows

$$\begin{aligned} \frac{e_l u_l - e_e u_e}{u_l - u_e} &= \frac{\int_{c_l}^{\infty} x dF(x) - \int_{c_e}^{\infty} x dF(x)}{F(c_e) - F(c_l)} \\ &= \frac{\int_{c_l}^{c_e} x dF(x)}{F(c_e) - F(c_l)} \\ &\leq \frac{c_e (F(c_e) - F(c_l))}{F(c_e) - F(c_l)} = c_e. \end{aligned}$$

The proof of the bound in (2.4) follows similarly.

- (ii) Let the vector $\mathbf{e}_{\mathcal{T}}$ denote the observed efficiencies that correspond to the usage rates $\mathbf{u}_{\mathcal{T}}$. Both of these vectors are observed so we know everything except θ in the mapping

$$\mathbf{e}_{\mathcal{T}} = E(\mathbf{u}_{\mathcal{T}}; \theta).$$

By our assumption of one-to-oneness, there can be only one value $\theta_0 \in \Theta$ that satisfies this relation. Given this, we can find the reservation value c_t as the $1 - u_t$ quantile of the opportunity distribution, that is

$$c_t = G(1 - u_t; \theta_0).$$

□

2.4.3 Parametric structural model

We now develop a parametric model to efficiently test the adherence to dynamic efficiency in each team's core lineup. The model links moments we can observe in the data, usage rates and average efficiencies, to the unobserved reservation thresholds in each lineup's decision rule.

We assume lineup l 's scoring opportunities are drawn from a uniform distribution along the interval $[B_l, A_l]$, that is $\eta_t \sim U(B_l, A_l)$. Since opportunities are unobservable, it's hard to have strong intuition on what constitutes an appropriate approximation. A more interpretable implication of uniformity is that it implies usage rate and average efficiency are both linear in the chosen reservation threshold. To see this, recall that a lineup uses a possession in period t if the draw (η_t) weakly exceeds the reservation threshold, that is if $\eta_t \in [c_{l,t}, A_l]$. Conditional on the parameters $\theta_l \equiv \{A_l, B_l, c_l\}$:

$$e_t = \frac{A_l + c_{l,t}}{2} \quad u_t = \frac{1}{A_l - B_l} (A_l - c_{l,t}).$$

Then to check identification of this model we refer to Theorem 2(ii) and compute the mapping:

$$E(u_t, \theta_l) \equiv e_t = A_l - \frac{A_l - B_l}{2} u_t. \quad (2.5)$$

It is clear that the observation of *only two distinct usage levels* (and their corresponding efficiencies) is sufficient for the one-to-oneness of E and thus the parametric identification of this model. This also shows that a uniform parameterization of the opportunity distribution is equivalent to the assumption of a linear relationship between chosen usage and resulting efficiency for a given lineup. This linear relationship also seems to be borne out in aggregates of our data. The scatterplot shown in Figure 2.10 shows combinations of usage and efficiency for all periods of the shot clock for an aggregate of all thirty of

our core lineups. This is hardly a formal test, but seems to concord with linearity.

Let N_t denote the number of possessions in which lineup l had the ball with t seconds remaining on the shot clock. Then we can write the likelihood function as:

$$Prob(\{\hat{e}_t, \hat{u}_t\}_{t \in \{0, \dots, 12\}} | \theta_l, N_t) = \prod_{t \in T} \phi \left(\frac{\hat{u}_t - \frac{A_l - c_t}{A_l - B_l}}{\sqrt{N_t \frac{A_l - c_t}{A_l - B_l} (1 - \frac{A_l - c_t}{A_l - B_l})}} \right) \phi \left(\frac{\hat{e}_t - \frac{A_l + c_t}{2}}{\frac{\sigma_l}{\sqrt{N_t}}} \right). \quad (2.6)$$

Since the parameters A and B fully describe a lineup, it's important to build a bit of intuition around how they govern offensive output. All else equal, higher values of A lead to greater offensive output. If A and B are close together, the lineup draws from a low variance distribution, which implies it can shoot more frequently without much deterioration in shot quality. Conversely, a lineup with a wide opportunity distribution has a steeper tradeoff between usage and efficiency (as represented by the E function in (2.5)) and must dip deeper into the shot quality distribution to use possessions more frequently. The latter lineup will find it risky to be forced to shoot with high probability in the last few periods of the shot clock and thus should adopt a lower, steeper reservation threshold—it is willing to accept worse opportunities early in the shot clock in order to avoid reaching these periods. In contrast, a lineup with a very flat usage curve is better off being patient by using a high, flat reservation threshold as they are likely to observe a relatively high quality shot if they are forced to shoot with high probability in one of the final periods.

It's essential that the imposed linearity captures reality to an acceptable degree so as not to induce bias. Fortunately, the model is overidentified because main estimation is conducted for the final thirteen shot clock periods (we'll discuss this sample restriction in more detail in the next section) and we are only estimating two parameters per lineup. In the Appendix, we plot model residuals against time remaining on the shot clock. Model

fit is good in all periods and does not appear to show any trends with time remaining on the shot clock. So while the linearity assumption is almost certainly “wrong”, it nonetheless provides a useful approximation of reality and is not statistically rejected in our data.

2.5 Results

Our main results are based on the parametric model described in the previous subsection. Since we wish to estimate this model only for the subset of possessions that it adequately describes, we first discuss the sample restrictions we have put in place.

2.5.1 Sample Restrictions

As discussed in Section 2.1, our model requires that a team is a risk neutral point maximizer. Toward the end of the game, this assumption fails to hold as the leading team will wish to use up time and the trailing team will wish to conserve time. To eliminate these cases we directly measure the value of time using the probit approach in Goldman and Rao (2013) (and duplicated in the Appendix for completeness), which models the probability of winning based on features of the game state such as score margin, time remaining and home court advantage. We define the “win value of time” (WVT) as the value of one second in terms of the probability of winning the game and “win value of a point” (WVP) analogously. For example, if WVT equals -0.001 and WVP equals 0.01, then the team, which in this case is leading, would give up one point to roll ten seconds off the clock. It is clear that we need to remove such cases where time is of first order importance. We thus restrict our main analysis to cases when the value of time is negligible relative to the value of points by removing possessions where $\left| \frac{WVT}{WVP} \right| > 0.01$.

This restriction removed 18.78% of the data. In supplementary analysis, we re-introduce these possessions to test if teams correctly respond to the value of time.

Second, our model requires that both team's have a chance to win and are competing over meaningful points. To this end we rule out possessions in which $WVP < 0.003$. Since this restriction rules out many of the same possessions as the first restriction, it only costs us an additional 1.13% of possessions.

Our final set of restrictions is to eliminate all possessions that originate after made baskets or that are used before the 12-second mark on the shot clock. These restrictions are discussed in Section 2.5 and are necessary to insure comparability between possessions and the accuracy of our shot clock measurement. These restrictions exclude an additional 50.46% of our data.

2.5.2 Main results

We estimate the parametric model described Section 4.3 for each of the thirty lineups in our data. Figure 2.4 provides aggregate results computed by taking (inverse-variance) weighted averages across our thirty core lineups. The left panel shows the average estimated reservation value and continuation value for the final twelve periods of the shot clock excluding the final period ($t = 0$). In the final period, all possessions are used so that the reservation threshold is exactly equal to the lower bound of their opportunity distribution (B) whereas the continuation value is zero, meaning optimal play does not require equality in this case. The right panel zooms in on periods 4-12 for more contrast.

The plots clearly show that NBA players use a monotonically declining threshold that is quite close to the continuation value of the possession. Impressively, the players match the concave shape of the continuation value (a property common in stopping

problems since continuation values decline rapidly as one approaches the terminal period). Deviations tend to occur in periods $t = 9$ to $t = 1$, with the reservation threshold slightly too low in these cases. This results in excess shooting in these periods, which can be interpreted as impatience.

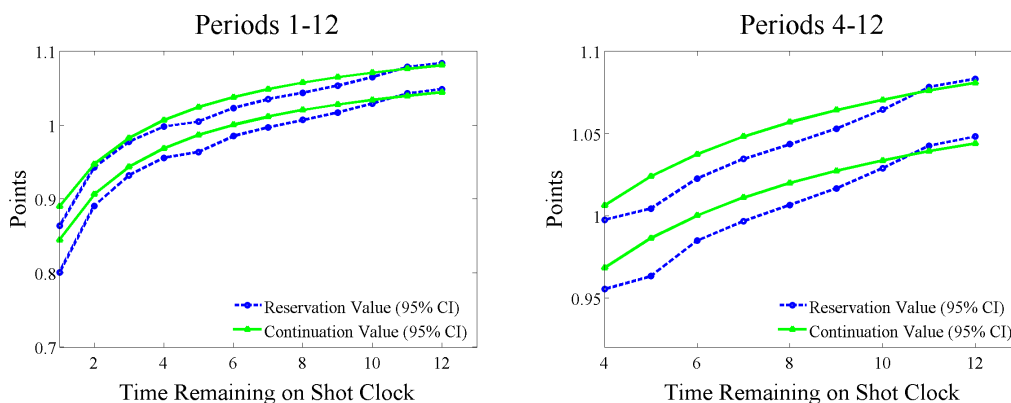


Figure 2.4: The left panel computes a simple average of estimated reservation thresholds and continuation values across our thirty lineups. The right panel zooms in on periods 4-12. These pictures reveal no aggregate violations of dynamic efficiency.

We now dig deeper by applying tests of optimal behavior at the lineup level. To improve the power of these tests, we focus on periods $T = \{4, \dots, 12\}$, which are estimated with less noise. A natural joint test of dynamic efficiency is given by the standard Wald statistic for the hypothesis that the reservation threshold in each period is equal to the continuation value: $\{V_{j,t} = c_{j,t}\}_{t=4}^{12}$, which is distributed $\chi^2(9)$. The left panel of Figure 2.5 gives the results of this test for each lineup. For comparison, the dashed line gives the distribution under the null hypothesis of optimal behavior. Although we cannot reject optimal play for the majority of lineups, it is clear the distribution is shifted to the right of the null. As indicated by the presence of the solid red line, we can reject joint optimality for 12 of the 30 lineups at a 5% level of significance. Four of these remain statistically significant after applying a multiple testing correction (represented by

the dark line). If all lineups adhere to dynamic efficiency, then we would expect only 5% of lineups ($.05 \times 30 = 1.5$) to exceed the first threshold and only a *single* lineup to exceed the multiple testing threshold only 5% of the time. As such, this is a strong rejection of perfect adherence to dynamic efficiency.

To measure the efficiency losses associated with sub-optimal shot choice we estimate dead weight loss (DWL), which can be understood graphically as the area between the blue and green curves in Figure 2.4. Formally, DWL is given by:

$$DWL_l = \frac{1}{2(A_l - B_l)} \sum_{t=1}^{12} p_{l,t} (c_{l,t} - V_{l,t-1})^2, \quad (2.7)$$

where $p_{l,t}$ represents the fraction of offensive possessions that have not ended before period t of the shot clock. This quantity is then de-biased so as not to erroneously report large values of DWL for small lineups with nosily estimated reservation values.¹⁰ Estimated DWL per 100 possessions (about a game's worth) is plotted in the central panel of Figure 2.5. No team in our sample lost more than 0.7 points per 100 possessions due to dynamic inefficiency. For purposes of comparison, the difference between the best and worst offense in the NBA in a given season is typically about 10 points per 100 possessions.

The low overall DWL could be a consequence of two qualitatively different strategic settings. It could be that these mistakes are actually quite large, relatively speaking, but that dynamic efficiency does not matter very much for overall point output. Alternatively, it could be that using a threshold that matches the continuation value of a possession does have meaningful consequences and that our estimated mistakes are indeed quite small. In order to tell the difference, we estimate the DWL generated by an offense incapable of *dynamic* decision making, with the caveat that we allow them to

¹⁰It is straightforward to show that a lineup adhering to dynamic efficiency still has $E(DWL_l) = \frac{1}{2(A_l - B_l)} \sum_{t=1}^{12} p_{l,t} \cdot \text{Var}(c_{l,t} - V_{l,t-1})$. This quantity is already estimated and used to debias DWL.

always shoot in the final period to avoid shot clock violations.¹¹ Specifically, we constrain the offense to use the same reservation value in all periods of the shot clock, conditional on reaching $t = 12$ (except the final period), but allow them to set this threshold to maximize total points. We will refer to this as the “optimal fixed threshold.”

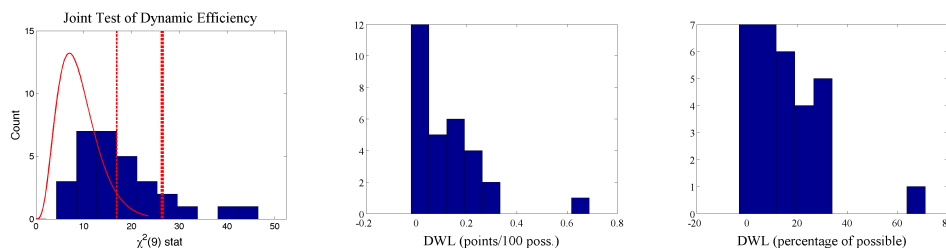


Figure 2.5: Chi-square statistics for the joint adherence to dynamic efficiency for each of our thirty lineups (left) and measures of DWL (center) as calculated according to equation (2.7). The right panel gives DWL as a percentage of total value gained by dynamic decision making.

An optimal fixed threshold produces DWL of roughly 1 point per 100 possessions—or about 10% of the difference between the best and worst offenses in the league.¹² Had we required the fixed threshold to apply to earlier periods of the shot clock as well, it would have fared more poorly. Nonetheless, put in context, this effect size means that a fifth or sixth (of eight) playoff seed would likely be knocked out of the playoffs if they were restricted to using an optimal fixed threshold and a higher seed would likely lose lucrative “home court advantage” for rounds past the first. These sort of performance shifts typically require signing players with annual salaries of roughly \$5 million dollars.¹³

¹¹It does not seem to be a meaningful baseline to for a lineup to have to choose a low fixed threshold just to avoid shot clock violations in the final period. Even an unsophisticated team could “just shoot” in the final second.

¹²The sub-optimality of a fixed threshold relates to the slope of a lineup’s usage curve, as discussed earlier. Lineups that can increase the probability in which they shoot and experience a relatively small deterioration in shot quality (a flat usage curve) suffer less from using a fixed threshold, since they can keep it quite high and still get decent returns if the terminal period is reached.

¹³Since most teams are constrained by the salary cap, the market value of these types of signings could be substantially higher.

The rightmost panel of Figure 2.5 plots the distribution of observed DWL estimates as a percentage of the DWL from an optimal fixed threshold. Impressively, many lineups capture more than 90% of the gains from dynamic decision making, the average lineup captures 84% and all but a single lineup is estimated to capture two-thirds.¹⁴ Later we'll see that a substantial fraction of the mistakes we observe are due to the reservation being shifted by a fixed constant (but possessing the right slope), which leads us to suspect that an optimal fixed threshold would likely fare far better than a fixed threshold would in practice, or in other words, we believe it's a competitive baseline. Taken together, these results indicate that dynamic decision making does indeed matter and NBA players (at least the commonly occurring lineups featuring good players) are up to the task.

Although the overall adherence to optimality is high, we do identify statistically significant deviations. We now examine these deviations in more detail at the lineup level. We start by computing the net number of “extra shots” that each lineup takes (positive values indicate shooting too frequently at a given point of the clock, negative values indicate under-shooting). In each period t a dynamically efficient offense should shoot $\frac{A_l - c_{l,t}}{A_l - B_l}$, which can be compared to the observed value $u_{l,t}$. The number of extra shots is simply the difference between these two quantities, multiplied by the percentage of the time an offense enters that period of the shot clock while still retaining the ball (denoted $p_{l,t}$ below).

$$eShots_l = \sum_{t=4}^{12} p_{l,t} \left(u_{l,t} - \frac{A_l - c_{l,t}}{2(A_l - B_l)} \right) \quad (2.8)$$

Figure 2.6 summarizes the results; the left panel provides a histogram of the

¹⁴This offending lineup was the Chicago Bulls core of Kirk Hinrich, Ben Gordon, Luol Deng, P.J. Brown, and Ben Wallace who's large estimated DWL is driven by their singular propensity to overshoot. We estimate (as described later) that 22.55% of their half-court possessions were used prematurely.

resulting t-statistics. The lighter dashed lines indicate significance for an individual lineup while the heavier lines provide the appropriate multiple testing correction. Consistent with the patterns shown in Figure 2.4, ten lineups have a statistically significant propensity to overshoot after applying the multiple testing correction and zero lineups significantly undershoot.

Panel 2 shows that impatient shot selection is concentrated in lineups that play relatively few possessions together, with the observed linear relationship between log possessions and extra shots statistically significant at the $p = .02$ level. Since the number of possessions a lineup plays together is presumably endogenous to lineup performance or “chemistry,” it is not clear if this is evidence for learning, merely reflects a factor important in determining a coach’s choice of the combinations of players to share the court or is some combination of both factors. Nonetheless, the results indicate that lightly used lineups would be less likely to show the tight adherence to optimality observed in our sample of commonly used lineups.

We also checked if adherence to dynamic efficiency was correlated with either the average salary of the lineup (to proxy for average player quality) or the standard deviation of salary (to proxy for quality variance) and found no evidence of a relationship (the plots can be found in the Appendix). This indicates that the positive correlation of possessions played together and dynamic efficiency is not driven just by the parts of player quality that is captured by average salary.¹⁵

A lineup can make two classes of mistakes in setting a reservation threshold. A fixed propensity to over or undershoot throughout the shot clock arises from a threshold with the correct slope but shifted by an additive constant. Alternatively, they could adopt

¹⁵With regards to the standard deviation for quality, one might have hypothesized that more unequal lineups in terms of player ability have a more difficult time working together on this dimension. Interestingly we find no evidence this matters.

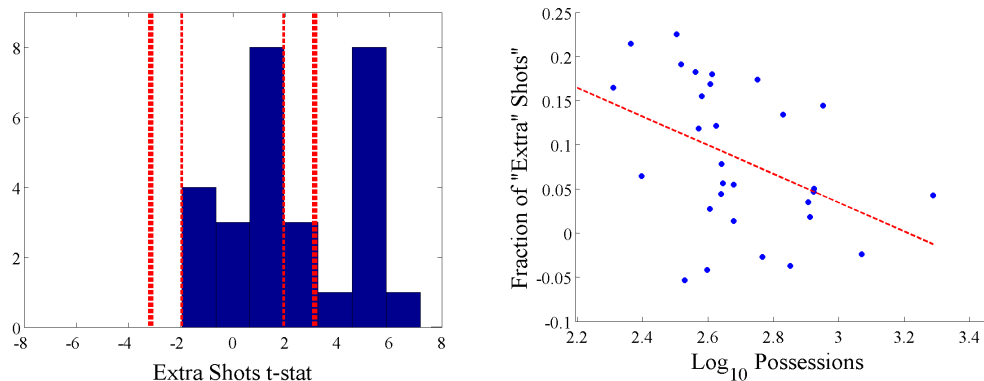


Figure 2.6: Player-by-player t-statistic for deviations from dynamic optimality. Positive values indicate “under-shooting.”

a reservation threshold that is either too steep or too flat. We separate these cases by allowing lineups to have a constant aversion (or affinity) for shooting at all periods of the shot clock by introducing a free parameter that shifts the reservation threshold by a constant. The resulting Wald test statistic for the hypothesis that $\{(V_{j,t} - c_{j,t}) - (V_{j,t-1} - c_{j,t-1}) = 0\}_{t=4}^{12}$ is distributed $\chi^2(8)$. The results of this test are displayed in the left panel of Figure 2.7. As compared to the baseline test in Figure 2.4, the distribution is shifted substantially to the left—we observe far fewer significant violations. Given that some mistakes occur as a level shift, our previous baseline of an

In the right panel of Figure 2.7, we investigate these results further by computing the excess flatness of each lineup’s reservation value curve. We define this as:

$$EF_l \equiv \sum_{t=4}^{12} [8-t](c_{l,t} - V_{l,t-1}) \quad (2.9)$$

To see how this test statistic captures excess flatness, let us first suppose a team over or undershoots by a fixed scalar. When t is large, the weight, $(8-t)$, is negative. When t is small, the weight is positive. Accordingly, the resulting sum is zero, indicating the team uses the correct slope. If $c_{l,t} < V_{l,t-1}$ tends to occur when t is large but not when

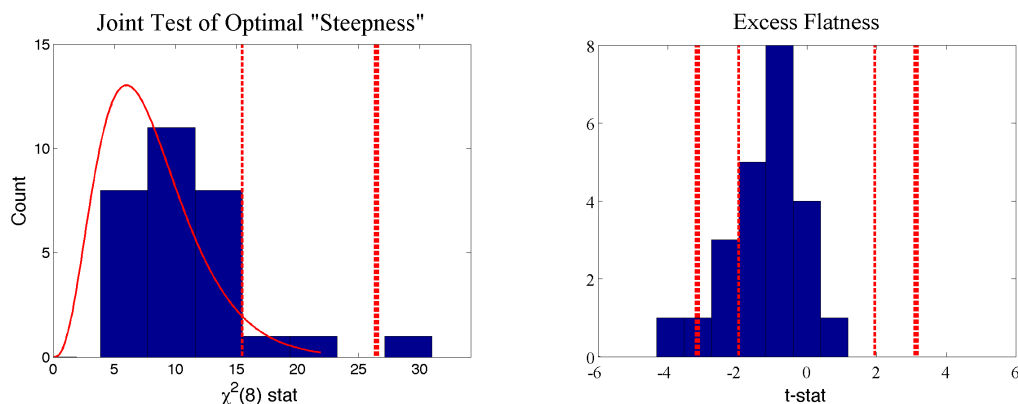


Figure 2.7: Chi-square statistics for optimal steepness” (left) and a t-statistic for excess flatness (right).

t is small, then a lineup experiences great overshooting at the start of the shot clock, but not at the end, meaning the threshold is too flat, and in this case, the test statistic is positive. Conversely, if $c_{l,t} < V_{l,t-1}$ occurs at the end of the shot clock, but not the start, this indicates excess steepness and the test statistic is negative.

The distribution of the associated t-statistics is shown in the right panel of Figure 2.7. Five lineups demonstrate statistically significant excess steepness in their decision making but none demonstrate excess flatness. Taken together, we see that while both classes of mistakes are present, they are seen to occur occur only in one of two possible directions. When the level of the reservation threshold is off it tends to be too low and when the slope is incorrect it errs towards excess steepness.

2.5.3 Value of time

So far, we have seen that in a risk-neutral, time-invariant setting the lineups we study capture most of the gains from dynamic decision making. Toward the end of a

game, the time invariance property is violated.¹⁶ This change in incentives provides an additional test of lineup decision making.

How should lineups respond? It is immediately obvious that the shooting hazard should shift up for the trailing team (they should shoot more quickly) and down for the leading team. The more subtle requirement is that the hazards should converge as the shot clock approaches zero. To see this, take the case of the leading team. With a relatively small amount of time remaining *on the shot clock*, delaying the usage of the possession offers a much smaller gain in time consumed than it does earlier in the shot clock. As the shot clock converges to zero, the expected point sacrifice the leading team should be willing to make to continue the possession should converge to zero as well. The same logic applies to the trailing team in its effort to conserve time.

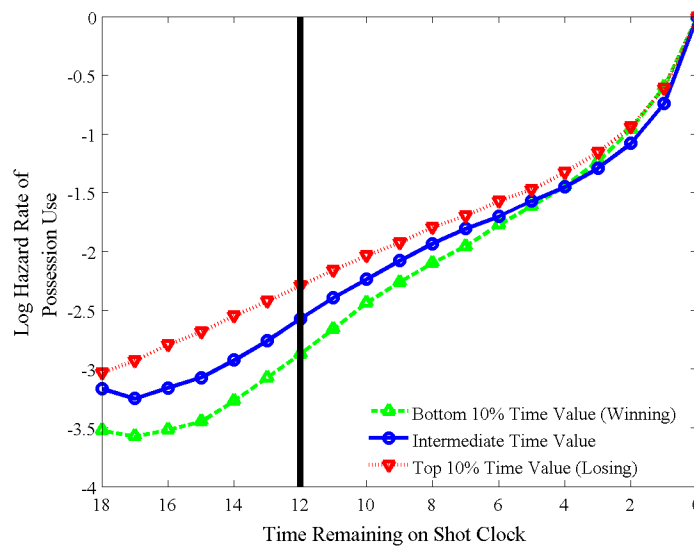


Figure 2.8: Shooting hazard broken down by the valuation a team should place on time.

We compare shooting hazards in the bottom 10% (team leading late in the game)

¹⁶Risk neutrality is also violated. Goldman and Rao (2013) study the risk tradeoff, which is outside the scope of this paper.

and the top 10% (team trailing late in the game) of time value to the time-invariant baseline (our main analysis), as shown in Figure 2.8 in log scale. We can see that both comparative static predictions go through. At early periods of the shot clock, the trailing team is much more likely to shoot (with the leading team much less likely), but with four or fewer seconds remaining, the value of time does not significantly impact shooting frequency.

2.6 Discussion and conclusion

Lab studies have documented that people have difficulty with fundamental elements of rational decision making, such as weighing marginal as opposed to average returns and properly estimating continuation values. Motivated by these shortcomings, economists have studied the decisions of highly paid professionals in games, such as penalty kicks in soccer and serving in tennis. In these relatively simple games, professionals do quite well, but questions remain as to whether rational play holds up in substantially more complex decision problems.

The unique decision environment we study allows us to extend a stylized stopping problem from the lab to a field setting. We study a huge volume of quick decisions made by NBA players observing the arrival of scoring opportunities. Overall, we find that lineups adhere quite closely to the theoretical requirement of dynamic efficiency, with the average lineup capturing 84% of the gains from using a dynamic reservation threshold as opposed to the optimal fixed threshold. Impressively, these lineups use a reservation threshold that matches both the shape and level of the continuation value function. Adherence to optimal stopping is strongly correlated with the number of possessions a lineup plays, indicating either learning or selection on “good chemistry”

by coaches. The mistakes we do observe occur in a consistent direction across lineups; erring thresholds tend to be too low and too steep.

A few caveats are worth keeping in mind. First, to obtain suitable sample sizes we focus on a team's favorite lineup (which presumably features their best players), meaning that we study players with substantial experience playing together. These lineups are natural to study as they are very important to team success, but the sample restrictions limit our ability to make claims about lightly used lineups or less experienced players, especially since we observe that even in this select sample cumulative time playing together predicts adherence to optimality. Second, our main results rely on a structural model that requires distributional assumptions on unobserved scoring opportunities. While we have been careful to use overidentification tests to be sure model misspecification is not driving our results, these types of assumptions always involve some degree of uncertainty and consequently some of the small mistakes we identify could be due to imperfect modeling. Third, we model the stopping problem as a group decision by the lineup, but individual incentives may diverge from team incentives. In particular, selfish play to boost one's own scoring output would be estimated as overshooting in our model. Mistakes of this nature thus could be due to a failure of incentives, not mistaken perception of the continuation value.

These caveats do impact the interpretation of the small "mistakes" we identify, but they would be very unlikely to spuriously create the alignment with optimal play found here. As such, we believe our results significantly expand the difficulty frontier of games that humans can play according to the predictions of economic theory.

2.7 Appendix

2.7.1 Parametric Model Specification Tests

Our parametric model assumes that a player faces a uniform distribution of scoring opportunities defined by only two parameters. However, our model is applied to thirteen different periods of the shot clock and is thus over-identified. As an over-identification test we compute the efficiency residual in each shot clock period for each of our thirty core lineups. These residuals have non-constant variance but they should all have mean and median zero under correct specification. We test this below and find general adherence to our parametric framework. Results are displayed in Figure 2.9.

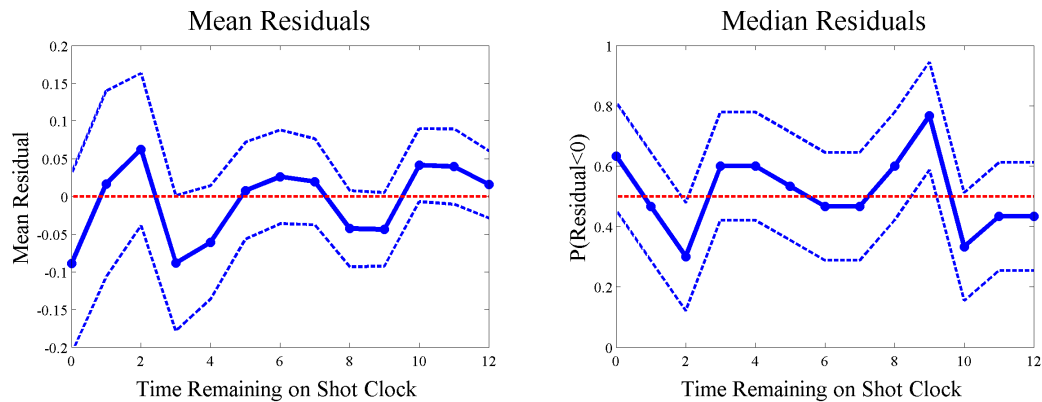


Figure 2.9: Efficiency residuals from our parametric specification.

As discussed in Section 4, the choice of a uniform specification of the opportunity distribution is equivalent to the specification of a linear relation between usage and efficiency. Figure 2.10 shows that this is roughly correct for an aggregation of our data across all thirty of our core lineups.

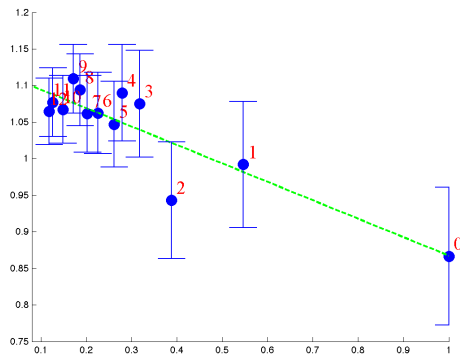


Figure 2.10: Scatterplot of the combinations of efficiency and usage observed aggregated across all of our core lineups. The blue brackets give the corresponding 95% CI for efficiency in each shot clock period. The dashed green line shows the linear fit from a GLS regression.

2.7.2 Computing the win value of time

The goal of the team is to score more points than the opponent in the entire game. Consider the two teams, home (h) and away (a). Let $S_{h,N}$ and $S_{a,N}$ denote the current scores for the home and away team with N offensive possessions (for each team) remaining in the game respectively. Let $P_{h,i}$ and $(P_{a,i})$ denote the number of points scored by the home (away) possession on the i^{th} possession from the end of the game. The home team wins if they have more points at the end of the game. This is equal to the current score, plus the points scored in subsequent possessions, as given by:

$$S_{h,0} > S_{a,0} \iff S_{h,N} + \sum_{i=1}^N P_{h,i} > S_{a,N} + \sum_{i=1}^N P_{a,i} \iff \sum_{i=1}^N P_{h,i} - P_{a,i} > S_{a,N} - S_{h,N} \quad (2.10)$$

To model how points are generated, let $\{\mu_h, \sigma_h^2\}$ and $\{\mu_a, \sigma_a^2\}$ represent the mean and variance of points per possession that each team is able to achieve in the match-up. If the number of remaining possessions (N) is large, the central limit theorem gives

probability of the home team winning as:

$$P(\text{Home Win}) = P(S_{h,0} > S_{a,0}) = P\left(\sum_{i=1}^N (P_{h,i} - P_{a,i}) > S_{a,N} - S_{h,N}\right) \quad (2.11)$$

$$= \Phi\left(\frac{S_{h,N} - S_{a,N} + N(\mu_h - \mu_a)}{\sqrt{N(\sigma_h^2 + \sigma_a^2)}}\right), \quad (2.12)$$

where Φ is the CDF of the standard normal distribution. Examining this expression, we see that if the score is currently tied and the teams are of equal quality, the game is a coin toss. Having an ability advantage (μ higher than opponent) matters proportional to the remaining possessions, which is intuitive. If you are the better team, it is more likely to be reflected in the game outcome when there are many periods remaining. The marginal impact on win probability for each factor can be easily obtained by differentiating equation (3). The following expression gives the impact on win probability of a point scored for the home team:

$$\frac{dP(\text{Home Win})}{dS_{h,N}} = \phi\left(\frac{(S_{h,N} - S_{a,N}) + N(\mu_h - \mu_a)}{\sqrt{N(\sigma_h^2 + \sigma_a^2)}}\right) \frac{1}{\sqrt{N(\sigma_h^2 + \sigma_a^2)}} \quad (2.13)$$

Expression (2.13) shows that points become increasingly impactful on the game outcome when the current score is close ($S_{h,N} - S_{a,N}$ small) and few possessions remain, while the impact becomes exceedingly small when the score margin is high; in a supplementary analysis (not included) we observed that when a point is worth less than 0.003 wins, teams tend to give up and pull their starters out of the game. Hence we use this as a sample restriction.

For the risk-neutral and time-invariance conditions to hold it must be the case that teams care much more about maximizing the efficiency of their possessions instead of trying to play fast (slow) or have high (low) variance possessions (ex. shooting more 3-pointers). By comparing ratios of marginal win effects we can determine the appropriate

indifference ratios. For ease of demonstration, we assume the teams are of approximately equal quality ($\mu_h = \mu_a$), the basic intuition is unaffected by this simplification, because the impact of score difference and ability difference (weighted by possessions remaining), is very similar.

The ratio of the value of possessions (extending or shortening game by one possession for each team) to the value of a point is given by:

$$\frac{\frac{dP(\text{Home Win})}{dN}}{\frac{dP(\text{Home Win})}{dS_{h,N}}} \approx \frac{1}{2} \left(\frac{(S_{a,N} - S_{h,N})}{N} \right) \quad (2.14)$$

To understand this equation, let's examine the case when the home team is trailing with N possessions remaining. Since $S_{a,N} > S_{h,N}$, the numerator is positive. Differentiating equation (3) with respect to N , one can show that the team values possessions intrinsically (it needs to catch-up). The relative values of points to possessions is given by equation 2.14. Indeed, as the number of possessions remaining grows, this ratio goes to zero at rate N . Given that an average possession cycle (one for each team) takes about 30 seconds, the number of remaining possessions need not greatly exceed the score difference before this quickly becomes a very unattractive tradeoff. Sacrificing a meaningful amount of points (on the order of 0.005) to save a few seconds is a losing proposition. In our empirical analysis we eliminate observations for which the terms of this tradeoff imply that seconds are relatively valuable as compared points. The threshold we use entails that if one team is trailing by 10 points, we eliminate all observations when less than 33 possessions (≈ 16 minutes) remain.

2.7.3 Additional plots and tables

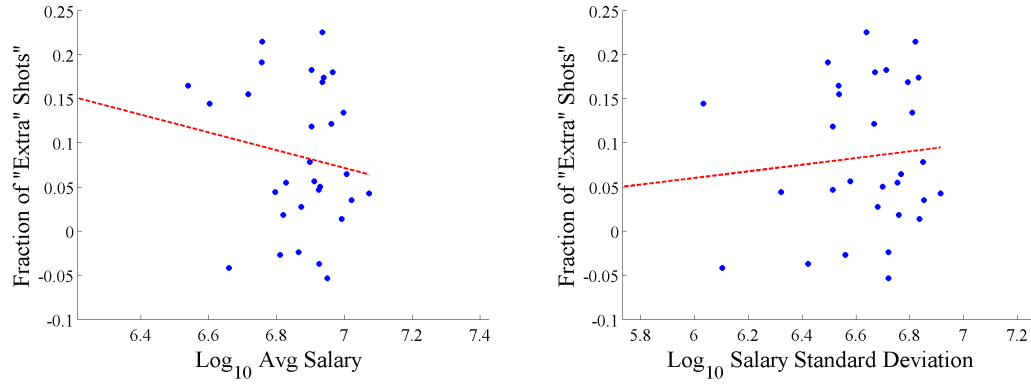


Figure 2.11: Lineup t-statistic for deviations from dynamic optimality scattered against lineup characteristics.

Table 2.3: Correlation Coefficients

	Extra Shots	Excess Flatness	Log Possessions	Log Avg Salary	Log Std Dev Salary
Extra Shots	X				
<i>p-value</i>	X				
Excess Flatness	0.2164	X			
<i>p-value</i>	(0.2508)	X			
Log Possessions	-0.4218	0.1853	X		
<i>p-value</i>	(0.0202)**	0.3270	X		
Log Avg Salary	-0.1509	-0.1779	0.3193	X	
<i>p-value</i>	(0.4261)	(0.3470)	(0.0854)	X	
Log Std Dev Salary	0.0959	-0.2728	0.0726	0.6721	X
<i>p-value</i>	(0.6141)	(0.1447)	(0.7030)	(0.0000)***	X

ACKNOWLEDGMENTS

The preceding chapter of this dissertation is coauthored with Justin Rao from Microsoft Research. This paper is being prepared for publication.

Chapter 3

Loss Aversion around a Fixed Reference Point in Highly Experienced Agents

3.1 Introduction

Employees in every line of work face tradeoffs between the exertion of costly effort and the rewards of better performance. There is considerable evidence that people adjust the effort they exert depending on whether they feel they are currently falling short of or exceeding some *internally meaningful standard*. This internal standard—referred to as a “reference point”—splits the space of outcomes into “losses” and “gains.” An agent that responds to a reference point is said to have reference dependent preferences, combining this with an increased sensitivity to losses results in loss aversion.

The predictive power of a reference-dependence depends critically on the ability to model how an agent forms and changes her reference point. If it merely affords

an additional free parameter in estimation, then it will beat a neoclassical model with uninteresting algebraic certainty. Accordingly, recent economic research, spearheaded by the theoretical work in Koszegi and Rabin (2006), has focused on pinning down reference points. Initial experimental findings indicate reference points are quite malleable. Chinese factory workers, public school teachers, and laboratory subjects all have been shown to adjust their reference point—in these cases over expected earnings—upward in response to re-framing their compensation as being docked for poor performance as opposed to rewarded for good Fryer Jr et al. (2012); Hannan et al. (2005); Hossain and List (2012). Similarly, laboratory subjects immediately lowered their reference point in response to the good performance of an opponent, exhibiting what the authors referred to as a “discouragement effect” Gill and Prowse (2012).¹

But does this malleability extend beyond experimental settings and into situations where agents have substantial experience with their effort/performance tradeoff? It is typically difficult to study reference points in the wild. A series of papers on cab drivers reveals substantial disagreement among prominent researchers Camerer et al. (1997); Crawford and Meng (2011); Farber (2008). Most relevant to our paper, Crawford and Meng (2011) argue that cab drivers’ effort provision is consistent with an income reference point fixed at the expected earnings at the start of the day and a static daily hours target. The disagreement surrounding this conclusion stems from the challenge in estimating the driver’s expected returns to effort (and the serial correlation therein) at any given time of day. And if a reference point cannot be pinned down, then it is difficult to make a reliable inference about loss aversion.²

¹Abeler et al. (2011) present what could be interpreted as a related finding. Post et al. (2008) find that reference points adjust quite fluidly for “Deal or No Deal” game show contestants.

²This has led some researchers to assert that the tight control of a field experiment, despite concerns about external validity, offers more reliable evidence on reference dependence Fehr and Goette (2007); Goette et al. (2004).

In this paper, we respond to this challenge with a unique observational environment in which nearly all payoff relevant factors can be precisely measured: the in-game performance of National Basketball Association (NBA) players. Importantly, the study of professional basketball requires few modeling assumptions to identify how reference points adjust over the course of a game. The neoclassical model (as presented in Section 3) has the strong prediction that, controlling for the quality of teams in a match-up, the marginal return (in terms of expected impact on the game outcome) of scoring one point should be a sufficient statistic for a team's chosen level of effort. Since the game is zero sum, for an even match-up the team trailing faces the exact same marginal returns to scoring a point as their opponent, who is leading by the same margin. This feature is particularly useful to study loss aversion since the neoclassical model will predict symmetry around zero. To study reference point adjustment, we make use of score fluctuation throughout the game. Throughout, we control directly for the quality of players on the court and use the betting line to capture time-varying quality differences.

Static reference points typically make different behavioral predictions than their dynamically adjusting counterparts; our setting is no exception. If within-game fluctuation of the score margin is immediately incorporated into expectations, then we would expect loss aversion to push the currently leading team to play better and the losing team to get discouraged. Conversely, if the reference point is static, for example around zero (a tie game), then we'd expect the losing team to play better and the winning team to play worse.

We find very strong evidence that NBA players are *loss averse* around a *fixed* reference point of a score margin equal to zero. Specifically, they play with progressively greater intensity and effectiveness as their team falls further behind. Controlling for the quality of all players on the court, a NBA team that falls 10 points behind will score

roughly three points more per 100 possessions. The estimate is robustly demonstrated in all four quarters of the game and with flexible methods of controlling for lineup quality. As a useful benchmark, in the last five seasons only 15 points per 100 possessions separated the absolute best offense from the absolute worst over the entire time frame. This means that an average NBA team plays like a playoff contender when 10 points behind and a bottom of the standings also-ran when 10 points ahead.

One might be concerned that this effect is driven by good teams letting bad teams mount a comeback, only to win the (now closer) game with a late push. This is not the case. The effect is robustly demonstrated to be very significant in subsamples of playoff, currently close and ex-ante evenly matched games. In most specifications, the magnitude is statistically indistinguishable in the playoffs and is only slightly smaller in competitive match-ups. For an interesting counterfactual, suppose a team could trick themselves into believing the game was tied when they were ahead (so they would now view a close win as a “loss”) but keep the reference point fixed at zero when losing. We estimate they would win 5% more of their regular season games and 10% more of their playoff series. An NBA team seeking a similar improvement would typically have add a player costing \$7–10 million per year to the roster, which is typically very difficult given the “luxury tax” penalties most teams would have to pay to make such an addition.

A final concern in interpreting these results is raised in Ariely et al. (2009): better performance in perceptual-motor tasks may not actually signal a heightened desire to win. In their study, subjects actually perform worse on routine tasks when monetary rewards are increased and there is evidence this effect has also been found in professional athletes Baumeister (1984); Dohmen (2008); Wallace et al. (2005). By studying outcomes, we might be making an unreliable inference about effort (the choice variable). To address this concern, we dig deeper into player behavior. The change in offensive output is

driven by the losing team's improved performance in the most effort-intensive aspects of basketball. Losing teams secure more rebounds, turn the ball over less and see their star players work to shoulder an even larger burden of the scoring load than usual. We do not however, observe significant changes in shooting accuracy (conditional on shot location) during game play. Interestingly, players shoot "free throws"—unguarded shots taken after a foul is called—with *lower* accuracy when trailing. We also observe lower free throw accuracy in the playoffs and when the game is nationally televised, both scenarios in which a player is presumably motivated to try harder. A consistent story thus emerges. When trailing, players perform better on effort-intensive tasks and worse on focus-intensive tasks (the opposite being true for leading teams), but in both cases the evidence supports loss averse preferences.³

Had we only studied free throws we might have been tempted to conclude that reference points adjust quickly and players exhibit disappointment aversion. In a related paper, Pope and Schweitzer (2011) study professional golfers and conclude, based on a lower accuracy on "birdie" putts (a gain) than similarly difficult "par" putts (avoiding a loss), that golfers exhibit loss aversion around the reference point of "par." A skeptic, however, could argue that the critique in Ariely et al. (2009) has bite; the worse performance on birdie putts is actually evidence of trying harder, but this effort is counter-productive. We guard against this critique by studying outcome metrics which vary from focus- to effort-oriented. We thus view our findings as some of the strongest evidence to date for both loss aversion and fixed reference dependence in highly experienced and incentivized agents. Fixed reference points, combined with robust loss aversion, in turn substantially increase the predictive power of reference-dependent models.

³The negative impact of effort on performance for focus tasks has been explained by what psychologists refer to as detrimental "self-focus" (thinking too much about how to perform the task rather than just doing it) Beilock et al. (2002); Gucciardi and Dimmock (2008); Neiss (1988); Wilson et al. (2007). It is sometimes referred to as "performance pressure."

3.2 Background and Preliminaries

An NBA basketball game lasts for forty-eight minutes, with ties being settled by five-minute overtime periods, repeated as necessary. Five players from each team, referred to as the “lineup,” share the court at any given time. Game play is a sequence of “possessions” in which one team has the ball and attempts to score a “basket,” which is worth two or three-points, depending on the distance the shot is taken from. If a player is fouled while attempting to score a basket, he is awarded “free throws,” unguarded 15-foot shots worth one point each, corresponding to the point value of the attempted shot. The large number of possessions and high frequency of scoring (roughly half of all possessions result in some points for the offense) are crucial features of basketball that allow for win probabilities and the implied incentives to be easily modeled—in the Appendix, we discuss exactly how this is done.

The magnitude of an action’s expected impact on the game’s outcome depends heavily on the score margin and time remaining. We’ll use the “win value of a point” (WVP)—the expected (per-point) increase in the probability of winning the game for making a two-point basket in a specified game state—to quantify the importance of any particular moment in the game.⁴ Formally, let

$$WVP \equiv \frac{PW(M_p + 2, t) - PW(M_p, t)}{2},$$

where PW is the win probability function estimated in the Appendix.

In Figure 3.1, we plot WVP for two evenly matched teams. There is considerable variation in the expected importance of a point throughout the course of a game. Since basketball is zero-sum, point importance is equal for both teams at any given moment;

⁴Just as in voting, either all points or pivotal or none are. The expectation effectively captures the probability the game will be decided by a single shot, which is naturally higher when the game is close and not much time is remaining.

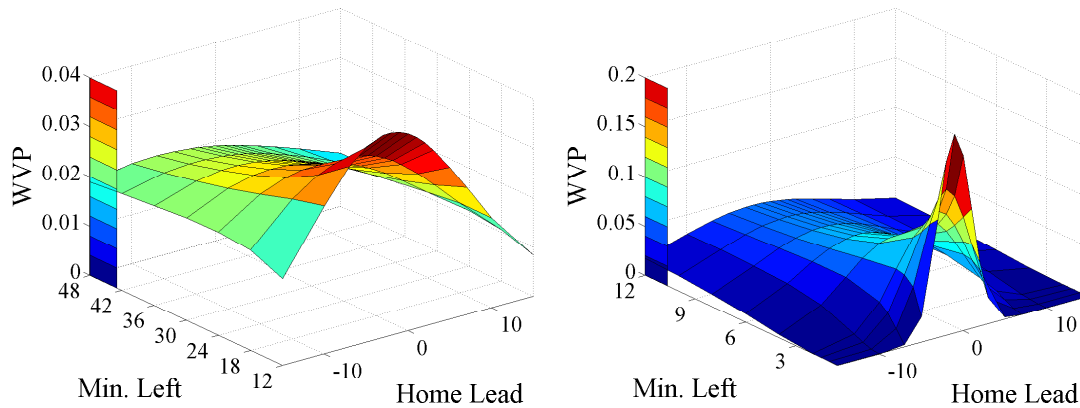


Figure 3.1: The marginal impact of a point on win probability, as a function of score margin and time remaining in the game for the first three quarters (left) and only the final quarter (right).

this can be seen graphically in the symmetry of the WVP function with a spine at a lead of zero.⁵ However, if the teams were of different quality, then the spine of the WVP function would be shifted. If, for example, the home team was superior, then WVP would be maximized (for a given amount of time remaining) when the away team had a sufficient lead to give them a 50% chance to win the game.

3.3 Theoretical Framework

Most of the outcomes metrics used in this paper are determined by the simultaneous exertion of effort by the home (e_h) and away (e_a) teams and these outcomes will have different interpretations depending on which team is on offense. As an example, our key performance measure is the average number of points scored by the offense on one possession. For a fixed matchup of players, define the performance mapping for this

⁵Since the win probability surface is non-linear, the leading and trailing team face opposing risk-return tradeoffs. See Goldman and Rao (2014) for more details.

metric as:

$$\rho(e_o, e_d) \equiv E[Pts_p | e_o, e_d].$$

Throughout, it will be assumed that ρ is concave. If the returns to effort are monotonic for both teams, then ρ is increasing in the first argument and decreasing in the second.

3.3.1 “Standard” Model of a Win-Maximizing Team

NBA teams face strong financial incentives to win. Winning has been linked to increased home gate revenues Berri et al. (2004) and helps qualify a team for additional, lucrative post-season games. Individual players may have incentives to deviate from a win-maximizing strategies in order to promote the appearance of their own productivity, but such behavior was not found to have a large impact on team-level shot selection in Goldman and Rao (2014). As such, we formulate our “standard” model as one of a team that maximizes utility *only* from winning games, while paying a separable cost for the exertion of effort on each possession. We explore below the ability of this benchmark to rationalize different patterns of performance.

Specifically, consider two competing basketball teams with preferences *only* for winning the game and facing a separable marginal cost of effort normalized to one ($c(e)=e$).⁶

Take an offense and defense competing on possession p where the offense is ahead by M_p with t_p minutes remaining in the game. Let PW and WVP be the win probability and win value of a point functions discussed in section 2. For notational convenience, let us assume that all NBA possessions end in zero or two-points.⁷ Recall

⁶This normalization is without loss of generality because we have not specified the functional form of ρ .

⁷This is only done for notational convenience in this section. All results could be easily generalized.

that WVP is defined as half the win probability return of making a two-point shot. Then on a given possession, the offense is seen to behave according to

$$\begin{aligned} \max_{e_o} \quad & U_{win_o} \cdot E[PW(M_{p+1}, t_{p+1}) | M_p, t, e_o, e_d] - e_o \\ & = U_{win_o} \cdot (PW(M_p, t_p) + WVP \cdot \rho(e_o, e_d)) - e_o. \end{aligned}$$

The first order condition is

$$U_{win_o} \cdot WVP(m_p, t_p) \cdot \frac{\partial \rho(e_o, e_d)}{\partial e_o} = 1. \quad (3.1)$$

The defense's problem is exactly parrallel and has solution

$$U_{win_d} \cdot WVP(m_p, t_p) \cdot \left(\frac{-\partial \rho(e_o, e_d)}{\partial e_d} \right) = 1. \quad (3.2)$$

Note that margin (m_p) and time (t_p) enter these first order conditions only through their dependence on WVP. This allows us to state a convenient result.

Result 1. *In a fixed matchup between two teams with utility functions given by the standard model, the score margin and time remaining can effect effort provision, and thus outcomes (ρ), only through their dependence on WVP.*

Intuitively, both the offense and defense are more motivated on possessions with high WVP, meaning they always scale effort in the same direction. If ρ is symmetric in offensive and defensive effort, then this model necessarily predicts equal performance across game states. See Figure 3.2 as an example.

Of course the offense and defense may have different returns to effort. For instance, if the defense had a more efficient technology to reduce scoring probability in high importance moments, then we'd expect scoring rates to drop at the end of close games for both teams. Alternatively, if one team has a higher utility of winning, for

See Appendix for further discussion.

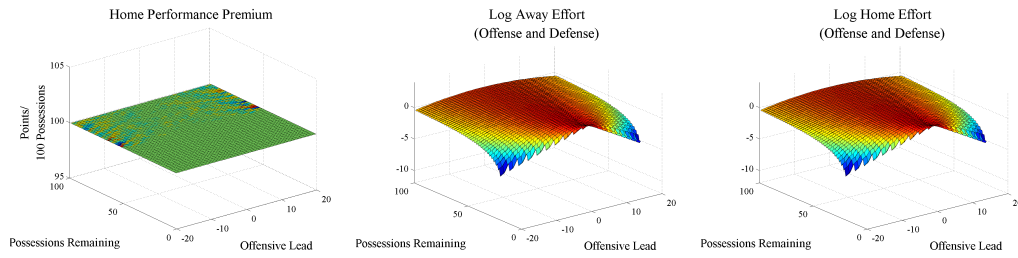


Figure 3.2: Prediction of the win-maximizing model with symmetric performance function, $\rho = 1 + .2(e_o^{1/2} - e_d^{1/2})$, specified identically for both the home and away team.

instance because they are playing in front of their home crowd, then $U_{win_h} > U_{win_a}$ and we will see them exert higher overall levels of effort throughout the game (particularly in high leverage moments) and thus better performance. See Figure 3.3 for an example.

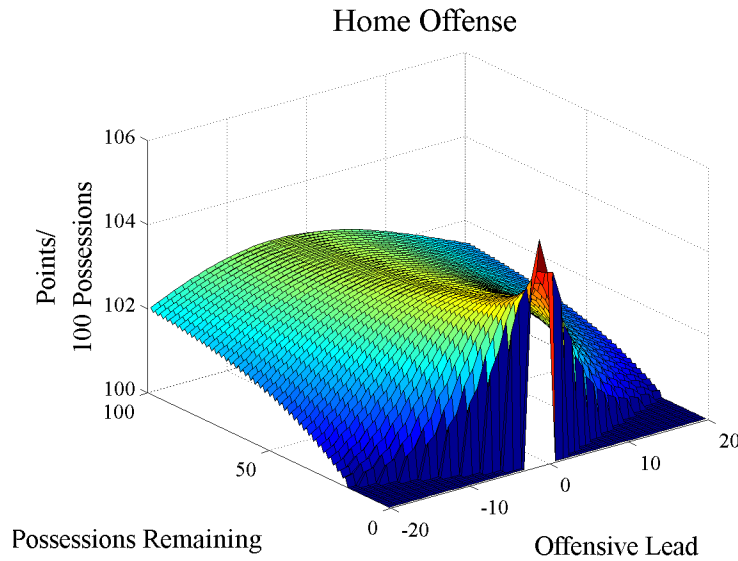


Figure 3.3: Predictions of the win-maximizing model with symmetric performance function, $\rho = 1 + .2(e_o^{1/2} - e_d^{1/2})$ and a stronger home preference for victory $U_{win,h} = 1.5 \cdot U_{win,a}$.

Finally, one might imagine that certain teams tend to have greater marginal returns to effort. These teams will tend to differentially outperform their opponents in high WVP

moments. If teams with high marginal returns to effort are also better teams, then WVP will generally be higher when these teams are behind. This is because when a good team is trailing, the game is more likely to end close (so that WVP is high) because they are expected to come back. When a bad team is trailing, the likelihood of a comeback is lower and thus WVP is lower as well. If good teams do indeed have higher marginal returns to effort, then a win-maximizing basketball team could appear to perform differentially better when behind. However, such a model must predict an even larger performance boost for the superior team at the very end of close games (when WVP is by far the highest). Figure 3.4 provides an example. As will be shown in our empirical section, such a “clutch” performance boost is not supported in the data.

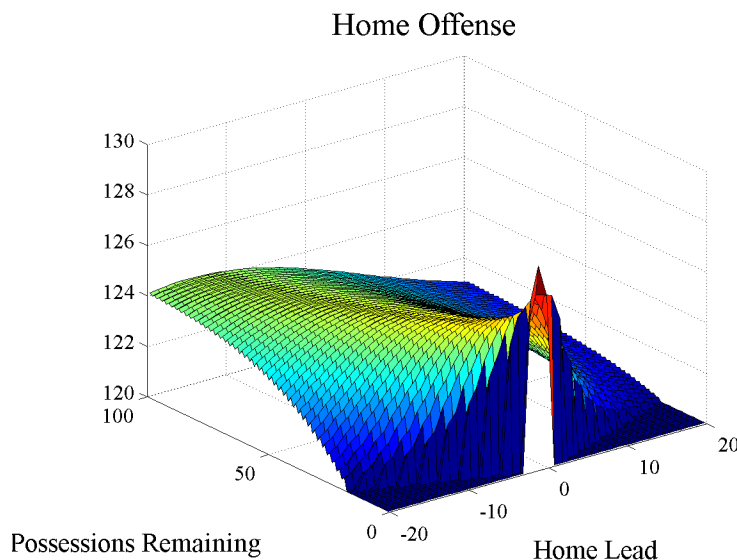


Figure 3.4: Predicted offensive performance for a win-maximizing team with performance functions for the home and away team (respectively) given by $\rho_H = 1.2 + .3e_o^{1/2} - .2e_d^{1/2}$ and $\rho_A = 1 + .2e_o^{1/2} - .3e_d^{1/2}$.

More generally, refer back to (3.1) and (3.2) and notice that the score margin and time remaining (m_p, t_p) enter the first order condition for effort *only* through their impact

on *WVP*. This is the key restriction of the “standard” model—for a given matchup, selected effort, and thus performance, can only be a function of incentives.

The next two subsections discuss extensions to the standard model that incorporate reference-dependent preferences. These models predict different patterns of team motivation and performance across game state, but on their own cannot predict patterns of choking observed on free-throws. Such discussion is reserved for the final subsection.

3.3.2 Reference-Dependent Preferences over Outcomes

We now maintain that teams get utility only from the outcome of the game, but allow that those outcomes are experienced relative to a reference point. It is immediate that loss aversion around a *fixed* reference point is not a meaningful construct since winning is a binary variable, hating to lose just means loving to win (and visa versa). Formally, it would rescale the value of U_{win} but would not effect the pattern of effort exertion across game states. Alternatively, the reference point could be *endogenous* to current game state and selected level of effort in order to represent the foreword looking probability of a win. Since losses around this reference point would be more painful than gains, both team’s would be motivated to limit their exposure to variance (relative to the standard model). This would lead to a greater level of effort chosen by the leading team (increasing PW towards 1 reduces variance) and a lesser effort by the trailing team (increasing PW moves it toward 0.5, increasing variance). This is exactly as in the *discouragement effect* found in Gil and Prowse (2012) and we direct an interested reader to the theory section of that paper.

3.3.3 Reference dependent preferences over game state

Now we allow that players are not motivated just by the final outcome of the game, but also get a discounted flow of utility based on the current score margin. The idea is that each team has a reference point in mind for how they expect to play against a given opponent and gain utility from their performance with respect to that benchmark. Let $v(\cdot)$ denote a standard loss averse utility function ($x > 0 \implies v'(x) < v'(-x)$) over the difference between the current score margin and the prescribed reference point. For example,

$$v(x) = 1\{x \geq 0\} \cdot x^{\gamma_A} - \lambda \cdot 1\{x < 0\} \cdot |x|^{\gamma_B},$$

where γ_A and γ_B denote separate utility exponents for teams that are *ahead* and *behind* their reference point. The idea is that players get more psychic utility from moving out of a losing state than moving into a winning state. For now, we assume the reference point (R) is fixed at a constant level and does not update in response to game-play. Then utility for the home and away, with T possessions remaining in the game, is respectively given by

$$\begin{aligned} V_{o,p} &= E \left(\sum_{t=0}^T \beta^t [v(M_{p+t} - R) - e_{o,p+t}] \right) \\ V_{d,p} &= E \left(\sum_{t=0}^T \beta^t [v(R - M_{p+t}) - e_{o,p+t}] \right), \end{aligned} \tag{3.3}$$

each of which can be represented recursively as (for example),

$$V_{o,p}(M_p) = v(M_p - R) + \beta E [V_{o,p+1}(M_{p+1})].$$

This recursive representation allows the model to be easily solved for specified parameter values (as is done in Figure 3.5). This parameterization implies that losing teams will

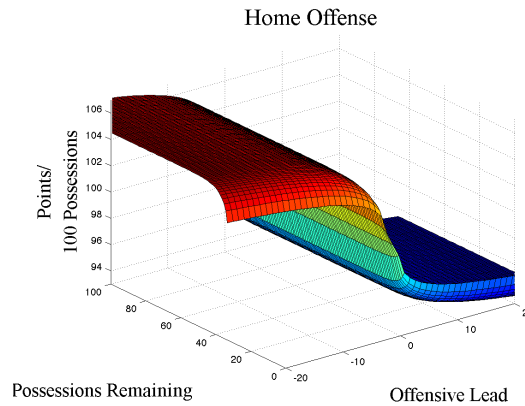


Figure 3.5: Predicted performance premium of a ref-dependent model over the score margin. $\gamma_A=.8$, $\gamma_B=.9$, $\lambda = 1$, $\beta = .8$. $\rho = 1 + .2(e_o^{1/2} - e_d^{1/2})$ is specified identically for each team.

outperform winning teams and that the amount by which they do so will grow with the score margin.

Of all the models considered, this is the only one capable of rationalizing the following factors of the data. (1) improved performance by the trailing team by an amount that is roughly independent of the amount of time left in the game (2) a performance premium function that is consistently steepest around a tie game (3) a relatively small performance boost for superior teams in high WVP moments.

3.3.4 Summary of Predictions

Our “standard” model is quite flexible in that it allows opposing teams to experience any possible relation between effort and performance, but requires that they have preferences only to win the game. As demonstrated, this allows for effort and performance to vary substantially over the course of the game, but only in such a way that WVP is a sufficient statistics for both. Thus, if we see that one team is systematically outplaying another when they are behind, we must conclude that WVP is actually high

in these moments and that this team is superior. We must then also expect to see this team have a significant performance boost in other high WVP moments—specifically, the final moments of close games. As discussed in the following section, the losing motivates effect is found quite broadly and we find no evidence of any kind of clutch performance boost for superior teams. Thus, we must look for alternative explanations for team behavior.

We consider two forms of loss averse preferences to help us rationalize the data. Loss aversion over game outcomes, with a reference point that immediately updates and is *endogenous* to chosen effort, predicts disappointment aversion as in Gill and Prowse (2012). This factor, induces both leading and trailing teams should seek to avoid variance in the final outcome. In particular, NBA teams that find themselves behind, should reduce effort in order to avoid simultaneously raising their reference point and exposing themselves to disappointment. Such a model unambiguously predicts the opposite of the “losing motivates” model we have observed.

Alternatively, loss aversion over the score margin is considerably more flexible. If the reference point is fixed over the course of the game, we have the unambiguous prediction that losing teams will outplay winning teams. If the reference point updates with a lag, then we may also find that lagged levels of a team’s lead may be positive predictors of a teams performance.

3.4 Main Results

We first estimate the aggregate impact of game state on team performance. We are mindful of two possible confounds. First, good teams are more likely to be ahead. If no attempt is made to control for player or team quality is made, then mechanically,

the offensive team's score margin and scoring efficiency should be positively correlated. Second, losing teams may be more likely to keep their better players in the game. Such behavior is also not rationalized by a standard model and may reflect the loss averse preferences of a team's coaching staff. We refer to this as the effect of *coach loss aversion* in order to distinguish it from the effect of *player loss aversion* which we define as the impact of a change in the score margin on the performance of a *fixed set of players*.⁸

3.4.1 Quantifying loss aversion

Our first specification does not control for lineup composition, it is designed to measure the combined effect of player and coach loss aversion on team performance. We use the closing Las Vegas betting spread for each NBA game as an additional regressor to control for differences in team quality.⁹ The spread is the score margin that a team must win by in order for an even odds bet on them to be declared a winner. Empirical research demonstrates it represents the expected median score differential for that game and is a strong (if not sufficient) control for differential team quality in each game Paul et al. (2004). Specifically, we estimate:

$$E[Pts_p] = \alpha + \delta_1 Home_p + \delta_2 Spread_p + \gamma Playoffsp_p + \beta_1 Lead_p \times (1 - Playoffsp_p) + \beta_2 Lead_p \times Playoffsp_p, \quad (3.4)$$

where β_1 and β_2 represent the gross impact of loss aversion on team performance in the regular season and playoffs respectively. We estimate (3.4) separately for each quar-

⁸Note that insofar as coaches call better plays when trailing, for a fixed set of players, we'll attribute this to the players.

⁹Obtained from covers.com.

ter of the game and use fixed effects to control for the way each possession originates.¹⁰

Table 3.1: Total impact of lead on offensive performance (Points/100 possessions)

		1 st Quarter	2 nd Quarter	3 rd Quarter	4 th Quarter	Avg Value
All Data	$\hat{\beta}_1$	-.3205 (.0396)***	-.4263 (.0255)***	-.3683 (.0250)***	-.6930 (.0344)***	-.4520 (.0159)***
	$\hat{\beta}_2$	-.1378 (.1826)	-.3060 (.0960)***	-.1645 (.1157)	-.6447 (.1634)***	-.3132 (.0719)***
Evenly Matched Teams	$\hat{\beta}_1$	-.2481 (.0559)***	-.4310 (.0327)***	-.3580 (.0315)***	-.6435 (.0440)***	-.4202 (.0211)***
	$\hat{\beta}_2$	-.0965 (.2655)	-.1608 (.1190)	-.0531 (.1440)	-.5699 (.2079)***	-.2201 (.0964)**
Close Games	$\hat{\beta}_1$		-.4101 (.0332)***	-.4049 (.0385)***	-.7426 (.0517)***	-.5192 (.0242)***
	$\hat{\beta}_2$		-.2852 (.1431)**	-.2522 (.1962)	-.7002 (.2516)***	-.4125 (.1166)***

Estimates of β_1 and β_2 presented in Table 3.1. The impact of of lead is strongly negative and easily statistically significant in all specifications and all four quarters of the regular season data. The overall average for regular season games indicates that trailing by 10 points increases output per 100 possessions by 4.5 points. As a frame of reference, 10 points per 100 possessions typically separates very good offenses from very bad offenses. The observed magnitudes are indeed very large. A playoff contender morphs into merely a below average team up 10 points and one of the best teams in the league when down by 10. Estimates of the impact in playoff games, which are economically more important, naturally have larger standard errors due to the smaller sample size, but are significantly negative in the second and fourth quarters and only slightly smaller in

¹⁰Possessions originating off steals or defensive rebounds tend to be more valuable than those originating off of dead balls or made baskets by the opposing team. Controlling for this soaks up some residual variability but does not significantly effect our results.

aggregate than the estimates obtained from regular season data. The global average for playoffs is a strongly significant 3.13 points per 100 possessions, or about 70% of the regular season effect. Interestingly, even though the magnitude of the effect is smaller in the playoffs it is actually twice as likely to impact the game outcome given how much closer playoff games are on average.

A natural concern is that the effect is driven by good team vs. bad team matchups in which the good team plays lackadaisically with a lead, only to “turn it on” late in the game to ensure the win. Uneven matchups are less likely to occur in the playoffs, so the fact that we see significant loss aversion there as well is our first clue this concern does not have bite. Panels 2 and 3 restrict the estimation to ex-ante evenly matched teams (games where the spread was less than 6 in magnitude) and close games (games which were within 10 points at the start of that quarter). Neither restriction significantly impacts the estimates and all specifications show a larger gross impact of losing on performance in the fourth quarter. As will be shown in future specifications, this is driven in part by an increased impact of coach loss aversion.

Player loss aversion

Our next specification allows us to isolate only the impact of player loss aversion on team performance.¹¹ In order to do this, we non-parametrically condition on the entire five-man lineup employed by both teams at any given moment of game play. Specifically, we estimate:

¹¹We discuss this later, this could also include play-calling by the coach.

$$\begin{aligned}
E[Pts_p] = & \alpha_{Off_p, Def_p} + \delta_1 Home_p + \delta_2 Playoffs_p \\
& + \beta_1 Lead_p^* \times (1 - Playoffs_p) + \beta_2 Lead_p^* \times Playoffs_p,
\end{aligned} \tag{3.5}$$

where Off_p and Def_p denote the unique five man offensive and defensive players employed on possession p and α_{Off_p, Def_p} is a unique fixed effect for each combination. Thus specification (3.5) is identified only by comparing performance of exactly equivalent match-ups across given different score margins, truly holding all else equal (unique 10 player combinations). Specification (3.6) is slightly less general. It achieves much more power by allowing for additively separable fixed effects for the offensive and defensive lineup, but may be confounded if match-up specific effects interact strongly with the score margin.

$$\begin{aligned}
E[Pts_p] = & \alpha_{Off_p} + \nu_{Def_p} + \delta_1 Home_p + \delta_2 Playoffs_p \\
& + \beta_1 Lead_p^* \times (1 - Playoffs_p) + \beta_2 Lead_p^* \times Playoffs_p
\end{aligned} \tag{3.6}$$

By incorporating so many fixed effects, specifications (3.5) and (3.6) reduce our data to a short, wide panel. Since the current score margin is endogenous to lagged performance, this can create substantial dynamic panel bias. Our solution is to again preform the entire analysis separately for each quarter and to use the score margin at the beginning of that quarter (denoted $Lead^*$ in (3.5) and (3.6)) as a proxy for the current score.

Estimates from our single fixed effect model (3.5) rely on a very limited set of comparisons and are thus noisy but nonetheless demonstrate a negative, statistically significant ($p < .01$) impact of lead on score margin for both regular season and playoff data. Our additively separable fixed effect model has much higher precision and produces

Table 3.2: Player (only) impact of lead on offensive performance (Points/100 possessions)

		1 st Quarter	2 nd Quarter	3 rd Quarter	4 th Quarter	Avg. Value
Double Fixed Effect (3.6)	$\hat{\beta}_1$	-.4413 (.0615)***	-.3152 (.0343)***	-.2481 (.0218)***	-.3064 (.0316)***	-.3277 (.0201)***
	$\hat{\beta}_2$	-.5913 (.2001)***	-.3127 (.0993)***	-.0781 (.0722)	-.2288 (.0785)***	-.3027 (.0619)***
Penalized Regression (3.7)	$\hat{\beta}_1$	-.2390 (0.0345)***	-.2763 (0.0208)***	-.2424 (0.0621)***	-.3536 (0.0145)***	-.2778 (0.0115)***
	$\hat{\beta}_2$	-.1625 (0.1303)	-.2871 (.0796)***	-.1798 (.0964)*	-.2866 (.0543)***	-.2290 (.0434)***

similar point estimates. Here we find strongly significant loss averse behavior in every quarter of the regular season and all but the third quarter of our playoff data. The aggregated estimates are shown in the final column; they are roughly two-thirds the magnitude of the overall loss aversion estimates shown in Table 3.1 indicating that substitution decisions—preferences of the coaching staff—contribute to the overall effect, particularly in the fourth quarter.

Our final specification further generalizes (3.6) to allow for each player to have a unique additively separable impact on his team’s performance. This type of regression model—often referred to as an adjusted plus/minus model—is a popular way of measuring the value of professional basketball players¹². We use a ridge penalization (chosen by cross-validation) to control the coefficient estimates for each player and then add additional, unpenalized regressors to estimate the impact of the score margin. Formally,

¹²Regularized regression has been used to evaluate NBA player performance since Sill (2010). Jeremias Englemann has worked to improve and popularize the technique and it is now the basis for the ‘real plus minus’ stat posted on ESPN.

we estimate

$$E[Pts_p] = \sum_{i \in Off_p} \alpha_i - \sum_{i \in Def_p} v_i + \delta_1 Home_p + \delta_2 Playoffs_p \quad (3.7)$$

$$+ \beta_1 Lead_p \times (1 - Playoffs_p) + \beta_2 Lead_p \times Playoffs_p,$$

where the penalization factor ($\lambda = 2,400$) is applied to the L^1 norm of α and v . This method imposes strong separability in the contribution of each player, but in returns provides a dramatic increase in power. The penalization also biases all coefficients on player effects toward zero. We do not penalize the coefficients on our lead variables, but some bias could be transmitted through to these.

Taking advantage of the increased power, we replace the estimates of β in (3.7) with a non-parametric function of the game state. Figure 3.6 demonstrates the results.

3.4.2 Fixed vs. quickly adjusting reference point

Recall that a quickly adjusting and fixed reference points make entirely different predictions in this setting. If a loss-averse team that is leading quickly adjusts the reference point up to the current score margin, then a close victory will now be viewed as a “loss” from this vantage point. We use in game fluctuations in score margin to get at this question. For example, if a team jumps out to a ten point lead at half time, will they adjust their reference point up? Will their opponent adjust down? Does that ten point differential now bifurcate the margin space into new gain and loss regions? If so, then we would expect an early lead to predict improved performance over the remainder of the game.

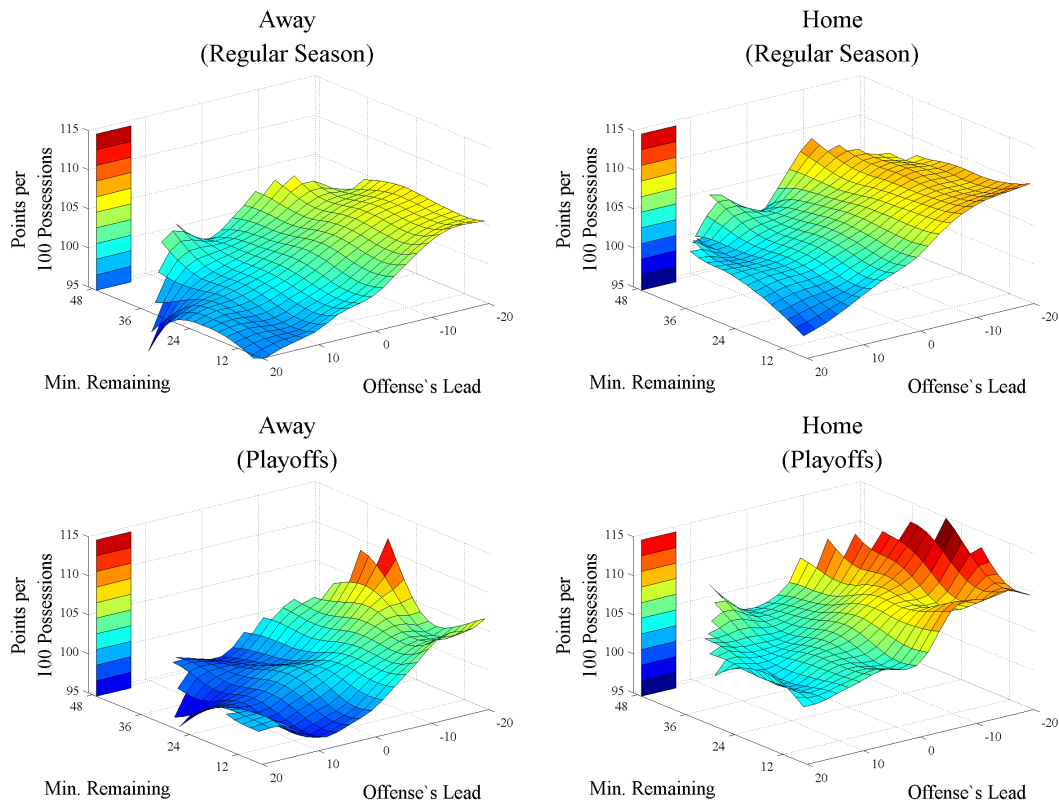


Figure 3.6: Offensive efficiency achieved by “replacement players” across game states as estimated according to (3.7). Holding player quality constant, both home and away teams improve dramatically as they fall behind and the effect is only slightly smaller in the playoffs.

Reduced form analysis

We test for this by including additional lags of the score margin our specifications given in Table 3.1. If basketball dynamically update their reference point over the course of a game, we should expect coefficients on these regressors to be positive. If they maintain a fixed reference point throughout the game then the current score margin should be a sufficient statistic for the entire history of the game and the coefficients should be zero.

The results for all four of our specifications are printed in Table 3.3. Nearly all

the coefficients on lagged lead lack statistical significance. Of the 24 estimates only 4 are statistically significant at the 10% level. In particular our most powerful specification, based on ridge regression, provides very tight estimates around zero for all coefficients, indicating that reference point updates over the course of a basketball game are neither statistically nor economically significant. Our double fixed effect specification shows an interesting pattern of significant positive coefficient on first quarter performance and a negative coefficient on second quarter performance. Identification here is dominated by the performance of common lineups and, in particular, each team's starting lineup. Starting lineups typically play the most in the first and third quarter. Then this pattern of coefficients indicates that starters play the best in the third quarter when they had built a lead in the beginning of the first quarter, but saw their teammates give up a lead at the end of that quarter. Further investigation may be warranted.

Table 3.3: Impact of lagged score margins on offensive performance (Points/100 possessions)

		Lag Margin at 42 min left	Lag Margin at 36 min left	Lag Margin at 24 min left
Penalized Regression (3.7)	Second Quarter	-.0291 (.0245)		
	Third Quarter	-.0163 (.0334)	-.0340 (.0240)	
	Fourth Quarter	-.0164 (.0321)	.0182 (.0293)	-.0025 (.0182)

To further complement this analysis, we break down our nonparametric estimates from Figure 3.6 into separate functions of game state depending on performance in the first quarter. Specifically, we define a team that was ahead (behind) by more than 5 at the end of the first quarter to have been 'Ahead Early' ('Behind Early'). This results in a roughly equal split of our data into thirds. We use our same ridge regression approach

to control for player quality, results for each combination of first quarter performance and home/away are presented below. Estimates are presented in Figure 3.7. There is no evidence that first quarter performance has an impact on the level or slope of the performance function in later quarters.

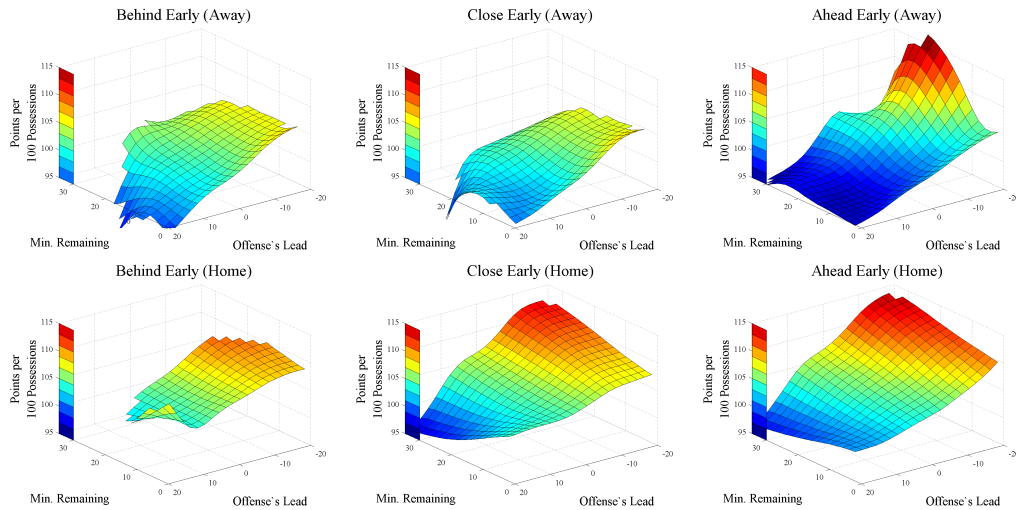


Figure 3.7: Nonparametric estimates of the impact of game state over the last 36 minutes broken up by performance over the first 12.

Estimating the reference point as a latent variable

In the last subsection we looked at evidence of drift in the RP, and found none. In this section we'll try to directly estimate it as a latent variable. The basic strategy is to estimate Figure 3.6 on various subsets of the data and look for areas of greatest steepness. Figure 3.8 does this for the aggregate data; it gives estimates the derivative in lead from the regression results presented in Figure 3.6. The first derivative is maximal at 0, providing suggestive evidence that players use a score margin of zero (or a tie game) as fixed reference point for effort allocation.

We now repeat this analysis splitting our sample into teams that are favorites

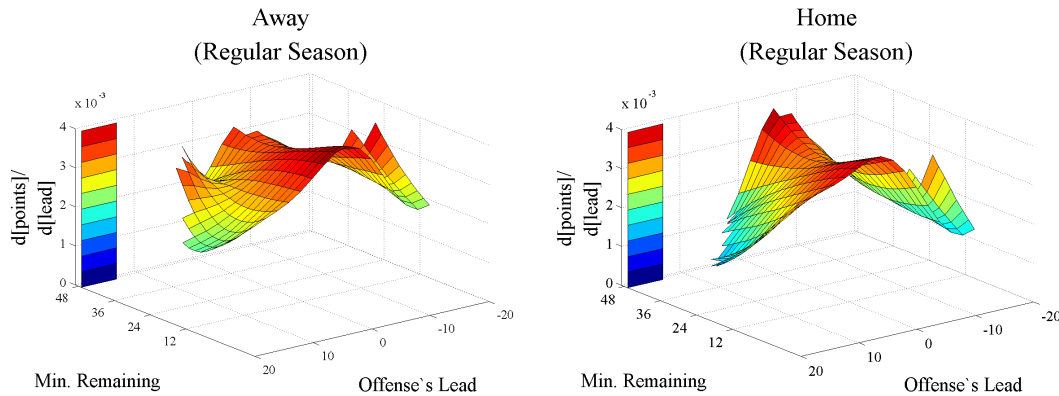


Figure 3.8: Nonparametric derivative of the estimates in Figure 3.6. Both Home and Away teams appear to show the sharpest changes in relative performance around a score margin of zero.

(spread > 5), underdogs (spread < -5), and evenly matched (anywhere in between) and estimating a nonparametric game-state function for each group. Results are presented in Figure 3.9. Note that going from favorite to underdog does not display a significant level effect because that has already been soaked up by our controls for player quality. The overall shape of the game function seems to be roughly identical in all six pictures (except that it is shifted up for the home team). More importantly, we do not observe a significant performance boost for a favored team in high WVP moments (at the end of close games). This was a prerequisite for the standard model to explain a pattern of apparent loss motivation in the data.

3.5 Additional Results and Robustness

In the introduction we discussed two potential critiques of our findings of strong, fixed reference point loss aversion in NBA players. We dispatched the first, that the effect is driven by behavior in meaningless games or from uneven matches, in the last

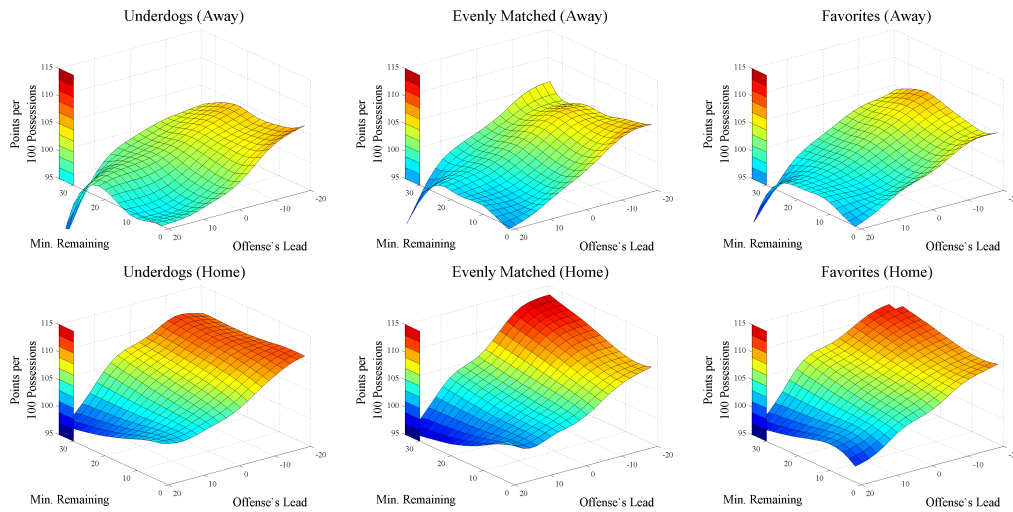


Figure 3.9: Nonparametric estimates of the impact of game state broken up by the pre-game spread.

section. The second is that effort/performance mapping is not monotonic. An additional concern is that it is in fact the referees driving the effect by favoring the trailing team (not this would presumably not apply to the coach loss aversion inferred through substitution patterns). We address these concerns here.

3.5.1 Breaking down loss aversion: Effort intensive tasks

In basketball, after a missed shot there is scramble for the ball known as rebounding. Securing a rebound involves physically “boxing out” and out jumping the opponent. We again apply our fixed effect estimation methodology, the results are given in Table 3.4.

A rebound is worth a possession and adds a little value to the resulting offensive possession. All together it is worth around 1.2 points. There is roughly one available rebound for every 2 possessions. +1 to rebounding rate is worth around +.6 to team efficiency. The LA effect on rebounding rate is worth around +.07. We now estimate the same model on turnover probability. The results are presented in Table 3.5.

Table 3.4: Impact of game state on Offensive Rebounding Rate (Offensive Rebounds /100 missed shots)

		Quarter 1	Quarter 2	Quarter 3	Quarter 4	Avg. Effect
Spread Regression	$\hat{\beta}_1$	-.1649 (.0252)***	-.1142 (.0152)***	-.0983 (.0124)***	-.0933 (.0114)***	-.1177 (.0085)***
	$\hat{\beta}_2$	.0558 (.1036)	-.1898 (.0629)***	-.0401 (.0503)	-.0435 (.0417)	-.0544 (.0344)
Double Fixed Effect	$\hat{\beta}_1$	-.2249 (.0429)***	-.0977 (.0171)***	-.1114 (.0133)***	-.0854 (.0116)***	-.1298 (.0124)***
	$\hat{\beta}_2$	.0386 (.1856)	-.1779 (.0735)**	-.0203 (.0573)	-.0151 (.0438)	-.0437 (.0531)

Table 3.5: Impact of game state on Turnover Rate (Turnovers /100 possessions)

		Quarter 1	Quarter 2	Quarter 3	Quarter 4	Avg. Effect
Spread Regression	$\hat{\beta}_1$	.1011 (.0126)***	.1058 (.0081)***	.0838 (.0065)***	.1315 (.0061)***	.1055 (.0044)***
	$\hat{\beta}_2$	.0732 (.0552)	.0775 (.0325)**	.0662 (.0274)**	.0640 (.0239)***	.0702 (.0184)***
Double Fixed Effect	$\hat{\beta}_1$	.1086 (.0221)***	.1104 (.0092)***	.0913 (.0070)***	.1287 (.0060)***	.1097 (.0064)***
	$\hat{\beta}_2$	.2306 (.0921)**	.0863 (.0382)**	.0704 (.0313)**	.0630 (.0234)***	.1126 (.0268)***

A turnover costs a possession which is worth slightly more than 1 pt. As such +1 to turnover rate is worth slightly more than +1 to points scored. The LA effect on turnovers is worth around +1.1.

Overall, effort intensive tasks can account for 80% of the loss aversion effects we observe.

3.5.2 Breaking down loss aversion: focus intensive tasks

The most focus intensive action in NBA basketball is the free throw. A free throw is an uncontested 15-foot shot taken while the play is stopped after the commission of a foul. We use a simple fixed effect specification to capture the impact of the score margin on the probability of making a free throw. Results are presented in Table 3.6. As before, $\hat{\beta}_1$ and $\hat{\beta}_2$ represent the coefficient in regular season and playoff data respectively.

Table 3.6: Impact of Game State on Free Throw Percentage (%)

	Quarter 1	Quarter 2	Quarter 3	Quarter 4	Average Value
	(1)	(2)	(3)	(4)	(5)
$\hat{\beta}_1$	.0610 (.0290)**	.0244 (.0167)	.0291 (.0135)**	.0187 (.0115)	.0333 (.0095)***
$\hat{\beta}_2$	.0177 (.1307)	.0961 (.0718)	.0299 (.0522)	.0286 (.0416)	.0431 (.0409)

† Player-season FE included. Standard errors clustered at game level.

Contrary to all our findings so far, we find evidence that players actually have *worse* free throw accuracy when trailing (the statistical significance is not overwhelming, however). A look at playoff and nationally televised games—those with higher economic and psychological stakes—sheds some light on this potential inconsistency (of our sample of 7,410 games, 332 occur in the playoffs and 646 are nationally televised by a cable (ESPN, TNT) or broadcast (NBC) network). In Table 3.7 we use the same fixed-effect estimator and add dummy variables for playoff and national TV status.

The estimates indicate an across the board decline in free throw shooting percentage in playoff and nationally televised games. The difference between the effects on home and away shooters in specification (1) is statistically insignificant for both cases. As such, specification (2) aggregates over home and away to show that both playoffs (coeff. $-.98, t = -2.92$) and national TV (coeff. $-1.4, t = -5.25$) induce significant drops in

Table 3.7: Impact of Game Type on Free Throw Percentage (%)

	Away Nat. TV	Home Nat. TV	Nat. TV	Away Playoffs	Home Playoffs	Playoffs
(1)	-1.3839 (.5144)***	-.6270 (.4820)		-.8687 (.3712)**	-1.8650 (.3843)***	
(2)			-.9882 (.3687)***			-1.3791 (.2796)***

free throw accuracy. This provides overwhelming statistical evidence of a performance decline when the player presumably “cares” more about the outcome. This provides an explanation as to why players shoot free throws better while leading and worse while losing. When they try harder, they do worse. When they are losing, they try harder. For most aspects of basketball this helps, but for a focus intensive task like free throw shooting, it hurts. In a related paper, we also find they shoot free throws worse in other “high pressure” moments, such as at the end of close games.¹³ This nuance of the results, instead of revealing an inconsistency, actually strengthens our primary finding.

3.5.3 Do Officials Favor The Losing Team

A final concern to address is that it is in fact the referees, not the players, driving loss aversion. NBA referees are not perfect, with past work documenting a bias for the home team and for players of the same race Price and Wolfers (2010). Perhaps referees are predisposed to “keep the game interesting.” We have already seen evidence that this sort of bias cannot account for all the loss aversion we observe as it is unlikely the it would inspire star players to shoot more or coaches to play their better players when losing, but not winning. Any analysis of referee bias is challenging because of the endogeneity of players’ actions to the purported bias. These concerns are present her as

¹³Put citation. Psychologists have linked this to increased “self focus.”

well, but what we will be able to show is that the magnitudes of foul rate differences are not of the same order as the point differentials we saw—it is simply not plausible they are driving the effect, but we cannot rule out a very small contribution.

Taking each whistled foul as a data point, we replicate specification 3.5 to predict whether or not it will be called on the offense. The baseline offensive foul rate is about 8% in our data. There is roughly one foul called for every four possessions of basketball and the value of the foul is typically to swing possession of the ball (worth around one point). Thus a one percentage point increase in the offensive foul rate is worth a quarter point of offensive efficiency. The results are presented in Table 3.8.

Table 3.8: Lineup fixed effect estimates of the impact of lead on offensive foul rate.

	One	Two	Three	Four	Simple Mean	Weighted Mean
Home	-0.0107 <i>t</i> =-2.43	-0.00774 <i>t</i> =-2.30	-0.0158 <i>t</i> =-3.25	-0.0165 <i>t</i> =-4.14	-0.0127 <i>t</i> =-6.06	-0.012 <i>t</i> =-5.95
Lead	0.000309 <i>t</i> =1.14	0.000604 <i>t</i> =3.33	-9.32e-05 <i>t</i> =-0.26	0.00174 <i>t</i> =3.09	0.000639 <i>t</i> =3.45	0.000496 <i>t</i> =3.68

We are able to confirm the standard results of a lower foul rate for the home team and also show statistically significant apparent favoritism for the trailing team. Both effects seem to stronger in later quarters. Averaged across course of the entire game we estimate a -1.2% absolute bonus to a teams offensive foul rate can be achieved either by playing at home or seeing a twenty point decrease in their lead (e.g. +10 to -10). Each of these effects represents a 15% relative decline in offensive foul rate, but only represents a roughly 0.3 point/100 change swing in offensive efficiency. In contrast, the loss aversion impact of moving from +10 to -10 was 9 points (27x higher).

3.6 Conclusion

Our study of highly experienced agents reveals reference dependent preferences resulting in a strong “losing motivates” effect. Moreover the reference point appears to be stable and exogenous—players respond to whether they are in the proverbial red or black, regardless of what would have been a reasonable expectation. Past experimental work has demonstrated that reference points can adjust quickly in inexperienced agents, leading to discouragement effects and opening up the possibility of using decision frames to motivate effort. Our results indicate that at least some of these findings do not extend to highly experienced agents. The stability of the reference point around the focal point of zero informs economic interactions beyond basketball. For instance, investors may respond to the returns of a single investment by ignoring the performance of the S&P 500 during the holding period. In our setting, there are huge incentives to eliminate the reference dependence for a good team (because they are leading frequently) but not only are coaches unable to overcome players’ psychology, the exhibit they appear to exhibit the same dependence themselves, indicating just how deep reference dependent preferences run in this setting.

Appendix

Win Probability Models

We will now compute how important actions are at a given game state in terms of the magnitude of the expected impact on the game outcome. Consider two teams, home (h) and away (a). Let $S_{h,N}$ and $S_{a,N}$ denote the current scores for the home and away team with N offensive possessions (for each team) remaining in the game. Let $P_{h,i}$ and

$P_{a,i}$ denote the number of points scored by the home/away team on the i^{th} possession from the end of the game. The home team wins if:

$$S_{h,0} > S_{a,0} \iff S_{h,N} + \sum_{i=1}^N P_{h,i} > S_{a,N} + \sum_{i=1}^N P_{a,i} \iff \sum_{i=1}^N P_{h,i} - P_{a,i} > S_{a,N} - S_{h,N}.$$

To model how teams generate points, let $\{\mu_h, \sigma_h^2\}$ and $\{\mu_a, \sigma_a^2\}$ represent the mean and variance of points per possession that each team is able to achieve in the match-up. If the number of remaining possessions, N , is large, the central limit theorem gives the probability of the home team winning as:

$$P(\text{Home Win}) = P(S_{h,0} > S_{a,0}) = P\left(\sum_{i=1}^N (P_{h,i} - P_{a,i}) > S_{a,N} - S_{h,N}\right) \\ \xrightarrow{N} \Phi\left(\frac{S_{h,N} - S_{a,N} + N(\mu_h - \mu_a)}{\sqrt{N(\sigma_h^2 + \sigma_a^2)}}\right), \quad (3.8)$$

where Φ is the CDF of the standard normal distribution. The marginal impact of a point scored on winning the game is easily obtained by differentiating equation (3.8) to get

$$\frac{dP(\text{Home Win})}{dS_{h,N}} = \phi\left(\frac{(S_{h,N} - S_{a,N}) + N(\mu_h - \mu_a)}{\sqrt{N(\sigma_h^2 + \sigma_a^2)}}\right) \frac{1}{\sqrt{N(\sigma_h^2 + \sigma_a^2)}}, \quad (3.9)$$

where ϕ is the standard normal PDF or by using the discrete analog given in the text. Given the normality, we estimate this equation with a Probit. We impute the number of remaining possessions using the team-specific paces-of-play in a given match-up and by adding one possession to the team currently holding the ball. The coefficient estimates of this model are given in the Appendix. The projections for the probability the home team wins again an evenly matched opponent is given are shown in Panel 1 of Figure 3.10.

One might be concerned that the parametric procedure we have employed relies too heavily on the normality afforded by the central limit theorem. Panel 2 depicts

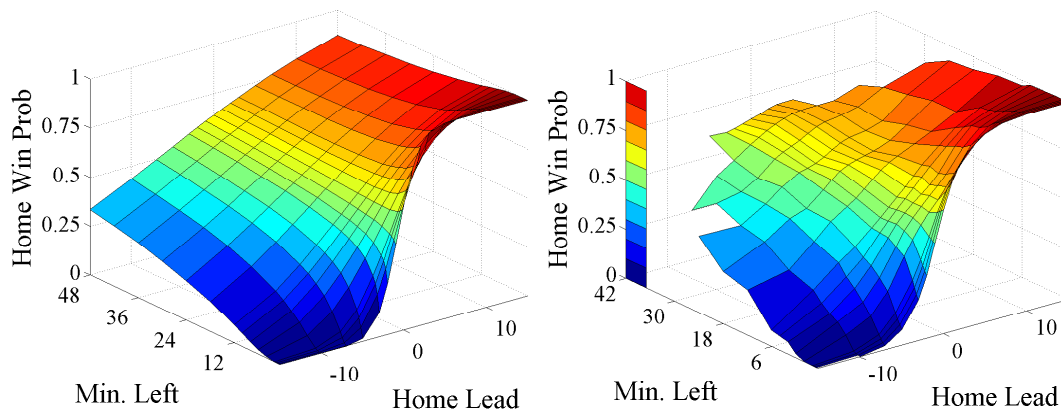


Figure 3.10: Parametric projections of win probability conditional on score margin and time remaining for the home team in an even match-up; Panel 2: non-parametric estimates of the same function. Non-parametric estimator is censored where its point-wise standard error exceeds .03 .

non-parametric estimates of the win probability function constructed from applying a two-dimensional Gaussian kernel to the 2×1 game state vector (margin, time remaining) with bandwidths $h_{margin} = 1.5$ and $h_{minutes} = 1$. Values of the nonparametric estimator are censored for game states rare enough that the associated standard error exceeds 0.03. The two approaches yield nearly identical win probability projections, even as the end of the game nears. Because of the well known difficulty of differentiating non-parametric estimators, we proceed with the values of WVP generated by the parametric model.

Using the parameter estimates given in Appendix Table 1, we apply equation (2) to get the marginal value of a point as a function of time remaining and score margin, which are plotted in Figure 3.1 for two evenly matched teams. Quarters 1–3 are shown in the left panel and fourth quarter in the right panel (note the big change in z-axis scale). The most important point in the first 3 quarters of the game is worth less than 0.05 wins in expectation. In the final minute of a close game, making two free throws can increase the chance the team will win by over thirty percentage points.

Appendix Table 1: Coefficient estimates for our probit model of win probability model.

Regressors are selected based on the model in Equation 3.8 of Section 2.

	Home Lead	Ball	Home Adv.	Off. Adv.	Def Adv
	$\left(\frac{S_{h,N}-S_{a,N}}{\sqrt{N}}\right)$	$\left(\frac{1\{Ball\}}{\sqrt{N}}\right)$	$(\sqrt{N} \cdot 1\{Home\})$	$(\sqrt{N} \cdot \Delta\hat{\mu}_O)$	$(\sqrt{N} \Delta\hat{\mu}_D)$
Coeff.	2.780	1.972	0.00623	0.00176	0.00180
	<i>t=43.61</i>	<i>t=26.69</i>	<i>11.72</i>	<i>15.79</i>	<i>15.23</i>

Games=5,254, Possessions=902,803

Standard errors are clustered at the game level.

ACKNOWLEDGMENTS

The preceding chapter of this dissertation is coauthored with Justin Rao from Microsoft Research. This paper is being prepared for publication.

Bibliography

- Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *The American Economic Review*, 101(2):470–492.
- 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*.
- Angrist, J., Imbens, G., and Krueger, A. B. (1995). Jackknife instrumental variables estimation.
- Arlotto, A., Chick, S. E., and Gans, N. (2013). Optimal hiring and retention policies for heterogeneous workers who learn. *Management Science*, 60(1):110–129.
- 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*.
- Ausubel, L. M. and Milgrom, P. (2006). The lovely but lonely vickrey auction. *Combinatorial auctions*, 17:22–26.
- Baumeister, R. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46(3):610.
- Bearden, J., Murphy, R., and Rapoport, A. (2005). A multi-attribute extension of the secretary problem: Theory and experiments. *Journal of Mathematical Psychology*, 49(5):410–422.
- Bearden, J., Rapoport, A., and Murphy, R. (2006). Sequential observation and selection with rank-dependent payoffs: An experimental study. *Management Science*, 52(9):1437.

- Beilock, S., Carr, T., MacMahon, C., and Starkes, J. (2002). When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. *Journal of Experimental Psychology: Applied*, 8(1):6.
- 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.
- Berger, J. and Pope, D. (2011). Can losing lead to winning? *Management Science*, 57(5):817–827.
- Berri, D. J., Schmidt, M. B., and Brook, S. L. (2004). Stars at the gate the impact of star power on nba gate revenues. *Journal of Sports Economics*, 5(1):33–50.
- Binmore, K., McCarthy, J., Ponti, G., Samuelson, L., and Shaked, A. (2002). A backward induction experiment. *Journal of Economic Theory*, 104(1):48–88.
- 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.
- Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997). Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 112(2):407–441.
- Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.
- Chiappori, P.-A., Levitt, S., and Groseclose, T. (2002). Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. *The American Economic Review*, 92(4):1138–1151.
- 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.
- Crawford, V. and Meng, J. (2011). New york city cab drivers’ labor supply revisited: Reference-dependent preferences with rational expectations targets for hours and income. *The American Economic Review*, 101(5):1912–1932.
- Dohmen, T. (2008). Do professionals choke under pressure? *Journal of Economic Behavior and Organization*, 65(3):636–653.
- Edelman, B. and Ostrovsky, M. (2007). Strategic bidder behavior in sponsored search auctions. *Decision support systems*, 43(1):192–198.

- 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.
- Farber, H. (2008). Reference-dependent preferences and labor supply: The case of new york city taxi drivers. *The American Economic Review*, 98(3):1069–1082.
- Fehr, E. and Goette, L. (2007). Do workers work more if wages are high? evidence from a randomized field experiment. *The American Economic Review*, pages 298–317.
- Fryer Jr, R. G., Levitt, S. D., List, J., and Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: a field experiment. Technical report, National Bureau of Economic Research.
- Ghose, A. and Yang, S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10):1605–1622.
- Gill, D. and Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *The American Economic Review*, 102(1):469–503.
- Goette, L., Huffman, D., and Fehr, E. (2004). Loss aversion and labor supply. *Journal of the European Economic Association*, 2(2-3):216–228.
- Goldman, M. and Rao, J. M. (2011). Allocative and dynamic efficiency in nba decision making. In *In Proceedings of the MIT Sloan Sports Analytics Conference*, pages 4–5.
- Goldman, M. and Rao, J. M. (2013). Motivational loss aversion: Theory and evidence. Working paper.
- Goldman, M. and Rao, J. M. (2014). Optimal stopping in the nba: An empirical model of the miami heat. Available at SSRN: <http://ssrn.com/abstract=2363709>.
- Gomes, R. and Sweeney, K. (2012). Bayes-nash equilibria of the generalized second-price auction. *Games and Economic Behavior*.
- Groves, T. (1973). Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631.
- Gucciardi, D. and Dimmock, J. (2008). Choking under pressure in sensorimotor skills: Conscious processing or depleted attentional resources? *Psychology of Sport and Exercise*, 9(1):45–59.
- Hall, A. R. (2000). Covariance matrix estimation and the power of the overidentifying restrictions test. *Econometrica*, 68(6):1517–1527.

- Hannan, R. L., Hoffman, V. B., and Moser, D. V. (2005). Bonus versus penalty: does contract frame affect employee effort? In *Experimental business research*, pages 151–169. Springer.
- Hossain, T. and List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167.
- Hsu, S., Huang, C., and Tang, C. (2007). Minimax play at wimbledon: Comment. *The American Economic Review*, 97(1):517–523.
- 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.
- Johnson, E. J., Camerer, C., Sen, S., and Rymon, T. (2002). Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory*, 104(1):16–47.
- Kahan, J., Rapoport, A., and Jones, L. (1967). Decision making in a sequential search task. *Attention, Perception, & Psychophysics*, 2(8):374–376.
- 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.
- Koszegi, B., Köszegi, B., and Rabin, M. (2006). A model of reference-dependent preferences. *Quarterly journal of economics*, 121(4).

- Lahaie, S. and Pennock, D. M. (2007). Revenue analysis of a family of ranking rules for keyword auctions. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 50–56. ACM.
- Lana, R. E. (1963). Interest, media, and order effects in persuasive communications. *The Journal of Psychology*, 56(1):9–13.
- Lee, M., O'Connor, T., and Welsh, M. (2004). Decision making on the full information secretary problem. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 819–824. Citeseer.
- Lee, M. D. (2006). A hierarchal bayesian model of human decision making on an optimal stopping problem. *Cognitive Science*, 30.
- 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.
- McAdams, D. and Schwarz, M. (2007). Who pays when auction rules are bent? *International Journal of Industrial Organization*, 25(5):1144–1157.
- 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.
- Murnighan, J. K. and Roth, A. E. (1983). Expecting continued play in prisoner's dilemma games a test of several models. *Journal of Conflict Resolution*, 27(2):279–300.
- Narayanan, S. and Kalyanam, K. (2014). Position effects in search advertising: A regression discontinuity approach. *Working Paper*.
- Neiss, R. (1988). Reconceptualizing arousal: psychobiological states in motor performance. *Psychological bulletin*, 103(3):345.
- Ochs, J. and Roth, A. E. (1989). An experimental study of sequential bargaining. *American Economic Review*, 79(3):355–384.

- 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.
- Pakes, A. (1986). Patents as options: Some estimates of the value of holding european patent stocks. *Econometrica*, 54(4):755–784.
- Palacios-Huerta, I. (2003). Professionals play minimax. *Review of Economic Studies*, 70(2):395–415.
- Palfrey, T. R. and Rosenthal, H. (1994). Repeated play, cooperation and coordination: An experimental study. *The Review of Economic Studies*, 61(3):545–565.
- Paul, R. J., Weinbach, A. P., and Wilson, M. (2004). Efficient markets, fair bets, and profitability in nba totals 1995–96 to 2001–02. *The Quarterly Review of Economics and Finance*, 44(4):624–632.
- Pope, D. and Schweitzer, M. (2011). Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes. *The American Economic Review*, 101(1):129–157.
- Post, T., Van den Assem, M. J., Baltussen, G., and Thaler, R. H. (2008). Deal or no deal? decision making under risk in a large-payoff game show. *The American economic review*, pages 38–71.
- Price, J. and Wolfers, J. (2010). Racial discrimination among nba referees. *The Quarterly Journal of Economics*, 125(4):1859–1887.
- Provencher, B. (1997). Structural versus reduced-form estimation of optimal stopping problems. *American Journal of Agricultural Economics*, 79(2):357–368.
- 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.
- Rapoport, A. and Tversky, A. (1970). Choice behavior in an optional stopping tasks. *Organizational Behavior and Human Performance*, 5(2):105–120.
- Romer, D. (2006). Do firms maximize? Evidence from professional football. *Journal of Political Economy*, 114(2).
- Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55(5):999–1033.

- Seale, D. and Rapoport, A. (2000). Optimal stopping behavior with relative ranks: The secretary problem with unknown population size. *Journal of Behavioral Decision Making*, 13(4):391–411.
- Sill, J. (2010). Improved nba adjusted+/-using regularization and out-of-sample testing. In *Proceedings of the 2010 MIT Sloan Sports Analytics Conference*.
- Varian, H. R. (2007). Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178.
- Walker, M. and Wooders, J. (2001). Minimax play at wimbledon. *The American Economic Review*, 91(5):1521–1538.
- Wallace, H., Baumeister, R., and Vohs, K. (2005). Audience support and choking under pressure: A home disadvantage? *Journal of Sports Sciences*, 23(4):429–438.
- Wilson, M., Chattington, M., Marple-Horvat, D., and Smith, N. (2007). A comparison of self-focus versus attentional explanations of choking. *Journal of Sport and Exercise Psychology*, 29(4):439.
- Zwick, R., Rapoport, A., Lo, A., and Muthukrishnan, A. (2003). Consumer sequential search: Not enough or too much? *Marketing Science*, pages 503–519.