Measuring Digital Advertising Effectiveness:

Solving the Count/Quality Dilemma

by

Bradley Fay

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2017 by the
Graduate Supervisory Committee:

Michael P. Mokwa, Co-Chair
Sungho Park, Co-Chair
Sang-Pil Han
Ranjit M. Christopher

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

Total digital media advertising spending of $72.5 billion surpassed total television Ad spending of $71.3 billion for the first time ever in 2016. Approximately $39 billion, or 54% of the digital media advertising spend, involved pre-programmed software that purchased Ads on behalf of a buyer in Real-Time Bidding (RTB) settings. A major concern for Ad buyers is sub-optimal spending in RTB settings owing to biases in the attribution of customer conversions to Ad impressions. The purpose of this research is twofold. First, identify and propose a novel experimental design and analysis plan for to handling a previously unidentified and unaddressed source of endogeneity: count/quality simultaneity bias (CQB). Second, conduct a field study using data for Ad response rates, cost, and observed consumer behavior to solve for the profit maximizing daily Ad frequency per customer. One large online retailer provided data for Ad impressions, bid costs, response rates, revenue per visit, and operating costs for 153,561 unique users over 23 days. Unique visitors were randomly assigned to one of seven treatment groups with one, two, three, four, five, and six impressions per day limits as well as a final condition with no daily impression cap. Ordinary least square models (OLS) were fit to the data and a non-linear relationship between Ad impressions and site visits demonstrating declining marginal effect of Ad impression on site visits after an optimal point. The results of the field study confirmed the existence of negative CQB and demonstrated how my novel experimental design and analysis can reduce the negative bias in the estimate of impression quantity on customer response. Second, managers interested in improving the efficiency of advertising spend should restrict display advertising to only the highest quality inventory through specific site targeting and by leveraging direct buys and private marketplace deals. This strategy ensures that

subsequent impressions are not of lower quality by restricting the pool of possible impressions from a homogenous set of high quality inventory.

To Baker – Persevere in your own way.

ACKNOWLEDGMENTS

This work is the collective accomplishment of a large group of people. Without the support of everyone and more, this project would not have been completed. I want to thank my wife, Daniele, for her constant support and willingness to bear more responsibility than necessary so I could have the time and energy to complete this work. I also want to thank my parents, Brian and Debbie, and my sister, Monica, and her family, Syed, Syed Brian, and Zayn for constantly and positively encouraging me to finish.

My sincerest thanks go to Dr. Michael Mokwa. He constantly helped me keep this whole crazy process in perspective, served as an amazing advisor and confidant to whom I could express all of my frustrations, and most importantly, has been and will continue to be an invaluable mentor and friend. Sincere thanks to Sungho for encouraging me to pursue this line of research and for serving as an incredible mentor. His patience and teaching helped me identify and cultivate a passion that's resulted in a fulfilling career in data science. An immense amount of gratitude to Ranjit for being a trusted thought partner and an incredible friend throughout the entire doctoral program. Thank you to Sang-Pil for joining and pushing me to think deeper and more theoretic about the general problem. I want to also think Diane for more than I could ever enumerate here. Through the MBA and Ph.D., she has helped me navigate through everything. Finally, I want to thank Rajiv. Although he is not here to see me finish, I hope he would be proud. I would not have even considered this journey if we had not crossed paths.

Finally, I want to thank my colleagues at Wayfair – Dave Drollette, Dan Wulin, Anvesh Sati, Mike Anselmi, Tim Hyde, Matt Herman, Ryan Milligan, Alyssa Siegmann, Scott Collins, and the rest of the Data Science team. I have learned an incredible amount

about Data Science and display advertising from all of you. You also make it enjoyable to tackle new and interesting problems.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

After a record growth of 22% between 2015 and 2016, total expenditure on digital media advertising stood at roughly $72.5 billion, surpassing total ad expenditures on television ($71.3 billion)[1]. Nearly four out of five of all digital advertisement transactions occur programmatically, and the growth trend in programmatic ad spending indicate that by the end of 2019, 84% of all digital ad spending will be through programmatic advertising[2]. In general, 'Programmatic Advertising' refers to any process that involves a pre-programmed software that undertakes the purchase of ads on behalf of a buyer in real time. There are two major variations in programmatic advertising, Real-Time Bidding (RTB) where ad buyers bid for advertisements in real-time second-price and Programmatic Direct (PD) where ad buyers negotiate prices directly with publishers such that the software guarantees advertisements when buyers bid above a fixed price. Currently, RTB accounts for 67% of total expenditures on digital display advertising. In an RTB system, ad buyers submit their bids in real-time via a Demand-Side Platform (DSP) that interacts with a Supply Side Platform (SSP) managing publishers' inventory through an Ad Exchange that facilitates the real-time purchase process electronically. The DSPs in the RTB framework provide the technology infrastructure to manage real-time bidding decisions, the ability to track and target customers of interest to the buyers, and customization of their bidding mechanisms to maximize the buyer-determined outcome (e.g., clicks, visits, conversions)[3].

[1]http://adage.com/article/digital/digital-ad-revenue-surpasses-tv-desktop-iab/308808/

[2]https://www.mediapost.com/publications/article/299332/forecast-nearly-80-of-us-display-spending-will.html

[3]Buyers who wish to gain more control over the purchase process can either opt for a 'programmatic direct' mechanism to directly negotiate a fixed-price deals with the publishers or instead opt for 'private market places' where the buyers that can bid in a platform or restricted thereby increasing the chances of winning a real-time auction. In either case, buyers lose some of the cost efficiencies associated with RTB to ensure that their ads reach the desired inventories.

The benefits of RTB notwithstanding, a major concern for buyers is the sub-optimal spending in automated buying owing to biases in the attribution of customer conversions to ad impressions. DSPs employ a machine learning approach to customizing bidding algorithms in real-time to maximize profits for the platform. Buyers currently have no way of ascertaining the efficacy of these algorithms in attributing user behavior to display ads, as they largely remain a black box from the perspective of the ad buyer. Thus, a wide variety of biases can drive the algorithm's decision to bid for individual pieces of inventory in real-time. The biases can be owing to selection issues driven by end-user/customer behavior, or targeting algorithms for performance optimization employed by DSPs.

There have been multiple attempts to quantify the impact of display advertising in the literature. Initially, the best available data was observational (Chatterjee, Hoffman, and Novak 2003; Drèze and Hussherr 2003; Manchanda et al. 2006), which, while providing novel insights, suffered from sources of endogeneity due to selection bias created by heterogeneous customer behavior. Previous literature identified and proposed novel experimental solutions to deal with customer activity bias (CAB) (Lavrakas, Mane, and Laszlo 2010). Hoban and Bucklin (2015) displayed PSA advertisements to their control group to account for customer activity bias when measuring the causal effect of the presence of advertising. Often, black-box algorithms purchase ads and optimize their decision process as they collect new information. In the case of PSA experiments, the algorithms potentially optimize to two different messages resulting in selective targeting bias. Johnson, Lewis, and Nubbemeyer (2016) suggest the use of "ghost ads" or predicted ads to combat selective targeting bias. Finally, Johnson, Lewis, and Reiley (2017) utilize a field experiment to generate exogenous variation in the number of impressions shown to quantify the marginal impact of advertising on a single site. To

the best of my knowledge, previous research has yet to address the measurement of the marginal impact of advertising with heterogenous inventory quality.

Statement of the Problem

While prior work has furthered the understanding of CAB and selective targeting bias, there was a gap in the literature regarding inventory quality as a source of customer-independent bias. In an RTB setting, the quality of inventory is unknown at the time of auction. As an ad buyer bids on each impression, he realizes the quality of placement at the same time as the quantity of advertisements served is determined. This count/quality simultaneity creates a previously unaddressed source of endogeneity in the state of the art experiments. Due to the simultaneous realization of count and quality, it is impossible to randomly distribute the quality of impressions independent of the quantity of impressions, creating an unobserved correlation between count and quality, or count/quality simultaneity bias (CQB). The omission of inventory quality in prior research confounds existing conclusions about the marginal value of display advertising.

For advertisers interested in quantifying the marginal impact of display ads, the experimental ideal randomly assigns customers to different frequency goal groups and programs the bidding algorithm to maximize the probability each customer receives her assigned a number of ads. Ignoring the role that customer activity bias plays in the algorithms ability to meet this goal, this design creates an additional source of bias, referred to as term goal incongruence bias. This bias occurs in situations where algorithms necessarily must bid on lower quality inventory with a higher probability of winning for customers in treatment groups with high-frequency goals. Alternatively, the algorithms can bid more strategically on potentially higher quality more expensive inventory for customers in lower frequency goal groups. In this situation, there is

another source of unobserved correlation between count and quality created by the algorithm bidding to achieve the stated goal.

Purpose of the Study

The purpose of this study is twofold. First, identify and propose a novel solution to handling the previously unaddressed source of endogeneity: count/quality simultaneity bias. Second, conduct a field study using the proposed experimental design to solve for the profit-maximizing daily ad frequency per customer. I employed a novel experimental design using daily frequency caps instead of frequency goals to generate exogenous variation in the number of impressions to which each customer is exposed. Solely applying frequency caps still results in an experiment that suffers from CAB and CQB. The goal of the frequency caps is not to create a perfect experiment, but to generate the ideal instrumental variables that enable me to handle both customer activity bias and count quality simultaneity. By using frequency caps instead of frequency goals, I remove goal incongruence as a possible source of bias. Within my experiment, I program the precise rules of the bidding algorithm such that they are entirely independent of any variation in customer characteristics, eliminating selective targeting bias. Partnering with a large online retailer, I implement the experimental framework in a large-scale field study and analyze data using two-stage least squares regression to obtain unbiased estimates of the marginal impact of display advertising. Finally, I demonstrate how my finding can support decision making when used in an optimization framework to identify the profit-maximizing daily frequency caps to program into an RTB algorithm.

Research Questions

*RQ1***.** In an RTB setting, what is the optimal number of daily display Ads per customer after adjusting for customer activity bias and count/quality simultaneity bias?

RQ2. In an RTB setting, what is the profit maximizing number of daily display Ads per customer after adjusting for customer activity bias and count/quality simultaneity bias for a range of economic values per visit?

Organization of the Study

The remainder of the document is organized as follows. In Chapter 2, I describe the display advertising landscape as it relates to RTB, and review the relevant literature on programmatic ad buying, the challenges to measurement, and the current state of the art in practice. I identify a previously undefined source of bias afflicting the measurement of ad effectiveness and propose a novel solution to remediate the bias. Chapter 3 outlines my empirical approach and describes how my novel experimental design uniquely solves both new and known sources of bias in the measurement of ad effectiveness. I describe the data and illustrate the presence of customer-independent sources of bias. Chapter 4 presents results and discussion of my two-stage least squares solution. Chapter 5 concludes with discussing the implications of my work for ad buyers in the RTB environment and demonstrating how this work can be used to tune bidding algorithms to utilizing information on the diminishing marginal benefit of increasing impression frequency independent of all other effects.

LITERATURE REVIEW

Programmatic Advertising, RTB, and DSPs

The emergence of real-time bidding (RTB) in programmatic advertising (PA) has ushered in a number of intermediaries such as SSPs, DSPs, Ad Exchanges (ADX), and data exchanges (DX). These agencies facilitate RTB auctions between a vast number of publishers and advertisers resulting in the seamless scaling of the auction process. Wang, Zhang, and Yuan (2016) provide an elaborate exposition of the RTB environment, its benefits, and challenges. From the perspective of Ad buyers, the use of RTB has two important benefits. First, Ad buyers no longer must deal with numerous disparate ad-networks to buy impressions in bulk. SSPs and ADXs have developed RTB auctions as a mechanism to sell the publisher surplus (Liu and Viswanathan 2014; Zhu and Wilbur 2011). PA ensures that Ad buyers delegate the bidding process to a DSP that interacts with an ADX[4] to buy real estate on the Advertiser's behalf, thus simplifying the buying process[5]. Second, Ad buyers can now rely on DSPs to develop an automated bidding strategy by using algorithms that consider inputs from Ad buyers and the general market characteristics from DXs that provide reports on customer behavior on websites across the Internet. This DSP service relieves the ad buyer of the need to design publisher specific bidding strategies to optimize bidding. (Skiera and Abou Nabout 2013; Yao and Mela 2011).

ADXs typically employ second price auctions under a 'pay per impression' mechanism. The optimal bidding strategies in these types of auctions rely on knowledge of the true value of an advertisement (Edelman and Ostrovsky 2007). Thus, running an

[4] The ADXs aggregate several ad networks.

[5] Publishers who work with SSPs to register with different ADXs can still employ a diverse set of rules to both filter and modify bids from ad buyer to increase the long-term value of their real estate. However, the ad buyers no longer manage the diversity of ad network rules directly

efficient ad campaign relies on one's ability to attribute profitable customer behavior to an advertiser. Herein lies the key challenge facing both ad buyers and DSPs in the RTB environment. DSPs employ machine-learning tools to associate ad impressions with customer response for performance optimization to the extent they have information from DX agencies and Ad buyers. Similarly, campaign managers of individual ad buyers' associate impressions with more granular aspects of customer response such as conversions and purchase amount with the intent of customizing the DSPs algorithms. Despite the benefit of having the DSP aid the development and execution of a bidding strategy, it is in the interest of the ad buyer to guard against inaccurate valuations of a marginal ad impression.

Effectiveness of Display Advertisements

Given the rapidly growing economic importance and the precise control over targeting and messaging provided to marketers, it is paramount that marketers understand how to measure the effectiveness of display advertising. The complexity of the display advertising ecosystem described above creates unique challenges to quantifying the causal impact of digital advertising. Despite these challenges, previous literature has made substantial progress in enumerating, quantifying, and solving many of the biases that plague the measurement process.

In one of the earliest attempts to measure the effectiveness of display advertising, Chatterjee, Hoffman, and Novak (2003) utilize observational data to link advertising exposure to consumer response. Using non-RTB impression data from a single site with multiple content pages and mandatory visitor registration, they model the consumer's probability of clicking on an advertisement. Their results show that there is a heterogeneous click response to advertisements with a decline in click-through-rate as

the number of impressions increases from one to eleven, followed by an increase in click-through-rate as the number of impressions increases beyond eleven.

While modeling click-through-rate is an important type of consumer response, ultimately firms are interested in how their ads influence buying behavior. Manchanda et al. (2006) attempted to make this link between display ad exposure and purchase behavior. Using observational data shared by a third party from a non-RTB setting, the authors model the consumer decision of if and when to purchase as a function of advertising exposure. They find a positive relationship between advertising and purchase incidence with the strongest effect found in those customers who had previously purchased.

It is reasonable to assume that targeted advertisements have, at worst, no effect on individual behaviors. Using a large-scale field experiment, Goldfarb and Tucker (2011) found that targeting has a negative effect when advertisements are highly obtrusive. Interestingly, they also found that less obtrusive, but highly targeted have a positive impact on customer response.

A unique feature of display retargeting is the ability to personalize the advertising message to the customer. While at first, it may seem reasonable to do this in all situations; prior work has shown this may not always be true. Lambrecht and Tucker (2013) use non-RTB impression data collected in a quasi-experimental setting to show that highly personalized ads outperform a generic equivalent only when consumers have narrowed their search. In a series of field experiments in non-RTB settings, Bleier and Eisenbeiss (2015) examine the role of timing and placement factors on the effect of personalization. They find that the effectiveness of personalization is related to the proximity in time from when a user was last on an advertiser's site and the exposure to a display ad. In a follow-up study, they show that personalization increases click-through-

rate regardless of the location of the advertisement, but there is only a positive relationship between personalization and view-through visits when the advertisement appears on a motive congruent site.

Several sources of bias increase the difficulty of measuring the precise causal impact of display advertising. In addition to exploring the impact of different aspects such as location and level of personalization, prior literature has identified interesting sources of bias that arise when attempting to execute and analyze a traditional A/B experiment in the field. Customer-driven selection bias resulting from heterogeneous browsing activity occurs because the ability of an advertiser to serve an advertisement to a customer is not a random process. An advertiser's ability to serve a customer an ad is dependent upon the customer actively browsing sites where an advertiser is able and willing to purchase impressions. I refer to this bias as customer activity bias (CAB). Heterogeneity in both the duration and intensity of browsing behavior creates confounding correlation with the probability of exposure and likely hood of response. The issue does not arise in the treatment group as advertisers know who was and was not exposed to advertisements. The issue lies is the control group. In the simplest experimental design, the advertiser withholds advertisements from a portion of the population serving as the control. If the advertiser can treat each person on the treatment, it is reasonable to assume he could have also treated everyone in the control. Since it is not possible to treat each person in the treatment as a result of CAB, it is also not possible to treat each person in the control if the advertiser would choose to do so. Without knowing who in the control an advertiser could have treated had those persons been in the treatment, the simple A/B experimental design results in two groups which are no longer identical and without an ability to remove those users in the control group who would have been shown an impression to create equivalence.

To address the issue of CAB, prior research and practitioners propose the use of PSAs. The simplest PSA or public service ad experimental design entails splitting the target population into two groups. The treatment groups receive the advertising treatment as normal. The nuance of this design occurs in the control group. Instead of completely excluding customers randomly assigned to the control from advertising, the advertiser attempts to serve ads to those customers just as they would if the customers were in the treatment group. When an advertising auction is won, the advertiser serves a creative with a completely unrelated message, often on behalf of a charitable organization. The advertiser pays the same price to serve the PSA as he would to serve a normal impression, and importantly, he receives a transaction log of all PSA advertisements served. This allows him to identify those users in the control group whom he would have served advertisements to had he chosen to serve them advertisements. At the time of analysis, the experimenter can use this information to remove the unexposed portion of both the treatment and control groups decreasing sample size, but increasing statistical power (Johnson, Lewis, and Reiley 2017). In a randomized field experiment Hoban and Bucklin (2015) leverage the PSA experimental design, and demonstrate a positive but diminishing marginal effect of display advertisements on the likelihood a user returns to site. Particularly, the returns diminish slower for customers higher in the purchase funnel suggesting the benefit of building awareness for less familiar customers.

In addition to CAB, the bidding algorithms designed and implemented on behalf of advertisers generate a second source of bias conflating results of display advertising experiments. A primary benefit of display advertising is the ability to precisely target individual users with personalized advertising. A challenge to executing this strategy is the necessity to make bidding decision in the RTB setting at scale. The only way to

achieve performance is to code the decision-making process into some bidding algorithm. Often, agencies or DSPs leverage their expertise to build bidding algorithms on behalf of advertisers. Just as humans incorporate new information into their decision process as it comes it, the bidding algorithms are built within machine learning frameworks such that they update the decision calculus as new information is consumed. Often, these algorithms are optimizing towards a specific goal such as achieving a target average metric linked to consumer response (e.g., cost-per-click or cost-per-action). When using these algorithms in PSA experiments, the algorithms in test and control begin optimizing to consumer responses to distinctly different creatives (Johnson, Lewis, and Nubbemeyer 2016). Over time, this results in the algorithms in the test and control groups targeting distinctly different populations. The bidding algorithm creates an additional source of bias, which I term selective-targeting bias (STB). To handle STB, Johnson, Lewis, and Nubbemeyer (2016) suggest using "Ghost Ads" as an alternative to PSAs. In their proposal, algorithms bid on customers in both treatment and control identically throughout the experiment. When an advertiser wins an auction in the "Ghost Ad" condition, the ADX awards the inventory to the second-place bidding. Instead of showing the winners ad, the "Ghost Ad" is logged for use in analysis. The key piece of this design is the experimenting firm receiving a log of every "Ghost Ad" served. In this design, firms do not spend advertising dollars on creatives unrelated to the business because they forfeit their right to serve "won" impressions. Additionally, the targeting algorithm in the control condition does not learn using behavioral information related to an orthogonal creative. Although theoretically sound, marketers cannot implement the "Ghost Ad" method on their own. A successful "Ghost Ad" experiment requires the

support of the Ad Exchanges[6] to collect the data and implement the logic necessary to execute this design.

In addition to CAB and STB confounding control populations, these biases also confound the number of ads shown to each user creating endogenous variation in the count of impressions shown. Johnson, Lewis, and Reiley (2017) conducted a field experiment on the Yahoo! Homepage to generate exogenous variation in the number of ads shown to browsers. In addition to using PSAs, they randomly assigned users to one of three conditions that varied the ratio of treatment to control impressions. While their analysis of the marginal effect of advertising does not yield significant results, they did show a directional return suggesting there is a positive marginal benefit to the number of impressions shown to a user. To this point, the literature has dedicated a fair amount of effort to sufficiently address customer-dependent sources of bias (Barajas et al. 2016; Hoban and Bucklin 2015; Johnson, Lewis, and Nubbemeyer 2016; Johnson, Lewis, and Reiley 2017), the literature has paid little attention to bias induced by inventory quality.

Impression Quality in RTB

In the RTB setting, when a customer browses to a page where the publisher has multiple pieces of inventory to sell, each available advertisement is sold simultaneously in independent auctions. If the effectiveness of impressions depends on other content on the page (Bleier and Eisenbeiss 2015; Goldfarb and Tucker 2011), then one can assume the quality of each piece of inventory is uncertain at the time of the auction as it must depend on the other advertisements displayed on the page. It is straightforward to imagine a situation in which a single piece of inventory becomes less effective purely because the creative message displayed in another advertising slot is so eye-catching that

[6] https://www.thinkwithgoogle.com/intl/en-gb/articles/a-revolution-in-measuring-ad-effectiveness-knowing-who-would-have-been-exposed.html

it decreases the likelihood that a customer notices any other impression on the page. By the same logic, it is also reasonable to imagine a situation where all other ads served on a page alongside an advertiser's creative are irrelevant to the user, increasing the likelihood the user responds to the advertiser's message. In either instance, the advertiser does not have full information on the quality of the impression on which they are bidding because they do not know what other advertisements will also be on the page and their influence on the browsing customer's attitudes. Only when the auction has finished, and an advertiser has won does he ascertain information on the quality of inventory where he just served his advertising message.

RTB auctions present advertisers with an unprecedented opportunity for marketing precision in budget allocation. The RTB environment enables ad buyers to purchase impressions one at a time. RTB auctions are second-price sealed-bid auctions in which the dominant strategy for the advertiser is to bid his max willingness to pay. A key component to his willingness to pay is inventory quality. All else being equal, his willingness to pay will correlate positively with inventory quality. If inventory quality is uncertain in real-time, the best strategy for an advertiser is shading his bid up or down based on some expectation of quality. Any variation in quality at auction time will necessarily affect the number of impressions served. If quality varies towards higher quality, bidding the mean expected quality decreases the chances of winning the impression. Conversely, if quality varies towards lower quality, bidding the mean quality increases the chances of winning an impression. In either case, the advertiser simultaneously realizes quality and quantity, which means that when an advertiser is interested in understanding the marginal impact of advertising, he must also account for variation in inventory quality independent of customer characteristics.

Quantifying the quality of an advertisement is a difficult challenge. One way to measure quality is to look at the historical clearing prices of a piece of inventory. Under an assumption of an efficient marketplace, I propose using the clearing price for a piece of inventory as a reasonable proxy for the quality of that piece of inventory at that moment in time. The clearing price reflects market demand accounting for variation in time of day, day of week, customer information, and historical understanding of the impression. When market clearing prices serve as a proxy for quality, then it is impossible for an ad buyer to know the quality of inventory prior to placing his bid. If an auction is won, the ad buyer simultaneously realizes impression count and inventory quality. This simultaneity creates a confounding issue in display field experiments by creating unobserved correlation between the number of impressions shown and the quality of those impressions. The simultaneous nature of this realization reveals a deficiency in previously proposed experimental design methods as it becomes impossible to ensure randomized quality throughout all treatment groups and impression frequencies. As such, this simultaneous realization of count and quality creates a previously unaddressed source of endogeneity to experiments, which I refer to as Count/Quality Simultaneity Bias (CQB).

CQB occurs merely as an artifact of advertisers needing to incorporate judgments of inventory quality into their bid in the RTB setting. If advertisers or agencies building bidding algorithms are unaware of this phenomenon, then a second type of inventory quality related bias may arise. A common practice in advertising is achieving a set frequency goal of impressions per person (Cheong, de Gregorio, and Kim 2010; Schmidt and Eisend 2015). Setting a frequency goal for an algorithm bidding on real-time display advertising tunes the underlying model hit a specific number of impressions per user within some budget and timing constraint. Before deciding to bid on any single

impression, the algorithm needs information on the number of previously served impressions to the user and how many opportunities there are likely to be left in the time window in which to achieve the goal. Assuming a fixed budget per person, a low-frequency goal provides the algorithm the freedom to bid high on a single impression or two without exhausting the budget. Early in the goal window and budget, the algorithm bids strategically on different types of inventory. If an agency or DSP designs an algorithm on behalf of an advertiser, there is the incentive to target the frequency goal set by the advertiser as well as full exhaust the budget allocated to the campaign[7]. As the number of expected bidding opportunities decreases, the algorithm must necessarily become more aggressive with lower quality inventory to meet the frequency goal. In a situation with a high-frequency goal, the algorithm is incentivized to only bid on cheaper inventory because the fixed budget may not enable it. I refer to CQB induced by the misalignment of goals between the advertiser and the algorithm as Goal Incongruence Bias (GIB).

Given the customer-independent nature of these biases, the literature has addressed neither CQB nor GIB. Additionally, existing methods such as PSA control groups cannot account for the simultaneous nature of realizing both quality and quantity of impressions served to a user. I contribute to the growing literature on display advertising measurement by identifying these unaddressed sources of bias, quantifying the robust negative impact of these biases on conventional measurement methods, proposing a novel identification strategy, and demonstrating the efficacy of my solution using a large-scale field experiment.

[7] At the end of every fiscal quarter, there is a spike in the average clearing price for all display inventory across the Internet. This is assumed to be a function of many agencies working to exhaust the total budget allocated to them by their clients.

METHODOLOGY

The ideal experiment for quantifying the marginal impact of digital advertising requires randomly assigning users to strict frequency goals and instructing the algorithm to bid in whatever manner necessary to achieve the exact goal. As discussed above, this exact situation creates both CQB and GIB where users randomly assigned to higher frequency goal groups would necessarily receive exposure to lower quality inventory. CQB occurs because there is a fundamental negative relationship between count and quality and GIB occurs because programming an algorithm to achieve higher quantity while holding the budget fixed across groups can exacerbate the CQB issue. In addition to CQB and GIB, an experiment of this nature would also be prone to CAB as part of the ability of an algorithm to fulfill the frequency goal is dependent upon the number of bid opportunities created by user activity. Finally, there are marketplace effects that influence the algorithms ability to meet a goal. Even if a user browses enough to provide the algorithm enough bid opportunities, the auction dynamics of RTB create a situation where winning an impression is always uncertain. If competitors systemically over value or under value certain customers, the probability of winning an impression for a user can change. These marketplace effects serve as a third source of bias to an ideal experiment.

Instrumental Variable

Given the inability to conduct a controlled, randomized experiment, I turn to a well-understood econometric approach – instrumental variable regression (Angrist, Imbens, and Rubin 1996). Traditionally, econometricians look to naturally occurring phenomenon to generate valuable instruments (Abadie, Angrist, and Imbens 2002; Angrist and Chen 2011; Levitt 1996, 1997). Instead of looking for naturally occurring instruments, I use a novel experimental design to generate an instrumental variable, which correlates with the independent variable of interest and uncorrelated with the

error term as the treatment occurs through exogenous assignment. Working with a large online retailer, I conducted a large-scale field experiment in which I randomly assigned users to one of seven different conditions manipulating daily frequency caps. A daily frequency cap incorporates a rule into a bidding algorithm that restricts the algorithm from bidding if a customer has already been served her allotted number of impressions during a time interval. While I could not guarantee that a user will get the exact number of impressions she would be assigned to receive in a frequency goal framework, I can guarantee that a user in an *n*-cap group will not receive *n+1* impressions. By implementing the frequency cap, I artificially restricted the bidding algorithm from bidding on a customer who has reached her cap even if the algorithm could have served the additional impression. While the number of advertisements served to each user in the experiment is still confounded by multiple sources of bias, my experimental design created exogenous variation the number of impressions served to each customer. My bidding algorithm specified to be independent of customer characteristics generates data which is free of selective targeting bias. The use of a daily frequency cap instead of a daily frequency goal eliminates the potential for goal incongruence bias. The use of the experimental condition as an intention to treat instrument eliminates customer activity bias as the first stage of my two-stage least square regression computes the mean number of impressions to use in the second stage eliminating any endogenous variation in impressions for individual differences in behavior. Most importantly, the intention to treat instrument eliminates count quality simultaneity because by controlling for any unobserved correlation between count and quality due to the random nature of group assignment. The characteristics of this approach which solve each of the four potential sources of bias create a distinct ability to generate truly unbiased estimates of the marginal causal effect of display advertising in an RTB setting.

Experimental Design

For 30 days prior to the start of the experiment, I randomly assigned all traffic to the retailer's site into one of the treatment conditions. The assignment was at the cookie-level, and each cookie identifies a unique device[8]. Prior literature has acknowledged the deficiency in treating devices as users (Chatterjee, Hoffman, and Novak 2003). To minimize the confounding nature of cross-device browsing behavior between mobile and desktop, I imposed a desktop-only device restriction. I assumed that cross-device browsers are randomly distributed throughout all treatment groups, and thus the impact of any mobile or other retargeting campaigns will not bias the results.

I created seven treatment groups with one, two, three, four, five, and six impressions per day limits as well as a final condition which was no daily impression cap. I utilized these caps to create as much variation in the instrument as possible while still maintaining reasonably sized subsets of customers. Additionally, I included an uncapped group to capture the upper bound of possible effect. Prior research studies assumed variation in impression frequency to be exogenous to quantify their claim of causal impact (Chatterjee, Hoffman, and Novak 2003; Hoban and Bucklin 2015; Manchanda et al. 2006). Without directly manipulating impression frequency, both activity bias and targeting bias suggest this variation is likely endogenous. Johnson, Lewis, and Reiley (2017) took advantage of a natural experiment on the Yahoo! Homepage where two ads were rotated based on the second in which server loads the page. This feature created exogenous variation within groups of individuals with identical browsing behavior (e.g.,

[8] A device refers to any unique browser, not physical device. There could be a situation where the same person using the same physical device visits the same site using both Google Chrome™ and Mozilla Firefox™. Cookie-level tracking treats these browsers as independent devices. Many companies are working diligently to create user-specific device graphs that match all known devices to their respective users. The limitations of developing a trust-worthy cross-device mapping of each user and the impact on advertising is beyond the scope of this work.

two individuals who visit Yahoo! exactly ten times), but it was limited to a single website, thus holding the quality of the website constant throughout the experiment. This method is reasonable for publishers who wish to quantify the impact of an ad on a single site, but for advertisers with customers who browse a variety of sites, this method does not generalize.

I began the experiment after the initial 30-day assignment period and focused only on those users who visited the partner's site during the pre-experimental window. I ran this experiment for 23 days and impressed 153,561 unique users (approximately 5% of the addressable population). To control for creative effects (Braun and Moe 2013), every user received the same general creative which consisted of a white background, six products they recently viewed, and a message related to pricing. There were slight variations due to different sizes of advertisements, but in general, the creative was identical throughout the experiment.

Bidding Algorithm

As noted by Hoban and Bucklin (2015), instrumental variable regression requires knowledge of the bidding algorithm. Targeting bias owing to black-box bidding algorithms is one of the largest sources of endogeneity in using field data for this type of causal research. The ability to fully program the bidding algorithm was a unique feature of my experiment. This key feature provides complete transparency into the "black box," and I was fully aware of any potential targeting biases induced by the algorithm and, more importantly, the biases avoided in the algorithm. Many bidding algorithms are either proprietary to the company or built by an agency. At a high level, these algorithms are designed to spend a higher proportion of the firms advertising dollars on high-value customer segments. Since an algorithm can only bid on behalf of a company, it is likely bidding higher on customers who are expected to be more responsive. This higher bid

correlates with a higher probability of winning a given auction. This outcome generates endogenous variation in the number of ads served to different users because customers who are more likely to respond are more likely to be served advertisements purely because they are more likely to respond.

Unlike previous studies, I had full transparency into the algorithm bidding on the display auctions. Specifically, I defined the precise rules used to bid on users in the experiment. I used this opportunity to eliminate all sources of STB and targeted all users equally.

The simplest rule to achieve this goal is setting a single bid for every impression such that the bid is unrelated to the user the algorithm is bidding on. Unfortunately, this simple rule does not quite achieve what the desired goal because the strategy is also unrelated to the piece of inventory on which the algorithm is bidding. Since the clearing prices of auctions vary widely and as a function of different sites (e.g., an ad on NewYorkTimes.com is more expensive than an ad on coolmathgames.com), a single bid is more likely to win a disproportionate number of auctions on lower value sites than on higher value sites. Even in the case where a single bid was set sufficiently high, the random nature of winning implies this strategy still over-index on lower value sites as the same high bid would have a higher probability of winning an auction on a lower quality site and a higher quality site. Given the heterogeneous nature of customer browsing, the single bid strategy would lead to another source of selection bias. Even randomly assigning daily frequency caps, the algorithm hits the daily caps for customers who tend to browse on lower priced sites than customers who tend to visit higher value sites.

To address the issue of selective targeting bias created by the bidding algorithm treating each customer differently, I designed the bidding algorithm to only vary with the

historical performance of each piece of inventory. My goal was to develop a bidding strategy in which the probability of winning an auction was identical across all users and all inventory. Working with the partner firm, I used historical impression data to model a bid for each inventory location independent of the customer characteristics. For each site, I looked at the historical distribution of clearing prices and computed the median clearing price. I programmed the algorithm to consider only inventory characteristics for each unique auction and based on the unique characteristics, bid the estimated median inventory-clearing price. Bidding the estimated median price means that for a random customer and a random auction, the bidding algorithm had approximately a 50% chance of winning an impression and showing customers an ad. Throughout the course of a day, if the number of bid opportunities for a given user was sufficiently high, there was a strong probability of filling their daily cap. This bidding strategy also serves as a nice pacing control in that it prevents the algorithm from filling the specified daily cap for each user too quickly, and provides a bit of random time interval between the impressions served to customers throughout a day. Unfortunately, I only received information on winning bids, so there is no way to validate how the bidding algorithm performed regarding winning an expected number of auctions.

Data Collection

My data is from a single large online retailer. Data are reported at the cookie-level and aggregated across the length of the experiment. While in practice it is possible to look at the data in a panel structure (e.g., Hoban and Bucklin 2015), I chose to examine the effects at the aggregate level as it best mirrors the general practice of my partner company when evaluating the effectiveness of campaigns and updating strategy moving forward.

Customers were randomly assigned one of the treatment conditions upon landing on site during the 30-day pre-experimental window. I leverage the proprietary tool my partner has built which splits users based on the customers universally unique identifier[9] and a proprietary algorithm which computes and assigns treatment conditions. This tool is used primarily for on-site testing, but my partner was able to adapt its use for external marketing campaign testing. Table 1 reports the descriptive statistics.

One of the interesting features of my data is the ability to quantify the heterogeneity and historical quality of sites on which advertisements are bought and displayed. During the experimental window, there were no restrictions placed on the inventory where the algorithm could purchase impressions. Additionally, I had access to historical impression logs, which contained information on the domain, size, and location along with clearing prices my partner has historically paid and click-through rates.

Although the retailer sells a variety of products, I cannot draw generalized conclusions on the effects of display advertising. Instead, I use the data to illustrate the presence of unaddressed sources of bias, implications of not addressing the bias, and the efficacy of my proposed solution in generating unbiased estimates of effect. Future research should explore this method in alternative contexts to ensure it is robust to industry and buying behavior.

[9] https://en.wikipedia.org/wiki/Universally_unique_identifier

Table 1

Descriptive Statistics

Treatment	Variable	Mean	Std. Dev.	*N*
Frequency Cap – 1/Day (Cap1)	Impressions	7.368	9.811	21,463
	Visits	0.749	2.171	
Frequency Cap – 2/Day (Cap2)	Impressions	13.741	15.594	23,605
	Visits	0.768	2.105	
Frequency Cap – 3/Day (Cap3)	Impressions	17.31	17.235	23,047
	Visits	0.78	2.354	
Frequency Cap – 4/Day (Cap4)	Impressions	20.259	20.863	23,155
	Visits	0.775	2.275	
Frequency Cap – 5/Day (Cap5)	Impressions	23.624	26.113	23,443
	Visits	0.862	2.493	
Frequency Cap – 6/Day (Cap6)	Impressions	26.148	28.607	23,498
	Visits	0.823	2.367	
No Frequency Cap (NoCap)	Impressions	91.414	195.67	15,350
	Visits	0.844	2.31	

Notes: Each group was assigned 15% of allocated traffic except for the NoCap group which was assigned the final 10% of traffic.

FINDINGS

Although I account for the two major sources of biases described by previous researchers (Hoban and Bucklin 2015, Johnson et al. 2015), it is important to note that I am only interested in customers that have already been served at least one impression for this work. Consequently, I do not speak to the effectiveness of the first exposure owing to the absence of a rigorous control group that ensures counterfactual evidence where customers with equal propensity to view an ad do not end up viewing an ad.

Ordinary Least Squares Approach

I first used OLS to understand if there is an expected relationship between impressions and visits. The results (see Table 2) suggest there is a strictly negative relationship between the number of impressions shown and the number of visits to the site. These results contradicted my expectation and previous research (Hoban and Bucklin 2015). Next, I considered various model specifications by taking the log of dependent (visits) and independent variables (impressions). The results of the alternative specifications provided further evidence of a highly consistent and robust negative effect of impressions on visits.

Table 2

OLS Results

	Visits		ln(Visits + 1)	
Constant	0.8059 (0.0063)	0.8630 (0.0122)	0.3182 (0.0122)	0.3382 (0.0032)
Impressions	-0.0003* (0.0001)	---	-0.0001 (0.0000)	---
ln(Impressions)	---	-0.0273 (0.0045)	---	-0.0097 (0.0012)
R^2	0.00007	0.00023	0.00016	0.00044

Notes: All coefficients significant at the p < .0001 level unless otherwise noted, N = 153,561.
* $p < .001$

Since the results sharply contrast my expectations and previous findings in the literature, I took further steps and performed additional robustness checks to ensure the results were not an unexpected consequence of the proposed experimental design. From Table 1, I observed that users in the No Frequency Cap (NoCap) group receive substantially more impressions than users in the other treatment conditions. To rule out any bias in the initial results driven by the vast number of impressions shown, I estimated the same models excluding the NoCap group and continued to find a negative relationship between impressions and visits.

Next, I divided the dataset by treatment condition into seven independent data sets and estimate the same models. Again, I found a significant negative relationship between impressions and visits. I report the results of the best-fitting log-log model for each group in Table 3. All in all, OLS results indicate that despite utilizing the customizable features available to me through the collaboration, I am unable to rule out the multiple sources of bias endemic to this type of experiment. The descriptive statistics illustrate the degree to which activity bias may be influencing the results. If there were no activity bias, I would expect the average number of impressions in each treatment group to be equal to the daily maximum frequency cap in each group times the length of the experiment.

Table 3

OLS Results for Each Treatment Group

	Treatment Group							
	All Capped	Cap1	Cap2	Cap3	Cap4	Cap5	Cap6	NoCap
Constant	0.337	0.317	0.314	0.337	0.344	0.381	0.361	0.376
ln(Imps)	0.010	-0.012	-0.001	-0.012	-0.014	-0.021	-0.014	-0.015
N	138,211	21,463	23,605	23,047	23,155	23,443	23,498	15,350
Adj. R^2	0.0004	0.0003	-0.0000	0.0005	0.0008	0.0018	0.0009	0.0017

Notes: Only results for log-log specifications. Results for additional specifications share the same pattern and can be provided upon request All coefficients significant at the $p < .0001$ level unless otherwise notes

Count Quality Simultaneity

I believe CQB drives a portion of the negative bias observed in the previous results. As described earlier, CQB occurs because of the uncertain nature of inventory before showing an impression. Since quality and quantity are realized simultaneously, it is impossible to design an experiment that randomly distributes inventory quality throughout the impressions served.

Using historical data, I computed the average clearing price of all unique pieces of inventory on which my partner firm has won an advertising auction starting 60 days before the experiment starting and continuing 60 days after the experiment concluded. I removed all impressions served as a part of my experiment to keep the quality estimate independent of the influence that might happen in my experimental data. Once I computed a historical quality score for each piece of inventory, I appended that information back to each impression served during the experiment. I then computed the average quality of all impressions served to each customer.

First, I wish to demonstrate that average quality declines as the number of impressions increases. I do this by regressing each customer's inventory quality score against the total impressions served during the experiment. I report the results of this analysis in Table 4. The results demonstrate a strong significant negative relationship between the number of impressions shown and impression quality. Table 5 illustrates the robustness of this effect across each of the individual treatment groups. Thus, I use this evidence to conclude the existence of a negative CQB.

Table 4

Count Quality Relationship

	Impression Quality
Constant	3.483 (0.007)
Impressions	-0.002 (0.000)
Nobs	153,561
Adj. R^2	0.002

Notes: All coefficients significant at p < .0001 level unless otherwise noted

Table 5

Count Quality Relationship by Treatment Group

	Treatment Group						
	Cap1	Cap2	Cap3	Cap4	Cap5	Cap6	No Cap
Constant	3.790	3.470	3.476	3.506	3.529	3.496	3.462
	(0.024)	(0.018)	(0.017)	(0.021)	(0.028)	(0.022)	(0.024)
ln(Imps)	-0.013	-0.006	-0.006	-0.005	-0.004	-0.004	-0.002
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
N	21,126	23,291	22,792	22,894	23,227	23,291	15,065
Adj. R^2	0.0021	0.0024	0.0029	0.0025	0.0015	0.0019	0.0066

Notes: I only report results for linear specifications. Results for additional specifications share the same pattern and can be provided upon request. All coefficients significant at the p < .0001 level unless otherwise noted.

Using the above evidence to illustrate the existence of negative CQB, I include impression quality as a covariate and re-estimate the initial OLS model. I expect there to be a positive relationship between quality and visits. I also expect the negative coefficient of impressions to move towards zero as quality should remove some of the unobserved negative correlation in the biased estimate. I report my results in Table 6. As expected, there is a significant and positive relationship between quality and the total visits observed. Additionally, when comparing with the results in Table 2, I show a decrease in

the magnitude of the negative coefficient of impressions. I conclude that including information on inventory quality can reduce the negative bias in the estimate of impression quantity on customer response, but I still have not addressed the remaining sources of endogeneity. In the next section, I utilize the experimental conditions as instrumental variables and estimate a two-stage least squares regression to get completely unbiased estimates of the true causal impact of advertising impressions on customer visits.

Table 6

OLS Accounting for Impression Quality

	Visits		ln(Visits + 1)	
Constant	0.7181 (0.01058)	0.7727 (0.0153)	0.2819 (0.0027)	0.3004 (0.0039)
Impressions	-0.0003* (0.0001)	--	-0.0001 (0.0000)	--
ln(Impressions)	--	-0.0251 (0.0046)	--	-0.0087 (0.0012)
Quality	0.0261 (0.0024)	0.0256 (0.0024)	0.0108 (0.0006)	0.0106 (0.0006)
R^2	0.0008	0.0010	0.0021	0.0024

Notes: All coefficients significant at the p < .0001 level unless otherwise notes; *N* = 153,561.
* indicates significance at p < .001

Instrumental Variable Approach

To estimate the true causal impact of an incremental Ad, I employ the treatment conditions as perfect intention to treat instrumental variables in a two-stage least squares framework. I formally specify the empirical models as follows:

$$Imps_i = \beta_0 + \beta_1 I(Cap2) + \beta_2 I(Cap3) + \beta_3 I(Cap4) + \beta_4 I(Cap5) + \beta_5 I(Cap6) + \beta_6 I(NoCap) + \varepsilon_i, \quad (1)$$

$$Visits_i = \gamma_0 + \gamma_1 Imps_i + \upsilon_i, \quad (2)$$

where *i* denotes individuals and $Imps_i$ and $Visits_i$ represent the counts of impression and site visits of *i* during the experimental period. *I("group")* is an indicator function which

takes the value of one if *i* belongs to "group" and otherwise zero. ε_i and v_i are mean-zero random shocks. Specifically, v_i captures the unobserved factors that influence individual *i*'s decision to visit the site. I identified two potential mechanisms that generate significant correlation between v_i and $Imps_i$ – CQB and CAB. My treatment condition is purely random and therefore must be uncorrelated with unobserved factors captured by v_i. The treatments influence $Imps_i$ as each condition directly manipulated the maximum number of impressions a customer was eligible to receive on any given day. As described earlier, the treatment conditions satisfy both the inclusion and exclusion restrictions necessary for a strong instrument.

Table 7 reports the results of Equation (1) using OLS. I consider two functional forms of the dependent variables – $Imps_i$ and $\ln(Imps_i)$. As expected, I observe that the experimental conditions are strong predictors of impressions. The coefficients of group indicators increase as the frequency cap is increased indicating that the average impressions increase with the frequency cap. These results confirm that my treatment condition satisfies the inclusion requirement and the remaining results are not subject to the weak instrument problem.

Table 7

First-Stage Estimation Results

IV	Impressions	ln(Impressions)
Constant	7.368 (0.443)	1.540 (0.008)
I(Cap2)	6.373 (0.613)	0.588 (0.012)
I(Cap3)	9.942 (0.616)	0.785 (0.012)
I(Cap4)	12.890 (0.615)	0.901 (0.012)

I(Cap5)	16.256 (0.614)	1.001 (0.012)
I(Cap6)	18.780 (0.613)	1.071 (0.012)
I(No Cap)	84.045 (0.686)	1.616 (0.0130)
Adj. R^2	0.109	0.108

Notes: All coefficients significant at p < .0001 unless otherwise notes; *N* = 153,561.

To estimate Equation (2), I use $\widehat{Imps_i}$ instead of $Imps_i$. The predicted impression counts serve as a source of exogenous variation to predict the marginal effectiveness of an advertisement. In this way, I can tackle the endogeneity problems arising from the correlation between v_i and $Imps_i$. In addition to Equation (2), I estimate the following models for robustness, and examine whether the marginal returns to advertising are linear, decreasing, or declining:

$$Visits_i = \gamma_0 + \gamma_1 \ln\left(\widehat{Imps_i}\right) + v_i, \quad (3)$$

$$Visits_i = \gamma_0 + \gamma_1 \widehat{Imps_i} + \gamma_2 \widehat{Imps_i^2} + v_i. \quad (4)$$

In addition to $Visits_i$, I also consider $\ln(Visits_i + 1)$ as a dependent variable. I use the same instrument, the treatment conditions, in my estimation of all models.

Table 8 reports the estimation results of each model specifications using instrumental variables. First, I observe that the effect of impressions on visits is now positive and significant (Model 1 and Model 4). This finding indicates that there is a strong negative correlation between $Imps_i$ and v_i. When this correlation is not properly handled, OLS results in biased estimates as I have already reported in Table 2. I conclude that the associated between $Imps_{i+1)}$ and v_i is substantial and when not properly

addressed, can lead to incorrect conclusions. Additionally, the results in Model 1 (Model 4) indicate that when impressions increase by one unit, the number of visits a customer makes increases by 0.001 (0.03%)

Models 3-4 and 5-6 examine the nature of the marginal effect of advertisement. In both dependent variables (*Visits* and ln(*Visits* + 1)), the results of Models 2 and 5 support a decreasing marginal effect of impressions. Using ln($Imps_i$) and adding a square term improves the model fit (R^2). Figure 1 and Figure 2 illustrate the estimated non-linear relationships between impression and visit of Model 5 and Model 6.

Table 8

Second-Stage Estimation Results

	Visits			ln(*Visits* + 1)		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Constant	0.7736 (0.00883)	0.631 (0.033)	0.68917 (0.0223)	0.3077 (0.0023)	0.268 (0.009)	0.2864 (0.0058)
Impressions	0.0010 (0.0003)	--	0.00687 (0.0015)	0.0003 (0.0001)	--	0.0018 (0.0004)
Impressions^2	--	--	-0.00006 (0.0000)	--	--	-0.00001 (0.00000)
ln(Imps)	--	0.071 (0.014)	--		0.021 (0.004)	--
R^2	0.00009	0.00017	0.00019	0.00012	0.00022	0.00024

Notes: All coefficients significant at the $p < .0001$ level unless otherwise notes, N=153,561

Figure 1. Decreasing Marginal Effect of Ad on Visit

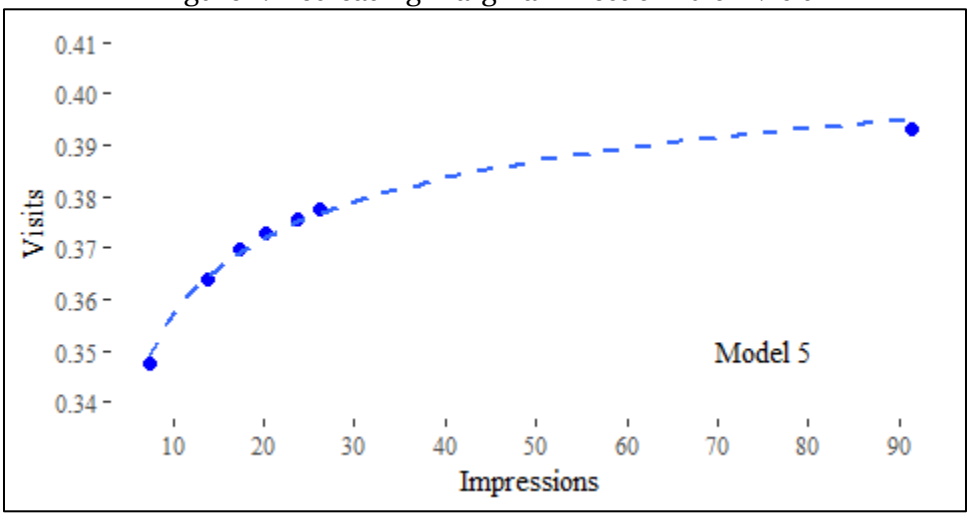
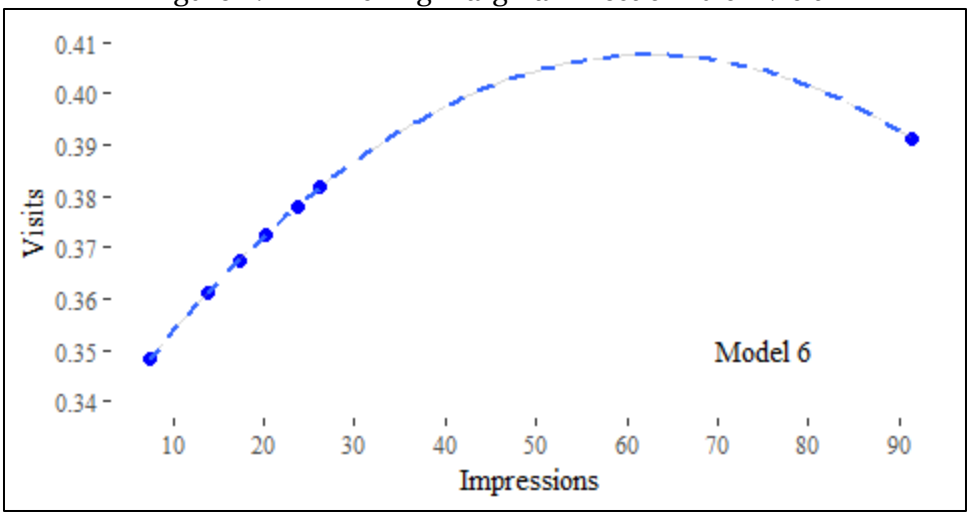


Figure 2. Diminishing Marginal Effect of Ad on Visit



Field Study

In this section, I use the response function estimated above in an optimization framework to solve the optimal daily frequency cap. I test the results of my findings in a follow-up field study. Using data collected during my previous experiment, I calibrate the relationship between total impressions served and dollars spent per person. I then pair the ad response function with the cost function to solve for the profit-maximizing daily frequency cap. Using information from my partner on the general value of an

incremental visit, I add a new rule to my partner's current best-performing bidding algorithm to dictate the optimal caps using a solution derived from the optimization. I report anonymized results from the follow-up field study and show that implementing the optimal frequency caps decreases both revenue and cost of the program, but slightly increases the return on investment.

Cost Function

Advertisements exchanged in RTB environments are typically bought and sold in CPM units. CPM is a metric common to media across various channels and stands for the cost per 1,000 impressions. Some DSPs allow advertisers to place cost-per-click or CPC bids, but subsequently translate those values into an effective CPM given some expected click-through rate as modeled by the DSP. While bids and costs are typically reported in CPM units, advertisers pay only for a single impression. For example, if an advertiser wins an auction at a $2.00 CPM, they are paying $.002 to display their advertisements. An alternative to CPM is pay-per-click (PPC). This payment structure is the common structure in paid search advertising where advertisers compete for priority in search rankings and are only responsible for paying the publisher when a customer clicks the sponsored search result(Rutz, Bucklin, and Sonnier 2012). Unlike paid search advertising (Skiera and Abou Nabout 2013), in a display advertising environment, the advertiser is responsible for paying for every impression won in the auction.

While it is evident that total advertising spend must increase as the number of advertising impression auctions won increases, the exact nature of that relationship is unknown. I use data collected during the initial field study to examine the relationship between the total cost and the number of impressions shown. Formally, I specify the cost model as:

$$Spend_i = \delta_0 \text{Impressions}_i + \epsilon_i \qquad (5)$$

where $Spend_i$ is the total spent and $Impressions_i$ is the number of impressions shown to individual i. ϵ_i is mean zero randomly distributed shock. I specified an intercept free model because intuitively, if zero impressions are shown, an advertiser spends nothing. I estimate Equation 5 using OLS and report my results in Table 9. A positive estimate of δ_0 demonstrates the expected positive relationship between total spend and impressions. The reported R^2 of .895 demonstrates an extremely strong fit. I am confident my cost-side model provides robust results to include in the subsequent optimization.

Table 9

Cost Function OLS

	Cost
Imps	.0027
	(0.0000)
R^2	.895

Notes: All coefficients significant at the p < .0001 level unless otherwise notes. *N* = 153,561.

Optimization

Firms can set many different types of goals for their advertising program. For example, a firm attempting to build brand awareness might invest to maximize reach and frequency given a budget constraint (Danaher, Lee, and Kerbache 2010). Alternatively, a firm with strong brand awareness may seek to maximize the profit of the direct response aspect of advertising. For this optimization, I take the latter approach and use the ad response and cost functions to solve for the profit-maximizing daily frequency cap.

Using Equation (3), the response model, and Equation (5), the cost model, I formally state the profit-maximizing optimum as:

$$\max_{Imps} \Pi(Imps) \tag{6}$$

I specify profit as the difference between the consumer response function given some constant *C* value of a visit and the cost as a function of impressions:

$$\Pi(Imps) = \gamma_0 + \gamma_1 \ln(Imps)C - \delta_0 Imps \tag{7}$$

The first-order conditions for Equation (7) are:

$$\frac{\partial}{\partial imps} = \frac{\gamma_1 C}{x} - \delta_0 \tag{8}$$

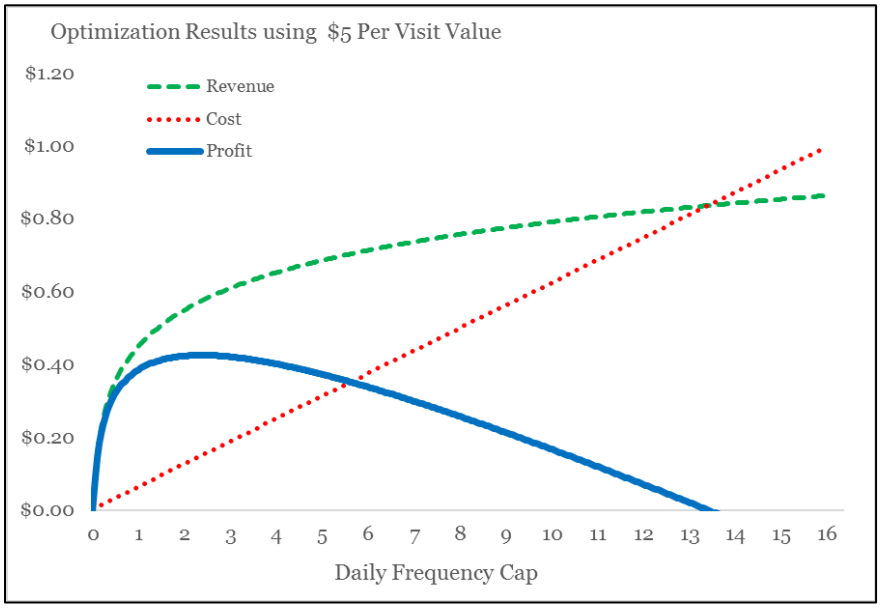
Table 10 contains the resulting daily optimal for a distribution of expected value per visits. The results in this table illustrate the profit-maximizing daily caps heavy reliance on the economic value of a visit. Since Equation 3 does not directly capture revenue, I assume the value of all visits are equal, and any incremental increase in the number of visits made by a customer will result in increased revenue. I solved this optimization using Excel Solver with an integer constraint on the solution. Figure 3 demonstrates there is a range of decisions, which will result in a profit in a situation where the expected value of a visit is, on average, $5. The profit-maximizing solution when visits are expected to be worth $5 is either two or three impressions per day. When choosing between these two solutions, a firm interested in maximizing revenue as a secondary goal should choose a three impression per day limit. Alternatively, if the firm wishes to maximize ROI, the final decision should be for two impressions per day as the expected profit is the same, but the total cost is greater for three impressions per day compared to two.

Table 10

Simulated Results of Optimization

Daily Cap	$2.50 Per Visit	$5 Per Visit	$10 Per Visit
1	$0.16	$0.38	$0.83
2	$0.15	$0.42	$0.97
3	$0.12	$0.42	$1.03
4	$0.08	$0.40	$1.05
5	$0.03	$0.37	$1.06
6	$(0.02)	$0.34	$1.05
7	$(0.07)	$0.30	$1.04
8	$(0.12)	$0.26	$1.01
9	$(0.17)	$0.21	$0.99
10	$(0.23)	$0.17	$0.96
15	$(0.51)	$(0.08)	$0.77
20	$(0.79)	$(0.35)	$0.55

Figure 3. Field Study Optimized Frequency Caps



The goal of my optimization was to identify the profit-maximizing daily frequency cap for my partner to improve channel performance. Using this framework, I tested the efficacy of my optimization in a follow-up study. Currently, my partner has implemented a ten impression per day cap for the channel, a solution they arrived at using experience-based heuristics. Comparing this cap to the results from my optimization, this ten impression per day limit along with their current targeting algorithm should and does result in measurable incremental revenue for the firm. However, a rule of ten impressions per day is not necessarily the profit-maximizing solution. Based on the results of the optimization, I believe the firm can improve channel performance from a profit and ROI perspective without sacrificing a significant portion of revenue by setting a daily individual-level impression cap using this framework.

I tested the optimized frequency caps against the current ten-impression per day limit by cloning the existing best performing campaign including the exact targeting algorithm used in practice. In addition to cloning the exact setup, I also added a layer of

decision-making to the algorithm to implement the optimal frequency cap. The only difference is that instead of a ten impression per day frequency cap, I imposed a model-derived frequency cap. I collected data for two weeks in the spring of 2017. I observed a 29.8% decrease in total profit for the channel. I do observe a 30.2% decrease in total spend, but this is offset by a 29.9% decrease in total revenue. So, while overall, the frequency cap decreased both revenue and profit, it decreased profit at a slower rate.

While these results are unexpected, I cannot draw any significant conclusions about the efficacy of my approach from this test for multiple reasons. First, I estimated the response model using data collected from the entire population of browsers whereas the current algorithm utilized in the test targets a finely tuned subset of the population determined by the partner firm. Additionally, while my model derived caps suggested a global ten impression per day cap was too high, the firm realizes an average a daily impression frequency much lower than ten. It could be that the algorithm was targeting a subset of customers who experience a slower diminishing marginal return than the model estimated. Furthermore, it could be that the bidding algorithm takes advantage of market forces by bidding low enough to decrease the probability of serving wasteful ads to much of the targeted population. To generate more interpretable validation results, I suggest the firm implement my proposed experimental design but restrict the population of customers included in that experiment to the segments they are currently targeting in their campaigns.

DISCUSSION, IMPLICATIONS, AND RECOMMENDATIONS

CQB presents a significant challenge to marketing managers looking to optimize their marketing spend for online display advertising. Prior research has suggested that managers take advantage of experimental techniques such as PSAs and Ghost Advertisements. I have shown that while these serve as an effective solution to bias that arises when trying to measuring the impact of the presence of a single advertisement, these strategies are not able to eliminate the impact of CQB when measuring the marginal impact of more than one advertisement. In this section, I propose two suggestions for managers who wish to address CQB and the impact of effective marketing and marketing measurement.

First, I propose that managers run similar field experiment at random intervals throughout the year. In this research, I propose and implement a straightforward experimental design to mitigate the influence of CQB on measuring advertising effectiveness. Managers should implement or work with their display-advertising partners to implement an identical design. Once the data is collected, managers can estimate the unbiased marginal impact of their display advertising campaign. This strategy is most applicable to managers who are primarily interested in proper attribution compared to efficient spending. Executing this experiment takes resources in terms of time and money as it necessitates a portion of advertising spending on ineffective advertisements served on low-quality inventory. However, without that information, managers cannot derive the true value of an incremental advertisement.

Second, for managers primarily interested in improving the efficiency of advertising spend, I recommend restricting their purchasing of display advertising to only the highest quality inventory through specific site targeting and by leveraging direct buys and private marketplace deals. This strategy ensures the subsequent impressions

are not of lower quality by restricting the pool of possible impressions to only high quality. It should remove the negative correlation between advertising quality and advertising quantity by restricting all advertisements to be purchased from a homogenous set of high-quality advertisements. Inopportunely, this strategy is not without its limitations as well. Primarily, by manually restricting inventory, marketing managers forego the ability to learn about new inventory or adapt to improving/worsening inventory. Additionally, this manual limitation of inventory necessarily restricts the population whom managers reach by worsening the customer-driven selection bias discussed earlier.

Unfortunately, there is no perfect solution to the issue of CQB. Managers need to choose the solution that best matches their strategic goals. The best solution is potentially a flighted combination of the above two proposals where managers trade off running rigorous field experiments to explore and test on new data and follow with campaigns on heavily restricted inventory. Most importantly, managers need to be aware of CQB and the impact it has on measuring the effectiveness of incremental digital advertisements. Without an awareness of this issue, managers may draw incorrect conclusions concerning the effectiveness of their digital marketing strategy and either over- or under-invest in the channel.

Recommendation for Future Research

To properly execute my experiment, I necessarily restricted the creative message shown to a single variation. With the opportunity to dynamically change the content of digital advertising on an impression by impression basis, modeling the impact of a single creative is potentially limiting. Future research in the area of digital advertising effectiveness should explore how creative sequencing can mitigate the diminishing marginal impact of consecutive advertisements. One might even imagine a scenario

where a proper sequence of messages increases the consumer response by telling a compelling story driving the customer back to site in a more purchase-oriented mindset. Executing the necessary experiment to measure this is tricky as it requires the ability to specifically control the number of advertisements a user sees and the creative that is shown as a function of the number of advertisements already shown. Emerging display platforms are testing this functionality, but at this point, there is not a widely available solution to execute this type of experiment.

In addition to creative sequencing, future research should examine how CQB is influenced by heterogeneity in the population. Prior research has shown that customer-driven activity bias happens through heterogeneous browsing behaviors. In my study, I looked at the entire population together to demonstrate the global existence of the phenomena as well as a simple experimental design to mitigate the impact. An important next step is to examine if CQB is systematic across all subpopulations or if it there are certain segments where the influence is more pervasive on measurement than others. Future research should replicate this same experiment, but focus the analysis on meaningful subsets of the population

Conclusion

In this dissertation, I measured the marginal causal impact of display advertising using a large-scale randomized field experiment. I identified two previously unaddressed sources of bias that confound the marginal measurement. First, Count/Quality Simultaneity bias is when both the number of impressions served and simultaneously realize the overall quality of those impressions. This bias owes to the nature of RTB and is driven by the uncertain nature of inventory quality before winning an ad auction. Second, Goal Incongruence bias occurs when the goals of the algorithm or third-party firm do not perfectly align with the advertiser. Specifically, if the goal is to meet a

specified frequency requirement, the algorithm working to achieve the goal can create CQB.

I proposed a novel experimental design to address these new sources of bias as well as the additional sources of biased discussed in the literature. By designing an experiment where customers are randomly assigned to various frequency caps and writing the exact rules of the bidding algorithm, I create perfect intention-to-treat instruments. To my knowledge, this is the first study on ad effectiveness leveraging an experiment to generate exogenous variation in the number of impressions shown to users across a vastly heterogeneous distribution of web pages. Using the treatment conditions as instruments, I model the unbiased marginal causal impact of display retargeting.

By comparing the results of the OLS to those utilizing the instrumental variable, I demonstrate the importance of properly identifying and addressing all sources of bias with measuring the causal impact of display advertising. I highlight CQB and GIB as two previously unaddressed sources of bias which can only be handled using my novel experimental framework.

In addition to arriving at an unbiased estimate of the true causal value of an advertisement, I demonstrate how the results can be extended to solve for the profit-maximizing daily frequency cap. I found evidence of the linear relationship between costs and the number of impressions served. Given there is no decrease in cost, limiting impressions can help ad buyers reduce their overall ad budget. The realized savings can either be reallocated to reach new customers or reinvested in other marketing channels.

The results of this work serve as a cautionary tale to ad buyers when attempting to model data at face value. If a buyer were to build a bidding algorithm based on the

initial OLS model, the buyer would not build an algorithm at all. In fact, the buyer would simply stop spending and reallocate all marketing dollars away from this channel.

REFERENCES

Abadie, Alberto, Joshua Angrist, and Guido Imbens (2002), "Instrumental Variables Estimates of Subsidized Training on the Quantile of Trainee Earnings," *Econometrica*, 70 (1), 91–117.

Angrist, Joshua D. and Stacey H. Chen (2011), "Schooling and the Vietnam-era GI Bill: Evidence from the draft lottery," *American Economic Journal: Applied Economics*, 3 (2), 96–118.

———, Guido W. Imbens, and Donald B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91 (434), 444.

Barajas, Joel, Ram Akella, Marius Holtan, Aaron Flores, Joel Barajas, and Ram Akella (2016), "Experimental Designs and Estimation for Online Display Advertising Attribution in Marketplaces Experimental Designs and Estimation for Online Display Advertising Attribution in Marketplaces," *Marketing Science*, 35 (3), 465–83.

Bleier, Alexander and Maik Eisenbeiss (2015), "Personalized Online Advertising Effectiveness : The Interplay of What, When, and Where," *Marketing Science*, 34 (5), 669–88.

Braun, Michael and Wendy W Moe (2013), "Online Display Advertising : Modeling the Effects of Multiple Creatives and Individual Impression Histories," *Marketing Science*, 32 (May 2015), 753–67.

Chatterjee, Patrali, Donna L. Hoffman, and Thomas P. Novak (2003), "Modeling the clickstream: Implications for Web-based advertising efforts," *Marketing Science*, 22 (4), 520–41.

Cheong, Yunjae, Federico de Gregorio, and Kihan Kim (2010), "The power of reach and frequency in the age of digital advertising: Offline and online media demand different metrics," *Journal of Advertising Research*, 50 (4).

Danaher, Peter J., Janghyuk Lee, and Laoucine Kerbache (2010), "Optimal Internet Media Selection," *Marketing Science*, 29 (2), 336–47.

Drèze, Xavier and François Xavier Hussherr (2003), "Internet advertising: Is anybody watching?," *Journal of Interactive Marketing*, 17 (4), 8–23.

Edelman, Benjamin and Michael Ostrovsky (2007), "Strategic bidder behavior in sponsored search auctions," *Decision Support Systems*, 43 (1), 192–98.

Goldfarb, Avi and Catherine Tucker (2011), "Online Display Advertising: Targeting and Obtrusiveness," *Marketing Science*, 30 (3), 389–404.

Hoban, Paul R. and Randolph E. Bucklin (2015), "Effects of Internet Display Advertising in the Purchase Funnel: Model-Based Insights from a Randomized Field Experiment," *Journal of Marketing Research*, 52 (3), 375–93.

Johnson, Garrett A., Randall A. Lewis, and Elmar I. Nubbemeyer (2016), "Ghost Ads : Improving the Economics of Measuring Ad Effectiveness," *Simon Business School Working Paper No. FR 15-21*, 1–56.

———, ———, and David H. Reiley (2017), "When Less Is More: Data and Power in Advertising Experiments," *Marketing Science*, 36 (1), 43–53.

Lambrecht, Anja and Catherine Tucker (2013), "When Does Retargeting Work? Information Specificity in Online Advertising," *Journal of Marketing Research*, 50 (5), 561.

Lavrakas, Paul J., Sherrill Mane, and Joe Laszlo (2010), "Does anyone really know if online ad campaigns are working? An evaluation of methods used to assess the effectiveness of advertising on the internet," *Journal of Advertising Research*, 50 (4), 354–74.

Levitt, S D (1996), "The effect of prison population size on crime rates: Evidence from prison overcrowding litigation," *Quarterly Journal of Economics*, 111 (2), 319–51.

Levitt, Steven D. (1997), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime," *American Economic Review*, 87 (3), 270–90.

Lewis, Randall A. (2010), "Measuring the effects of online advertising on human behavior using natural and field experiments," 1–133.

———, Justin M. Rao, and David H. Reiley (2011), "Here, there, and everywhere," *Proceedings of the 20th international conference on World wide web - WWW '11*, (January), 157.

Liu, De and Siva Viswanathan (2014), "Information Asymmetry and Hybrid Advertising," *Journal of Marketing Research*, 51 (5), 609–24.

Manchanda, Puneet, Jean-Pierre Dubé, Khim Yong Goh, and Pradeep K. Chintagunta (2006), "The Effect of Banner on Internet Advertising Purchasing," *Journal of Marketing Research*, 43 (1), 98–108.

Rutz, Oliver J., Randolph E. Bucklin, and Garrett P. Sonnier (2012), "A latent Instrumental variables approach to modeling keyword conversion in paid search advertising," *Journal of Marketing Research*, 49 (3), 306–19.

Schmidt, Susanne and Martin Eisend (2015), "Advertising Repetition: A Meta-Analysis on Effective Frequency in Advertising," *Journal of Advertising*, 44 (4), 415–28.

Sherman, Lee and John Deighton (2001), "Banner advertising: Measuring effectiveness and optimizing placement," *Journal of Interactive Marketing*, 15 (2), 60–64.

Skiera, Bernd and Nadia Abou Nabout (2013), "PROSAD: A bidding decision support system for profit optimizing search engine advertising," *Marketing Science*, 32 (2), 213–220.

Wang, Jun, Weinan Zhang, and Shuai Yuan (2016), "Display Advertising with Real-Time

Bidding (RTB) and Behavioural Targeting," *arXiv preprint arXiv:1610.03013*.

Yao, Song and Carl F. Mela (2011), "A Dynamic Model of Sponsored Search Advertising," *Marketing Science*, 30 (3), 199–214.

Zhu, Yi and Kenneth C. Wilbur (2011), "Hybrid Advertising Auctions," *Marketing Science*, 30 (2), 249–73.