An Empirical Meta-analysis of E-commerce A/B Testing Strategies

Alex P. Miller, Kartik Hosanagar
{alexmill,kartikh}@wharton.upenn.edu
The Wharton School, University of Pennsylvania

Abstract

In this project, we attempt to provide a rigorous, empirical study of e-commerce A/B testing strategies. We perform a meta-analysis on 2,732 A/B tests conducted by 252 e-commerce companies across seven industries over the course of three years. While there is much interest in the field of digital experimentation generally, little is known empirically about the testing strategies of firms in real-world environments and how these strategies are related to business outcomes. Our dataset gives us unique insight into what firms are experimenting with on their websites and which of these strategies are associated with larger experimental effect sizes. We develop a framework for quantifying the effect of two different experimental factors on an intervention's ultimate effect size: the type (or content) of an experiment and its location within a website's conversion funnel. After providing a descriptive analysis of A/B testing practices among the firms in our sample, we exploit the metadata in our dataset to classify the experimental interventions using this framework. We find that experiments involving price promotions and those targeted on category or product listing pages are associated with the largest effect sizes, relative to other experiment types in our sample. We then attempt to identify heterogeneity in the effectiveness of different types of interventions at different stages of the conversion funnel. We find evidence that consumers' response to different types of promotions depends on where those promotions are targeted within a website's architecture. In particular, we find that promotional interventions on product prices are most effective early in the conversion funnel, whereas shipping-related promotions are most effective late in the conversion funnel (on product and checkout pages). As a unique, large-scale, cross-firm meta-analysis of empirical experimentation practices, this project not only provides practical insight for managers, but also makes a theoretical contribution to the e-commerce literature by documenting and quantifying how multiple dimensions of website design shape online shopping behavior.

1 Introduction

Recent technological solutions have dramatically lowered the cost of conducting digital experiments. The enterprise software market is now awash with low-cost, easy-to-install testing tools from companies such as Optimizely, HubSpot, Adobe, and Google (among many others). While large companies with significant resources have been conducting online experiments for years, the advent of these new tools has dramatically increased the availability and popularity of A/B testing among firms of all sizes. As such, managers are increasingly turning to A/B tests to make objective decisions backed by statistical theory.

Though a large number of researchers have used the proliferation of these technologies to test their own hypotheses, there has been much less focus on how firms natively use A/B testing platforms in the course of everyday operation. But as A/B testing makes its way into mainstream business practice, there is a growing demand for insight into how to exploit these technologies and an increasing need for researchers to examine these tools from a strategic—rather than purely technical—perspective. In its current state, academic research provides little insight into basic questions about real-world A/B testing practices: What kinds of experiments do companies run? What is the distribution of effect sizes in online experiments? Which types of experiments have the largest effect sizes? How can firms better target their experiments to increase conversion rates? While there is much interest in the field of digital experimentation generally, little is known empirically about the testing strategies of e-commerce companies, let alone which strategies may be most effective.

In this project, we attempt to answer these questions by providing a rigorous, empirical study of e-commerce A/B testing practices. To accopmlish this, we perform a meta-analysis on 2,732 A/B tests conducted by 252 e-commerce companies across seven industries over the course of more than three years. We are able to exploit the metadata in our sample and analyze experimentation strategies from multiple perspectives. We first use this metadata to characterize the content of the experiments in our sample. We then classify the interventions associated with each experiment into several high-level categories that are applicable to the majority of e-commerce websites. We attempt to measure the effectiveness of these various types of interventions on customer conversion rates by comparing the average absolute effect sizes of the most common interventions used in online experiments. We find that experiments involving price promotions and those targeted on category or product listing pages are associated with the largest effect sizes, relative to other experiment types in our sample. We then investigate the treatment heterogeneity of different types of interventions at different stages of the conversion funnel. and find that promotional interventions on product prices are most effective early in the conversion funnel, whereas shipping-related promotions are most effective late in the conversion funnel (on product and checkout pages).

This work not only provides key insight for both describing and informing the A/B testing strategies of a diverse set of companies, but it also has important implications for understanding online consumer behavior and optimizing the customer conversion funnel of e-commerce websites.

2 Background & Related Literature

As the practice of A/B testing has grown in popularity over the last decade, there has emerged academic interest in both the methodology and empirical practice of digital experimentation. However, this literature largely focuses on *how* to do digital experiments, with relatively little to say about *what* to experiment on. This bias toward theoretical know-how has left a gap in our understanding of how the characteristics of real-world interventions connect to business objectives. We attempt to provide some insight into the nature of empirical e-commerce experiments and also study which factors have the largest influence on test outcomes. This motivates the development of a classification framework, in which we categorize e-commerce interventions along two dimensions: experiment type and experiment location.

To contextualize and inform this analysis, we will review several related bodies of existing research. We first summarize related literature on digital experimentation and A/B testing and then turn to the existing research on various factors that are known to drive purchasing behavior in e-commerce. We also highlight the importance of how the placement of an intervention affects its performance by discussing the existing research on how consumer behavior can vary throughout the marketing conversion funnel. Given the diverse body of work that has been done on these topics, we see our project building upon—and hopefully contributing to—existing research in statistics, information systems, human computer interaction, and marketing.

2.1 Digital experimentation & A/B testing

While randomized controlled trials have been used by firms and researchers for more than a century, experimentation in the online environment presents a distinct set of opportunities and challenges that has motivated several recent developments related to A/B testing. Extremely large sample sizes, the presence of rich demographic and technographic data on participants, instantaneous data collection, and the ability to deploy many different experiments simultaneously are all features that distinguish modern A/B testing from older forms of experimentation. New methodological research has emerged to exploit these novel features of the online testing context, including papers on identification of heterogeneous treatment effects (Taddy et al., 2016, Wager and Athey, 2017), targeted experimentation (Liu and Chamberlain, 2018), sequential testing (Johari et al., 2015), and experimentation at large scales (Kohavi et al., 2013, Xu et al., 2015).

Another unique aspect of A/B testing that has drawn the interest of some researchers is how inexpensive and easy it is to test many different interventions in a short period of time. This characteristic is in stark contrast to the context in which classical randomized controlled trials were historically developed; in agriculture, medicine, and even most academic research, the hypothesis is given and an experiment is conducted to answer a very clear question. However, the cost of *digital* experimentation is so low that marketers have a preponderance of interventions that *could* be tested with little guidance on which of these interventions *should* be tested. These questions, as opposed to being concerned with statistical methodology, seek to guide firm's experimentation *strategy*. There is a small but growing stream of research on this topic. The most closely related work in this literature includes a recent study that develops a theoretical framework for addressing the challenge of an experiment-rich regime, in which the number of potential hypotheses is so abundant that it is the observations themselves that are actually more expensive—that is to say, there are more potential experiments to run on a website than could ever be tested with enough observations to yield statistical significance (Schmit et al., 2018). In a different study, the strategic problem of determining an optimal experimentation strategy in terms of sample size is considered in (Azevedo et al., 2018). By developing a theoretical model of the distribution of effect sizes and analyzing over

1,500 experiments at Microsoft Bing, the researchers find that the platform can expect higher returns by running a higher number of low-powered experiments rather than a low number of high-powered experiments.

2.2 E-commerce & Website Design

Apart from the growing literature on digital experimentation, this project is also related to the existing body of work in information systems, marketing, and human computer interaction on how various factors in website design and marketing affect consumer behavior. Researchers have shown that price promotions (Zhang and Krishnamurthi, 2004, Zhang and Wedel, 2009) and shipping fees (Lewis, 2006, Lewis et al., 2006) can have significant effects on customer conversions in e-commerce settings. The literature on price partitioning has also shown that consumers respond differentially to changes in product price vs. shipping price (Chatterjee, 2011, Hamilton and Srivastava, 2008). Aside from price-based interventions, there is an existing body of work in information systems and human-computer interaction that studies how various non-marketing aspects of website design affect online behaviors. Indeed, "usability" is known to be a major factor in how users assess the quality of a firm's website (Agarwal and Venkatesh, 2002, Venkatesh and Agarwal, 2006). Several studies attempt to analyze how specific website characteristics affect user behavior; this includes research on page loading times (Galletta et al., 2006), presentation flaws (Everard and Galletta, 2005), and image characteristics (Hausman and Siekpe, 2009, Zhang et al., 2016). A large-scale study soliciting user comments about factors that affect website credibility found that the "design look" was the most prominent (Fogg et al., 2003). In this study, elements such as overall aesthetic, spacing, sizing, colors, and fonts were all coded as reflecting the design characteristics of a website.

2.3 Marketing Conversion Funnel

In addition to studying how various types of interventions affect online shopping behavior, our project will also examine the role of an intervention's *location* within a website's architecture. To motivate this analysis, we build upon the "conversion funnel" framework, which is a ubiquitous concept in both the academic and industrial literature on digital marketing. There are many ways this concept has been operationalized in existing research, but a number of studies across different contexts have demonstrated that the effectiveness of marketing interventions depends on where individuals are within their customer journey. One technique for studying the marketing funnel is to use observable customer outcomes as proxies for their position in the funnel. In a meta-analysis of many online advertising experiments, Johnson et al. (2017) uses site visits and conversions as proxies for middle and late funnel stages; a similar approach is taken in Braun and Moe (2013). In a B2B setting, Jansen and Schuster (2011) uses the number of customer interactions, quotes, and orders to capture the progress of a lead down the funnel. In a grocery store setting, Seiler and Yao (2017) operationalizes different funnel stages by using aisle visits and purchases as dependent variables.

An alternative approach for modeling the conversion funnel is to use individuals' observable characteristics as proxies for their latent psychographic funnel state. This method is based on the notion that a customer's position in the funnel is determined by their internal thought processes and intentions. Traditionally, words used to characterize different funnel stages across both academic and industrial literatures describe an individual's internal state: "awareness", "consideration", "decision", "loyalty" (Court et al., 2009, Lavidge and Steiner, 1961). Studies in this paradigm are often designed to segment or target individuals in different conversion states with various marketing interventions (Abhishek et al., 2012, Moe, 2003, Netzer et al., 2008).

Though our study related to this body of work, the unique nature of our context requires us to introduce an alternative way of operationalizing the conversion funnel. While prior literature has analyzed the conversion funnel using different outcomes or individual-level characteristics, our meta-analytic approach means the primary unit in our study will be interventions themselves. Rather than asking how an intervention or a customer's latent state affects intermediate outcomes, our dataset allows us to investigate how an experiment's intrinsic characteristics affect the primary outcome of interest in A/B testing (purchase behavior). Based on the way firms label their experiments, we will motivate a way of thinking about the location of an experiment within a website's architecture as a marker of its "location" within the conversion funnel. Analyzing the experiments in our dataset this way also maps onto the user experience of most A/B testing platforms, which requires firms to specify a page (or set of pages) on which an intervention will take place. We will explore this phenomenon more formally in our analyses below.

In sum, this project is able to shed new light on experimentation strategy, website design, and consumer behavior in the e-commerce conversion funnel. We build on existing research on these subjects from several disciplines, but also believe this project represents a unique contribution to the existing literature. In contrast with prior studies, this project will be comparing multiple types of marketing interventions simultaneously and comparatively. Furthermore, insight into the exact nature of the interventions in A/B testing datasets has been identified as a key limitation in the previous large-scale analyses of online experiments (Peysakhovich and Eckles, 2017). However, in our research context, we have access to a set of metadata that provides us information about the nature of the interventions being investigated in our sample. This allows us to analyze A/B testing practices at a granular level, which gives us the unique opportunity to connect these various streams of research. Additionally, one characteristic of this study that distinguishes it from the previously cited literature is its use of data from thousands of experiments by hundreds of different firms. As such, we believe the findings of this study and the associated managerial insights can be expected to generalize quite broadly to a large number of e-commerce companies in a way that micro-analyses of individual firms often fail to do.

3 Data & Descriptive Analysis

3.1 Data Collection

Provenance. Our data has been collected from a SaaS-provider of A/B testing technology and services ("the platform"). Like many other platforms in this space, our partner is a third-party technology service that allows websites to conduct randomized controlled trials on their online customer base. To use the service, firms go through a relatively minor integration process that involves inserting a Javascript snippet into their website's code that (1) makes it possible for the testing platform to manipulate what customers see in real-time and and (2) measure customer responses (time on site, pageviews, whether something was purchased, etc.). After this snippet is installed, firms can log into the testing platform's website where they will see a dashboard that allows them to create new experiments and see the analytics associated with ongoing and past experiments. To create a new experiment, firms can either use the platform's point-and-click editor or custom Javascript that allows them to manipulate essentially any element on their website. As we will describe below, these manipulations are often new promotions or small changes to a site's visual elements and layout. Using the platform, firms will then choose how much of their web traffic to experiment on and specify any technographic targeting conditions that can limit experiments to particular segments of their customer base (e.g., mobile device users, returning customers). Once an experiment begins, the platform automatically allocates incoming website visitors to a randomly selected treatment arm and metrics about each visitor's behavior are reported back to the platform. Using the platform's reporting dashboard, firms can then see the analytics associated with an experiment and the results of standard statistical tests comparing the treatment groups on various outcomes of interest.

Inclusion Criteria. We have collected the results of all experiments conducted on the platform by US-based firms between January 2014 and February 2018. We limited our analysis to experiments with only two treatment groups. This was motivated by the fact that two-condition experiments are the most common type of intervention on the platform and that requirement allows us to cleanly identify the intended intervention being tested in a given experiment. We also selected experiments for which the firm specified "conversion rate" as their primary outcome variable. A "conversion" in this context occurs whenever a visitor completes a purchase (of any amount) on the site. This is easily the most common outcome firms specify as their primary dependent variable, as nearly 90% of the experiments in our population have this set as their goal metric. Lastly, we have only included experiments that have at least 30 observations in each treatment group and at least 10 observations for each outcome (conversion, no conversion). This last requirement matches the testing platform's minimum data requirements before they report the results of any statistical tests to the firm. The resulting dataset contains 2,732 experiments from 252 unique firms. An important feature of our sample is that 100% of the websites have some type of e-commerce checkout process. Thus all conversions in our dataset involved a monetary transaction for some good or service. While this limits the generalizability of our results to other sectors (e.g., digital media sites, whose primary conversion metric may be email sign-ups), this maximizes the value of our insights for e-commerce companies.

3.2 Firm Level Characteristics

Within the population of e-commerce companies, our data contains a significant degree of heterogeneity in the size and type of firms being investigated. It includes many modestly-sized firms with daily traffic of less than 100 visitors, but it also includes some of the largest brands and e-commerce websites in the United States (with daily traffic exceeding 100,000 visitors); 16.7% of firms in our dataset are publicly listed. We are also able to use a business intelligence service[1] to identify the industrial sector of firms in our sample, which we have shown in Table 1. The largest sector in our dataset by far is consumer discretionary websites (149), which mostly includes websites selling fashion and clothing accessories. We also have 51 websites in the consumer staples space (e.g., food and household items) and 27 classified as information technology (mostly firms that sell software and technology services). A portion (21) of firms are in typically B2B industries such as the healthcare, telecommunications, industrials, and financials; 29 firms in our sample could not be matched.

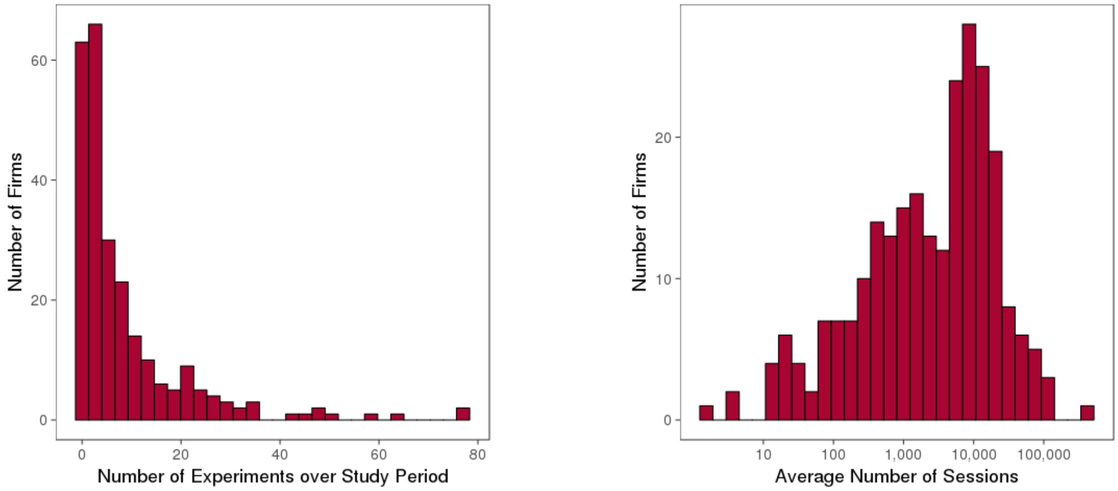
We have generated plots for some high-level statistics to help visualize some of the most important features of our dataset. Figure 1a displays the distribution of experiment counts by firm. The majority of firms in our sample (58%) have fewer than 5 experiments; the firm with the most number of experiments accounts for 75 of the observations in our dataset. We do not directly observe the volume of traffic associated with each website in our sample. However, we are able to approximate the daily traffic of each website by using the number of sessions (i.e., customers or observations) throughout the duration of each experiment to calculate an imputed velocity of web traffic over a 24-hour time period. We then average this value across all experiments by each firm to arrive at an estimation of how much daily traffic each website receives; this distribution is plotted in Figure 1b (log scale). As can be seen, the vast majority of firms have daily traffic between 1,000 and 100,000, with the primary mode of the distribution near 10,000 daily visitors.

3.3 Experiment Level Characteristics

We have also calculated some summary statistics to better understand the nature of the individual experiments in our sample. The average number of sessions in an experiment (i.e.,

[1] We used a database developed by Clearbit to match the domains of the firms in our sample to existing public records to obtain these data.

Figure 1: Histograms of Firm Level Characteristics


(a) Distribution of Experiments Counts per Firm

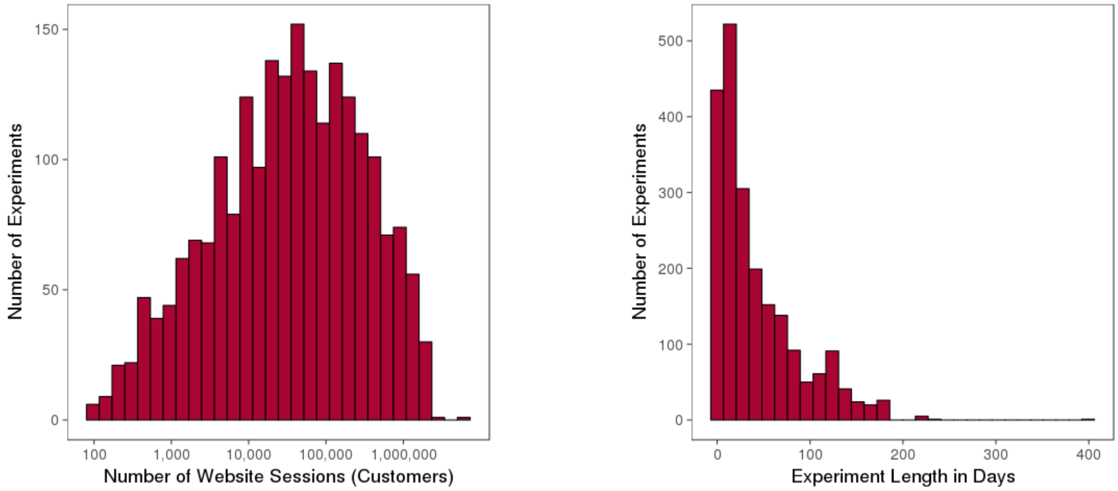
(b) Distribution of Imputed Daily Traffic by Firm

the number of website visitors for which an observation was made during the test period) is 185,540; the distribution of session counts is extremely skew with a standard deviation of 365,099 sessions (see Figure 2a, log scale). The average experiment in our sample runs for 42.4 days, with a sample standard deviation of 44.4 days (Figure 2b).

3.4 What types of experiments do e-commerce firms run?

3.4.1 Language of A/B Testing: Unstructured Text Analysis. We now turn to offer a partial answer to the question, "What are firms experimenting with on their websites?" To do this, we will examine the textual metadata firms use to describe the nature of their intervention. The

Figure 2: Histograms of Experiment Level Characteristics


(a) Distribution of Session Counts by Experiment

(b) Distribution of Experiment Durations

Table 1: Top Words in A/B Test Descriptions

Word	Frequency	Number of Firms Using Word
banner	556	95
free	212	57
offer	206	46
sale	196	49
top	172	54
nav	167	42
homepage	167	54
shipping	150	42
countdown	149	23
email	142	41
promo	142	41
show	126	32
checkout	119	27
cta	117	26
split	116	55
mobile	111	48
cart	111	44
product	110	41
day	101	40
header	91	27

testing platform allows firms to give both titles to their experiments and descriptions for each treatment group. A representative example of the text firms provide in these fields would be, “Top Banner Shipping Test” for the experiment title and (“Control”, “Free Shipping”) for the description of the two groups. To provide some insight into the language firms use to describe A/B tests in our entire sample, we calculated word frequencies in the entire corpus of experiment titles and descriptions. We removed common English language stopwords, numbers, any company-identifying words, and the most common non-descriptive words in our sample: “test”, “control”, “new”, “version”, “page”, “html”, “css”, “javascript”. We then counted the number of times the remaining words appeared in our sample and also the number of unique firms that used each word. We have displayed the top 20 words by frequency in Table 1.

We highlight two observations about the words that commonly appear in our sample. First, it appears that the majority of A/B tests are fairly incremental changes to a website’s existing design. By far the most common word in our sample is "banner" which, in the language of web design, most frequently refers to an image placed above or to the side of a website’s navigation that contains seasonal messaging or sales information. Second, we can see firms often use words to describe the *location* of their intervention (“homepage”, “nav”, “checkout”) or the *nature* of the intervention (“free”, “promo”, “cta”). We build on this observation and formalize this distinction in the following section.

3.4.2 Feature Extraction & Classification: Structured Text Analysis. To facilitate a meaningful quantitative analysis of our dataset, we now attempt to provide a more structured classification of the most common experiment types in our sample. In particular, we set out to exploit the textual metadata described above and categorize the experiments in our sample into high-level groups that are meaningful to compare from a theoretical and practical standpoint. We will draw on the marketing and website design literature referenced earlier to analyze the effectiveness of various intervention types that are found in our dataset. As referenced in Section 2.2, both marketing/promotional and design/usability aspects of e-commerce websites are known to play significant roles in influencing user behavior. At the same time, prior literature has highlighted the importance of funnel stage when evaluating the effectiveness of various interventions. As such, we will classify experiments in our dataset along these two high-level dimensions: experiment *type* and experiment *location*.

Experiment Type. Building on the literature cited in Section 2.2, we distinguish between three primary types of experiments our analysis. Given the centrality of price as a driver of economic behavior, we believe it is important to distinguish between interventions that affect purely aesthetic parts of the e-commerce experience and those that affect prices. Thus our first category of experiments will be *non-promotional design* interventions. To identify experiments of this type, we cross-referenced the metadata of each experiment with a list of design-related keywords that are commonly used in the online user experience literature (see Table 2). Additionally, we wanted to study the effects of interventions that affect pricing which are quite common among our sample. However, the final price consumers pay can be affected by both adjustments to the list price of a given product or adjustments to the shipping costs. Since we know from the price partitioning literature that consumers often respond differently to these interventions, we separated them out in our analysis. Thus our second and third experiment type categories are those involving *promotional* (i.e., list price adjustments) and *shipping*-related interventions.

Table 2: Experiment Type Classification

Class	Incidence	Sample Keywords
Design	1,542	"button", "cta", "hero", "image", "text", "color", "layout", "show", "hide"
Promotion	415	"promo", "sale", "deal", "X% off", "discount"
Shipping	268	"shipping", "delivery", "FS" (abbrevation for "free shipping")

We have applied this coding scheme to our data in way that allows an experiment to belong to one (and only one) of these three categories. Any experiment that matched for both "design" and "promotion" keywords was counted as "promotion". Furthermore, any experiment that matched for both "promotion" and "shipping" keywords was coded as "shipping". (This is because an experiment titled "50% off shipping" would match for both "promotion" and "shipping" categories, but the promotion is clearly tied to the shipping costs. We could find no examples of an experiment in which a price promotion and shipping promotion were targeted simultaneously.)

Experiment Location. As described in 2.3, we know that different marketing interventions can have differential impact depending on where they are targeted in the conversion funnel. A useful way for thinking about the conversion funnel in the context of e-commerce design is to map different aspects of a website to different stages in the online conversion process. Perhaps the most natural phase of the conversion funnel on any website is the homepage, where almost all website visitors start their interactions with online merchants. As such, homepage interventions will serve as our baseline "early funnel" class of experiments. In coming up with other ways of mapping website elements onto funnel stages, we draw upon the work of Moe (2003) and Song and Zahedi (2005). In particular, both papers distinguish between behaviors having to do with browsing and searching and those concerning purchase deliberation, facilitation, and checkout. Thus, we define one funnel phase as experiments targeting "product listing pages" or "landing pages" (as they are commonly referred to in the web design industry); these are pages that list many products at once within a given category (e.g., "men's pants" and "women's accessories" are common landing pages in fashion retail). Interventions on these pages affect consumers who are in the process of searching and filtering the product listings on a website. We then operationalize experiments targeting the last phase in the e-commerce funnel as those that manipulate elements on individual product listing pages or in the checkout process. These interventions are interpreted as affecting customers who are

Table 3: Experiment Location Classification

Class	Incidence	Sample Keywords
Sitewide	418	“sitewide”, “navigation”,
Homepage	354	“homepage”, “HP”
Category	148	“listing page”, “landing page”, “plp”
Purchase	248	“details page”, “pdp”, “cart”, “checkout”

actively deliberating about or committing to the purchase of a given item. Lastly, there are a number of web design elements that appear on every page; these include the website header at the top of the page, the navigation menu, and the footer at the bottom of the page. Because it doesn’t make sense to think of these as targeting users at any particular phase of the conversion funnel, we consider manipulations of these sitewide elements as their own category of experiments. In total, we have four distinct experiment locations: *sitewide, homepage, listing, and purchase*. The incidence of experiments in each of these funnel categories and the keywords used to identify them in our dataset are shown in Table 3.[2]

Having developed this classification scheme and labeled the experiments in our dataset using their titles and descriptions, we will now ask if different experiment types systematically vary in their effectiveness. We will try to quantify the average effect sizes of experiments in our sample across both experiment *type* and *location*. Importantly, we will attempt to identify how the effectiveness of different types of interventions varies across different stages in the e-commerce conversion funnel. However, there are several important subtleties in our dataset that are important to model correctly to ensure we are identifying the right effects in our analysis. We will address some of these challenges in our model setup below and then report on our findings about how experimental effects vary across the conversion funnel.

4 Meta-Analysis of Experimental Outcomes

4.1 Aggregate Analysis

As mentioned earlier, the stated primary objective of all A/B tests in our sample is to increase a website’s conversion rate, i.e., the proportion of customers who buy something out of the entire population of website visitors over a fixed period of time. As such, the main

[2] A careful reader will notice we have not mapped the most common word in our dataset, “banner”, to any category. This is because, by itself, an experiment testing a banner could be changing the promotional information contained in the banner or changing the design of the banner without adding any new information. A banner could also be something that is placed throughout the entire website or only a subset of pages. Thus, by itself, the word “banner” does not resolve much uncertainty in placing an experiment in either the type or location dimension.

dependent variable associated with each experiment is its measured *effect size*—or conversion rate "lift", as its known colloquially in digital marketing. To define this outcome concretely, consider a given experiment (indexed by i) and its two associated treatment conditions, $t \in a, b$. In our dataset, we observe both the number of conversions (represented by c_i^t) and total sessions (represented by n_i^t) in each of the two treatment conditions for each experiment in our sample. The observed "effect size" of an experiment is then defined as the difference in mean conversion rates between the two treatment conditions: $\hat{\delta}_i := c_i^a/n_i^a - c_i^b/n_i^b$.

One limitation in our dataset is that we cannot consistently identify the control condition in many of our experiments. That is, while we know whether individuals are either in treatment arm a or b, we do not always know which treatment arm represents the intervention and which represents the control group. Furthermore, an intervention may not even have a meaningful "control" group; this is because interventions in one treatment condition can be positive, negative, or lateral changes compared to interventions in the other treatment condition. For example, an experiment with a title of "20% promo test" may be adding or removing promotional information, relative to the status quo version of the website at time of the intervention. So while we will know this experiment has something to do with promotions, we usually cannot identify the exact intervention being tested. This causes the distinctions between treatments a and b to vary arbitrarily across experiments; as such, the sign of $\hat{\delta}_i$ also varies arbitrarily. This can be resolved by considering the *absolute effect size* of each experiment, $|\hat{\delta}_i|$, rather than the signed effect size. Given that the distribution of effect sizes is extremely skew (tightly clustered around zero), we will also be working with the log-transform of absolute effect size in subsequent analyses: $y_i = \log|\hat{\delta}_i|$. We have plotted the distribution of observed effect sizes in our sample in Figure 3 (left panel), along with the distribution of absolute (middle panel) and log-absolute effect sizes (right panel).

Before proceeding to examine heterogeneity across different experiment types, there are several characteristics of the aggregate effect size distribution that are worth remarking on. For one, the typical effect size observed in an A/B test is very small: the median (absolute) effect size is just 0.1%, with the mean slightly higher at 0.7%. It is notable that for half the experiments in our sample, the interventions failed to move the conversion rate (in any direction) by more than one tenth of one percent. These figures suggest a large degree of skew associated with the distribution of effect sizes. Indeed, the shape of the effect size distribution is a central theme in the work of Azevedo et al. (2018), cited earlier. Using data from Microsoft, the authors find a very similarly skewed distribution of effect sizes; they use the shape of their distribution to argue that it is better to run a large number of low-powered experiments to find the small number interventions with outsized returns. Our sample, which has data from more than 250 different firms, appears to be consistent with this finding. Indeed, the distribution of effect sizes in our sample appears to almost perfectly follow the classic “Pareto principle”: 20% of the experiments in our sample account for 81% of the lift aggregated across experiments. This relationship can be seen in Figure 4, in which we have ordered the effect sizes from largest to smallest (left panel) and then calculated the cumulative sums by percentile (right panel).

4.2 Heterogeneity Across Experimental Types

Having documented the heterogeneity of experimental interventions in our dataset in Section 3.4, we now turn to the question of whether these different strategies can be linked to a test’s outcome. This analysis will not only provide practical insight for managers when determining their testing strategies, but it also makes a theoretical contribution to the e-commerce literature by documenting and quantifying how different factors drive online shopping behavior. To investigate this topic, we will be performing a meta-analysis on the effect size of experiments in our sample. Specifically, we will ask how the intervention types identified earlier are related to an experiment’s absolute effect size. We already documented why we discard the sign of the effect sizes in our sample, but—before proceeding to the model definition itself—it is worth taking the time to clearly delineate how this definition of our dependent

Figure 3: Distribution of experimental effect size and its transforms in our sample

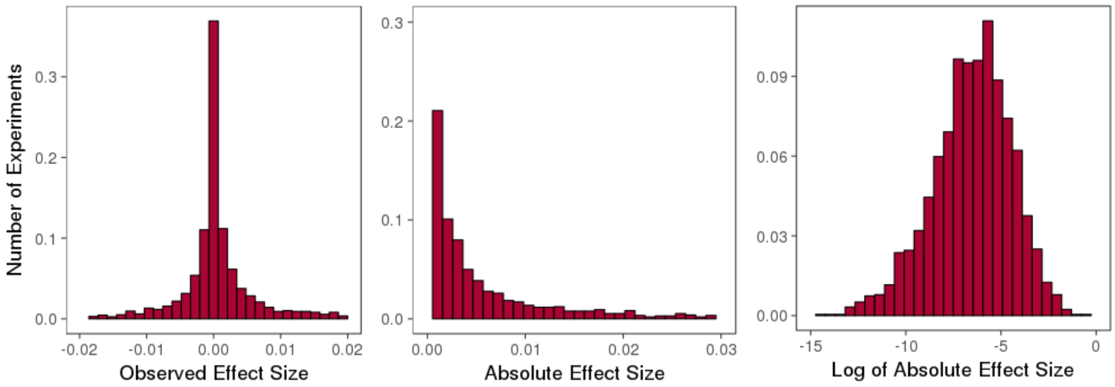
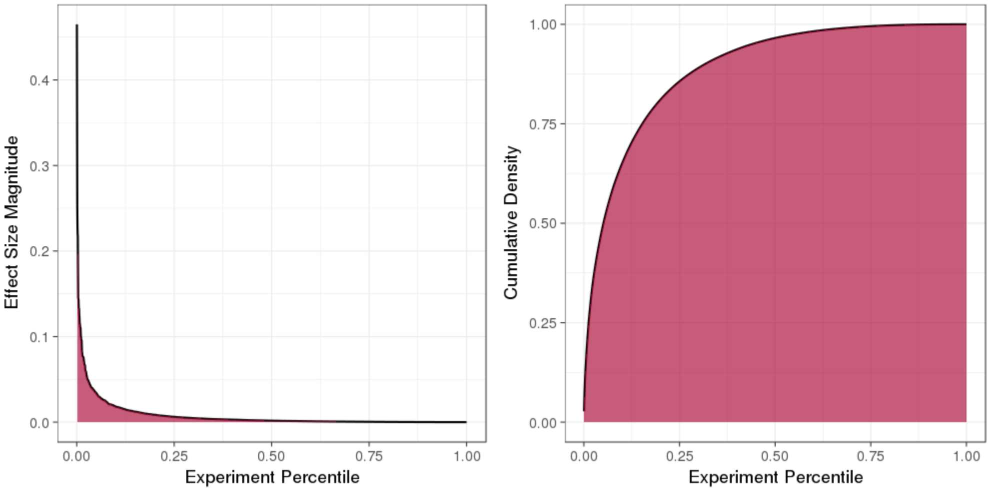


Figure 4: Distribution of Absolute Effect Sizes by Percentiles



variable affects the interpretation of our model and its parameter estimates.

While we lose some information in our data by discarding the direction of each experiment's effect size, we also gain the ability to aggregate data across experiments by intervention type. This is particularly useful in the context of a cross-firm, cross-experiment meta-analysis. To consider this concretely, consider a scenario in which two websites are testing an intervention that increases the size of their product images. Relative to a baseline *status quo* control condition, suppose larger product images lift conversions for Website A (positive effect size), while smaller product images may be better for Website B (negative effect size). If we attempted to simply average these signed effects (in this over-simplified, stylized model), we would find that manipulating product images has no effect on conversion rates. But by taking the absolute values, we are instead able to say that image manipulations *do* have a significant effect on conversion rates, but the best way to implement that manipulation will vary from website to website. As such, by taking the absolute value, we abstract away from answering the question of whether a particular intervention increases or decreases conversion rate, but rather answer the question of how generic types of interventions affect conversion rates. In this sense, instead of studying which interventions *improve* conversion rates, our analysis answers a slightly different question: "Which types of interventions have the largest average impact on moving a website's conversion rate (either positively or negatively)?" Note that given the discussion around the shape of the distribution of effect sizes—in which experiments with large effect sizes are actually quite rare—identifying broad classes of interventions that have larger effects is indeed a valuable exercise. As with any generalizable insights, firms will need to experiment within a given class of interventions to identify which particular manipulations result in positive effect sizes, but this analysis can at least guide this search process toward the fatter tail of the effect size distribution.

4.3 Top-Level Models

We are now in a position to precisely define our primary regression model. We define our dependent variable to be the log of the absolute effect size associated with an experiment: $y_{ij} = \log|\hat{\delta}_{ij}|$; in this setup, index i represents the i-th experiment associated with firm j in our sample. Our main research questions are about how intervention *type* and *location* varies with effect size in e-commerce experiments. Before analyzing any interaction effects between these factors, we first identify the main effects of these variables. We run two separate regressions modeling experiment outcomes y_{ij}—the log absolute effect size of each experiment—on either a Type_{ij} variable or a Location_{ij} variable that includes dummies for each of the experimental categories identified in section 3. We also include a set of control variables, Controls_{ij}:

$$y_{ij} = \text{Type}_{ij}\boldsymbol{\theta} + \text{Controls}_{ij}\boldsymbol{\gamma} + \varepsilon_{ij} \qquad (M1)$$

$$y_{ij} = \text{Location}_{ij}\boldsymbol{\theta} + \text{Controls}_{ij}\boldsymbol{\gamma} + \varepsilon_{ij} \qquad (M2)$$

The controls in this regression are included to address potential concerns about endogeneity in our model. The largest likely source of endogeneity is unobserved heterogeneity across firms. In particular, it is plausible that firms from different industries exhibit systematically different testing strategies across experiment types. Even within an industry, it is possible that—for unobservable reasons—some firms are simply more likely to make interventions with large effect sizes. The extent to which this is also correlated with our main independent variables—the types of tests firms choose to perform—will bias the estimates of our primary parameters of interest. Thus, we include firm fixed effects variables in our set of controls. Furthermore, it is likely that different types of experiments (say, promotions) are conducted with varying frequency through the year. If observed effect sizes also systematically vary throughout the year, this would be a source of omitted variables bias in our estimation. Indeed, we know from conversations with the testing platform that both firms and consumers exhibit atypical behavior around certain time periods throughout the year (e.g., those near Thanksgiving, Christmas). For this reason, we also include separate time-specific dummy variables for each week of the year.

Imperfectly Observed Data. There is one challenge with how we have defined our dependent variable that presents a non-trivial challenge for obtaining unbiased estimates of the parameters in our model. In particular, we must consider the fact that our dependent variable is necessarily observed with noise. Considering first the raw effect size that our dependent variable is based on, recall that $\hat{\delta}_{ij}$ does not represent the *true* effect size associated with an experiment, but rather an *estimate* of this value (we have used the "hat" notation to distinguish this estimate from the true parameter, δ_{ij}). This is, of course, why statistics are necessary in A/B testing in the first place: to quantify the uncertainty around this estimated effect size and determine if it is significantly different from zero. Standard practice in A/B testing is to use the proportional means Z-test with pooled variance. In this test, a Z-score is computed by first calculating the standard error of the mean:

$$\hat{\nu}_{ij} = \sqrt{\hat{p}_{ij}(1-\hat{p}_{ij})(1/n^a_{ij} + 1/n^b_{ij})}, \text{ where } \hat{p}_{ij} = (c^a_{ij} + c^b_{ij})/(n^a_{ij} + n^b_{ij})$$

This is then divided into the estimated mean, $\hat{\delta}_{ij}$, and cross-referenced with the standard normal CDF to obtain a p-value: $Z = |\hat{\delta}_{ij}/\hat{\nu}_{ij}|; p = \Phi(Z)$.

To minimize the influence of spurious results and the incidence of false positives, this p-value is typically checked against a pre-determined significance level α (almost universally 0.05). Firms usually only consider experiments that reach this significance level to be of value. We are in a similar position, in which we want to minimize the impact of spurious correlations on our parameter estimates. As such, we may be tempted to exclude any experiments in our dataset with p-values above a designated threshold, α. However, this also has significant downsides that require us to discard the vast majority of our data. Another problem with this approach is that, despite the prominence of the conventional α level of 0.05 throughout the history of statistics, any value of α is essentially an arbitrary modeling choice (Gelman and Stern, 2006).

However, there is an alternative: a common approach to address imperfectly observed variables is to weigh the observations by the inverse of their variance. This causes outcomes that are observed more precisely to have higher weight in determining the model outcome. Not only that, but this method—inverse variance weighted least squares, or WLS—is actually the best, linear, unbiased estimator of average partial effects in a regression model (Hartung et al., 2011). In most empirical research projects, the effects of noise on dependent variables are often difficult to deal with, since one rarely has estimates of both the observation mean *and* its variance. Our dataset is unique in that we are able to calculate both the point estimate, $\hat{\delta}_{ij}$, and the standard error, $\hat{\nu}_{ij}$, of our dependent variable.

Lastly, because we are ultimately working with the log-norm transform of $\hat{\delta}_{ij}$, suitable care must be taken to calculate the proper variance of the resulting y_{ij} variables. By the Central Limit Theorem, the sampling distribution of $\hat{\delta}_{ij}$ will be asymptotically Gaussian, centered around the true effect size with standard deviation $\hat{\nu}_{ij}$. If we let $u_{ij} = |\hat{\delta}_{ij}|$ be defined as the absolute value of the observed effect size in a given experiment, then the sampling distribution of u_{ij} will follow what is known as the *folded normal distribution*. Finally, by setting $y_{ij} = \log u_{ij}$, we can see that the exponential of y_{ij} will be a folded normal random variable. Using these facts with the density function of a folded normal variable and the integral definition of variance, we can parameterize the variance of y_{ij} purely in terms of $\hat{\delta}_{ij}$ and $\hat{\nu}_{ij}$ using the formula below. In this study, we evaluate this integral using Monte Carlo simulation.[3]

$$\begin{aligned}
\hat{\sigma}_{ij} = \mathbf{Var}[y_{ij} \mid \hat{\delta}_{ij} = \delta, \hat{\nu}_{ij} = \nu] &= \mathbf{E}\left[(y_{ij} - \mathbf{E}[y_{ij}])^2\right] \\
&= \int_{\mathbb{R}} (y - \mathbf{E}[y_{ij}])^2 p_y(y)\, dy \\
&= \int_{\mathbb{R}} y^2 p_y(y)\, dy - \mathbf{E}[y_{ij}]^2 \\
&= \int_{\mathbb{R}} y^2 p_u(e^y)\, dy - \mathbf{E}[y_{ij}]^2 \\
&= \int_0^{\infty} y^2 \left[\frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(e^y-\delta)^2}{2\nu^2}} + \frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(e^y+\delta)^2}{2\nu^2}}\right] dy - \mathbf{E}[y_{ij}]^2
\end{aligned}$$

Having obtained an estimate for the variance, $\hat{\sigma}_{ij}^2$, in our sample, we then define each

[3] Specifically, given that we have closed form estimates for $\hat{\delta}_{ij}$ and $\hat{\nu}_{ij}$ for each experiment, we first generate 100,000 draws (per experiment) from a normal distribution parameterized by these values. We then calculate the empirical variance of the log-norm transform of the simulated values to obtain a reliable measure of the variance associated with each observation in our sample.

observation's weight as its inverse variance: $w_{ij} = 1/\hat{\sigma}_{ij}^2$. We then estimate the parameters of our model, $\boldsymbol{\beta} = (\gamma\ \boldsymbol{\theta})$, using the weighted least squares estimator: $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{y}$, where $\mathbf{X}$ is the stacked matrix of data vectors $\mathrm{x}_{ij} = (\text{Controls}_{ij}\ \text{Type}_{ij})$ and $\mathbf{W}$ is a diagonal matrix of observation weights w_{ij}. This estimator minimizes the *weighted* squared error of the model's residuals, resulting in estimates of our regression coefficients that will be unbiased (assuming conditional mean independence).

4.4 Main Effects Models

We now turn to the results of our regression analysis, summarized in Table 4. Columns (1) and (2) correspond to the main effects models described earlier for experiment type and location (respectively). In the first model, we have used the *design* category as the baseline class. We see in column (1) that the coefficient on "Promotion" is positive and significant, with a point estimate suggesting that the average promotional experiment in our dataset has an effect size that is $100e^{1.26-1} \approx 127\%$ that of the average "Design" experiment (without controlling for intervention location). The coefficient on "Shipping" is negative, but not significantly distinguishable from zero. Calculating the contrast between the "Promotion" and "Shipping" coefficients (equivalent to changing the baseline class in our regression design) results in a significant T statistic of 2.02 (p=0.04). The fact that there does appear to be a differential effect between price and shipping promotions is consistent with the literature on price partitioning.[4]

Turning to the location model in column (2), we have set the "Homepage" class as our baseline experiment location. Without controlling for intervention type, the "Sitewide" experiments (those manipulating elements that appear on every page of a website) appear significantly smaller than the average "Homepage" experiment ($\beta = -1.65, p < 0.001$); experimental effects in the "Category" funnel stage are significantly larger ($\beta = 1.24, p < 0.01$). Perhaps surprisingly, the coefficient on the last funnel stage ("Purchase") is not distinguishable from that of "Homepage" experiments.

[4] However, note that we are not controlling for the level of discount across interventions; this means we are not directly comparing the effect of a fixed amount of price change across listing and shipping prices, as would be required for a proper analysis of price partitioning.

Table 4: Statistical models

	(1)	(2)	(3)
Intervention Type			
Design (Baseline)	—		—
Promotion	1.26^{*}		2.10^{*}
	(0.64)		(1.00)
Shipping	-1.08		-1.03
	(0.96)		(0.72)
Intervention Location			
Homepage (Baseline)		—	—
Sitewide		-1.65^{***}	-0.06
		(0.48)	(0.48)
Category		1.24^{**}	1.79^{***}
		(0.40)	(0.38)
Purchase		0.06	0.84
		(0.52)	(0.68)
Interaction Effects			
Sitewide x Promotion			-1.49
			(1.12)
Sitewide x Shipping			1.38
			(0.84)
Category x Promotion			-6.20^{***}
			(1.84)
Category x Shipping			0.29
			(1.12)
Purchase x Promotion			-2.30^{+}
			(1.25)
Purchase x Shipping			2.59^{**}
			(0.99)
Firm Fixed Effects	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes
Num. obs.	2203	1085	899

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$. Dependent Variable: Log of absolute effect size. Heteroskedasticity robust (White-Huber) standard errors are reported in parentheses.

4.5 Interaction Effects: Heterogeneity by Funnel Depth

Among our primary research questions is to investigate how various types of e-commerce interventions may differentially affect purchase behavior throughout the online conversion funnel. Having run two main effects regressions above—in which we looked at the effect of experiment type and location separately—we will now specify an interaction model that will allow us to determine if treatment effectiveness does depend on where an intervention is targeted in the funnel. In particular, we include the two categorical variables from the prior specifications (Type$_{ij}$ and Location$_{ij}$), as well as the interaction between them.

$$y_{ij} = \text{Type}_{ij}\boldsymbol{\theta}_1 + \text{Location}_{ij}\boldsymbol{\theta}_2 + \left(\text{Type}_{ij} \times \text{Location}_{ij}\right)\boldsymbol{\theta}_3 + \text{Controls}_{ij}\boldsymbol{\gamma} + \varepsilon_{ij} \tag{3}$$

This model allows us to not only control for treatment heterogeneity in the identification of our main effects coefficients, but the interaction coefficients will also give us the ability to detect how different types of interventions vary in their effectiveness by funnel depth. The results of this model (estimated with WLS and the same control variables as in prior specifications) are shown in column (3) of Table 4.

Looking at the results, we first note that the main effects coefficients among the intervention type variables are qualitatively similar to the main effects regression in column (1). Comparing model (3) with the location regression in model (2), we see an attenuation of the coefficient on the "Sitewide" location; this suggests controlling for intervention type is an important factor to consider when looking at how intervention location affects consumer purchasing behavior. Turning to the interaction effects, we find that there does exist heterogeneity in the effectiveness of various interventions across funnel depth. In particular, we find evidence that "Promotion" interventions—those offering discounts or purchase incentives on product list prices—become less effective at later stages in the funnel. The differential effect across the funnel between "Promotion" and the baseline "Design" class appears to be strongest at the "Category" phase, with the largest interaction coefficient being found between "Category x Promotion" ($\beta = -6.2, p < 0.001$). A marginally significant effect is also seen at the "Purchase" funnel stage as well ($\beta = -2.30, p < 0.10$). While our results do not suggest promotions are *ineffective* at later stages of the online conversion funnel, they appear to be most effective if advertised earlier in the customer's journey through an e-commerce website. While this would need to be confirmed with further research, these results are consistent with the hypothesis that promotions on a website do more to convince shoppers that they want to buy *something* rather than affecting their decision about whether to buy a particular product they are already considering.

Turning to the interaction coefficients on the "Shipping" category, we find a evidence for a positive moderation between funnel depth and the effect size of shipping-related interventions. While the interactions between "Shipping" and both "Sitewide" and "Category" funnel locations are not statistically distinguishable from zero, there is a significant and positive coefficient on the "Purchase x Shipping" interaction ($\beta = 2.59, p < 0.01$). This is perhaps not surprising by itself, as it would make sense that customers are not as sensitive to logistical costs like shipping if they have not yet decided to purchase anything. On the other hand, rational consumers would be expected to consider *total costs*—including both list price and shipping costs—when making a purchase decision at all phases of the consideration process.

We have known from the literature on price partitioning cited in 2.2 that consumers deviate from rationality in the form of a first-order effect between list and shipping prices—i.e., that consumers react more strongly to a $1 change in shipping price more than an equivalent change in listing price. However, our research suggests the existence of a second-order effect of price-partitioning on consumer purchase behavior. When we consider the differential effects of funnel depth on both "Promotion" and "Shipping" interventions—that list price promotions become less effective later in the funnel while shipping promotions become more effective later in the funnel—our analysis suggests that both the type and location of marketing interventions are important factors for understanding consumers' response to promotions. In addition to providing novel insight into how online consumers deviate from rational consumption behavior, this research is also of consequential practical significance to managers of e-commerce firms, as it suggests how firms can maximize the impact of both their A/B tests and unilateral marketing interventions by factoring both the type of intervention and where it is advertised within a website's architecture.

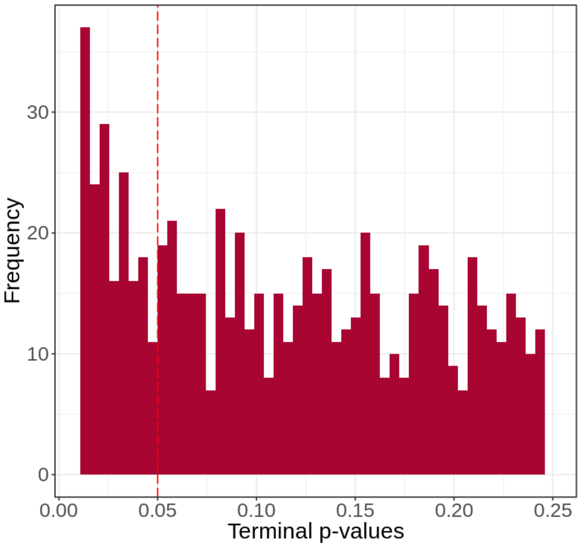
5 Robustness Checks

5.1 Have our data been p-hacked?

As a meta-analysis, our findings will only be as valid as the data we put into the aggregate model. Said differently, if the experiments in our sample were subject to some form of data corruption, then any findings derived from inferential statistics about the levels and uncertainty about effect size estimates may reasonably be called into question. Some recent findings on other experimentation platforms have suggested that firms using A/B testing tools engage in a practice of "continuous monitoring", whereby they actively watch test statistics as data arrive (Berman et al., 2018). This allows firms to choose when to stop an experiment if results appear significant (early stopping), or gather additional data if results do not appear significant (extra data collection). These behaviors—a form of so-called p-hacking—are known to inflate the empirical false discovery rate above the nominal α level (Simmons et al., 2011). While our meta-analytic methods do not directly make use of p-values, there is an implicit assumption about the validity of the effect size and uncertainty estimates being used to form our dependent variable and weighting coefficients. Thus, an investigation into whether this phenomenon can be detected in our dataset would provide some evidentiary value to the hypotheses described earlier in this report.

5.1.1 How would we detect p-hacking?. In our context, the primary concern is that experimenters have manipulated the results of their experiments to achieve "statistical significance". On the platform studied in this project, the interface was designed to prominently highlight experiments with p-values less than 0.05. Specifically, for each experiment in the interface, the platform displays a "confidence" value, calculated as $1-p$, which is continuously updated as new data arrive. Once this confidence value reaches 95%, the results from that experiment are highlighted and are visually distinguished from results with confidence levels below this threshold. If firms were indeed responding to this threshold effect, we would expect there to be a disproportionate amount of experiments with confidence values just above the 95% threshold. Said differently, if we believe this p-hacking behavior is prominent in our sample, it leads to a prediction that there would be a discontinuity in the distribution of p-values near the 0.05 significance threshold. We will investigate whether such a discontinuity exists

Figure 5: Distribution of p-values near 0.05 significance threshold



in our data.

5.1.2 Detecting a discontinuity in the density of p-values. A histogram of terminal p-values in our experiment is shown in Figure 5, with a dotted line plotted at the 0.05 threshold. While there is no visually-striking discontinuity near this threshold, this could be due to the histogram bin-width in this particular plot. To provide formal statistical evidence about the presence (or absence) of a discontinuity at the 0.05 threshold, we will apply the methods of Cattaneo et al. (2018b). Note a simple regression discontinuity design is inappropriate for testing discontinuities in density functions, since one must account for the sampling uncertainty associated with any given point in the estimated density. The method developed by Cattaneo et al. (2018b) accounts for this by first using low-order, local polynomial regression estimates of the empirical *cumulative* distribution function, and then calculates estimates of the *density* function as the derivative of this estimated CDF. Calculating the variance around a point estimate of the density function then becomes equivalent to calculating the variance of a slope coefficient, which is a long-standing and well-established practice in econometric theory.

To derive a test statistic for our specific context, we specify a null hypothesis that assumes continuity of the underlying density function f at the $c = 0.05$ threshold: $H_0 : f(c_-) = f(c_+)$. Cattaneo et al. (2018b) provide an asymptotically Gaussian test statistic, T, that compares independent estimates of the density function on either side of the threshold value (i.e., one estimate for data below 0.05 and a separate estimate for data above 0.05.[5] As is common in the RD literature for testing robustness to bandwidth selection (McCrary, 2008), we first calculate an optimal bandwidth $\hat{h}$ by minimizing asymptotic mean squared error (MSE), and then provide test statistics of the discontinuity using the optimal bandwidth ($\hat{h}$), half the optimal bandwidth ($\hat{h}/2$), and twice the optimal bandwidth ($\hat{h} \times 2$).

We note that of the 2,782 experiments in our sample, there are 414 with terminal p-values below 0.05 and 2,368 experiments with p-values above this threshold. We used the companion software provided by Cattaneo et al. (2018a) to calculate our data's test statistics. The MSE-optimal bandwidth for our data is $\hat{h} = 0.012$. For each of the three bandwidth values described above, we calculate a robust T-statistic, a 2-sided p-value (testing for *any* discontinuity at the 0.05 threshold), and a 1-sided p-value (testing specifically for the expected form of the discontinuity, that there are more experiments with p-values below 0.05 than above 0.05).

5.1.3 Results & Interpretation. Turning to the results of these tests in Table 5, we see only null results for the test of a discontinuity using both the optimal bandwidth (first column) and twice the optimal bandwidth (third column). When using half the optimal bandwidth (middle column) there appears to be some evidence of a discontinuity, but in the opposite direction of what we had expected. That is to say, the 2-sided test resulted in a p-value of 0.025, but the 1-sided test—that specifically tests for whether there are more experiments with p-values below 0.05—resulted in a p-value of 0.988.

We argue that the 2-sided p-value should be interpreted cautiously, as we are performing 6 hypothesis tests, any one of which could provide evidence for a discontinuity. Most weight should be given to the tests based on the optimal bandwidth, which resulted in a 2-sided p-value of 0.31. In totality, there appears to be vanishingly weak evidence that there are fewer p-values below the 0.05 threshold than there are above this threshold. More importantly, given the lack of any significant results testing for a preponderance of p-values *below* the 0.05 threshold, it is reasonable to conclude that there is no evidence of p-hacking in our sample

[5] This statistic is also a function of the order of the polynomials used in the underlying regressions, p, a kernel bandwidth parameter h, and (implicitly) a kernel function $K(\cdot)$ that controls the smoothing of the data. We implement this test using the recommended second-order polynomial fit ($p = 2$), a triangular kernel, and jackknife estimator for variance estimation. Other choices of these parameters yield qualitatively similar results.

Table 5: Tests for discontinuity in density of p-values at 0.05

Bandwidth	$\hat{h}$	$\hat{h}/2$	$\hat{h} \times 2$
	0.012	0.006	0.024
Robust T-statistic	1.015	2.243	1.480
2-sided p-value			
$H_a : f(c_-) \neq f(c_+)$	0.31	0.025^{*}	0.139
$P(\lvert T \rvert > t)$			
1-sided p-value			
$H_a : f(c_-) > f(c_+)$	0.845	0.988	0.931
$P(T < t)$			

of experiments. Though we cannot prove the null hypothesis, this robustness check suggests that if any of the experiments in our sample were subject to p-hacking, the overall impact of such data manipulation on our main results would be quite minimal.

5.2 Alternative Funnel Specification

The results from the primary regression model in Section 4—that models the interaction between experiment type and funnel stage—provides suggestive evidence for two findings: relative to non-marketing design interventions (a) promotional interventions are *less* effective at later stages in the e-commerce conversion funnel; and (b) shipping interventions are *more* effective later in the conversion funnel. An important data requirement in this model was that experiments needed to have metadata that allowed us to identify both independent variables: the experiment type (*design, promotion, shipping*) and location on the website (*sitewide, homepage, category, purchase*). This limited our analysis to experiments meeting both these criteria, which left us with a selected, lower-powered sample to detect our effects in. One may also be reasonably concerned with measurement error in how well our labeling scheme accurately captured the location of a given experiment on a website (due to firms using the same terminology differently or lack of specificity in the keywords we matched against).

In this section, we augment the prior analysis using an alternative operationalization of each experiment's funnel stage. Specifically, we will use what we call an experiment's *baseline conversion rate* as a proxy for how "deep" in the conversion funnel a given experiment is targeted. To motivate this concept, consider two experiments; one offering a promo code on the website's homepage and another offering a promo code when a user begins the checkout process (say on the "cart" page of the website). Note that sessions are only counted in the results of an experiment if the user was exposed to the manipulation. Thus, while essentially all customers will see the homepage promotion, only customers that click through to the cart page will see the cart promotion. A non-converting customer will be much less likely to see the second experiment, whereas a converting customer would be counted in the results of both hypothetical experiments. If we define the baseline conversion rate of a website element as the conversion rate among the set of users exposed to that website element in their session, it is obvious that the homepage experiment will have a lower baseline rate than the cart experiment.

To capture this notion quantitatively, we will use the conversion among all sessions recorded in a given experiment, (independent of which treatment each session was exposed to):

$$BCR_i = \text{BaselineConversionRate}_i = \frac{c_i^a + c_i^b}{n_i^a + n_i^b}$$

While the quantity defined above is by no means the "correct" proxy to use, using any other reasonable proxy yields nearly identical results to those we will report below. Other proxies could be the minimum (or maximum) of the two conversion rates between treatment conditions, the conversion rate of all A arms or all B arms, or a random choice between the conversion rates of each condition—the results presented in this section do not depend on any of these operationalizations.

The goal of defining this quantity is to be able to make comparisons about the funnel stage of different experiments. We are making the case that an experiment with a baseline conversion rate of 10% is "deeper" in the e-commerce conversion funnel than one with a conversion rate of 1%. This seems reasonable since, almost by definition, a section of a website that has a high conversion rate is affecting the behavior of users that are closer to completing a purchase. Thus, we will interpret experiments with high baseline conversion rates as those targeting later stages of the e-commerce conversion funnel (and vice versa).

5.2.1 Modeling Treatment Heterogeneity. By using an experiment's baseline conversion rate as a proxy for its position in the checkout funnel, we can look for heterogeneity between experiments that are targeted at different stages in the funnel. In this analysis, we will divide experiments into "Early Funnel" and "Late Funnel" groups and look for heterogeneous effects between funnel stages among the different experiment types identified earlier. In particular, we choose a conversion rate $\tau \in (0,1)$ as our binarization threshold; experiments with baseline conversion rates above this threshold will be considered "Late Funnel":

$$\text{LateFunnel}_i^\tau = \mathbb{I}(BCR_i > \tau)$$

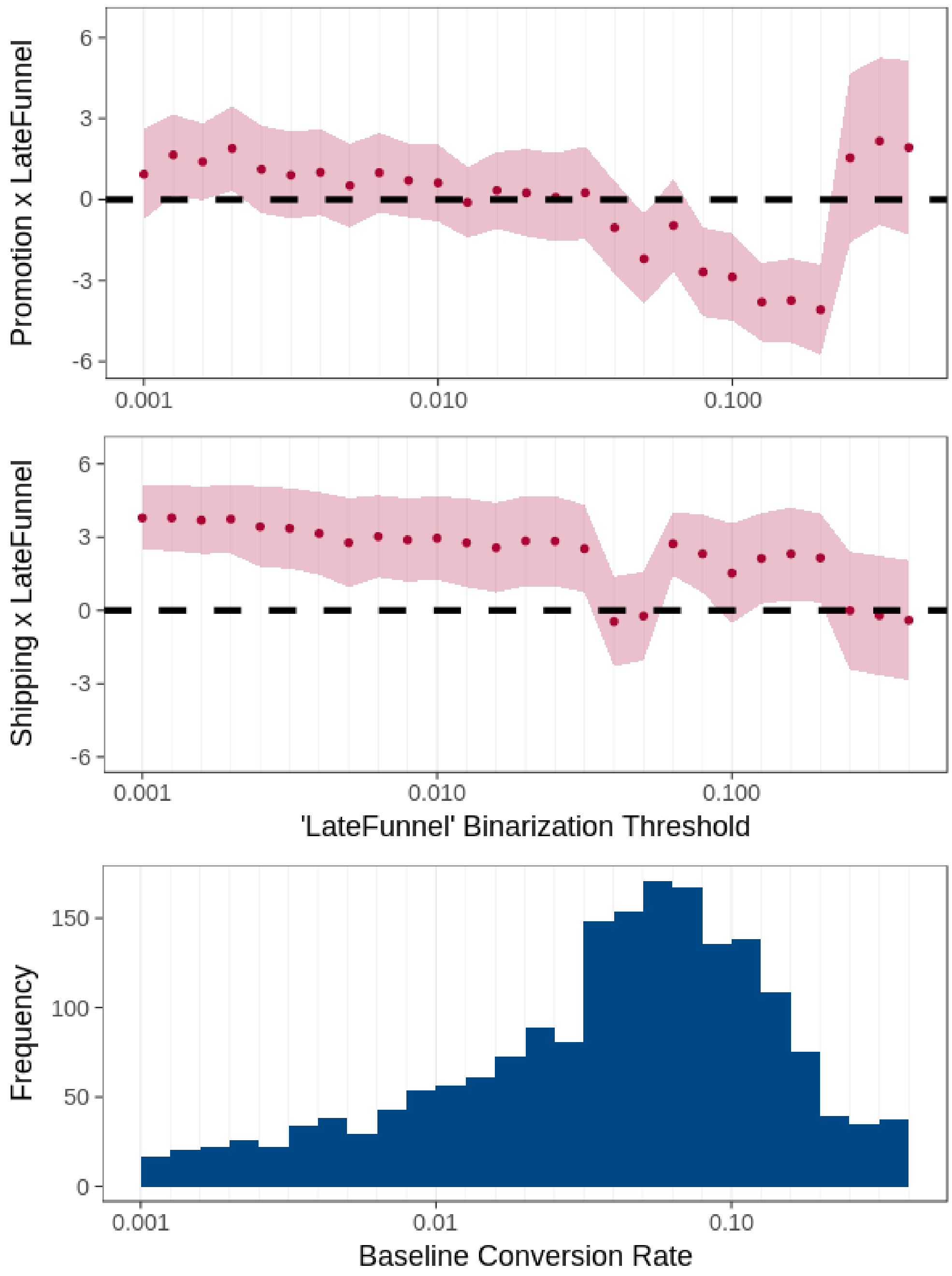
We can then run a similar regression as in Equation (3), except rather than interacting the intervention types with a categorical location variable, we will use the binary LateFunnel variable defined here:

$$y_{ij} = \text{Controls}_{ij}\boldsymbol{\gamma} + \text{Type}_{ij}\boldsymbol{\theta} + \text{LateFunnel}_{ij}^\tau + \left(\text{Type}_{ij} \times \text{LateFunnel}_{ij}^\tau\right)\boldsymbol{\beta}^\tau + \varepsilon_{ij} \qquad (4)$$

Note the conversion rate threshold τ (at which we consider an experiment to go from being "early" or "late" funnel) is entirely arbitrary. Rather than choosing any particular value of τ, we will calculate the regression coefficients in Equation (4) for many levels of τ across the support of the distribution of conversion rates in our dataset.

Because we are primarily interested in how intervention effectiveness varies throughout the conversion funnel, we calculate the main interaction effects between $\text{LateFunnel}_{ij}^\tau$ and Promotion_{ij} and Shipping_{ij} variables (*design* interventions are our baseline intervention type). The top two panels in Figure 6 have both point estimates and (heteroskedastic robust) 95% confidence intervals for this coefficient (y-axis) plotted for different specifications of the binarization threshold τ (x-axis). The bottom panel contains a histogram of baseline conversion rates, which aids in visualizing how many experiments fall above or below a given binarization threshold level. Finally, because we are not using keyword filtering to define our funnel variable in this specification, the analyses below are conducted on the sample of 2,201 experiments in our dataset that have been coded for intervention type. As we turn to the results, recall that since *design* experiments serve as our baseline class, the interaction effects reported here should be interpreted as how conversion rates among *promotion* and *shipping* experiments vary across the conversion funnel *relative* to *design* interventions.

Figure 6: Robustness Test for Alternative Funnel Specification



6

Looking at Figure 6, we see in the “Promotion × LateFunnel” panel that for experiments very early in the conversion funnel (near the 0.001 baseline conversion rate), the coefficient on this interaction may be positive, indicating that effect sizes are larger later in the funnel. However, these results rely on a small number of experiments with very small conversion rates below the binarization threshold. Closer to the central mass of the conversion rate distribution (near 0.08), the coefficients (both point estimates and confidence bands) on the "Promotion × LateFunnel" variable are negative. At the tail end of the distribution (the last three points on the right), the confidence intervals become much larger and the effects become indistinguishable from zero. To summarize, the results of this analysis show mostly null or negative effects, which is largely consistent with the results reported in Section 4.5, where we found promotional experiments become less effective later in the conversion funnel.

Turning to the "Shipping × LateFunnel" panel in Figure 6, we see that the coefficients on this variable are almost always positive and significantly different from zero. There are some regions where the point estimates appear to be closer to zero, but—on the whole—this analysis also appears to be consistent with our previous findings: that shipping interventions are more effective later in the conversion funnel.

In sum, the results of this analysis—using a different operationalization of funnel depth and a larger set of experiments—are largely consistent with the evidence presented in columns 3 of Table 4. Relative to design interventions, shipping interventions are typically more effective later in the conversion funnel. And across both specifications, we see evidence that promotions are less effective at some later stages of the conversion funnel. That fact that these findings are directionally consistent in both of our interaction analyses—using text-based operationalization or conversion rate operationalization of funnel stage—provides stronger evidence of our conclusions than either analysis alone.

6 Conclusion

The goals of this project have been to provide insight into the content of real-world e-commerce experimentation. We have investigated the types of experiments firms run and provided a robust analysis of how varying intervention types vary throughout different phase of the e-commerce conversion funnel.

The results described in this project provide both meaningful managerial insight and contribute to the broader literature on the marketing conversion funnel. We have shown that, consistent with prior work, the distribution of effect sizes in digital experimentation is extremely skew. To help firms find those interventions with larger effect sizes, we then performed a meta-analysis across the experiments in our sample. This revealed that the largest effect sizes (in absolute terms) for the average experiment in our sample are achieved by focusing on promotional and shipping-related interventions. Naturally, profit-maximizing firms will need to weigh the costs of price and shipping discounts against the changes in conversion rates associated with these interventions. Nonetheless, there are times when maximizing conversion rates specifically can be of strategic value to firms; in these cases, this work provides evidence-based guidance on the most effective ways for influencing customer conversions.

Furthermore, even without altering the *types* of interventions firms experiment with (i.e., by increasing or decreasing the number of promotions they test), our research provides evidence on where firms can best target those interventions in their website architecture to maximize their impact. This is because we were able to use two independent operationalizations to analyze the effectiveness of various promotional interventions throughout the typical e-commerce conversion funnel. In particular, results of both our analyses suggest that price promotions are best advertised early in the website conversion funnel, whereas shipping-related interventions are best targeted towards customers that are in the later stages of the buying process. Lastly, because we have arrived at these insights through an aggregate meta-analysis of many different websites, we can make reasonable claims about the generalizability of our findings among the population of e-commerce companies. We believe this research provides a unique insight to the factors that influence online shopping behavior in a quite general way across the on-site conversion funnel.

These findings and the framework laid out in this project provide several promising avenues for future lines of both industrial and academic inquiry. For one, our results underscore the importance of testing not just the right types of interventions in e-commerce experiments, but also making sure those interventions are targeted at the right position in the marketing funnel. Even in cases where our results are not immediately generalizable (e.g., for SaaS or media companies), this framework provides a useful template to individual firms for optimizing online conversions along multiple dimensions. Determining whether our findings do generalize to these other industries would be a valuable direction for future research. Furthermore, our research can provide a starting point for targeted lab studies or field experiments to provide more specificity about the interaction between intervention types and locations. In addition to expanding the scope of our top-level findings, natural hypotheses to test in follow-up work are whether different sub-types of design interventions (e.g., generic banners vs. pop-ups) and different magnitudes of promotions (e.g., 10% vs. 50%) affect consumers heterogeneously across the conversion funnel.

Though there is room for future work in this area, this research represents an important contribution to our understanding of the increasingly important practice of A/B testing. By determining which types of experiments firms are running on their websites and quantifying the relative effect sizes of these experiments, we have provided meaningful managerial and theoretical insight into how firms use and can optimize their use of digital experiments. Lastly, we reiterate how this project represents a shift in perspective among how researchers can think about the topic of digital experimentation. Along with a small but growing body of literature on this topic, we advocate that business researchers use modern A/B testing technologies to not only test their own hypotheses, but also to investigate how companies themselves use these technologies. We hope future research can build on this project to further our understanding of the practice of digital experimentation in real-world business environments.

References

Vibhanshu Abhishek, Peter Fader, and Kartik Hosanagar. 2012. Media exposure through the funnel: A model of multi-stage attribution. (2012).

Ritu Agarwal and Viswanath Venkatesh. 2002. Assessing a firm's web presence: a heuristic evaluation procedure for the measurement of usability. *Information Systems Research* 13, 2 (2002), 168–186.

Eduardo M Azevedo, Deng Alex, Jose Montiel Olea, Justin M Rao, and E Glen Weyl. 2018. A/B testing. (2018).

Ron Berman, Leonid Pekelis, Aisling Scott, and Christophe Van den Bulte. 2018. p-Hacking and False Discovery in A/B Testing. *Available at SSRN: https://ssrn.com/abstract=3204791* (June 2018).

Michael Braun and Wendy W Moe. 2013. Online display advertising: Modeling the effects of multiple creatives and individual impression histories. *Marketing Science* 32, 5 (2013), 753–767.

Matias D Cattaneo, Michael Jansson, and Xinwei Ma. 2018a. Manipulation testing based on density discontinuity. *The Stata Journal* 18, 1 (2018), 234–261.

Matias D Cattaneo, Michael Jansson, and Xinwei Ma. 2018b. Simple local polynomial density estimators. *arXiv preprint arXiv:1811.11512* (2018).

Patrali Chatterjee. 2011. Framing online promotions: Shipping price inflation and deal value perceptions. *Journal of Product & Brand Management* 20, 1 (2011), 65–74.

David Court, Dave Elzinga, Susan Mulder, and Ole Jørgen Vetvik. 2009. The consumer decision journey. *McKinsey Quarterly* (June 2009).

Andrea Everard and Dennis F Galletta. 2005. How presentation flaws affect perceived site quality, trust, and intention to purchase from an online store. *Journal of Management Information Systems* 22, 3 (2005), 56–95.

Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*. ACM, 1–15.

Dennis F Galletta, Raymond M Henry, Scott McCoy, and Peter Polak. 2006. When the wait isnt so bad: The interacting effects of website delay, familiarity, and breadth. *Information Systems Research* 17, 1 (2006), 20–37.

Andrew Gelman and Hal Stern. 2006. The difference between significant and not significant is not itself statistically significant. *The American Statistician* 60, 4 (2006), 328–331.

Rebecca W Hamilton and Joydeep Srivastava. 2008. When 2+ 2 is not the same as 1+ 3: Variations in price sensitivity across components of partitioned prices. *Journal of Marketing Research* 45, 4 (2008), 450–461.

Joachim Hartung, Guido Knapp, and Bimal K Sinha. 2011. *Statistical meta-analysis with applications*. Vol. 738. John Wiley & Sons.

Angela V Hausman and Jeffrey Sam Siekpe. 2009. The effect of web interface features on consumer online purchase intentions. *Journal of Business Research* 62, 1 (2009), 5–13.

Bernard J Jansen and Simone Schuster. 2011. Bidding on the buying funnel for sponsored search and keyword advertising. *Journal of Electronic Commerce Research* 12, 1 (2011), 1.

Ramesh Johari, Leo Pekelis, and David J Walsh. 2015. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922* (2015).

Garrett Johnson, Randall A Lewis, and Elmar Nubbemeyer. 2017. The Online Display Ad Effectiveness Funnel & Carryover: Lessons from 432 Field Experiments. (2017).

Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1168–1176.

Robert J Lavidge and Gary A Steiner. 1961. A model for predictive measurements of advertising effectiveness. *The Journal of Marketing* (1961), 59–62.

Philip Leifeld. 2013. texreg: Conversion of Statistical Model Output in R to LaTeX and HTML Tables. *Journal of Statistical Software* 55, 8 (2013), 1–24. http://www.jstatsoft.org/v55/i08/

Michael Lewis. 2006. The effect of shipping fees on customer acquisition, customer retention, and purchase quantities. *Journal of Retailing* 82, 1 (2006), 13–23.

Michael Lewis, Vishal Singh, and Scott Fay. 2006. An empirical study of the impact of nonlinear shipping and handling fees on purchase incidence and expenditure decisions. *Marketing Science* 25, 1 (2006), 51–64.

CH Liu and Benjamin Paul Chamberlain. 2018. Online Controlled Experiments for Personalised e-Commerce

Strategies: Design, Challenges, and Pitfalls. *arXiv preprint arXiv:1803.06258* (2018).

Justin McCrary. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics* 142, 2 (2008), 698–714.

Wendy W Moe. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology* 13, 1-2 (2003), 29–39.

Oded Netzer, James M Lattin, and Vikram Srinivasan. 2008. A hidden Markov model of customer relationship dynamics. *Marketing science* 27, 2 (2008), 185–204.

Alexander Peysakhovich and Dean Eckles. 2017. Learning causal effects from many randomized experiments using regularized instrumental variables. *arXiv preprint arXiv:1701.01140* (2017).

Sven Schmit, Virag Shah, and Ramesh Johari. 2018. Optimal Testing in the Experiment-rich Regime. *arXiv preprint arXiv:1805.11754* (2018).

Stephan Seiler and Song Yao. 2017. The impact of advertising along the conversion funnel. *Quantitative Marketing and Economics* 15, 3 (2017), 241–278.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22, 11 (2011), 1359–1366.

Jaeki Song and Fatemeh Mariam Zahedi. 2005. A theoretical approach to web design in e-commerce: a belief reinforcement model. *Management Science* 51, 8 (2005), 1219–1235.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. 2016. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics* 34, 4 (2016), 661–672.

Viswanath Venkatesh and Ritu Agarwal. 2006. Turning visitors into customers: a usability-centric perspective on purchase behavior in electronic channels. *Management Science* 52, 3 (2006), 367–382.

Stefan Wager and Susan Athey. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* just-accepted (2017).

Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2227–2236.

Jie Zhang and Lakshman Krishnamurthi. 2004. Customizing promotions in online stores. *Marketing science* 23, 4 (2004), 561–578.

Jie Zhang and Michel Wedel. 2009. The effectiveness of customized promotions in online and offline stores. *Journal of marketing research* 46, 2 (2009), 190–206.

Shunyuan Zhang, Dokyun Lee, Param Vir Singh, and Kannan Srinivasan. 2016. How much is an image worth? An empirical analysis of propertys image aesthetic quality on demand at AirBNB. (2016).