Cross Channel Effects of Search Engine Advertising on Brick & Mortar Retail Sales: Meta Analysis of Large Scale Field Experiments on Google.com

Kirthi Kalyanam
Santa Clara University

John McAteer
Google Inc.

Jonathan Marek
Applied Predictive Technologies

James Hodges & Lifeng Lin
Division of Biostatistics, University of Minnesota *

This Version: September 2015

Abstract

We investigate the cross channel effects of search engine advertising on Google.com on sales in brick and mortar retail stores. Obtaining causal and actionable estimates in this context is challenging: Brick and mortar store sales vary widely on a weekly basis; offline media dominate the marketing budget; search advertising and demand are contemporaneously correlated; and estimates have to be credible to overcome agency issues between the online and offline marketing groups. We report on the meta-analysis of a population of 15 independent field experiments, in which 13 well-known U.S. multi-channel retailers spent over $4 Million in incremental search advertising. In test markets broad keywords were maintained in positions 1-3 for 76 product categories with no search advertising on

*Corresponding author: kkalyanam@scu.edu. Kirthi Kalyanam is the J.C.Penney Research Professor and Director of the Retail Management Institute at Santa Clara University. John McAteer is Vice President, US Sales at Google.com. Jonathan Marek is Senior Vice President at Applied Predictive Technologies. James Hodges and Lifeng Lin are respectively Professor and a graduate student in the Division of Biostatistics at the University of Minnesota. We are grateful to the retailers and Google for participating in the field experiments and for sharing the data for the analysis, irrespective of the outcomes; to APT for implementing the field experiments; and to Google teams for data collection and project management support of our analysis. We also thank Sridhar Narayanan, Navdeep Sahni, Don Lehmann, Gary Lilien, Wes Hartmann, Harikesh Nair, Arvind Rangaswamy, Chris Nosko, Gunter Hitsch, Randall Lewis, Garrett Johnson, Puneet Manchanda, Stephan Seiler, Carl Mela, Gerard Tellis, Ken Wilbur, Sonika Singh, Hal Varian, Eric Anderson, Florian Zettlemeyer, Joe Golden, participants at the 2014 Marketing Dynamics Conference, seminar participants at Penn State University, the 2015 FORMS conference at UT-Dallas, the 2015 SICS conference, executives from Wal-Mart's eCommerce Division, The Clorox Company, Tesco Inc, Mattel Inc, Samsung Brazil and Waitrose (UK) for their comments. Kirthi Kalyanam, James Hodges and Lifeng Lin thank Google for research support. All remaining errors are our own.

these keywords in the control markets. We estimate an average effect of each outcome for this population of experiments using a Hierarchical Bayesian (HB) model. The estimate from the HB model provides causal evidence that increasing search engine advertising on broad keywords on Google.com had a positive effect on sales in brick and mortar stores for the *advertised categories* for this population of retailers. There also was a positive effect on total store sales. Hence the increase in sales in the advertised categories was incremental to the retailer net of any borrowed sales from non-advertised categories. The total store sales increase was a meaningful improvement compared to the baseline sales growth rates. The posterior density for Return on Ad Spend (ROAS) shows that several retailers achieved or exceeded break-even based only on brick and mortar sales . Counterfactually, we estimate that if this population of retailers were to extend the test to their entire network of stores the incremental sales opportunity from increased search advertising would be $1.94 Billion. We also examine the heterogeniety of the effects across retail formats and find that specialized retailers get higher sales lift but ROAS on brick and mortar sales favors the generalist. We examine the robustness of our findings to alternative assumptions about the data specific to this set of experiments. Our estimates suggest that media planners should account for the offline effects in the planning and execution of search advertising campaigns and that these effects should be adjusted by category and retailer type.

1 Introduction

Search engine advertising, which refers to paid listings on search engines such as Google, Bing, and Yahoo, is a relatively new but important and growing part of the advertising market. Figure 1 shows an example of the results of a search for the phrase "toto toilets" on Google. The sponsored ads on the top and on the right side of the page are examples of search engine advertising. The ads on the top right of the page (top of the "right rail"[1]) are from Home Depot, a multi-channel home improvement retailer. Home Depot sells online via homedepot.com and *offline* through a national network of brick and mortar stores. The consumer can click on the search ad from Home Depot, browse the web site, but purchase the product in a brick and mortar store. The focus of this paper is on this cross channel impact of search engine advertising on offline sales in brick and mortar stores.

The *offline* impact is important for practical and substantive reasons. First and foremost, the vast majority of retail sales occur offline. For example according to the U.S. Census[2] in 2012 eCommerce sales accounted for only 5.2% of total retail sales. However it is believed that the influence of the web on retail shoppers is disproportionate to its share of sales. Forrester Research[3] reports that more than 50% of U.S. offline retail sales will be influenced by the web by 2017. This suggests that shoppers search and shop online even when they plan to buy offline. Some recent reports suggest that the trend of researching online and buying in the store has accelerated and is impacting store traffic patterns.[4] These reports note that since it is easy to search online, comparison shopping online is an efficient way to pre-shop before visiting a brick and mortar store.

Brick and mortar retailers can advertise to these cross channel shoppers on search engines. However, although search engine advertising is known to be a direct response medium that gets consumers to buy online, it's offline effects are less clear. Brick and mortar retailers spend heavily on offline media such as television and weekly free-standing inserts in newspapers (FSI or circulars) and these offline media might generate awareness and search queries related to

[1] For a definition of "right rail" see http://www.sempo.org/?page=glossary&hhSearchTerms=%22right+and+rail%22

[2] http://www2.census.gov/retail/releases/historical/ecomm/12q4.pdf

[3] http://www.forrester.com/Forrester+Research+Online+Retail+Forecast+2012+To+2017+US/fulltext/-/E-RES90661

[4] According to data reported by Shopper Trak, foot traffic to retailers over the holiday season dropped by over 50% over the three holiday seasons spanning 2011 to 2013 (Banjo and FitzGerald [2014]).

the category or the retailer.[5] For example Joo et al. [2013] show that television advertising increases searches on Google.com for both category and branded[6] search queries for financial services firms. If television advertising has already persuaded the consumer about purchasing from a particular retailer, then the search engine advertising might not have an incremental effect on offline sales. Another reason why search engine advertising might not be incremental is that consumers may already know or have prior shopping experiences with well known brick and mortar retailers.[7] In such contexts it is conceivable that offline shoppers might type a query into a search engine, click on a search ad to simply *navigate* to the retailer's web site and pre-shop before a trip to the brick and mortar store. In this scenario, the shopper had a prior propensity to shop with this retailer and the impact of search engine advertising might not be incremental. Search advertising is delivered when consumers search. Hence advertising is contemparaneously correlated with demand (Berndt [1991]) and organic links might be a substitute for search ads (Blake et al. [2013]). Retail sales can very considerably from week to week making it difficult to establish causality. Figure 2 provides an illustrative example. Even if search advertising increases sales in the advertised categories, the incremental sales might be borrowed from other categories or from the future. So the incremental net impact on total store sales is an important consideration for retailers. Finally, online and offline marketing are typically managed by different teams and estimates have to be credible to overcome agency issues between these teams.

An alternative viewpoint is that offline shoppers might use a search engine to search for information (Ratchford et al. [2003]). Even if there was prior exposure to say television advertising, if this advertising provided objective knowledge, it can increase consumers' ability to obtain additional information (Brucks [1985]) and hence trigger search. Exposure to advertising can also influence the formation of consideration sets (Shapiro et al. [1997], Sahni [2013]), or

[5]For example Table 1 provides information on the media spending of each retailer in our data set. The media spending is rank ordered from highest spend to lowest spend for each retailer. As this table shows the retailers in our experiment spent heavily on newspapers, TV, free standing inserts, radio and direct mail.

[6]For example, "Savings account" is a broad search term and is also referred to as top of the funnel search term or a generic search term. "Citi Bank savings account" is a branded search term since it includes the brand name of the retailer.

[7]For an illustrative example, consider a well known retailer–Walmart. According to Investor FAQ's on Walmart's web site Walmart serves more than 245 million customers and members weekly worldwide (http://stock.walmart.com/faqs/, accessed on 1/27/2015). According to Wal-Mart's web site "Every month more than 60 percent of Americans shop at Walmart", (http://news.walmart.com/news-archive/2013/05/04/walmart-launches-national-advertising-campaign-to-show-the-real-walmart, accessed on 1/27/2015).

motivate comparison shopping (Zettelmeyer et al. [2006]). With respect to comparison shopping, media coverage of holiday shopping behavior suggest that nowadays shoppers are less likely to do comparisons in person "bouncing from store to store because they've made their decisions ahead of time [online]" (FitzGerald [2013]). Under this scenario, search advertising might influence shoppers as they comparison shop online and generate incremental offline sales even for well-known retailers.

This discussion raises the following first order questions: (1) Is there a causal effect of search engine advertising of a category on *offline* sales of that category at brick and mortar retail stores? (2) Is the effect incremental to the store, net of borrowing from other categories and future sales? (3) What is the return on ad spend (ROAS)? (4) How generalizable is the effect? (5) Do outcomes differ between retailers that are more specialized in a few categories (specialists) versus retailers that carry a broad assortment (generalists)? To the best of our knowledge we do not have causal evidence regarding these first order questions about the cross channel effects of this new form of advertising.

This paper obtains causal estimates of the cross-channel $effect$ of search engine advertising on brick and mortar sales using field experiments. We report on the results from 15 *independent* field experiments representing 76 categories. The experiments were conducted for 13 well-known U.S. multi-channel retailers. We used the field experimentation software provided by Applied Predictive Technologies (APT) and this allowed us to automate the execution of this population of experiments.[8] The experiments represent a variety of categories including apparel, baby products, electronics, toys, cosmetics, sporting goods, furniture, pet food, and home improvement. Collectively these retailers represent $236.39 Billion in annual sales. In each experiment, the retailer increased search advertising spending ("heavied up"), so that broad keywords appeared in positions 1 to 3, for a randomly selected set of test markets that were representative of the entire store network of the retailer. In the control markets these keywords were not advertised. There was no change in policy in terms of search advertising spending on any other keyword in the test or the control markets. Incremental effects were estimated for each test store relative to a matched set of stores in the control markets. The inference is based on store level estimates as opposed to market level or individual level estimates. For each experiment, based

[8] http://www.predictivetechnologies.com/

on a data sharing agreement between Google, the retailer and Applied Predictive Technologies, *only* the estimates of incremental effects,[9] and p-values or more commonly p-value intervals for these effects were released to the researchers. The data released also includes estimates of incremental sales to incremental advertising (return on ad spend or ROAS), a measure of return on investment.

We estimate the overall average $effect$ across field experiments for all outcomes by adapting Hierarchical Bayesian (HB) random effect models used in evidence-based medicine (Babapulle et al. [2004], Berry et al. [2003], DerSimonian and Kacker [2007], Sutton and Higgins [2008]), which allow for heterogeneity in treatment effects both within and across retailers and categories. We find causal evidence that an increase in search engine advertising at the category level incrementally increased brick and mortar retail sales in the advertised categories. The estimate of overall average incremental sales increase in the advertised categories is 1.27%. The posterior density for the overall estimate of sales increase has very little mass below zero. In the HB analysis all but two of the categories showed positive sales increases. We also find that an increase in online search advertising incrementally increased total store sales[10] for these well-known retailers. Investing in search ads, even at a modest experimental level, increased total brick & mortar store sales by an average of 1.18% in these experiments. Since the U.S. retail industry grew at a compound annual rate of 1.55%[11] during the duration of the experiments, this sales increase, which is 77.5% of the base growth rate, is very meaningful. The sales increase for the total store indicates that the increase in the sales of advertised categories was net of any inter-category substitution. All estimates incorporate a two week post experimental period to measure the impact on future sales. We find that all experiments showed estimated sales increases and ten of the 14 experiments had the entire posterior 95% interval for sales increase above zero. Counterfactually, we estimate that for this population of retailers the total store sales increase projected to their total retail network would generate $1.9 billion in incremental annual sales due to increased search advertising.

Our data set does not contain the standard error or p-value of ROAS, and we develop a

[9] In some experiments total store refers to total sales of all the categories in the store, in others it refers to a top level aggregation of categories that include the advertised categories.

[10] Defined as the total store sales or total sale for an aggregation of advertised and related non-advertised categories.

[11] Calculated based on data reported by the National Retail Federation.

method to estimate it. Our method is likely to be useful in other contexts where advertising spending is not available at the individual store level. We incorporate the ROAS estimates and standard errors in the HB analysis. The posterior distribution of ROAS (excluding online sales) has a mean of 2.50, which is significantly different from zero and has very little mass that is less than zero. The posterior distribution shows that several retailers achieved break-even. Directionally, we find that the incremental sales are higher for specialized retailers who have narrow assortments and appeal to a narrow audience whereas ROAS favors the generalist retailer who has a broad assortment and appeals to a broader targer audience. We examine the robustness of our findings to alternative assumptions about the p-values of the estimates.

This paper makes several contributions. First, it makes causal inferences on offline effects of search advertising, a new form of advertising that is growing rapidly, whereas the prior literature has examined the online effects (Narayanan and Kalyanam [2015], Blake et al. [2013], Rutz and Bucklin [2011], Yang and Ghose [2010], Agarwal et al. [2008], Varian [2007]). Second, it provides causal estimates of the cross channel effects of advertising per se, an area of growing importance, where causal estimates have been hard to find in the literature (Joo et al. [2013], Lewis and Reiley [2009], Naik and Peters [2009]). Third, although it focuses only on the offline impact, it provides evidence regarding the profitability of advertising. Fourth, it examines whether the effects are generalizable using a population of field experiments that include a variety of advertisers. Very few papers in the literature report on more than one field experiment (Eastlack Jr and Rao [1989], Lodish et al. [1995]). Fifth, it examines whether these effects vary across specialists and generalists. While there is a robust literature focusing on the differences in the performance of specialists versus generalists (Carroll [1985], Swaminathan [1995, 2001]), differences in advertising have not received attention. Sixth, this paper demonstrates the use of Hierarchical Bayesian models to obtain estimates for a population of field experiments, accounting for both within- and between-experiment variation, an approach that is quite common in biostatistics but is less prevalent in marketing in the analysis of estimates obtained from multiple studies (Sethuraman et al. [2011], Farley et al. [1995], Tellis [1988], Farley and Lehmann [1986]). The modeling framework also enables us to combine different types of information such as p-value intervals and p-values when experiments differ in the information they report. Seventh, this paper demonstrates an approach to obtaining standard errors for ROAS in field experiments involving

retail stores when limited information is available. Lastly, while this paper does not directly investigate the mechanisms underlying these effects, it indirectly contributes to the literature on online information search and cross channel shopping (Zettelmeyer et al. [2006], Ratchford et al. [2003]), by investigating whether consumers are open to advertising influences as they traverse channels.

The next section provides background on search advertising, the relevant literature and motivation for key research questions. Section 3 describes the participating retailers and their motivations. Section 4 describes the design of the field experiments. Section 5 presents the results from one field experiment to ease our understanding of the results of the next section. Section 6 presents the Hierarchical Bayesian meta-analysis followed by results including robustness analysis. Section 7 presents conclusions, limitations, and opportunities for further research. Section 8 is an appendix that presents our derivations.

2 Background

2.1 Search Engine Advertising

Search advertising is a large and growing market. According to the Internet Advertising Bureau, $9.1 billion was spent in the United States alone on search advertising in the first half of 2014. Search advertising, with 39% of all online advertising revenues in the first half of 2014, is the largest component of the online advertising market.[12] Although it is a relatively new medium for advertising, search is the third-largest medium after TV and print, and surpassed radio in 2012.[13]

Several features of search advertising have made it a popular online advertising format. Search ads are triggered by specific keywords (search phrases). Presumably shoppers search when they are ready to make a purchase and search advertising seems well positioned to produce an immediate response (direct response). Search advertising has important targeting capabilities (Narayanan and Kalyanam [2015]). For example consider an advertiser who is selling health insurance for families. Some of the search phrases related to health insurance could include

[12] http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_HY_2014_PDF.pdf

[13] http://www.iab.net/media/file/IABInternetAdvertisingRevenueReportFY2012POSTED.pdf (last accessed on October 31, 2013).

"health insurance" or "family health insurance". An advertiser can target families by advertising on the phrase "family health insurance". Further, these ads can be targeted by geography with different ads showing in different locations. Search ads are sold on a per click basis. Advertisers can match the clicks on the ad to online outcomes such as visits, conversions, and sales using cookies.

2.2 Search Engine Advertising and Online Outcomes

Prior research has investigated the impact of search advertising on *online* outcomes. Ghose and Yang [2009] and Agarwal et al. [2008] showed that the top positions are not necessarily the most profitable positions. Yang and Ghose [2010] examined whether search advertising and organic links are complements or substitutes. Rutz and Bucklin [2011] investigated whether advertising on category level keywords creates spillovers by generating incremental subsequent searches for the brand name. Using a large scale field experiment, Blake et al. [2013] find that advertising on branded keywords did not incrementally impact sales for eBay. Narayanan and Kalyanam [2015] obtain causal estimates of position effects on online outcomes using a Regression Discontinuity approach. They showed that selection effects vary by position, that position effects are stronger for advertisers who are not well known and for broad keywords. Sahni [2011] obtained causal estimates of the effect of repetition of ads on a restaurant search engine by randomizing ads at the individual level. His results showed benefits to spacing of ads. In a follow up paper, Sahni [2013] showed that advertising effects spill over to non-advertised restaurants.

This literature has not investigated the impact of search engine advertising on brick and mortar sales. However, several key insights from the literature are relevant for our study. First, the literature shows that position effects do exist and that the top positions generate more clicks. Second, the literature shows differences between branded and non-branded keywords. In particular branded keywords might be used by consumers who are already aware of the advertiser and in the absence of the ad might click on an organic link. In this study we focus on the impact of broad (non branded) keywords since it is less likely that these keywords are used by searchers who have a prior propensity for the advertiser. Further, since broad keywords have position effects, in our experiment we maintain these keywords in positions 1-3 to increase the number of clicks that the advertiser can obtain.

2.3 Cross Channel Advertising Effects

An emerging literature focuses on the impact of advertising in digital media on outcomes in non-digital media and vice versa. This literature is also relevant to the present study. This literature provides opposing points of view for our investigation. For example, Fulgoni and Morn [2009] showed that a significant fraction of consumers who clicked on a search ad purchased offline, suggesting offline effects of search advertising. Joo et al. [2013] find that television advertising increases broad and branded search queries in the financial services category. To the extent that the television advertising is persuasive, consumers might simply be clicking on a search ad to navigate to the advertiser and the impact of the search ad might not be incremental. However if the television advertising increases objective knowledge then it might make consumers seek additional information via search (Brucks [1985]) and the impact of search advertising might be incremental. Lewis and Reiley [2009] examine the impact of *display* advertising on sales in brick and mortar stores and highlight the difficulty in obtaining estimates of its offline impact. The implication is that there are numerous influences on brick and mortar sales and proper controls are required to get credible estimates.

2.4 Cross Channel Shopping

Research on cross-channel shopping from both industry and academia is relevant to our study. As mentioned in Section 1, Forrester Research reports that over 50% of U.S. retail sales are web influenced. Web influence is a broad construct that provides retailers with some directional evidence of the importance of web influenced offline shopping. However, it does not provide specific causal evidence about the cross channel effectiveness of a specific medium such as search advertising. Recent industry reports suggest that the trend of researching online and buying in the store is accelerating (Banjo and FitzGerald [2014]). These reports note that since shopping costs are lower online, shopping online is an efficient way to pre-shop before visiting a brick and mortar store.

According to Zettelmeyer et al. [2006] prior to the Internet, automotive shoppers tended to visit 4 dealers whereas after the Internet they tend to visit 1 dealer. Ratchford et al. [2003] investigate how consumers use the Internet to search for information in the context of shopping for automobiles. Verhoef et al. [2007] investigate the tendency of shoppers to obtain informa-

tion in one channel and buy in another channel, the so-called 'research shopper' phenomenon. Collectively these studies suggest that even if a consumer intends to buy in a brick and mortar store, it is efficient to research online, form consideration sets, narrow the choices and then finish the shopping at a brick and mortar store. However these studies do not establish whether there is an opportunity to persuade the consumer who is in this cross channel shopping journey with advertising on search engines. Our study adds to this literature by investigating the advertising opportunity.

2.5 Specialists versus Generalists

Some retailers are generalists and offer a broad range of product categories whereas others specialize in a narrower set of categories. For example Costco offers low prices on a broad assortment spanning many categories but does not offer depth of selection in each category. A retailer such as The Gap is more *specialized* on a narrower set of categories (Levy et al. [1998]). A robust literature (Carroll [1985], Swaminathan [1995, 2001]) has focused on the differences between specialists and generalists. In particular the literature notes that generalists pursue economies of scale and occupy the center of the consumer preference distribution. Since specialists focus on a narrower set of categories and perhaps a narrower set of customer needs they tend to provide a better offering in these categories compared to generalists (Hannan and Freeman [1977]). For example a specialized retailer such as The Gap vertically integrates to develop its own proprietary line of apparel that are edited to provide a unified look. The specialist/generalist dichotomy has implications for Internet retailing and search advertising.

An important aspect of Internet retailing is the technology platform. So technology constraints and tradeoffs shape the features of Internet retailing. The technology platform of a generalist has to provide features for a broad range of categories. This typically forces a tradeoff towards a web site design that provides a set of basic features common to a large set of categories. In contrast, the specialist, due to a focus on a narrower set of categories, can provide a richer feature set that serves these categories and hence provides a better retailing experience for these categories compared to a generalist. This suggests that the web site of a specialist might be able to convert traffic better compared to a generalist and it is plausible that these effects carry over to offline sales. In conclusion, in an Internet retailing, setting a narrower focus favors

the specialist both in terms of superior merchandise and the presentation of this merchandise.

This study focuses on broad keywords and this has implications for the specialist/generalist dichotomy. Broad keywords might be used by searchers who are new to the category and have more basic needs (Narayanan and Kalyanam [2015]). This traffic might be a better match for the assortment offered by a generalist and hence lead to higher sales for a generalist. Broader awareness might also favor a generalist in terms of the pricing mechanism of search advertising. Since the generalist sells a broader range of categories compared to a specialist it is more likely that consumers are more aware of the generalist because they have purchased in *some* category from the generalist. This might result in higher awareness and hence a higher click thru rate on search advertising on broad terms. In Google's auction mechanism, a retailer that has a higher click thru rate pays a lower cost per click. The implication is that even though the clicks for a specialized retailer might convert better due to better products, better merchandising, and better online presentation, the generalist might obtain a lower cost per click due to higher brand recognition. If the lower cost per click overcomes the lower conversion rate, the generalist will get a better ROAS compared to a specialist. Our population of field experiments include some retailers who are specialists and some who are generalists, allowing us to directionally investigate these potentially divergent effects on sales and ROAS.

3 Participating Retailers

Table 1 provides information about the multi-channel retailers who participated in the field experiments and provided data for this study. The retailer names are disguised due to disclosure policies. However, these retailers are very well known and are considered "household names" in the U.S. retail market, a virtual "who's who" of retailing.[14] Thirteen participating retailers conducted fifteen field experiments. Study A and Study E were conducted by the same retailer; Study D and Study H were conducted by the same retailer. The second column in the table provides information on the advertising repertoire of each retailer. Television had the highest share of budget for 9 retailers. Our retailers also advertised in newspapers and magazines. Retailer A, E and J emphasized direct marketing. Almost all of our retailers advertised on the

[14]Due to the disclosure policy we do not provide descriptives such as annual sales, type of retail format (e.g., department store versus specialty store), focal categories, or number of stores.

Internet, but Internet advertising was focused on driving online sales as opposed to offline sales.

The third column in Table 1 summarizes what motivated the retailers to conduct field experiments. A general theme that emerges from the table is that the retailers had indirect measures of the impact of search advertising on offline retail sales but needed direct measures to inform spending decisions. For example, the retailer in experiment K conducted exit interviews of visitors on their web site and found that several customers indicated that they visited a brick and mortar store after the web site visit. The retailer could try to trace back the source of this web traffic to a search ad and link it to the offline sale. These types of trace backs do not produce causal estimates because there is no counterfactual or control to measure incrementality. Some retailers were concerned with declining newspaper circulation and the potential decline in the efficacy of spending on newspaper inserts. They were interested in new forms of advertising that could provide an alternative to drive in store traffic. One of the retailers was interested in shaping new co-operative advertising programs with manufacturers for search engine advertising. Retailers who spent heavily on television and newspapers simply wondered if any of the traffic driven by search advertising was incremental to the awareness generated by television and newspapers. Promotional retailers who executed several marketing instruments such as coupons or loyalty card offers were interested in isolating the impact of search advertising to determine the incremental effects. Senior executives such as the Chief Financial Officers at our retailers were accustomed to using field experiments to obtain causal estimates to inform significant capital allocation decisions.

4 Field Experiments: Design, Execution and Measures

4.1 Design and Execution

Each field experiment was designed and executed independently of the other experiments. Each experiment was a "heavy up" experiment in which search advertising was increased in test markets to maintain broad keywords in positions 1-3. These keywords were not advertised in control markets. All other aspects of the search advertising policy were unchanged in the test and control markets. The field experiments were executed as a collaborative effort between the multi-channel retailer and Google using the Applied Predictive Technologies (APT) Test and Learn^{TM} soft-

ware. APT's software is designed to conduct field experiments in retail settings. The software is licensed to over 100 consumer facing organizations including 35 of the top 100 U.S. retailers. It is used by senior decision makers such as the Chief Financial Officer in retail organizations to evaluate the return on investment for capital allocation decisions around marketing, merchandising, store operations and store improvements.[15] The APT platform also allowed us to use an automated approach to scale and execute a population of field experiments. Only specific organizations who explicitly agreed to participate in this study are included in this analysis. Further, even for those organizations the limited data that was needed for this specific analysis was used.

Each field experiment was designed and executed as a sequence of steps. The retailer provided APT with store sales data, category sales data, and geo-demographic correlates for all stores in their network for the preceeding 2+ years. From the retailer, a list of other known major activities was obtained, including remodel activities, new store builds, and any major local marketing efforts. Designated market areas (DMAs) that were highly impacted by the known activities were excluded from the potential test pool. Based on historical sales data, APT determined the tradeoff between the number of test DMAs, the minimum number of weeks, and the minimum sales lift %[16] that the experiment could detect with 95% significance. Required budget for the experiment was calculated, using the principle that the budget should enable the experiment to be read with statistical significance, if the sales lift is high enough to produce a breakeven-or-better gross profit return from the search spend. Google provided estimates of available search inventory by product category to show that enough search inventory was available to fill the experimental budget. This analysis was brought together in a chart which the retailer then used to make decisions on test categories and budgets. If sufficient search inventory was not available for a given category, or if the required budget was greater than the retailer was willing to spend, then that product category was not used for the test.

APT then randomly selected DMAs that had the required number of stores and matched the characteristics of the sample of stores to the population of stores on multiple criteria including sales, demographics, category size as a percentage of sales, store size and geographies. The

[15] More details about APT and its clients can be obtained from the firm's web site at www.predictivetechnologies.com

[16] Defined in Section 4.2.1.

software matched each test store to 10 stores in the control condition based on sales history, store size, store format, geo-demographic correlates of each store obtained from Pitney Bowes and data on primary competitors. The matching was verified so that the average sales of the test stores and control store trended similarly in the pre-test period. Figure 3 provides an example. The figure shows that the averages for the group of test and control stores track each other in the pre-test period, while the sales diverge in the treatment period. Google teams set bids on broad keywords[17] in the test markets so that the average position on the page was maintained between 1 and 3. The ads were spaced out so that the test budget was allocated evenly over the entire duration of each experiment.

Table 2 provides information about the field experiments. The number of test markets and test stores varied across experiments. For example experiment F involved 405 test stores, the largest number of test stores in our analysis. Field experiment B on the other hand had 25 test stores, the smallest number. The last line in the table provides averages. On *average* each experiment had 29 test markets, 167 test stores, a duration of about 4 weeks and involved 5 categories. While the average test budget (ADS) was $276,826, the incremental test budget varied across retailers. For example Retailer A spent $139,000 whereas Retailer B spent $685,000.

To provide perspective on the collective scale of this population field experiments, the second last row in Table 2 reports totals for key metrics. In total the participating retailers spent $4,153,393 in test budgets on category keywords. The retailers and Google incurred additional manpower costs to execute the experiments. The total number of test stores across field experiments was 2506. To put this number is perspective, some of the largest US chain retailers would have over 3000 stores in their retail network. When estimating the overall average effect across experiments this large sample size of test stores should be useful in improving the precision of the overall estimates. There were a total of 76 categories across experiments. This large number of categories should provide insight into whether the search advertising effects are widespread or limited to a narrower set of categories.

Once the experiment was complete, the APT software used built-in capabilities to detect outliers, extreme events and other anomalies to improve the robustness of the tests. In the final step APT estimated and reported incremental sales for the advertised categories, non advertised

[17]For example "Sunglasses" is a category keyword, as opposed to Rayban Sunglasses or Rayban Aviator Sunglasses at Sunglass Hut. Rutz and Bucklin [2011] refer to these category keywords as generic keywords.

categories and the total store for a duration that included the test period and additional weeks beyond the test period. Table 3 provides an example of the estimates reported by APT for each field experiment. Under the data sharing agreement, only these reports were released to the researchers. The data set for this study was assembled from these reports.

4.2 Measures

4.2.1 Sales lift

Each retailer's experiment targeted specific categories, so the retailers are naturally interested in results for the targeted categories. But they are also interested in whether their ad campaign produced positive or negative spillover effects outside the category, i.e., in effects at a more aggregated level than category, either the total store or a larger grouping of categories that includes the target category. We present results for both the more aggregated level ("total store") and for categories.

For a given experiment, APT used the random sampling of test stores to estimate the sales lift for the total store and for categories. We present a formula for the total store computation, but the same formula was used for category-specific lifts. For test store s, the lift was estimated as

$$L_s = \frac{AS_s - ES_s}{ES_s} \tag{1}$$

where L_s is test store s's estimated lift expressed as a fraction (not as percent), AS_s is its actual sales, and ES_s is its expected sales based on its matched control stores. Further,

$$ES_s = TPre_s \frac{CPost_s}{CPre_s} \tag{2}$$

where $TPre_s$ is sales of test store s in the pre-treatment period, and $CPost_s$ and $CPre_s$ are the average sales of the matched control stores in the post-treatment and pre-treatment measurement periods respectively.

The combined estimate over all n test stores, estimating average sales lift for the whole retail chain, weighted test stores by expected sales:

$$\overline{L} = \sum_s \frac{ES_s}{\sum ES_s} L_s = \frac{1}{n\overline{ES}} \sum_s (AS_s - ES_s). \tag{3}$$

where $\overline{ES}$ is the average expected sales of the test stores. The method of sampling test stores, matching control stores to them, and weighting test stores in the combined estimate implies a particular standard error for the experiment's estimated sales lift. APT estimated standard errors that account for the fact that some control stores were matched to more than one test store and that there are multiple test stores from the same DMA. When the experiments were conducted, the estimate and standard error were used to compute P-values to test whether the total store or category sales lift differed from zero.

The dataset released for this study did not include the standard error for the total store or category tests.[18] Rather, it included either a one-sided P-value for the t-statistic estimate, computed from the average effect (available) and the standard error (unavailable), or a range within which the one-sided P-value lay [19]. When the exact P-value was available, we could infer the standard error from the estimate and P-value (subject to round-off error). When only a range of P-values was available, our baseline Bayesian analysis treated the P-value as being equally likely to take any value in the range, i.e., we put a uniform prior on the reported range. Further details are given in Section 6.1.2 below. Section 6.3.1 conducts a robustness analysis for this assumption.

For category-level results, few exact P-values were reported and the reported ranges of P-values were often wide. Thus, our category-level analysis is conservative in that it treats each experiment as if it produced less information than, in fact, it did.

4.2.2 Incremental Annualized Cross Channel Sales Opportunity for the Population of Stores

The next outcome variable of interest to retailers is the incremental cross channel sales opportunity if the increased advertising spending was expanded to the entire population of stores on a year round basis. We call this the incremental annualized sales opportunity for the population of stores and estimate it as follows:

$$ISO = \frac{Base}{t} N \, 52 \, \overline{L}, \tag{4}$$

[18] This was simply due to a data capture omission that occured in the time lag it took to get the permissions obtained for this study.

[19] APT expresses these as respectively, 0 to 0.05, 0.05 to 0.3, and 0.3 to 0.5. Note that 0.5 is the largest possible P-value in a one-sided test.

where ISO is the incremental annualized sales opportunity for the population of stores, $Base$ is the average baseline sales per store for the duration of the test period, t is the duration of the test period in weeks, N is the total number of stores for this retailer, and $\overline{L}$ is the average sales lift for the test stores as estimated in (3).[20] We note that this estimate does not adjust the baseline sales for seasonality and assumes that the sales lift obtained in the test period will generalize to other time periods.

4.2.3 Incremental Return on Ad Spend (ROAS)

The next outcome variable of interest to retailers (including decision makers in finance) is the return on investment due to the increased advertising. A common return on ad spending (ROAS), used by our retailers, was incremental sales due to incremental ad spending, defined as follows:

$$ROAS = \frac{1}{ADS} \sum_{s} (AS_s - ES_s), \tag{5}$$

where ADS is the incremental ad spending for the whole experiment (Note that this measure is implicitly weighted by the sales of individual test stores, analogous to the weighting in the combined estimate of sales lift.).[21]

In our experiments APT did not measure online sales impact so the ROAS is based only on offline sales. The data available for this study included estimated ROAS but did not report a P-value or a standard error for ROAS. We developed a method to estimate standard errors for experiment-specific ROAS estimates from the standard errors for sales lift (technical details are in Section 8.2 of the appendix).

5 Results from Field Experiment G

5.1 Background

This section presents the results from field experiment G, the results of which are very similar to the overall average. As Table 2 shows, this experiment involved 10 test markets consisting of 126 test stores with a test duration of 4 weeks. Figure 4 lists the test markets and provides a

[20] Table 2 presents some of this information. Column 2 provides t, the duration of the test in weeks, Column 3 provides $Base$ sales for the test period. N, the total number of stores is not reported as per disclosure policies.

[21] Ad spending data was not disaggregated to individual test stores.

breakdown of the number of stores in each test market. The figure shows that the test markets are distributed across the US. The green triangles show test stores with sales above the average of the group of test stores. Red triangles show test stores with sales below the average of the group of test stores. The marker 'c' refers to control stores. Figure 3 shows that the average sales of the test stores and the matched control stores trended togeather in the pre-treatment period.

The retailer selected three categories for the test and spent $466,000 in incremental search spending. The Google team set maximum bids on broad keywords to obtain positions 1-3 in the test markets. These bids achieved an average position of 2.07 and the click thru rate (CTR) for the campaign was 1.13%. This incremental test budget achieved a 57.88% impression share in the test markets. To put it differently, for 42.22% of the search queries in these markets, this retailer's ads did not appear in broad searches due to the constraint on the test budget.

5.2 Estimates

Table 3 presents the estimates for this field experiment. Of direct interest is the sales lift from the advertised categories, which are presented in the first row of the first panel. These estimates are based on six weeks of measurement, four weeks for the test period and two weeks for the post test period. The first row in the table presents the estimate of $\overline{L}$ in equation 3 here expressed as a percent. The estimate of sales lift in the advertised categories was 5.9% and statistically significant. This corresponds to a weekly sales increase per store of $6361.0 and a $4.8M total sales increase for the test stores over a six week period. The next two lines show that sales in the non-test categories in the same department decreased by 0.9% and 0.7% but these decreases are not statistically significant.

The fourth row presents the department level estimates which are an aggregate of the advertised and non advertised categories. The sales lift at the department level is 2.2% and is significant. The smaller estimated lift percent, compared to the advertised category, is due to dividing the estimate of sales increase for the advertised categories by a denominator that includes sales for both the advertised *and* non-advertised categories. The estimate of weekly sales increase per store is $5339.6 and the six week total impact for all of the test stores is $4.04 M.

The second panel in the table shows the test and non-test departments and their aggregation

into a total store impact estimate. The first row in this panel simply repeats the estimates of the test department to facilitate exposition. The second row shows the impact on the non-test departments. The estimate of 0.4% is positive but not significant. Finally, the last line reports the total store estimates of lift percent, average incremental sales per store per week, and 6 weeks total estimate, which are all significant. Once again the smaller magnitude of the lift estimate is primarily due to dividing the $7355.0 estimate for the advertised department by a larger denominator that includes *both* the advertised and non-advertised departments. The incremental annualized sales opportunity for the population of stores for this retailer is estimated to be $1.116 Billion. This estimate and the corresponding confidence intervals are reported in Table 5.

Finally dividing the six week total incremental sales estimate of $5.56 Million by the incremental advertising spending of $466,000 provides a ROAS of 12. So, an incremental dollar in search advertising spend yielded $12 in incremental sales at the store level. APT does not report significance levels for ROAS; we discuss this issue in more detail in the next section. If we assume a Gross Margin of 35% the break-even ROAS is 2.85. So this particular experiment exceeded breakeven.

5.3 Discussion

The statistically significant estimates of sales lift in advertised categories and the magnitude of these effects have important implications. First they offer experimental evidence of these effects, which to our knowledge have not been reported in the literature. Sometimes these cross channel effects are referred to as "spillover" effects with perhaps the unintended connotation of secondary importance. The magnitude of these estimates suggest that it might be worthwhile to actively manage this "spillover" aspect of search advertising. The magnitude of these effects is consistent with the fact that there are more offline shoppers than online shoppers. Although we do not find a significant negative or positive cross category or cross department effect for Retailer G, the results suggest the importance of designing these measurements into a field experiment. However, these are results from one retailer. This retailer is focused on a narrower grouping of categories. The categories studied are considered "big ticket items" requiring considerable expenditure, and are potentially high involvement. On the other hand, the impression share in

this experiment was only 57.88% and the click thru rate was 1.13% which suggests additional incremental sales opportunity with additional impression share and click thru. Results from multiple field experiments will help with generalizability. The next section develops models to combine results across experiments.

6 Hierarchical Bayesian Meta Analysis: Methods and Results

We are interested in obtaining an overall average estimate of say sale lift across this population of field experiments that takes into account within study variation and heterogeneity in treatment effects across studies and categories. This overall average estimate and its 95% confidence interval address whether there is causal evidence of a cross channel effect of search advertising for this population of retailers. This section describes Hierarchical Bayesian (HB) models to obtain the overall estimates.

6.1 Methods

6.1.1 Analysis of heterogeneous treatment effects using hierarchical Bayesian methods.

In psychology, health-care, and other social sciences, combining randomized experiments to estimate an overall treatment effect for a population of interest is called meta-analysis. Early in the development of meta-analytic methods it became clear that differences between experiments were often too large to be explained by chance or sampling variation within experiments, and that a simple combination of such heterogeneous experiments might be ill-advised. Meta-analysis of experiments has come to be understood by many as a way of doing regression analyses in which the units of analysis are experiments, reduced to summaries instead of individual measurements (Greenland [1994]), and allowing for heterogeneous treatment effects across experiments.

This approach uses the inherently hierarchical structure of a meta-analysis: the primary unit of analysis is an experiment, but each experiment's results are aggregates of its subjects' individual results (in our case, the results from individual test stores). This hierarchy is captured using hierarchical models with "error terms" at two levels, one describing variation between experiments and a second level describing variation between subjects within experiments (in

our case, the standard errors associated with an experiment's lift or ROAS estimate). The earliest popular meta-analysis method using a hierarchical model was the DerSimonian-Laird (1986) method, which assumes the true treatment effect varies between experiments and models those effects as draws from a normal distribution with mean μ and some variance, the object being to estimate μ and attach a suitable standard error. An obvious elaboration of this model is to add explanatory variables describing differences between experiments, thus reducing the unexplained between-experiment variation. In the analyses below, we have used the experiment level predictors retailer type, impression share, click-through rate, and (for categories only) category strength, a measure of the category's importance to the retailer.

Meta-analyses using hierarchical models lend themselves naturally to Bayesian methods, which also provide a simple way to handle some complications of the data from this group of experiments. For example, two retailers are represented in our dataset by two experiments, and presumably a retailer's two experiments will tend to be more similar to each other than they are to other retailers' experiments. In a Bayesian analysis, this presumed correlation is captured by treating retailer as another layer of the hierarchy and adding a random effect for that layer. Also, as noted, in inferring an experiment's standard error from its estimate and P-value, when P-values were reported as ranges we could specify a prior distribution on the reported range of P-values.

6.1.2 Models and computing

The models for total store sales lift percent and ROAS are very similar; in this section we present the model for sales lift and then state the changes needed for ROAS. All models are presented with experiment-level predictors (covariates) included. Recall that two retailers had two experiments each. In the model description below, retailers are indexed by i and experiments within retailers by j. The combined lift estimate from experiment (i, j) is called Lift_{ij}. The model includes an error term capturing sampling and other variation *within* each experiment, with variance σ_{ij}^2. It also includes two random effects, one each for retailers and experiments within retailers, with variances λ^2 and τ^2 respectively. The model includes three experiment-level predictors (covariates), "Impression share" ($x_{1,ij}$), "Click Thru Rate" ($x_{2,ij}$), and "Retailer type" ($z_{1,ij}$, $z_{2,ij}$), which are all entered linearly.

Note that "Retailer type" is a categorical variable, which can be "big box generalist", "big box category killer", and "specialist" (which is the reference category). $z_{1,ij}$ takes value 1 if the experiment store belongs to "big box generalist"; otherwise, it is 0; $z_{2,ij}$ takes value 1 if the experiment store belongs to "big box category killer"; otherwise, it is 0. Following the discussion in Section 2.5, big box generalists are retailers who operate large sized stores and offer a broad assortment that covers a number of categories (e.g. Costco). Big box category killers also operate a large sized store but are more focused on a particular need (e.g. office products) and offer a deep selection for this need. Specialists focus on a narrower set of categories compared to the other two types (e.g. Zara). Note that the big box category killer is 'in-between' a generalist and a specialist.

Formally, the model and prior distributions are (with sales lift expressed as a fraction, i.e., a 1% sales lift is 0.01):

$$
\begin{aligned}
&\text{Mean structure:} && \text{Lift}_{ij}|\theta_{ij},\sigma_{ij}^2 \sim N(\theta_{ij},\sigma_{ij}^2) && \\
& && \theta_{ij} = \eta + \alpha_i + \beta_{ij} + \xi_1 x_{1,ij} + \xi_2 x_{2,ij} + \xi_3 z_{1,ij} + \xi_4 z_{2,ij} && \\
&\text{Random effects:} && \alpha_i|\lambda \sim N(0,\lambda^2) && \text{Retailer} \\
& && \beta_{ij}|\tau \sim N(0,\tau^2) && \text{Experiment within retailer} \\
&\text{Prior distributions:} && \lambda \sim U(0,10) && \\
& && \tau \sim U(0,10) && \\
& && \eta \sim N(0,1000) && \\
& && \xi_1,\xi_2,\xi_3,\xi_4 \sim N(0,1000) && \\
&\text{For inferring standard errors } \sigma_{ij} = |\text{Lift}_{ij}|/t_{n_{ij}-1}(1-p_{ij}) && && \\
&\quad \text{when P-value} \in \text{interval: } p_{ij} \sim U(p_{\text{lower},ij}, p_{\text{upper},ij}). && &&
\end{aligned}
\tag{6}
$$

In Equation (6), Lift_{ij} is observed and thus treated as known in a Bayesian analysis, and the inferred standard error σ_{ij} is known if the P-value is exact. Also, $t_{n_{ij}-1}(1-p_{ij})$ is the $100\times(1-p_{ij})$ percentile of the t distribution on $n_{ij}-1$ degrees of freedom, where $p_{ij}\in[0,1]$ and smaller p_{ij} indicate "more significant". (Section 8.1 discusses how we selected n_{ij}.)

The model for total store ROAS is identical except for one change: replace the line in

Equation (6) for inferring the standard error with (see section 8.2 for the derivation),

$$\sigma_{ij} = \frac{n_{ij}\overline{\text{ES}}_{ij}}{\text{ADS}_{ij}} \times \frac{|\text{Lift}_{ij}|}{t_{n_{ij}-1}(1-p_{ij})}, \tag{7}$$

where $\overline{ES}_{ij}$ is the average of the expected sales for the test stores, and ADS denotes advertising spending.

The model for category-specific sales lift builds on the preceding model by adding more structure. It includes another covariate, "Category strength" ($x_{3,ijkl}$) which reflects whether the retailer was considered strong in the category. Also, the categories have a hierarchical structure, with each of the 76 categories being placed in one of 18 "broad categories". Broad categories are indexed by k; categories within broad categories are indexed by l. The model has an additional random effect for broad categories, with variance δ^2, which induces similarity between the categories included in a broad category. The combined lift estimate for category l

in broad category k from experiment (i, j) is Lift_{ijkl}. The full statement of the model is

$$
\begin{aligned}
&\text{Mean structure:} && \text{Lift}_{ijkl}|\theta_{ijkl}, \sigma^2_{ijkl} \sim N(\theta_{ijkl}, \sigma^2_{ijkl}) && \\
& && \theta_{\text{L},ijkl} = \eta + \alpha_i + \beta_{ij} + \gamma_k + \epsilon_{ijkl} + \xi_1 x_{1,ijkl} + \xi_2 x_{2,ijkl} + \xi_3 x_{3,ijkl} && \\
&\text{Random effects:} && \alpha_i|\lambda^2 \sim N(0, \lambda^2) && \text{Retailer} \\
& && \beta_{ij}|\tau^2 \sim N(0, \tau^2) && \text{Experiment within retailer} \\
& && \gamma_k|\delta^2 \sim N(0, \delta^2) && \text{Broad category} \\
& && \epsilon_{ijkl}|\rho^2 \sim N(0, \rho^2) && \text{Category w/in broad category} \\
&\text{Prior distributions:} && \eta \sim N(0, 1000) && \\
& && \xi_1, \xi_2, \xi_3 \sim N(0, 1000) && \\
& && \lambda \sim U(0, 10) && \\
& && \tau \sim U(0, 10) && \\
& && \delta \sim U(0, 10) && \\
& && \rho \sim U(0, 10) && \\
&\text{For inferring standard errors} && \sigma_{ijkl} = |\text{Lift}_{ijkl}|/t_{n_{ijkl}-1}(1 - p_{ijkl}) && \\
&\quad \text{when P-value} \in \text{interval:} && p_{ijkl} \sim U(p_{\text{lower},ijkl}, p_{\text{upper},ijkl}). &&
\end{aligned}
\tag{8}
$$

All analyses were performed by Markov chain Monte Carlo (MCMC) implemented in the R system (version 3.0.2) using the rjags package (JAGS version 3.4.0, rjags version 3-11). Each analysis used three chains of length 200,000 iterations each, with starting values chosen by the rjags package (using specified seeds), with the first 100,000 draws discarded as burn-in and retained draws thinned by taking every second draw. Point estimates reported below are posterior medians; 95% posterior credible intervals are the 2.5^{th} and 97.5^{th} percentiles of the MCMC samples.

6.2 Results

6.2.1 Sales Lift-Total Store

Results for sales lifts are presented as percents (in Section 6.1.2, the model and priors used lifts expressed as fractions). The first two columns in Table 4 present the reported sales lift and p-value or p-value interval for total store sales, which is the data available to us from each experiment. The rest of the columns present results from the Bayesian analysis. Figure 5 is a forest plot for total store sales lift, showing the original non-Bayesian estimates and intervals (gray boxes and lines) and estimates and intervals from the Bayesian analysis (blue boxes and lines). The overall estimate and interval are shown at the bottom.

Figure 5 and Table 4 show the reported sales lift estimates are all positive and range from 0.01% to 8.80% with a simple average of 1.87%. The 95% intervals for experiments C, D, F, G, H, I, J, K, L, M do not contain zero. Thus we have multiple experiments providing repeated causal evidence for a positive sales effect. These results for individual experiments describe the effects of these ad campaigns for individual retailers, but if they are combined in an overall estimate, with a proper accounting of variation within and between studies, they can tell us about the causal effect of search advertising in this population of retailers, and borrow strength across experiments. The estimated overall average sales lift from the Bayesian analysis accounting for within and between study variation, is 1.18%. The 95% credible interval of this overall estimate is 0.63% to 1.82% and does not contain zero. The posterior density of the overall estimate of sales lift, shown in the left panel of Figure 7 has negligible mass below zero. The average sales lift from the population of experiments shows that there is a causal cross channel impact of search engine advertising on total store sales offline. Since the U.S. retail industry grew at a compound annual rate (CAGR) of 1.55%[22] during the time period in which the experiments were run, the magnitude of these sales increases is meaningful. A simple back of the envelope calculation shows that with search engine advertising campaigns executed as reported in this paper, our population of retailers would have had a CAGR of 2.75%, a 77.5% increase over the base growth rate.[23]

[22] Calculated based on data reported by the National Retail Federation.

[23] The base line CAGR is 1.55%. If sales in the base year is 100, then one year later sales would be 101.55. Applying the incremental lift of 1.18% onto 101.55 we get $101.55 \times 0.0118 = 1.20$. So adding the incremental growth, sales one year later would be 101.55+1.20=102.75. So the new growth rate post search advertising on broad keywords is 2.75%. Compared to the pre-period growth rate of 1.55% the increase in growth was 77.5%.

In all of our models, the random effects can be interpreted as the difference between the average measure (sales lift or ROAS) in one specific experiment or category and the average measure in the group of experiments or categories respectively. In Table 4, the between-retailer random effects are the α_i in the model (6), and the within-retailer random effects are the β_{ij}. Referring to the forest plot in Figure 5, we can take the total store estimates in Table 4 as an example. Both between- and within-retailer random effects for experiment G are estimated as nearly zero, and this means that the difference between average sales lift in experiment G and the overall sales lift is almost zero. This inference can be confirmed by the estimate in the sales lift column in Table 4: the average sales lift in experiment G is estimated as 1.17, and the overall sales lift over all of the experiments is 1.18. As for the experiments A to F, almost all of the random effects are estimated as negative values, which leads to the sales lift estimates being smaller than the overall sales lift; for the experiments H to N, the opposite holds. The estimated standard deviation describing variation between retailers is 0.32 percentage points (95% CI 0.02 to 1.36), which suggests considerable heterogeneity between retailers.

Experiments C and F are from the same retailer so the estimated between retailer random effect for both experiments is identical with a value of -0.15. The 95% CI ranges from -0.95 to 0.45. The estimate of the within-retailer random effect for experiment C is -0.16 with a 95% CI of -0.97 to 0.39. For experiment F the within retailer random effect is 0.01 with a 95% CI of -0.70 to 0.69. A considerable portion of the 95% CI for these estimates do not overlap. The wide interval for the within-retailer random effect estimate suggests considerable heterogeneity within the experiment, presumably due to differences between categories. We offer this interpretation cautiously since the experiments were not conducted in the same time period. Experiments H and M are also from the same retailer and we get similar insights based on the between and within retailer estimates and 95% CIs.

For most of the other experiments the estimates of the two random effect are quite similar. This arises because only two retailers have more than one experiment, and they have just two each, so that the Bayesian machinery will have difficulty allocating variation between these two sources of variation. (Non-Bayesian machinery would have at least as much difficulty.) The variances of the two random effects were exchangeable *a priori*, so they are similar *a posteriori*

This back of the envelope analysis estimates growth based on a single year.

and the machinery splits an experiment's deviation from the overall average into two roughly equal pieces, attributed to the two random effects. However, the weak identification of the two random-effect variances has a negligible effect on the posterior distribution of the overall sales lift, because the sum of the two random effects, and thus the sum of the two variances, is well identified.

For the covariate *Impression share*, the estimated average increase in lift is 2.5 percentage points per 1-unit increase in impression share (95% CI -2.0 to 6.9). The sign is positive – so higher impression share implies that the ads were served in higher number of searches and hence had a greater "reach" of online shoppers – but the 95% CI contains zero. A positive relationship between reach and sales lift seems plausible. For the covariate *Click-through rate*, the estimated average increase in lift is 1.71 percentage points per increase of 0.1 in click-thru rate (95% CI -4.66 to 7.87). The sign for this covariate is again positive – higher click thru rate means more clicks and more traffic to the retailer web site – but again the 95% interval contains zero. Some of these shoppers might be gathering information and pre-shopping (clicking thru to the retailer's web site) prior to a visit to a brick and mortar store, so a positive relationship between click thru rate and sales lift seems plausible.

The last part of Table 4 reports the estimates for the covariate *Retailer type*. The lower part of Figure 5 graphs them. The lift estimates are respectively 0.95, 1.25 and 1.19 for the big box generalist, the big box category killer and specialty retailers. These estimates are consistent with our expectations that the generalist should get lower lifts, but the highest sales lift for the big box category killer suggests an interesting nuance. Note that the big box category killer is 'in-between' a generalist and a specialist: more specialized than a generalist but not as much as a specialist, with less scale than a generalist but more compared to a specialist. The higher sales lift for this type of retailer suggest that in our data it is plausible that the benefits of moderately increased specialization more than offsets any loss of scale. In the context of search advertising this can be restated as a tradeoff between clicks and conversion. Higher scale is consistent with higher brand awareness and higher clicks on a search ad. Higher specialization is consistent with a deeper offering, web site experience and higher conversion. To put it differently, it looks like this type of retailer, with moderate specialization, is able to get higher clicks while maintaining conversion. Given the wide intervals around the estimates, this evidence is directional in nature.

6.2.2 Incremental Annualized Sales Opportunity for the Population of Stores

Table 5 presents the estimates of incremental annualized sales opportunity for the population of stores. The computation follows (4) but we use the posterior density for $\bar{L}$ obtained from (3) to calculate the estimate and to obtain the 95% credible intervals. The estimates from *all* experiments are positive. The 95% credible intervals for experiments C, D, F, G, H, I, J, K, L, M and N do not contain zero. The last line in the table provides the total opportunity for this population of retailers, obtained by summing across the individual experiments. The mean is estimated to be \$1.948 Billion and the median estimate is quite similar. We note that this estimate was obtained without adjusting for seasonality.

6.2.3 Total Store ROAS

The first column in Table 6 presents the reported ROAS for the total store in each study. This was the data available to us from each experiment. The Bayesian analysis is presented in the rest of the table. The standard errors or P-values for the top level ROAS were not calculated or reported. Figure 6 is a forest plot for total store ROAS, showing the reported non-Bayesian estimates (gray boxes). The intervals (gray lines) were estimated using Equation 7 and the approach described in Section 8.2. The blue boxes and lines show the estimates and intervals from the Bayesian analysis. The overall estimate and interval are shown at the bottom.

Figure 6 and Table 6 show the reported ROAS estimates are all positive and range from 0.01 to 14 with a simple average of 4.41. The 95% intervals for experiments L, J, C, N, K, G, and H do not contain zero, so we have multiple experiments that provide repeated causal evidence for a positive and statistically significant ROAS. The estimated overall average ROAS from the Bayesian analysis is 2.5. The 95% credible interval of this overall estimate is 1.03 to 4.48 and does not contain one. The right panel of Figure 7 shows the posterior density of the overall estimate of ROAS, which has negligible mass below zero. This shows that the ROAS of search engine advertising on brick and mortar sales is positive for this population of retailers. Retailers expect a ROAS of 4:1 (Holmes [2014]) from certain direct marketing activities such as catalog mailings. The posterior density of our overall estimate has considerable mass around 4. This suggests that the ROAS from search advertising on offline sales is comparable to other types of direct response marketing.

On average the retailers in our experiments had a Gross Margin of 35%[24]. This implies that the break-even ROAS is 2.85. The 95% CI of the overall estimate and the posterior density (Figure 7) indicate that a considerable probability mass is greater than 2.85. The key implication is that for this population of retailers there is a considerable probability of breaking even.

The between-retailer random effect standard deviation (in equation 6) is estimated as 1.26 (95% CI 0.07 to 4.01). The between-experiments-within-retailer standard deviation (in equation 6) is estimated as 1.15 (95% CI 0.06 to 3.74). For the two retailers who had two experiments each, the insights regarding the between- and within-retailer random effects are similar to those for sales lift. For the covariate impression share, the estimated average increase in ROAS is 0.69 per 1 unit (10%) increase in impression share (95% CI -0.34 to 2.08 and includes zero). For the covariate click-thru rate, for an increase of 0.1 (10%), the estimate is 0.97 and the 95% CI is -5.00 to 6.78, which includes zero.

The last three rows in Table 6 provide the estimates for the covariate *Retailer type*. The estimates are graphed in the bottom part of Figure 6. The ROAS estimates are 3.2, 2.69 and 2.4 for the big box generalist, the big box category killer and specialist respectively. All three retailer types have positive ROAS and the confidence intervals do not contain zero. The Big box generalist has the highest ROAS and this is consistent with the high awareness enjoyed by generalists. Although the Big Box category killer type had the higher sales lifts, the increased focus did not translate into a better ROAS. The confidence intervals for these estimates are wide implying that the evidence about differences between retailer types is only directional. We once again note that these ROAS estimates do not include the online impact of the increase in search advertising.

6.2.4 Category-level sales lift

The first two columns in Table 7 present the reported sales lift for the category sales and the p-value or p-value interval in each experiment. These two columns represent the data available to us. The other columns present the Bayesian estimates. The between-retailer random effects are the α_i in model (8), the within-retailer random effects are the β_{ij}, the between-category random effects are the γ_k, and the within-category random effects are the ϵ_{ijkl}. Table 8 groups

[24]This estimate was obtained by the authors based on an analysis of the Profit and Loss statements of the retailers who participated in the experiments.

these categories into 18 broad categories.

Table 7 shows that a few category level lift estimates (category # 2, 3, 8, 13, 14, 15, 25, 28, 46, 51, 56, 62, 63, 67, 68, 69, 70, 71) are negative. However, none of these negative estimates are significant and they typically have very wide confidence intervals. The rest of the estimates are positive and for many of them the lower bound of the 95% CI does not contain zero. So we have repeated evidence for a positive effect of search engine advertising on *category sales* in offline stores. The evidence is from multiple retailers and categories. Figure 8 is a forest plot of the Bayesian estimates of sales lift for all 76 categories.[25] The last row of Table 7 shows that the estimated overall average sales lift from the Bayesian analysis is 1.27%. The 95% credible interval of this overall estimate is 0.30% to 2.34% and does not contain zero. Figure 9 shows the posterior density of the overall estimate of sales lift for the category level data. This density shows negligible mass below zero. The average sales lift from the population of experiments shows a causal cross channel impact of search engine advertising on category level sales in brick and mortar stores.

The between-retailer random effect shows how far the average category effect for each retailer is from the overall category average across all retailers. Since this is a retailer effect, all the categories for the same retailer will have an identical estimate and 95% CI. The within-retailer random effect shows how far the average category effect for this experiment is from the average across all experiments for this retailer. As noted in the discussion of sales lift estimates, only two retailers had two experiments each so these effects may not be well identified. The between-category random effect shows how far the effect for this broad category is for this retailer from the overall average across all retailers for this broad category. The within-category (category within a broad category effect in Equation 8) shows how far this specific category is from the average of other categories within this broad category for this retailer. The estimates of random-effect standard deviations describe the relative sizes of these various sources of heterogeneity. The between-retailer standard deviation estimate is 0.66 (95% CI 0.03 to 2.35). The between-experiments-within-retailer standard deviation estimate is 0.55 (95% CI 0.03 to 2.06). The between-broad-categories standard deviation estimate is 0.57 (95% CI 0.03 to 1.82). The between-categories-within-broad-category standard deviation estimate is 0.57 (95% CI 0.03 to

[25]Since only p-value intervals are reported for many estimates it is not possible to generate a forest plot of the reported estimates.

1.30). Note that these four components of variation all show significant magnitude and uncertainty.

To obtain insights from these random effects in a more intuitive manner Table 8 presents the category level estimates aggregated into 18 broad categories. The estimates are positive for all 18 broad categories. This suggests that the cross channel impact of search engine advertising on sales in brick and mortar stores is widespread across these different broad categories. Some of the broad categories with higher estimates such as furniture, home furnishings, kitchen and bath, small appliances, and large appliances are categories where it is plausible that consumers might search more online before they purchased in a store.

For the covariate Impression share, the estimated average increase in lift is 0.06 percentage points per 0.1 (10%) increase in impression share (95% CI -0.91 to 0.88). For the covariate Click-through rate, the estimated average increase in lift is 4.48 percentage points per 0.1 (10%) increase in click-thru rate (95% CI -6.29 to 14.6). For the covariate Category Strength, the estimated average increase in lift is 0.032 percentage points per 0.1 (10%) increase in category strength (95% CI -0.024 to 0.103). The latter interval does contain zero but we note that the 95% CI for this covariate is narrower than the intervals for the other covariates. While the signs of these coefficients are intuitive, the confidence intervals suggest that the evidence is directional.

6.3 Robustness

6.3.1 Various p-value scenarios

In the category-level data, most p-values were reported as an interval $(p_{\text{lower}}, p_{\text{upper}})$ for $0 < p_{\text{lower}} < p_{\text{upper}} < 1$. Our main analysis represented this uncertainty about the P-value using a uniform prior on this interval, for the purpose of deriving a standard error for each category/experiment's sales lift. To check sensitivity of the results to this choice, we considered these alternatives:

- **Pessimistic:** Use p_{upper} for this p-value; this gives the maximum standard error consistent with this interval.
- **Optimistic:** Use p_{lower} for this p-value; this gives the minimum standard error consistent with this interval.

- **Midpoint:** Use the midpoint of the interval, i.e., $(p_{\text{lower}} + p_{\text{upper}})/2$.
- **Sampling actual p-values (SAPV):** Collect all the p-values reported as exact values (not intervals) from both the original total store and category-level results. For each p-value reported as an interval, draw one sample from the subset of these exact P-values that is in the target interval.

Table 9 shows the resulting overall sales lift estimates and intervals. The 95% CI of overall sales lift excludes zero under all alternatives, and the overall sales lift estimate (posterior mean) is very similar under the uniform prior, mid-point, and SAPV alternatives. These results imply the main result is robust to our handling of p-values reported as intervals. These results also confirm that our handling of the p-values is conservative. Most of the other scenarios except the mid-point produce a higher estimate.[26]

It may seem counterintuitive that the posterior SD for the overall effect is *larger* under the optimistic alternative than under the pessimistic alternative. However, the observed data, i.e., the lift estimates for individual category/experiments, have a fixed amount of variation which the model allocates to five components of variation. If we reduce the part of that variance allocated to within-experiment variation — as we did by switching from the pessimistic to the optimistic alternative — then a larger fraction of the observed variation must be allocated to the other components of variation (between-retailer, between experiment within retailer, etc.). Because these other components of variation are not suppressed by replication to the same extent as within-experiment variation, the net effect is a larger posterior standard deviation.

6.3.2 Top-level ROAS: Correlation between test stores of ES_i

To derive a standard error for an experiment's total store ROAS (Appendix, Section 8.2, Equation 14), we needed to specify the correlation r between expected sales ES_s for pairs of test stores. This correlation arises from matching individual control stores to more than one test store to derive the test stores' expected sales. Our main analysis ignored this correlation because it had a tiny effect. This section briefly considers the robustness of the results to this assumption.

[26]For the midpoint alternative the p-values are fixed, treated as a known quantity. For the uniform prior alternative, there is an additional layer in the hierarchical model, and p-values are drawn from those prior. Therefore, the results in these two scenarios will not be identical.

The correlations are not directly available from the reported data, however, they can be estimated from the number of test stores and total stores in network, and their estimations are listed in Table 10.

The resulting overall ROAS estimate considering the correlation was 2.49, with a 95% credible interval $(1.02, 4.47)$. In our main analysis, ignoring the correlation, the estimate was 2.50 with 95% CI $(1.03, 4.48)$. These results support that our analysis is robust for the correlation issue. If the assumed values of r were doubled or tripled, the results would still change negligibly.

7 Conclusions, Limitations and Future Research

Retailers need to understand if advertising in one channel can generate sales in another channel. Since retailers sell an assortment consisting of multiple categories they also need to know how widespread these effects are, whether a category orientation is beneficial, and the ROI from this type of advertising. Our paper takes a first step at investigating these first order questions.

We report the results of a population of field experiments conducted on Google.com. In each experiment we increased advertising spending to maintain broad keywords in positions 1-3 in the experimental stores. In the control stores these keywords were not advertised. We used a hierarchical Bayesian model wit heterogeneous treatment effects to estimate the average treatment effect for this population of experiments.

Across the experiments we found causal evidence that an increase in search advertising on Google.com in targeted categories caused an incremental increase in brick-and-mortar sales in the advertised categories. We also found that top line sales increased, providing evidence that the category sales increases are incremental net of any inter-category substitution effects. these effects are widespread and heterogeneous across a large number of categories and retail formats. We estimate the counterfactual that the total sales increase opportunity for the total population of stores of the participating retailers is \$1.9B. We find some directional evidence that more specialized retailers get a higher sales lift, whereas ROAS favored the generalist. Finally we also found that the ROI from this type of advertising has a positive probability of break-even and compared well to acceptable benchmarks. We find that these effects are economically meaningful. Media planners should co-ordinate for the offline effects in the planning and execution of search advertising campaigns and for the fact that there are benefits to adjusting these effects by

categories and retailer type.

Our analysis also provides information about the likely sales lift and return on ad spend that other retailers would obtain in similar experiments. For a retailer considering an increase in search engine advertising, the predictive distributions for $\overline{L}$ and ROAS, with new draws from their respective random-effect distributions for retailers (in Equations 6 and 8 respectively), describe the information available about the likely lift and return. Retailers can simulate from these distributions to inform their advertising spending decisions. This paper also provides a proof of the principle of informational gains from meta-analyses that combine randomized experiments.

The analyses presented here are subject to several limitations arising mainly because of data sharing agreements. This required a variety of workarounds, including estimating standard errors. Our results are robust to these workarounds. Our method for estimating the standard errors of ROAS might be useful more generally in contexts where incremental advertising is not measured at the level of an individual store, which precludes generating an empirical distribution of ROAS.

Our analysis has other limitations. We only estimate the treatment effect of broad keywords in positions 1-3. Also, our results reflect the spending levels in these experiments. Higher spending levels could change the lift and ROAS estimates. While we have estimated and documented heterogeneous treatment effects and conducted some preliminary analysis of variables such as campaign impression share, campaign average position, category strength, and retailer type, further work is needed to obtain causal estimates and understand the impact of these and other variables. Finally while this paper provides causal estimates of cross channel effects of search advertising, and shows the extent to which they generalize to a population of retailers, further work in needed to understand the mechanism that drives these effects.

8 Appendix

8.1 Degrees of freedom of the t-test, for inferring standard errors

We inferred standard errors from reports of the estimated mean and P-value from a one-sided t-test. The degrees of freedom (df) of the t-test was {number of test stores + number of unique

control stores – 2}, which was not available. (Recall that control stores could be matched to more than one test store.) Lacking the actual df, we used for n_{ij} the total number of stores in the chain, which is necessarily larger. The inferred standard error is not sensitive to this choice for df in the range used here, as we now show.

We show this for the combined lift estimate $\overline{L}$. Note that

$$\begin{aligned} & \text{P-value} = \alpha \\ \text{if and only if} \quad & \text{Probability}(\overline{L}/\text{SE} > t_{\alpha,df}) = \alpha, \text{ because the test was one-sided} \\ \text{if and only if} \quad & \text{SE} = \overline{L}/t_{\alpha,df}, \end{aligned} \tag{9}$$

where $t_{\alpha,df}$ is the $100(1-\alpha)$ percentile of the standard t distribution with df degrees of freedom. As df is increased, $t_{\alpha,df}$ decreases, so SE increases. Thus, by using a df value larger than the correct but unknown df, we infer standard errors that are larger than the true values, which is conservative in the sense of acting as if each experiment was less informative than it actually was.

Further, $t_{\alpha,df}$ is insensitive to df in the range we've used; thus, the inferred standard errors are insensitive. For example, $t_{\alpha,df}$ is 1.708 for df = 25; 1.671 for df = 60; 1.658 for df = 120; 1.645 for df = ∞ (i.e., the normal distribution). Over this range of df, $t_{\alpha,df}$ decreases by 3.7%, and the ends of this range differ by far more than our conservative df differs from the actual but unknown df.

8.2 Standard errors for return on ad spending (ROAS)

The data available for this study did not include standard errors for individual experiments' ROAS estimates, and the corresponding P-values were reported in ranges. To avoid excessive conservatism in inferring standard errors for the ROAS estimates, we computed standard errors for ROAS as follows. The derivation has two steps. The first conditions on each experiment's average (over test stores) expected sales (i.e., acts as if average expected sales is known), while the second step removes that conditioning (i.e., treats average expected sales as a random variable).

Conditioning on an experiment's average expected sales $\overline{ES}$, from Section 6.1,

$$\begin{aligned} \overline{L} &= \frac{1}{n\overline{ES}} \sum_s (AS_s - ES_s) \\ \text{and } ROAS &= \frac{1}{ADS} \sum_s (AS_s - ES_s). \end{aligned} \tag{10}$$

Assume that $\sum_s (AS_s - ES_s) \sim N(\mu, \tau^2)$; then

$$\begin{aligned} \overline{L} &\sim N\Big(\frac{\mu}{n\overline{ES}}, \frac{\tau^2}{n^2\overline{ES}^2}\Big) \text{ and} \\ ROAS &\sim N\Big(\frac{\mu}{ADS}, \frac{\tau^2}{ADS^2}\Big). \end{aligned} \tag{11}$$

We have $\hat{\sigma}_L^2$, the estimated variance (square of standard error) for an experiment's lift estimate; we want an estimate for the variance of ROAS, $\hat{\sigma}_R^2$. From Equation (11),

$$\begin{aligned} \hat{\sigma}_L^2 &= \frac{\hat{\tau}^2}{n^2\overline{ES}^2} \\ \text{so that } \hat{\tau}^2 &= n^2\overline{ES}^2\hat{\sigma}_L^2, \\ \text{Therefore } \hat{\sigma}_R^2 &= \frac{\hat{\tau}^2}{ADS^2} \\ &= \frac{n^2\overline{ES}^2}{ADS^2}\hat{\sigma}_L^2. \end{aligned} \tag{12}$$

Conditional on $\overline{ES}$, then, we can estimate the standard error of ROAS. However, $\overline{ES}$ is itself a random variable because test stores were randomly sampled. To remove this conditioning, note that by Equation (10), conditional on the vector of expected sales for the test stores, $\boldsymbol{ES}$,

$$ROAS|\boldsymbol{ES} \sim N\left(\frac{n\overline{ES}}{ADS}\theta, \frac{n^2\overline{ES}^2}{ADS^2}\sigma_L^2\right). \tag{13}$$

where $\theta = E(\tau)$. The variance of $ROAS$ is thus

$$\begin{aligned}
\mathrm{Var}(ROAS) &= \mathrm{Var}[\mathrm{E}(ROAS|\boldsymbol{ES})] + \mathrm{E}[\mathrm{Var}(ROAS|\boldsymbol{ES})] \\
&= \mathrm{Var}\left(\frac{n\overline{ES}}{ADS}\theta\right) + \mathrm{E}\left(\frac{n^2\overline{ES}^2}{ADS^2}\sigma_L^2\right) \\
&= \frac{n^2\theta^2}{ADS^2}\,\mathrm{Var}(\overline{ES}) + \frac{n^2\sigma_L^2}{ADS^2}\,\mathrm{E}(\overline{ES}^2) \\
&= \frac{n^2\theta^2}{ADS^2}\,\mathrm{Var}(\overline{ES}) + \frac{n^2\sigma_L^2}{ADS^2}\{\mathrm{Var}(\overline{ES}) + [\mathrm{E}(\overline{ES})]^2\} \\
&= \frac{n^2\sigma_L^2}{ADS^2}[\mathrm{E}(\overline{ES})]^2 + \frac{n^2(\theta^2+\sigma_L^2)}{ADS^2}\,\mathrm{Var}(\overline{ES}).
\end{aligned} \tag{14}$$

In the last line above, the first term is the conditional estimate of σ_R^2 with $\mathrm{E}(\overline{ES})$ substituted for $\overline{ES}$. Thus the second term is the primary consequence of removing the conditioning on $\overline{ES}$.

The data available for this study included, for each experiment, the sample mean and standard deviation of the ES_s, so we can estimate $\mathrm{Var}(\overline{ES})$ if we can make a plausible assumption about the correlation between ES_s and ES_t for test stores s and t. To do this, we assumed the ES_s were exchangeable with correlation r between each pair of test stores. Therefore,

$$\begin{aligned}
\mathrm{Var}(\overline{ES}) &= \frac{1}{n^2}\,\mathrm{Cov}\left(\sum_{s=1}^{n} ES_s, \sum_{t=1}^{n} ES_t\right) \\
&= \frac{1}{n^2}\big(n\sigma_{ES}^2 + n(n-1)r\sigma_{ES}^2\big) \\
&= \sigma_{ES}^2\Big(\frac{1}{n} + \frac{n-1}{n}r\Big).
\end{aligned} \tag{15}$$

We roughly estimated the correlation coefficient r as $n_{st}/10$, where n_{st} is the number of control stores shared by the s^{th} and t^{th} test stores. Simulations in which control stores were assigned at random to test stores gave the estimates of r in Table 10.

A standard error for each experiment's ROAS was obtained by substituting Table 10's estimated r's and other known quantities (e.g., standard errors for sales lift) into Equation (14). The unconditional and conditional estimates of σ_R^2 (with $\mathrm{E}(\overline{ES})$ substituted for $\overline{ES}$) were very similar. Meta-analyses using the two different estimates gave very similar results, so Section 6.2.3 presents the simpler analysis using the conditional estimates of σ_R^2.

References

Ashish Agarwal, Kartik Hosanagar, and Michael Smith. Location, location, location: An analysis of profitability of position in online advertising markets. *Journal of Marketing Research*, 46 (6):1057–1073, 2008.

Mohan N Babapulle, Lawrence Joseph, Patrick Bélisle, James M Brophy, and Mark J Eisenberg. A hierarchical bayesian meta-analysis of randomised clinical trials of drug-eluting stents. *The Lancet*, 364(9434):583–591, 2004.

Shelly Banjo and Drew FitzGerald. Stores confront new world of reduced shopper traffic. *Wall Street Journal*, January 2014.

Ernst R Berndt. *The practice of econometrics: classic and contemporary*. Addison-Wesley Reading, MA, 1991.

Donald A Berry, Scott M Berry, John McKellar, and Thomas A Pearson. Comparison of the dose-response relationships of 2 lipid-lowering agents: a bayesian meta-analysis. *American heart journal*, 145(6):1036–1045, 2003.

Thomas Blake, Chris Nosko, and Steven Tadelis. Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *NBER Working Paper*, pages 1–26, 2013.

Merrie Brucks. The effects of product class knowledge on information search behavior. *Journal of consumer research*, pages 1–16, 1985.

Glenn R Carroll. Concentration and specialization: Dynamics of niche width in populations of organizations. *American journal of sociology*, pages 1262–1283, 1985.

Rebecca DerSimonian and Raghu Kacker. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary clinical trials*, 28(2):105–114, 2007.

Joseph O Eastlack Jr and Ambar G Rao. Advertising experiments at the campbell soup company. *Marketing Science*, 8(1):57–71, 1989.

John U Farley and Don R Lehmann. Generalizing about market response models: Meta-analysis in marketing. *Lexington, MA: Lexington Books*, 1986.

John U Farley, Donald R Lehmann, and Alan Sawyer. Empirical marketing generalization using meta-analysis. *Marketing Science*, 14(3_supplement):G36–G46, 1995.

Drew FitzGerald. Retail sales on thanksgiving, black friday rose 2.3reports., November 2013. URL `http://online.wsj.com/news/articles/SB10001424052702304017204579230801763930942`.

Gian M Fulgoni and Marie Pauline Morn. Whither the click? how online advertising works. *Journal of Advertising Research*, 49(2):134, 2009.

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

Sander Greenland. Invited commentary: a critical look at some popular meta-analytic methods. *American journal of epidemiology*, 140(3):290–296, 1994.

Michael T Hannan and John Freeman. The population ecology of organizations. *American journal of sociology*, pages 929–964, 1977.

Elizabeth Holmes. Why online retailers like bonobos, boden, athleta mail so many catalogs. *Wall Street Journal*, 2014.

Mingyu Joo, Kenneth C Wilbur, Bo Cowgill, and Yi Zhu. Television advertising and online search. *Management Science*, 60(1):56–73, 2013.

Michael Levy, Barton A Weitz, and Dhruv Grewal. *Retailing management*. Irwin/McGraw-Hill New York, 1998.

Randall Lewis and David Reiley. Retail advertising works! measuring the effects of advertising on sales via a controlled experiment on yahoo!, 2009. *White Paper*, 2009.

Leonard M Lodish, Magid Abraham, Stuart Kalmenson, Jeanne Livelsberger, Beth Lubetkin, Bruce Richardson, and Mary Ellen Stevens. How tv advertising works: A meta-analysis of 389 real world split cable tv advertising experiments. *Journal of Marketing Research*, pages 125–139, 1995.

Prasad A Naik and Kay Peters. A hierarchical marketing communications model of online and offline media synergies. *Journal of Interactive Marketing*, 23(4):288–299, 2009.

Sridhar Narayanan and Kirthi Kalyanam. Position effects and their moderators in search engine advertising: A regression discontinuity approach. *Marketing Science (Forthcoming)*, 2015.

Brian T Ratchford, Myung-Soo Lee, and Debabrata Talukdar. The impact of the internet on information search for automobiles. *Journal of Marketing Research*, pages 193–209, 2003.

Oliver J Rutz and Randolph E Bucklin. From generic to branded: A model of spillover in paid search advertising. *Journal of Marketing Research*, 48(1):87–102, 2011.

Navdeep Sahni. Effect of temporal spacing between advertising exposures: Evidence from an online field experiment. 2011.

Navdeep Sahni. Advertising spillovers: Field experimental evidence and implications for the advertising sales-response curve. 2013.

Raj Sethuraman, Gerard J Tellis, and Richard A Briesch. How well does advertising work? generalizations from meta-analysis of brand advertising elasticities. *Journal of Marketing Research*, 48(3):457–471, 2011.

Stewart Shapiro, Deborah J MacInnis, and Susan E Heckler. The effects of incidental ad exposure on the formation of consideration sets. *Journal of consumer research*, 24(1):94–104, 1997.

Alexander J Sutton and Julian Higgins. Recent developments in meta-analysis. *Statistics in medicine*, 27(5):625–650, 2008.

Anand Swaminathan. The proliferation of specialist organizations in the american wine industry, 1941-1990. *Administrative Science Quarterly*, pages 653–680, 1995.

Anand Swaminathan. Resource partitioning and the evolution of specialist organizations: The role of location and identity in the us wine industry. *Academy of Management Journal*, 44(6): 1169–1185, 2001.

Gerard J Tellis. The price elasticity of selective demand: A meta-analysis of economic models of sales. *Journal of Marketing Research (JMR)*, 25(4), 1988.

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

Peter C Verhoef, Scott A Neslin, and Björn Vroomen. Multichannel customer management: Understanding the research-shopper phenomenon. *International Journal of Research in Marketing*, 24(2):129–148, 2007.

Sha Yang and Anindya Ghose. Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science*, 29(4):602–623, 2010.

Florian Zettelmeyer, Fiona Scott Morton, and Jorge Silva-Risso. How the internet lowers prices: Evidence from matched survey and automobile transaction data. *Journal of marketing research*, 43(2):168–181, 2006.

Figure 1: An Example of Search Engine Advertising on Google.com

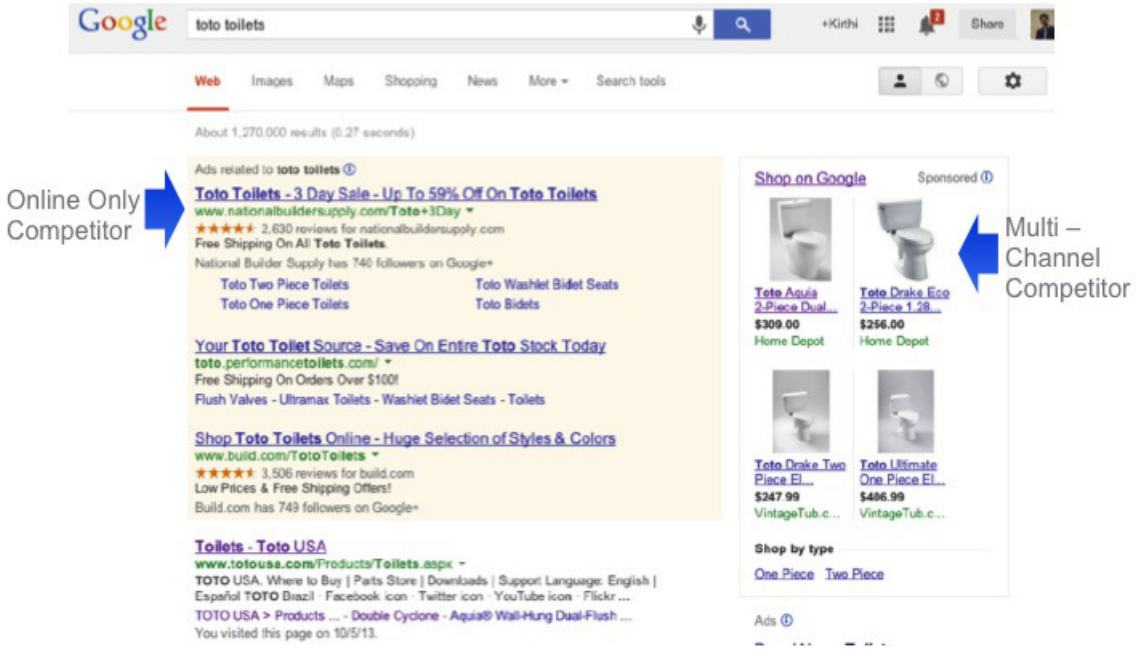

Figure 2: Brick and Mortar Store Sales Variability: Illustrative Example

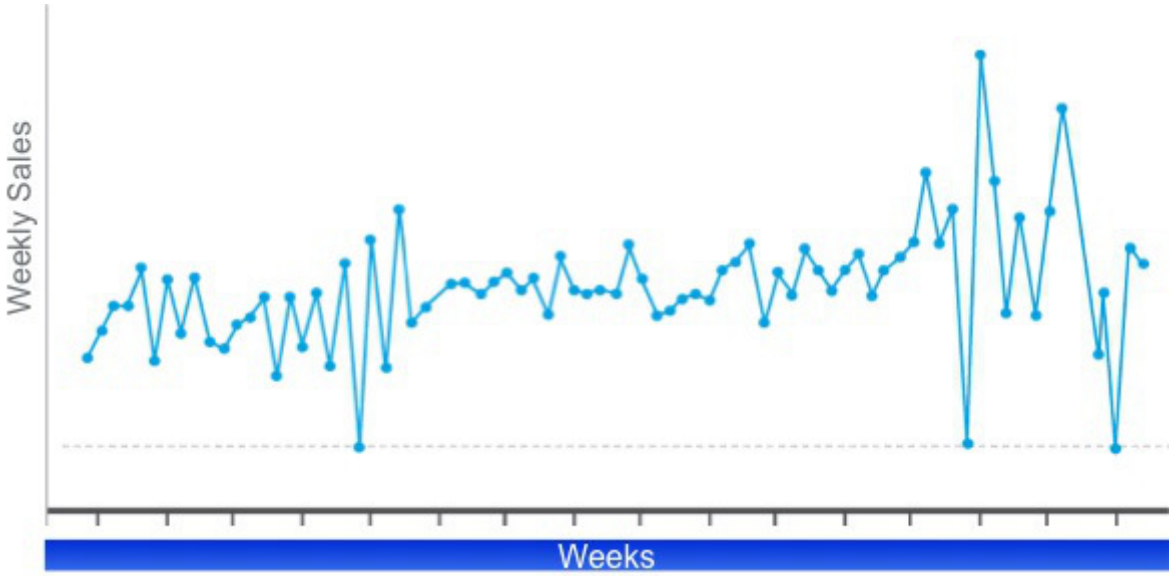

Figure 3: Example of Test and Control Performance

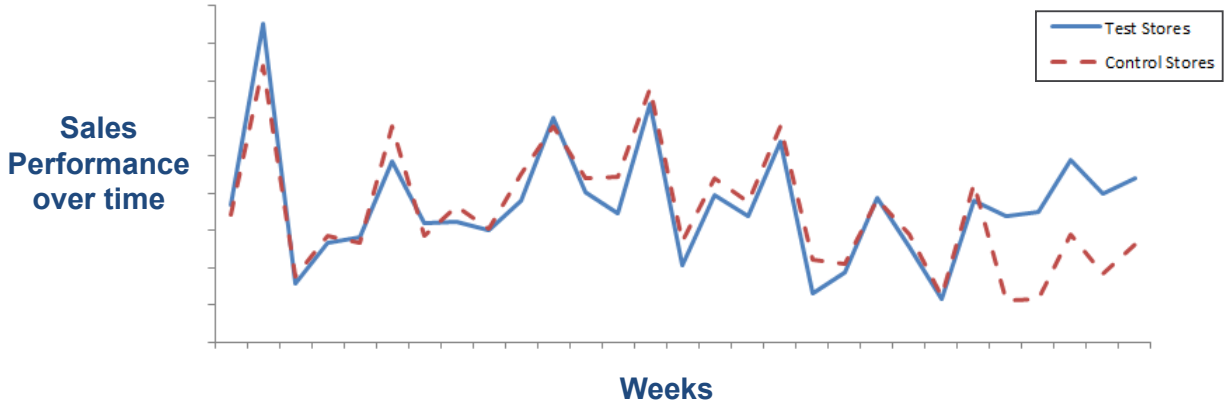

Figure 4: Map of Test and Control Stores for Field Experiment G

DMA	# Sites in Market
Albuq. / Santa Fe	4
Des Moines	4
Santa Barbara	5
Rochester NY	6
Colorado Springs	7
Milwaukee	13
Providence	14
San Diego	20
Baltimore	23
Tampa / St Pete	30

Figure 5: Forest plot: Top-level sales lift. Blue boxes and lines for each study describe the Bayesian estimate and 95% credible interval. Gray boxes and lines describe *reported* estimates and intervals (i.e., estimates ±1.96 × standard error used as inputs to the meta-analysis).

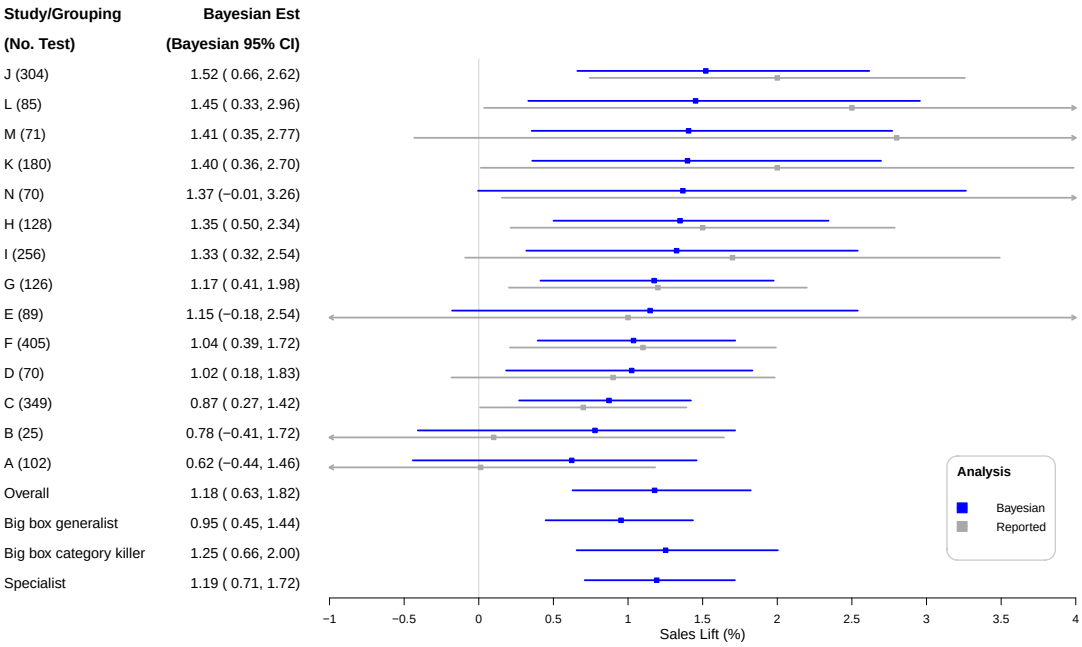

Figure 6: Forest plot: Top-level ROAS. Blue boxes and lines for each study describe the Bayesian estimate and 95% credible interval. Gray boxes describe *reported* estimates and imputed intervals (i.e., estimates ±1.96 × imputed standard error used as inputs to the meta-analysis).

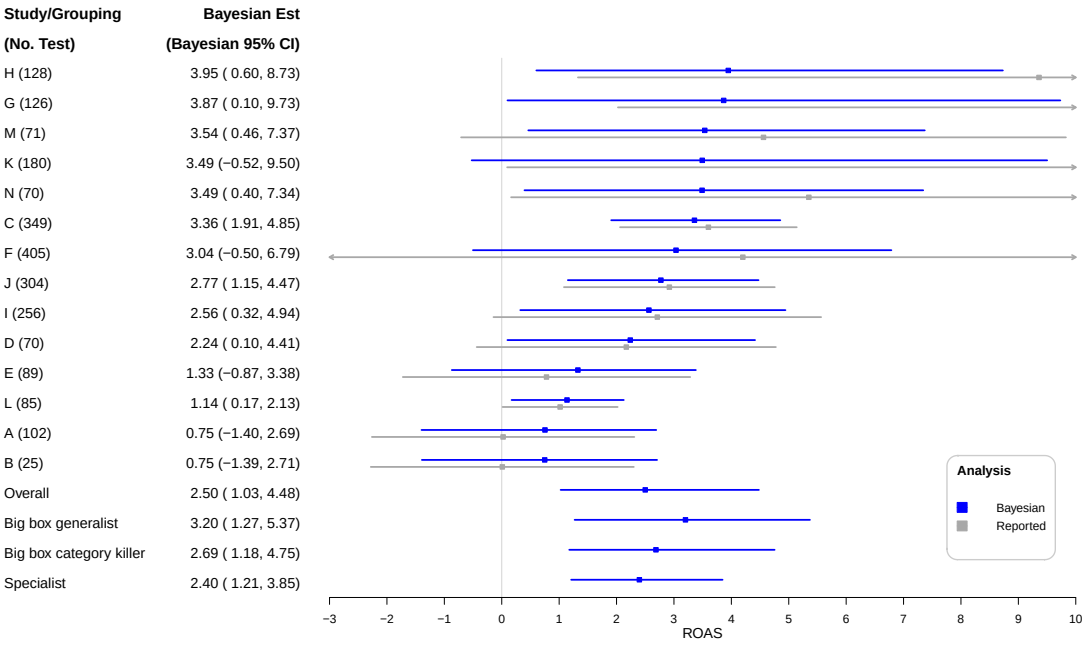

Figure 7: Posterior density plots of overall sales lift (left panel) and overall ROAS (right panel) for top-level data set.

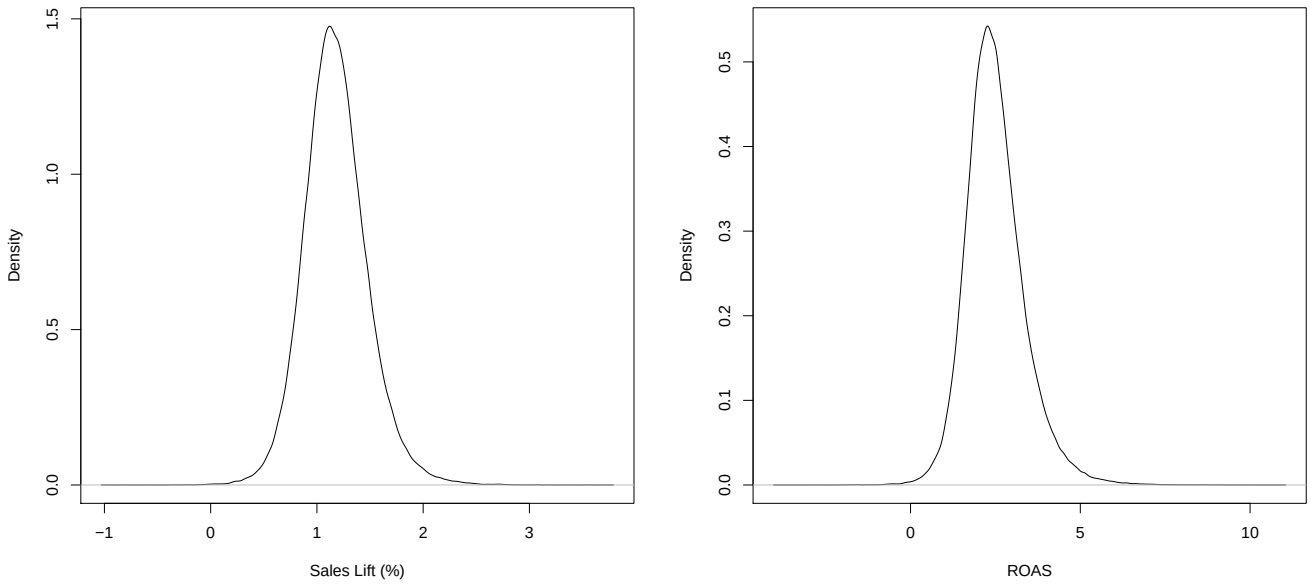

Figure 8: Forest plot: Sales lift for categories, Bayesian estimates and 95% credible intervals. The blue diamond at the bottom is the overall estimate from the meta-analysis with 95% credible interval.

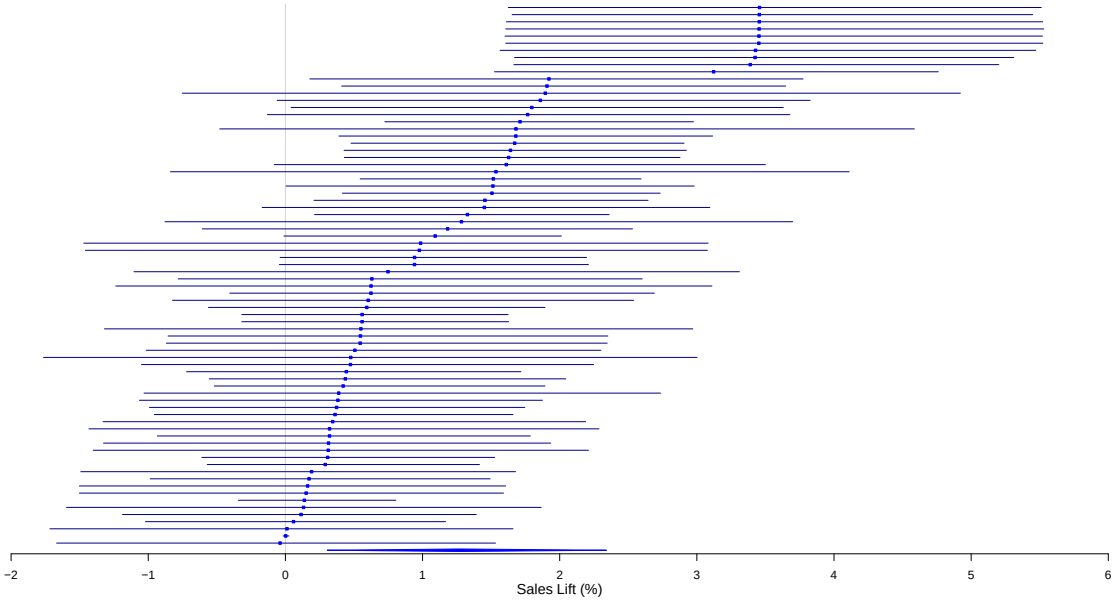

Figure 9: Posterior density plot of overall sales lift for the category-level data set.

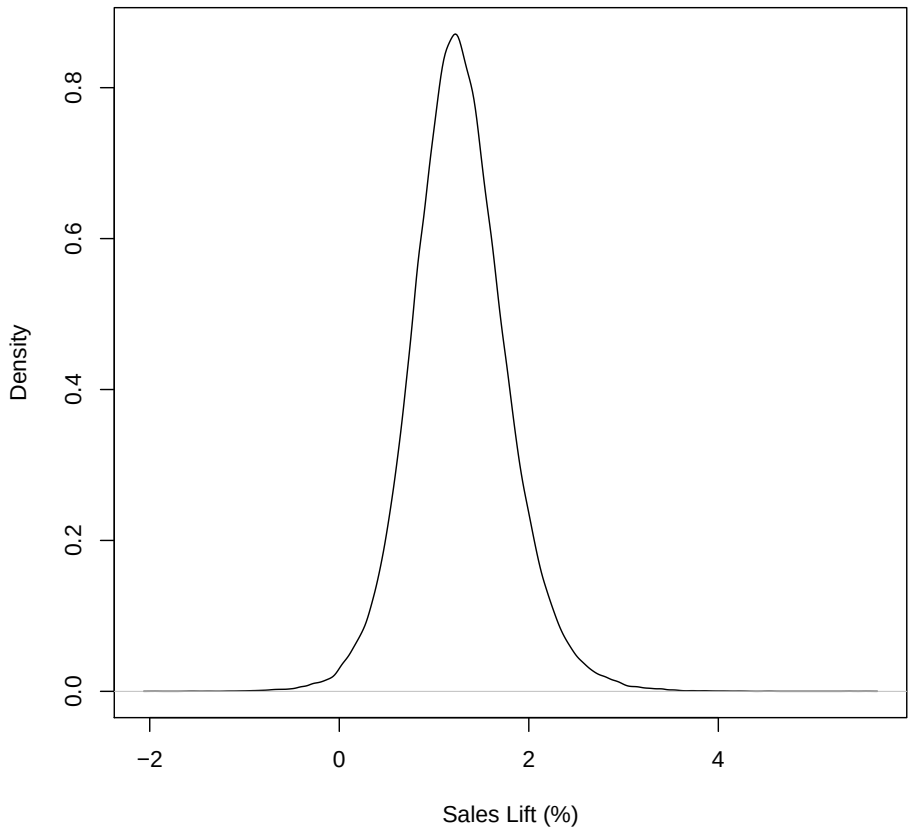

Table 1: Motivations of Participating Retailers

Experiment	Retailer Media Mix *	Retailer Motivations
A	Direct Marketing, Internet.	Rising direct marketing costs. Interested in alternate media to drive offline store traffic.
B	TV, Internet, Magazines, Newspapers.	Needed validation that search advertising can drive offline store traffic.
C	TV, Internet, Newspapers, Magazines.	Validate internal directional measure of web influence on offline sales.
D	TV, Newspapers, Digital, Radio, Other.	Format under competitive pressure. Needed to improve cost effectiveness of marketing. Validate ability of search advertising to drive offline store sales.
E	Direct Marketing, Internet.	Rising direct marketing costs. Interested in alternate media.
F	TV, Internet, Newspapers, Magazines.	Validate internal directional measure of web influence on offline sales.
G	Free Standing Inserts, TV, Radio.	Moving from print to digital marketing. Interested in establishing the value of digital media to the entire organization as opposed to only eCommerce sales.
H	TV, Newspapers, Internet, Radio, Other.	Format under competitive pressure. Needed to improve cost effectiveness of marketing. Validate ability of search advertising to drive offline store sales.
I	TV, Newspapers, Internet.	Retailer focused on new customer acquisition. Currently only giving credit to online sales in search advertising. Interested in expanding credit to offline sales.
J	Direct Marketing, Newspapers, TV, Internet.	Search advertising was driving traffic to web site but no evidence accumulated regarding offline store sales.
K	Media mix data not available.	Exit surveys indicated that significant number of web site visitors were subsequently purchasing in the offline store. Macro data suggest web influenced offline sales is very high in footwear. Need causal estimates of search advertising effectiveness on offline store sales.
L	TV, Free Standing Inserts, Internet.	Retailer needed multiplier derived from causal estimates for offline impact of search advertising.
M	Media Mix data not available.	Under competitive pressure. Needed to improve cost effectiveness of marketing.
N	TV, Free Standing Inserts, Radio, Internet.	Retailer spending on search with some assumptions of offline multiplier but needed causal estimate.
O	TV, Magazines, Newspapers, Internet.	Pursue Co-Op advertising opportunities with manufacturer. Manufacturer initiated the test.Retailers needed offline multiplier estimate for budgeting and purposes and planning.

* Presented in declining order of budget share.

Table 2: Overview of Field Experiments

Experiment*	No. of Test Markets/Stores	No. of Test Weeks/Categories	ADS†($)	Impression Share (%)	Average Position§	Average CTR¶(%)
A	40/102	4/3	$139,000	31.82	3.17	1.64
B	10/25	3/4	$685,000	56.33	2.16	1.31
C	45/349	4/1	$258,000	30.67	2.19	2.51
D	14/70	3/6	$161,000	55.39	1.37	2.41
E	49/89	6/6	$425,000	19.51	3.40	1.23
F	47/405	3/5	$223,000	43.71	2.88	2.83
G	10/126	4/3	$466,000	57.88	2.07	1.13
H	48/128	5/3	$207,000	66.51	1.73	3.69
I	26/256	4/3	$184,000	50.45	2.12	0.72
J	60/304	3/7	$273,879	59.92	2.51	3.96
K	6/180	4/3	$100,000	62.77	1.67	1.99
L	25/85	4/10	$399,514	39.11	2.63	1.75
M	14/71	3/5	$244,000	64.14	1.73	2.41
N	12/70	4/3	$140,000	57.97	1.57	3.36
O	30/246	4/2	$248,000	58.86	1.80	1.48
Total	436/2506	58/76	$4,152,393	–	–	–
Average	29/167	3.9/5.1	$276,826	50.34	2.20	2.16

* Experiment C and F are from the same retailer; Experiment H and M are from the same retailer. The sales lift in Study O is not reported.

† Incremental spending on search ads on category keywords.

§ Average position in test markets.

¶ Average click through rate (CTR) in test markets.

Table 3: Results from Field Experiment G

Level of Aggregation	Test or Non Test	Sales Lift (%)	Incremental sales per store, per week	Total Incremental Sales [1] for 6 weeks (Millions USD)
Test Categories and Department				
Category	Test	5.9%*	$6361.0*	$4.8*
Category	Non Test	−0.9%	NR[3]	NR
Category	Non Test	−0.7%	NR	NR
Department[2]	Test	2.2%*	$5339.6*	$4.04*
Total Store				
Department	Test	2.2%*	$5339.6*	$4.04*
Department	Non Test	0.4%	$1584.0	$1.2
Total Store[4]		1.2%*	$7355.0*	$5.56*

[1] Test duration was 4 weeks. APT measured impact for the test duration plus two additional weeks. Total for the test stores. Numbers reported are in the millions of dollars.

[2] Department is sum of test and non-test categories.

[3] NR = Not reported. Some non-significant estimates were not reported.

[4] Total store is the sum of test and non-test departments. In some field experiments what is reported is not the total store but the total impact for a subset of departments ("Top Level") in the store.

* Significant at 95% or above. APT reports results as significant at 95% or above, 80% or above and 50% or above. In some cases APT reports actual p values.

Table 4: Total store data set results: analysis for sales lift

Experiment/	Reported Sales Lift (%)		Sales Lift (%)	Estimated Random Effect (95% CI)	
Grouping	Estimate	p-value or range*	Bayesian Est (95% CI)	Between-retailer	Within-retailer
A	0.01	0.492	0.62 (−0.44, 1.46)	−0.31 (−1.53, 0.37)	−0.25 (−1.36, 0.37)
B	0.10	0.450	0.78 (−0.41, 1.72)	−0.22 (−1.42, 0.50)	−0.18 (−1.27, 0.49)
C	0.70	0.024	0.87 (0.27, 1.42)	−0.15 (−0.95, 0.45)	−0.16 (−0.97, 0.39)
D	0.90	0.054	1.02 (0.18, 1.83)	−0.09 (−1.00, 0.65)	−0.07 (−0.91, 0.62)
E	1.00	0.271	1.15 (−0.18, 2.54)	−0.02 (−1.09, 1.00)	−0.01 (−0.96, 0.90)
F	1.10	0.008	1.04 (0.39, 1.72)	−0.15 (−0.95, 0.45)	0.01 (−0.70, 0.69)
G	1.20	0.010	1.17 (0.41, 1.98)	0.00 (−0.83, 0.78)	0.00 (−0.75, 0.73)
H	1.50	0.012	1.35 (0.50, 2.34)	0.13 (−0.59, 1.07)	0.04 (−0.71, 0.85)
I	1.70	0.032	1.33 (0.32, 2.54)	0.08 (−0.76, 1.09)	0.07 (−0.71, 0.99)
J	2.00	0.001	1.52 (0.66, 2.62)	0.19 (−0.53, 1.23)	0.16 (−0.51, 1.11)
K	2.00	0.000–0.050	1.40 (0.36, 2.70)	0.12 (−0.70, 1.21)	0.10 (−0.66, 1.08)
L	2.50	0.000–0.050	1.45 (0.33, 2.96)	0.15 (−0.70, 1.36)	0.12 (−0.64, 1.21)
M	2.80	0.047	1.41 (0.35, 2.77)	0.13 (−0.59, 1.07)	0.10 (−0.71, 1.18)
N	8.80	0.000–0.050	1.37 (−0.01, 3.26)	0.11 (−0.88, 1.46)	0.08 (−0.80, 1.24)
Overall			**1.18 (0.63, 1.82)**		
Big box generalist			0.95 (0.45, 1.44)		
Big box category killer			1.25 (0.66, 2.00)		
Speciality			1.19 (0.71, 1.72)		

* p-value for the one-sided t-test in each study.
Note: The Bayesian estimate of sales lift for big box category killer is highest among the three retailer types, while big box generalist is estimated to have the lowest sales lift.

Table 5: Incremental Annualized Sales Opportunity Estimates for The Population of Stores ($M)

Experiment	Mean	SE	2.5%	Median	97.5%
A	5.8	4.6	−4.1	6.3	13.7
B	59.1	40.2	−30.0	63.2	129.9
C	32.7	10.9	10.2	33.1	53.2
D	157.4	62.7	29.3	158.4	280.6
E	7.4	4.2	−1.1	7.3	16.4
F	240.9	77.0	91.7	239.6	399.7
G	1116.1	369.8	395.8	1107.4	1873.4
H	36.7	12.6	13.6	35.8	63.7
I	42.1	17.6	10.0	40.6	80.6
J	33.4	11.0	14.7	32.4	57.3
K	141.3	58.9	36.7	135.5	271.6
L	16.9	7.6	3.9	16.0	34.4
M	35.3	15.0	8.9	33.7	69.3
N	23.5	13.7	0.0	21.8	55.9
Total	**1948.8**	**431.9**	**1116.7**	**1942.1**	**2816.2**

Table 6: Total store data set results: analysis for ROAS

Experiment/ Grouping	Reported ROAS	ROAS Bayesian Est (95% CI)	Estimated Random Effect (95% CI) Between-retailer	Within-retailer
A	0.02	0.75 (−1.40, 2.69)	−0.92 (−4.05, 1.10)	−0.83 (−3.94, 1.12)
B	0.01	0.75 (−1.39, 2.71)	−0.91 (−4.04, 1.11)	−0.84 (−3.97, 1.12)
C	3.60	3.36 (1.91, 4.85)	0.48 (−1.62, 2.74)	0.38 (−1.68, 2.57)
D	2.17	2.24 (0.10, 4.41)	−0.13 (−2.70, 2.11)	−0.14 (−2.64, 2.00)
E	0.78	1.33 (−0.87, 3.38)	−0.61 (−3.53, 1.47)	−0.56 (−3.46, 1.43)
F	4.20	3.04 (−0.50, 6.79)	0.48 (−1.62, 2.74)	0.06 (−2.99, 3.25)
G	12.02	3.87 (0.10, 9.73)	0.73 (−1.90, 5.06)	0.64 (−1.82, 4.75)
H	9.36	3.95 (0.60, 8.73)	0.81 (−1.38, 4.20)	0.63 (−1.78, 4.54)
I	2.71	2.56 (0.32, 4.94)	0.04 (−2.44, 2.43)	0.02 (−2.43, 2.33)
J	2.92	2.77 (1.15, 4.47)	0.16 (−2.11, 2.37)	0.12 (−2.10, 2.26)
K	14.00	3.49 (−0.52, 9.50)	0.53 (−2.31, 4.78)	0.46 (−2.27, 4.46)
L	1.02	1.14 (0.17, 2.13)	−0.71 (−3.26, 1.20)	−0.66 (−3.21, 1.19)
M	4.56	3.54 (0.46, 7.37)	0.81 (−1.38, 4.20)	0.22 (−2.34, 3.15)
N	5.35	3.49 (0.40, 7.34)	0.53 (−1.92, 3.83)	0.46 (−1.90, 3.64)
Overall		**2.50 (1.03, 4.48)**		
Big box generalist		3.20 (1.27, 5.37)		
Big box category killer		2.69 (1.18, 4.75)		
Speciality		2.40 (1.21, 3.85)		

Note: The Bayesian estimate of ROAS for big box generalist is highest among the three retailer types, while speciality is estimated to have the lowest sales lift.

Table 7: Non-aggregated category results: analysis for sales lift

Category	Reported Sales Lift (%)		Sales Lift (%)	Estimated Random Effect (95% CI)			
	Estimate	p-value or range*	Bayesian Est (95% CI)	Between-retailer	Within-retailer	Between-category	Within-category
1	3.90	0.080	1.89 (−0.75, 4.92)	0.17 (−1.34, 2.17)	0.15 (−1.24, 1.99)	0.22 (−1.44, 2.32)	0.08 (−1.01, 1.36)
2	−1.00	0.325	0.98 (−1.46, 3.08)	0.07 (−1.23, 1.45)	0.07 (−1.16, 1.38)	−0.31 (−2.36, 1.25)	−0.13 (−1.45, 0.91)
3	−1.00	0.328	0.99 (−1.47, 3.08)	0.07 (−1.23, 1.45)	0.07 (−1.16, 1.38)	−0.30 (−2.38, 1.26)	−0.12 (−1.46, 0.91)
4	2.40	0.059	1.80 (0.04, 3.63)	0.07 (−1.23, 1.45)	0.07 (−1.16, 1.38)	0.31 (−1.06, 2.08)	0.08 (−0.96, 1.28)
5	2.70	0.033	1.92 (0.18, 3.77)	0.07 (−1.23, 1.45)	0.07 (−1.16, 1.38)	0.40 (−0.94, 2.24)	0.12 (−0.89, 1.36)
6	2.40	0.000–0.500	1.77 (−0.13, 3.68)	0.07 (−1.23, 1.45)	0.07 (−1.16, 1.38)	0.31 (−1.06, 2.08)	0.05 (−1.06, 1.25)
7	2.70	0.000–0.500	1.86 (−0.06, 3.83)	0.07 (−1.23, 1.45)	0.07 (−1.16, 1.38)	0.40 (−0.94, 2.24)	0.05 (−1.05, 1.29)
8	−0.40	0.150–0.500	0.44 (−0.72, 1.72)	−0.15 (−1.42, 0.88)	−0.17 (−1.55, 0.92)	−0.23 (−1.59, 0.98)	−0.27 (−1.54, 0.63)
9	0.30	0.150–0.500	0.56 (−0.32, 1.63)	−0.15 (−1.42, 0.88)	−0.17 (−1.55, 0.92)	−0.23 (−1.59, 0.98)	−0.16 (−1.19, 0.67)
10	0.30	0.150–0.500	0.56 (−0.32, 1.62)	−0.15 (−1.42, 0.88)	−0.17 (−1.55, 0.92)	−0.23 (−1.59, 0.98)	−0.16 (−1.19, 0.67)
11	1.70	0.000–0.050	0.94 (−0.04, 2.21)	−0.15 (−1.42, 0.88)	−0.17 (−1.55, 0.92)	−0.23 (−1.59, 0.98)	0.22 (−0.60, 1.36)
12	1.70	0.000–0.050	0.94 (−0.04, 2.20)	−0.15 (−1.42, 0.88)	−0.17 (−1.55, 0.92)	−0.23 (−1.59, 0.98)	0.23 (−0.60, 1.36)
13	−2.12	0.200–0.500	0.55 (−1.32, 2.97)	−0.21 (−1.88, 1.11)	−0.18 (−1.75, 1.04)	−0.27 (−1.99, 1.21)	−0.05 (−1.34, 1.09)
14	−0.91	0.200–0.500	0.48 (−1.76, 3.00)	−0.21 (−1.88, 1.11)	−0.18 (−1.75, 1.04)	−0.29 (−2.34, 1.33)	−0.11 (−1.44, 0.94)
15	−0.30	0.200–0.500	0.39 (−1.03, 2.74)	−0.21 (−1.88, 1.11)	−0.18 (−1.75, 1.04)	−0.27 (−1.99, 1.21)	−0.22 (−1.58, 0.73)
16	3.33	0.200–0.500	0.62 (−1.24, 3.11)	−0.21 (−1.88, 1.11)	−0.18 (−1.75, 1.04)	−0.27 (−1.99, 1.21)	0.02 (−1.17, 1.27)
17	7.58	0.000–0.050	0.75 (−1.10, 3.31)	−0.21 (−1.88, 1.11)	−0.18 (−1.75, 1.04)	−0.27 (−1.99, 1.21)	0.14 (−0.92, 1.54)
18	9.70	0.000–0.050	1.28 (−0.88, 3.70)	−0.21 (−1.88, 1.11)	−0.18 (−1.75, 1.04)	0.30 (−0.79, 1.58)	0.11 (−0.98, 1.48)
19	2.20	0.169	1.54 (−0.84, 4.11)	0.08 (−1.50, 1.87)	0.07 (−1.40, 1.72)	0.08 (−1.34, 1.61)	0.04 (−1.08, 1.25)

Continued on next page

* p-value for the one-sided t-test in each category.

Table 7 – continued from previous page

Category	Reported Sales Lift (%)		Sales Lift (%)	Estimated Random Effect (95% CI)			
	Estimate	p-value or range*	Bayesian Est (95% CI)	Between-retailer	Within-retailer	Between-category	Within-category
20	0.40	0.272	1.09 (−0.01, 2.01)	−0.15 (−1.42, 0.88)	0.01 (−1.17, 1.17)	0.30 (−0.79, 1.58)	−0.33 (−1.52, 0.4)
21	1.00	0.137	1.33 (0.21, 2.36)	−0.15 (−1.42, 0.88)	0.01 (−1.17, 1.17)	0.30 (−0.79, 1.58)	−0.10 (−1.18, 0.79)
22	1.80	0.013	1.52 (0.54, 2.59)	−0.15 (−1.42, 0.88)	0.01 (−1.17, 1.17)	0.30 (−0.79, 1.58)	0.09 (−0.80, 1.10)
23	1.90	0.041	1.51 (0.42, 2.73)	−0.15 (−1.42, 0.88)	0.01 (−1.17, 1.17)	0.30 (−0.79, 1.58)	0.08 (−0.87, 1.18)
24	2.60	0.001	1.71 (0.72, 2.98)	−0.15 (−1.42, 0.88)	0.01 (−1.17, 1.17)	0.30 (−0.79, 1.58)	0.28 (−0.50, 1.50)
25	−0.50	0.200–0.500	0.32 (−0.93, 1.79)	−0.07 (−1.48, 1.19)	−0.06 (−1.37, 1.15)	−0.62 (−2.28, 0.49)	−0.20 (−1.52, 0.76)
26	2.00	0.200–0.500	0.55 (−0.87, 2.35)	−0.07 (−1.48, 1.19)	−0.06 (−1.37, 1.15)	−0.62 (−2.28, 0.49)	0.03 (−1.14, 1.25)
27	2.00	0.200–0.500	0.55 (−0.86, 2.35)	−0.07 (−1.48, 1.19)	−0.06 (−1.37, 1.15)	−0.62 (−2.28, 0.49)	0.03 (−1.13, 1.25)
28	−0.70	0.050–0.200	0.06 (−1.02, 1.17)	−0.07 (−1.48, 1.19)	−0.06 (−1.37, 1.15)	−0.62 (−2.28, 0.49)	−0.46 (−1.94, 0.32)
29	1.00	0.050–0.200	0.59 (−0.56, 1.89)	−0.07 (−1.48, 1.19)	−0.06 (−1.37, 1.15)	−0.62 (−2.28, 0.49)	0.08 (−0.90, 1.08)
30	2.40	0.000–0.050	1.61 (−0.08, 3.50)	−0.07 (−1.48, 1.19)	−0.06 (−1.37, 1.15)	0.32 (−1.08, 2.06)	0.15 (−0.82, 1.39)
31	5.90	0.000–0.050	1.68 (−0.48, 4.59)	−0.07 (−1.48, 1.19)	−0.06 (−1.37, 1.15)	0.39 (−1.21, 2.73)	0.15 (−0.89, 1.55)
32	9.80	0.000–0.050	0.63 (−0.78, 2.60)	−0.07 (−1.48, 1.19)	−0.06 (−1.37, 1.15)	−0.62 (−2.28, 0.49)	0.11 (−0.97, 1.48)
33	13.00	0.000–0.050	0.60 (−0.82, 2.54)	−0.07 (−1.48, 1.19)	−0.06 (−1.37, 1.15)	−0.62 (−2.28, 0.49)	0.09 (−1.04, 1.42)
34	1.50	0.000–0.050	1.67 (0.48, 2.91)	0.09 (−1.20, 1.50)	0.09 (−1.12, 1.46)	0.30 (−0.79, 1.58)	−0.08 (−1.19, 0.87)
35	3.60	0.000–0.050	1.91 (0.41, 3.65)	0.09 (−1.20, 1.50)	0.09 (−1.12, 1.46)	0.30 (−0.79, 1.58)	0.15 (−0.84, 1.47)
36	2.10	0.000–0.050	3.12 (1.52, 4.76)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	−0.24 (−1.63, 0.66)
37	3.70	0.000–0.050	3.39 (1.67, 5.20)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	0.02 (−1.10, 1.18)
38	4.30	0.000–0.050	3.42 (1.67, 5.31)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	0.06 (−1.04, 1.27)
39	5.70	0.000–0.050	3.46 (1.65, 5.45)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	0.09 (−0.99, 1.38)

Continued on next page

* p-value for the one-sided t-test in each category.

Table 7 – continued from previous page

Category	Reported Sales Lift (%)		Sales Lift (%)	Estimated Random Effect (95% CI)			
	Estimate	p-value or range*	Bayesian Est (95% CI)	Between-retailer	Within-retailer	Between-category	Within-category
40	8.70	0.000–0.050	3.46 (1.63, 5.51)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	0.09 (−1.01, 1.42)
41	9.30	0.000–0.050	3.46 (1.61, 5.52)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	0.09 (−1.03, 1.43)
42	9.30	0.000–0.050	3.45 (1.61, 5.53)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	0.09 (−1.03, 1.41)
43	9.40	0.000–0.050	3.45 (1.60, 5.52)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	0.09 (−1.03, 1.43)
44	9.60	0.000–0.050	3.45 (1.61, 5.52)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	0.09 (−1.03, 1.41)
45	16.90	0.000–0.050	3.43 (1.57, 5.47)	0.66 (−0.53, 2.92)	0.56 (−0.54, 2.73)	0.87 (−0.53, 3.10)	0.06 (−1.08, 1.36)
46	−0.80	0.200–0.500	0.17 (−0.99, 1.49)	−0.31 (−1.92, 0.85)	−0.27 (−1.80, 0.83)	−0.40 (−1.91, 0.85)	−0.13 (−1.40, 0.88)
47	0.00	0.200–0.500	0.00 (−0.02, 0.03)	−0.31 (−1.92, 0.85)	−0.27 (−1.80, 0.83)	−0.40 (−1.91, 0.85)	−0.30 (−1.16, 0.24)
48	0.10	0.200–0.500	0.14 (−0.34, 0.81)	−0.31 (−1.92, 0.85)	−0.27 (−1.80, 0.83)	−0.40 (−1.91, 0.85)	−0.16 (−1.07, 0.49)
49	0.40	0.200–0.500	0.29 (−0.57, 1.41)	−0.31 (−1.92, 0.85)	−0.27 (−1.80, 0.83)	−0.40 (−1.91, 0.85)	−0.01 (−1.00, 0.91)
50	0.50	0.200–0.500	0.31 (−0.61, 1.53)	−0.31 (−1.92, 0.85)	−0.27 (−1.80, 0.83)	−0.40 (−1.91, 0.85)	0.01 (−1.01, 0.98)
51	−3.00	0.098	0.11 (−1.19, 1.39)	−0.31 (−1.92, 0.85)	−0.27 (−1.80, 0.83)	−0.40 (−1.91, 0.85)	−0.18 (−1.60, 0.80)
52	1.40	0.142	0.42 (−0.52, 1.89)	−0.31 (−1.92, 0.85)	−0.27 (−1.80, 0.83)	−0.40 (−1.91, 0.85)	0.12 (−0.83, 1.28)
53	2.10	0.113	0.44 (−0.55, 2.04)	−0.31 (−1.92, 0.85)	−0.27 (−1.80, 0.83)	−0.40 (−1.91, 0.85)	0.14 (−0.85, 1.40)
54	8.10	0.001	0.62 (−0.41, 2.69)	−0.31 (−1.92, 0.85)	−0.27 (−1.80, 0.83)	−0.40 (−1.91, 0.85)	0.33 (−0.61, 1.99)
55	1.60	0.200–0.500	1.51 (0.00, 2.98)	0.06 (−1.21, 1.38)	0.05 (−1.16, 1.32)	0.13 (−1.12, 1.48)	0.00 (−1.17, 1.18)
56	−1.70	0.050–0.200	1.18 (−0.61, 2.53)	0.06 (−1.21, 1.38)	0.05 (−1.16, 1.32)	0.13 (−1.12, 1.48)	−0.33 (−1.90, 0.58)
57	1.20	0.050–0.200	1.45 (0.21, 2.64)	0.06 (−1.21, 1.38)	0.05 (−1.16, 1.32)	0.13 (−1.12, 1.48)	−0.05 (−1.12, 0.92)
58	1.40	0.050–0.200	1.45 (−0.17, 3.09)	0.06 (−1.21, 1.38)	0.05 (−1.16, 1.32)	0.08 (−1.34, 1.61)	−0.01 (−1.10, 1.06)
59	2.10	0.000–0.050	1.63 (0.43, 2.88)	0.06 (−1.21, 1.38)	0.05 (−1.16, 1.32)	0.13 (−1.12, 1.48)	0.12 (−0.80, 1.25)

Continued on next page

* p-value for the one-sided t-test in each category.

Table 7 – continued from previous page

Category	Reported Sales Lift (%)		Sales Lift (%)	Estimated Random Effect (95% CI)			
	Estimate	p-value or range*	Bayesian Est (95% CI)	Between-retailer	Within-retailer	Between-category	Within-category
60	2.20	0.000–0.050	1.64 (0.43, 2.92)	0.06 (−1.21, 1.38)	0.05 (−1.16, 1.32)	0.13 (−1.12, 1.48)	0.13 (−0.79, 1.29)
61	2.90	0.000–0.050	1.68 (0.39, 3.12)	0.06 (−1.21, 1.38)	0.05 (−1.16, 1.32)	0.13 (−1.12, 1.48)	0.17 (−0.77, 1.46)
62	−4.80	0.200–0.500	0.31 (−1.33, 1.93)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.24 (−1.75, 1.02)	−0.02 (−1.28, 1.18)
63	−1.30	0.200–0.500	0.13 (−1.60, 1.86)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.37 (−2.10, 0.91)	−0.07 (−1.33, 1.03)
64	0.50	0.200–0.500	0.36 (−0.96, 1.66)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.24 (−1.75, 1.02)	0.03 (−0.99, 1.07)
65	0.60	0.200–0.500	0.37 (−0.99, 1.75)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.24 (−1.75, 1.02)	0.04 (−1.00, 1.11)
66	0.90	0.200–0.500	0.38 (−1.06, 1.87)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.24 (−1.75, 1.02)	0.05 (−1.04, 1.20)
67	−3.10	0.050–0.200	0.19 (−1.49, 1.68)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.24 (−1.75, 1.02)	−0.15 (−1.53, 0.89)
68	−2.50	0.050–0.200	0.16 (−1.50, 1.61)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.24 (−1.75, 1.02)	−0.17 (−1.57, 0.83)
69	−2.30	0.050–0.200	0.15 (−1.50, 1.59)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.24 (−1.75, 1.02)	−0.19 (−1.58, 0.80)
70	−2.10	0.050–0.200	0.01 (−1.72, 1.66)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.37 (−2.10, 0.91)	−0.19 (−1.60, 0.78)
71	−1.50	0.050–0.200	−0.04 (−1.67, 1.53)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.37 (−2.10, 0.91)	−0.24 (−1.64, 0.65)
72	2.40	0.050–0.200	0.34 (−1.33, 2.19)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.37 (−2.10, 0.91)	0.14 (−0.89, 1.47)
73	3.60	0.050–0.200	0.31 (−1.40, 2.21)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.37 (−2.10, 0.91)	0.11 (−0.98, 1.43)
74	6.50	0.000–0.050	0.51 (−1.01, 2.30)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.24 (−1.75, 1.02)	0.17 (−0.86, 1.62)
75	8.20	0.000–0.050	0.47 (−1.05, 2.25)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.24 (−1.75, 1.02)	0.14 (−0.92, 1.54)
76	9.70	0.000–0.050	0.32 (−1.43, 2.29)	−0.38 (−2.06, 0.73)	−0.32 (−1.93, 0.72)	−0.37 (−2.10, 0.91)	0.12 (−0.97, 1.50)
Overall			**1.27 (0.30, 2.34)**				

* p-value for the one-sided t-test in each category.

Table 8: Aggregated (broad) category results: analysis for sales lift

label	broad categorization	Bayesian Est (95% CI)
1	Small appliances	1.49 (−0.40, 3.72)
2	Men's apparel	0.96 (−1.27, 2.78)
3	Footwear	0.98 (−1.25, 2.79)
4	Cosmetics	1.58 (−0.01, 3.48)
5	Women's apparel	1.66 (0.08, 3.64)
6	Consumer electronics	1.05 (−0.34, 2.51)
7	Toys	1.00 (−0.80, 2.75)
8	Video games	0.99 (−1.22, 2.87)
9	Baby products	1.57 (0.37, 2.92)
10	Furniture	1.35 (−0.23, 3.03)
11	Home improvement	0.66 (−1.04, 2.18)
12	Large appliances	1.59 (−0.01, 3.46)
13	Kitchen & bath	1.66 (−0.19, 4.13)
14	Pets	2.13 (0.43, 4.33)
15	Service activation	0.89 (−0.60, 2.45)
16	Home furnishings	1.40 (0.02, 2.86)
17	Sporting goods	1.05 (−0.54, 2.60)
18	Sports wear	0.92 (−0.91, 2.53)

Table 9: Overall sales lift (%) for category-level experiments under the four alternatives considered for p-values reported as intervals.

	Posterior quantities			
Alternative	Mean	SD	2.5%	97.5%
Uniform prior	**1.27**	**0.51**	**0.30**	**2.34**
Pessimist	1.77	0.55	0.77	2.96
Optimist	2.07	0.73	0.63	3.54
Mid-point	1.23	0.50	0.29	2.27
SAPV	1.32	0.53	0.30	2.42

Table 10: Estimated correlation

	A	B	C	D	E	F	G	H
$\hat{r}$	0.0596	0.1216	0.0079	0.0065	0.0903	0.0073	0.0054	0.0758
	I	J	K	L	M	N	O	
$\hat{r}$	0.0053	0.0151	0.0055	0.0261	0.0530	0.0097	0.0131	