Experimentation and Performance in Advertising: An Observational Survey of Firm Practices on Facebook[1]

Julian Runge, Steve Geinitz & Simon Ejdemyr
(*Facebook, Marketing Science Research*)

FORTHCOMING IN *EXPERT SYSTEMS WITH APPLICATIONS*

ABSTRACT

It is widely assumed that firms experiment with their online advertising to identify more profitable approaches to then increase their investment in more profitable advertising, increasing their overall performance. Generalizable evidence on the actual use of such experiment-based learning by firms is sparse. The study herein addresses this shortcoming – detailing the extent to which large advertisers are utilizing experimentation along with evidence on the benefits of doing so. The findings are gleaned from firms' marketing and experimentation practices on a large online advertising platform and indicate that, while experimentation is utilized by some, adoption is far from perfect. Among the few firms making use of experiments, even fewer invest a significant share of their advertising spend in experimentation. This finding is surprising in light of broadly assumed regular experimentation by firms. Experimenting firms further experience higher concurrent and subsequent performance, suggesting that leading firms indeed successfully use experiment-based learning to improve their advertising policies – and that many firms may fall short of their potential by not (yet) using experiments in advertising.

KEYWORDS: Firm experimentation; exploration; online advertising; organizational learning; reinforcement learning

[1] The authors wish to thank Brett Gordon, Harikesh Nair, Robert Moakler, Sophie Macintyre, Hannah Pavalow, Tom Cunningham and several anonymous referees for valuable comments and suggestions. The authors further acknowledge data and financial support by Facebook Inc. to be able to conduct this study. All thoughts, opinions and errors are the authors own. Please direct comments and questions at jrunge@fb.com.

1 INTRODUCTION

Online advertising is a main pillar of web-based economic activity and opens up new frontiers for the study of firms' use of data-driven decision-making (Brynjolfsson and McElheran 2016; Jankowski et al. 2016; Rao and Simonov 2019). Many online advertising platforms offer "experimentation-as-a-service" (Lin et al. 2019) that commonly allows advertisers to run experiments to assess the causal impact of advertising treatments on performance measures (Johnson, Lewis and Nubbemeyer 2017; Kalyanam et al. 2018; Gordon et al. 2019). In an organizational reinforcement learning framing (March 1991; Sutton and Barto 2018), experimentation enables firms to explore – to test new and innovative advertising approaches – to then use the generated insights to reinforce their advertising policies, i.e., to adopt innovative approaches and allocate advertising investment accordingly and more profitably (Fischer et al. 2011; Schwartz, Fader and Bradlow 2017; García-Galicia, Carsteanu and Clempner 2019). Profit-maximizing firms should hence use experimentation and reinforce their advertising policies with experimental results.

While this perspective seems to be widely embraced by scholarship (Kohavi et al. 2009, 2013; Schwartz, Fader and Bradlow 2017; Gordon et al. 2019; Kolsarici, Vakratsas and Naik 2020; Thomke 2020), generalizable evidence on the use of experimentation by firms in advertising is sparse. Seminal studies investigate advertising experiments, but a survey of a wide representative set of firms seems to be missing.[2] On this background, the present paper studies the advertising practice of thousands of large firms across 15 industries on a leading advertising platform in two recent years. Using detailed observations of firms' experimentation and advertising behavior in

[2] Studies leaning in this direction are meta analyses using several experiments by different advertisers, e.g., Lodish et al. (1995) analyze close to 400 TV split tests, Kalyanam et al. (2018) and Gordon et al. (2019) each analyze 15 experiments by different advertisers, and Lin et al. (2019) investigate 16 large-scale advertising experiments.

addition to select meta variables, the authors aspire to answer the following questions: What share of firms uses experimentation? How much of their budget do experimenting firms allocate to experimentation? Are there patterns of positive association between the use of experimentation and performance?

Results show that, while some firms use experimentation, there is variation across industries and, overall, its adoption is not as widespread and regular as suggested by scholarly accounts of firm behavior (Schwartz, Fader and Bradlow 2017; Gordon et al. 2019; Kolsarici, Vakratsas and Naik 2020; Thomke 2020). They further show that concurrent and subsequent performance on the platform is higher among firms using experimentation, suggesting that leading firms successfully use experiment-based learning to improve their advertising policies and that many firms may fall short of their potential by not (yet) using experiments in advertising.

The paper proceeds by providing further conceptual background, before presenting the data, method and results. Finally, it discusses and concludes.

2 CONCEPTUAL BACKGROUND

2.1 Learning through exploration

Computational accounts of organizational learning emphasize how an organization must explore new innovative approaches while exploiting existing certainties to sustainably succeed (March 1991). For the case of marketing organizations, Kolsarici, Vakratsas and Naik (2020) suggest that managers employ a dual control approach. Specifically, they "envision a manager who finds the control policy (i.e., ad spending over time) and learns about the uncertain model parameters simultaneously" (p. 4). Similarly, reinforcement learning as one of the key subfields of machine learning highlights how an artificial agent needs to explore to identify successful policies for

"exploitation" (Schwartz, Fader and Bradlow 2017; Sutton and Barto 2018). A crucial issue related to these accounts of learning is how to trade off exploration and exploitation for best longer-term performance of the agent or organization (Mc Namara and Baden-Fuller 2007; Sutton and Barto 2018; Kolsarici, Vakratsas and Naik 2020). One widely adopted method to control the trade-off is epsilon-based learning, specifically epsilon greedy (Sutton and Barto 2018): With probability $1-\in$ exploitation, i.e., the best-known policy to date, is chosen in newly arriving decision situations. And with probability $\in$, one of the available actions is chosen at random to enable the decision-making entity to learn about (explore) its environment and reinforce its decision policy.

In the empirical study of organizational learning, such an epsilon-based method is most often too simplistic to reflect how organizations solve the trade-off (March 1991; Mc Namara and Baden-Fuller 2007; Kolsarici, Vakratsas and Naik 2020). A key tenet however pertains to exploration being a necessary condition for any effective learning to take place (March 1991; Kane and Alavi 2007). Speaking to this rationale, the present study zeroes in on firm experimentation as an indicator of exploration and explicit learning about the environment and studies its associations with firm performance in the environment.

2.2 Information technology and firm learning

Digitization and the proliferation of data-driven decision-making have opened up new frontiers for the empirical study of firm learning and related outcomes (Brynjolfsson and McElheran 2016). Along these lines, Brynjolfsson, Hitt and Kim (2011) find the adoption of data-driven decision-making to increase firm profits, and a very recent study by Koning, Hasan and Chatterji (2019) suggests that explicit learning and exploration by means of experimentation is linked to startups' economic performance. Using data from the same source, Ghosh, Thomke and Pourkhalkhali

(2020) consider the effects of hierarchy in explaining the association between experimentation and performance. Rana and Oliveira (2015), Schwartz, Fader and Bradlow (2017) and García-Galicia, Carsteanu and Clempner (2019) study lower-level reinforcement learning systems that are used by firms to learn about a concrete environment and become more profitable in said environment.

In any of the described accounts of learning, the firm – either through managers or information technology, or a mix of both – executes some notion of exploration to learn about more profitable ways to interact with the environment. A long stream of literature studies how such endeavors by the firm can be executed through decision support and expert systems (Turban and Watkins 1986; Liao 2005). Vis-à-vis this literature, Figure 1 summarizes a reinforcement learning process that firms can use in their advertising operations: The firm continuously innovates and improves its advertising policies by experimenting with new ideas (Kolsarici, Vakratsas and Naik 2020; Thomke 2020). The process can be described as an expert and decision support system. Some firms may execute the process by means of the former in a fully automated fashion (see, e.g., Schwartz, Fader and Bradlow 2017), others may execute it by providing decision support to human decision-makers (Kolsarici, Vakratsas and Naik 2020), yet others may use a combination of both (Basu and Batra 1988; Zia and Rao 2019). In any case, the firm will experiment with a new policy in each period t (where t can be a day, a week, a month or another suitable amount of time), adopt the most profitable policy in period $t+1$, innovate, and experiment again.

[INSERT FIGURE 1 ABOUT HERE]

2.3 Experimentation in advertising

On online advertising platforms, advertising space is usually allocated to advertisers in an auction, based on their respective bid (Johnson, Lewis and Nubbemeyer 2017; Gordon et al. 2019; Lin et al. 2019). Experiments are implemented in different flavors, e.g., through PSA-style or "ghost ad" holdout treatments (see Johnson, Lewis and Nubbemeyer (2017) for an overview), generally providing the ability to learn with high precision due to full randomization and a holdout group without any exposure to the advertising policy (Johnson, Lewis and Nubbemeyer 2017; Gordon et al. 2019; Lin et al. 2019). This style of experimentation is equivalent to random exploration (also called off-policy learning) of (contextual) bandit approaches in reinforcement learning (Sutton and Barto 2018). Experiments allow advertisers to obtain an unbiased estimate of the incremental performance effect of an advertising policy (Zantedeschi, Feit and Bradlow 2016; Gordon et al. 2019), enabling them to learn accurately and to adjust prior advertising allocation policies towards higher performance (Manchanda, Rossi and Chintagunta 2004; Lavrakas 2010; Mason 2019). Firms who run experiments hence can be expected to better identify profitable advertising approaches, allocate advertising budget accordingly, and achieve superior advertising performance (Fischer et al. 2011; Schwartz, Fader and Bradlow 2017), similar to an applied reinforcement learning system (García-Galicia, Carsteanu and Clempner 2019).

Rational profit-maximizing firms should hence adopt experimentation to learn effectively and improve their advertising policies using an expert system as shown in Figure 1. It has been reported however that, in reality, many advertisers are slow to adopt experimentation (Lavrakas 2010; Kim, Kwon and Chang 2011; Blake, Nosko and Tadelis 2015; Gordon et al. 2019; Rao and Simonov 2019). One key reason for such slow adoption may be organizational inertia (Hannan and Freeman 1984). Many advertisers historically operated on – non-digital – channels where it was costly to run experiments for exploration (Abraham and Lodish 1990; Lodish et al. 1995; Kim, Kwon and

Chang 2011). While, in such settings, media mix models (MMMs) and last-click attribution have established themselves as powerful tools for marketing decision support (Manchanda, Rossi and Chintagunta 2004; Lavrakas 2010; Kim, Kwon and Chang 2011), they can regularly lead to biased measurements of advertising effectiveness (Gordon et al. 2019). Especially on "newer" online channels, such approaches are hence often not effective for learning about the environment (Kim, Kwon and Chang 2011; Blake, Nosko and Tadelis 2015).

A further inhibiting factor for the adoption of learning by means of experiments in advertising may be the perceived cost associated with holdout groups (Jankowski et al. 2016; Berman et al. 2018). Managers want advertising treatments to reach the full set of potential customers (Lavrakas 2010). This consideration leads back to the exploration-exploitation dilemma discussed above: While experimentation can indeed be costly in the short-term when the tested approach is profitable, it can yield crucial insight on non-profitable or more profitable advertising approaches (Eastlack and Rao 1989; Blake, Nosko and Tadelis 2015; Schwartz, Fader and Bradlow 2017). Precise experimental estimates can further be used to calibrate and validate observational decision support models such as MMMs or to evaluate the appropriateness of attribution models, increasing overall measurement precision (Mason 2019), in turn making existing organizational information technology more performant (Kane and Alavi 2007; Fischer et al. 2011).

Based on this account, the present study contributes to a large body of literature on the evaluation of different advertising policies and firm learning in advertising (Basu and Batra 1988; Zia and Rao 2019; Kolsarici, Vakratsas and Naik 2020). While a learning-based positive association between experimentation and advertising's economic performance has been proposed (Eastlack and Rao 1989; Schwartz, Fader and Bradlow 2017; Kolsarici, Vakratsas and Naik 2020), such a link has not been established more widely. Large-scale advertising platforms that provide

"experimentation-as-a-service" solutions (Lin et al. 2019) are a unique study ground for this assertion as a large number of advertisers can use experiments at their discretion and measures of their comparative performance are observable (Gordon et al. 2019).

2.4 Research questions

Speaking to literature on firm exploration (March 1991; Kane and Alavi 2007), reinforcement learning and related expert and decision support systems (Turban and Watkins 1986; Liao 2005; Rana and Oliveira 2015; García-Galicia, Carsteanu and Clempner 2019; Sutton and Barto 2018), data-driven decision-making (Brynjolfsson and McElheran 2016; Brynjolfsson, Hitt and Kim 2011), and the use of experimentation in innovation (Kohavi et al. 2009, 2013; Koning, Hasan and Chatterji 2019; Ghosh, Thomke and Pourkhalkhali 2020) and advertising (Eastlack and Rao 1989; Schwartz, Fader and Bradlow 2017; Gordon et al. 2019; Kolsarici, Vakratsas and Naik 2020), this paper studies the adoption of experimentation in advertising and its association with firm performance in the environment. We consider the highly digitized setting of online advertising and, on one of the largest platforms for digital advertising, fully observe experimentation behavior and performance for thousands of large firms across 15 industries. This unique dataset allows us to study how firms actually solve the exploration-exploitation trade-off as related to their advertising on the platform, specifically: What share of firms runs an experiment and hence chooses a learning rate (epsilon, see Section 2.1) greater than zero? Is this share similar or different across different industries? What share of their overall budget on the platform do experimenting firms allocate to exploration?

Further, assuming that firms indeed use learning from experiments to reinforce their advertising policies towards more performant approaches (Kolsarici, Vakratsas and Naik 2020; see Figure 1),

they should see higher performance than their counterparts who do not use exploration and purely rely on their existing exploitation policies (March 1991; Schwartz, Fader and Bradlow 2017; García-Galicia, Carsteanu and Clempner 2019). Our related empirical question is: Is it true that experimenting firms experience higher performance than their counterparts who do not use experimentation? How large is the performance differential between experimenting and non-experimenting firms?

[INSERT FIGURE 2 AND TABLE 1 ABOUT HERE]

3 MATERIAL AND METHOD

3.1 Sampling and measurement

The data used in this study originate from Facebook's advertising platform that has been active since November 2007 (Facebook 2007) and reaches close to three billion people per quarter (Statista 2020), with over eight million active advertisers. Large advertisers tend to operate using several accounts on the platform. We sample activity across accounts for all advertisers managed by dedicated Facebook sales and measurement partners and active during a select twelve-month period and observe their advertising outcomes in the subsequent twelve-month period (see Figure 2). Behavioral and meta data are observed both prior to the end of the first twelve month-period ("prior-to-outcome-period" in Figure 2 and Table 1) and for the second twelve-month period ("outcome period" in Figure 2 and Table 1). We choose this sampling approach:

- To ensure that sampled advertisers are continuously active on the platform, and operate at sufficient scale;

- To select a set of advertisers that has comparable and reasonably high sophistication in their marketing operations to successfully leverage experimental insights;
- To sample based on a condition that is observed prior to the outcome period, ensuring that all selected advertisers have similar information about the available measurement toolset prior to the study period (hence, we select advertisers who had a dedicated Facebook sales and measurement partner providing information on available measurement options);
- To be able to time-separate predictors and outcome measures;
- To account for seasonality of advertiser behavior throughout the calendar year by considering a twelve month-period.

For each advertiser, we observe a set of meta, experiment use,[3] and outcome variables as outlined in Table 1, with Table 2 showing sample size and mean for select behavioral variables aggregated at the vertical level. Verticals with less than 100 sampled advertisers are excluded from the table. For confidentiality reasons, we only identify the e-commerce vertical that we use for detailed analysis as described in Section 3.4. Generally, all data is anonymized prior to analysis, no user-level data is used throughout, and all results are shown at an aggregated level. Precise data points are only included where deemed essential to the development of the analysis.

[INSERT TABLE 2 ABOUT HERE]

3.2 Performance measures

[3] Commonly, advertising experiments on the platform are called "conversion lift studies" (Facebook 2019; Lin et al. 2019).

To investigate possible associations between the use of experimentation and performance, we study purchase conversions which closely mirror the economic success of advertising (Jankowski et al. 2016). We propose three measures:

- Purchase conversions, as obtained from last-click attribution (Li et al. 2016), per 1,000 USD of advertising spend;
- Incremental purchase conversions per 1,000 USD (experiments), as obtained from experiments and hence only observed for advertisers who run experiments;
- Incremental purchase conversions per 1,000 USD (DDA), as obtained from a data-driven attribution (DDA) model that provides estimates of incremental conversions for all campaigns, including the ones without holdout conditions, and all advertisers, including those who have not run any experiments.

The first measure's strength is its ability to address success across all advertisers, it is our main measure of performance. The second measure is the direct output of advertisers' experiments and is only observable for the set of advertisers who use experimentation. The last measure addresses incremental conversions as identified through a DDA model (Du et al. 2019), trained and validated on the experiments conducted on the platform (Li and Kannan 2014; Zantedeschi, Feit and Bradlow 2016; Kalyanam et al. 2018; Facebook 2019a). Using the experiments as training data, along with the proportion of observed conversions from each experiment that are deemed incremental (i.e., would not have occurred without the advertising campaign), the resulting model is able to estimate how many incremental conversions occur for all advertising campaigns, including those that are not running an experiment (Facebook 2019b; Lewis and Wong 2018). The DDA model is trained on far more than 10,000 experiments, continuously updated based on new

experiments, and has been found to be unbiased and provide highest winner prediction accuracy in holdout validation (Facebook 2019a, 2019b).[4] As we set a very conservative minimum number of experiments that needs to be available for training before relying on the model's output, the measure is only observed for a subperiod of one quarter during 2019. Seasonal effects are thus not as well accounted for with this outcome measure relative to the other two outcomes. However, the months observed are generally representative of advertisers' behavior throughout the year. Additionally, all of the advertisers in the sample are sufficiently large such that they display consistent activity across all observed months.

As with any trained model, the resulting DDA outcome is a prediction not of incremental conversions but rather mean incremental conversions. This fact, together with the unbiased nature of the DDA model, has an ameliorating effect on the variance when compared to the other outcomes. The manifestations of this are subtly present in the results below, e.g., higher linear regression R-squared and more significant parameter estimates.

Our main performance measure is the first one, i.e., last-click conversions per advertising spend. Results obtained under the other two measures should be regarded as robustness checks for results under the last-click conversions per advertising spend measure.

3.3 Predictors: Meta and experiment use variables

As shown in Table 1, we consider two sets of variables to predict advertising performance: meta variables that account for advertiser behavior on the platform more widely and experiment use-

[4] For confidentiality reasons, the authors are not able to provide more detail on the model than outlined in Facebook (2019a, 2019b, 2019c). To summarize: "The data-driven attribution model assigns fractional credit for a conversion to Facebook touchpoints based on their estimated incremental impact. This is a statistical model developed by Facebook and is updated periodically. The data-driven attribution model only measures campaigns on Facebook, Instagram, Audience Network, and Messenger." (Facebook 2019c) The model is validated against a large set of holdout experiments both in terms of providing unbiased estimates and having high accuracy in correctly predicting the winning treatment arm of an experiment.

related variables that address how advertisers use experimentation. Many experiments are concerned with non-purchase related outcomes such as brand perceptions, mobile app installs, or page/content views and clicks. Event-related experiments are colloquially called conversion lift tests (Facebook 2019; Gordon et al. 2019). Out of this type of experiment, we focus on the subset of experiments with a stated objective of measuring advertising's impact on online purchase conversions. As our performance measures relate to purchase conversions, we only consider experiments where advertisers explicitly stated learning on purchase conversions as the experiment goal.

3.4 Assumptions and vertical focus

We use regression to discern associations between holdout experiment use and advertising outcomes. This approach is exploratory and does not allow for causal interpretation. As we are interested in online advertising and its economic performance, the study's focus is on the objective of achieving online purchase conversions. Further, out of possible conversion objectives that also include clicks, content or page views, purchase conversions are most closely associated with advertisers' economic success.

To use purchase conversions per advertising spend as a performance measure, the value of such a conversion should be reasonably similar for the set of advertisers being compared. To ensure similarity of conversion value, we focus the analysis within verticals. To state the related assumption explicitly: The presented analysis assumes the value of a conversion to differ across verticals, but to be reasonably similar within verticals.

Finally, with 66% of overall ad spend going towards purchase objectives (see eighth column in Table 2), the e-commerce vertical spends the most on this type of advertising among all observed

verticals. Furthermore, it is the largest vertical in terms of overall advertising spend and hence best suited for development of our multivariate analysis.

3.5 Model specification

Towards a multivariate assessment of the association of experimentation with performance, two specifications are estimated:

(1) $Y_{outcome\_period} = f(meta, experiments_{outcome\_period}, experimentation\_adoption_{outcome\_period})$,

(2) $Y_{outcome\_period} = f(meta, experiments_{prior\_to\_outcome\_period}, experiment\_adoption_{prior\_to\_outcome\_period})$,

where Y is one of the outcome variables listed in Table 1 and described in Section 3.2. We narrow down the set of experimentation use-related variables due to high collinearity and consider the number of experiments that advertisers run and how early they adopted experimentation. We apply log-transformation on the outcome measures to address high dispersion in their distributions.

The first specification allows for a more direct test of testing and learning during the same year as Experimentation use and performance outcomes are observed simultaneously. The second specification investigates longer term associations between advertiser experimentation and performance. While endogeneity is still a concern in this specification, behavior observation temporally precedes outcome observation.

Both specifications are estimated using simple linear regression and a more flexible random forest (RF) regressor (Ho 1995; Breiman 2001) that can capture potential non-linearities, leading to a total of twelve estimated models. The regressors are set to minimize root mean squared error (RMSE) and tuned using ten-fold cross-validation. For the second outcome measure (experimental incremental conversions per 1,000 USD) that is only observed for advertisers who ran at least one

experiment during the outcome year, the sample is restricted to such advertisers for model estimation. Finally, the bottom 20% of advertisers in terms of advertising spend are excluded from the sample as firms in this segment do not run any experiments, leading to a final sample of 776 advertisers used for estimation in the e-commerce vertical.

[INSERT TABLE 3 ABOUT HERE]

4 RESULTS

4.1 Firm exploration in online advertising

Advertisers in all verticals explore, i.e., use experimentation with purchase-related objectives (see Table 2, eighth column). In the e-commerce vertical, the largest by revenue, where 66% of advertising spend goes towards purchase objectives (seventh column in Table 2), 22.2% of advertisers run at least one experiment during the outcome year. The advertisers who run experiments spend 9.4% of their overall advertising spend on exploration, i.e., on campaigns that are experimentally measured, and run 14.8 experiments on average throughout the year. In the other 13 verticals, between 1.4 to 29.4% of advertisers engage in exploration and run an experiment. The share of advertisers running a purchase objective-related experiment (exploring) does not seem to associate strongly with the share of overall advertising spend going towards such objectives (investment in exploitation) across verticals.

Advertiser behavior further differs among the set of advertisers who run purchase-experiments within each vertical: Not only do advertisers run different numbers of experiments on average, but they also set different holdout group sizes. While there is also some variation in terms of average

experiment duration in days (last column in Table 2), it is lower. Notably, verticals that choose smaller holdout groups run experiments for longer on average – which is sensible if experiments are to be similarly powered.

4.2 Associations between the numbers of experiments and advertising performance

To understand if and how exploration through experiments relates to the performance of exploitation policies we look at associations between the number of experiments and the advertising firm's performance. Bearing in mind that the outcome variables are log-transformed, advertisers (in the e-commerce vertical) who run one more experiment during the outcome period have an approximately 2% higher ($p = 0.0737$ for last-click and $p = 0.0224$ for DDA-incremental purchase conversions per 1,000 USD advertising spend) advertising performance on average, as measured both by last-click and DDA-incremental conversions per 1,000 USD in the same year (see Table 3 that shows estimation results for both specifications presented in Section 3.5, with linear regression, in the e-commerce vertical). The association between pre- and during-outcome period measures appears to be more positive with an approximately 3% higher ($p = 0.1080$ and $p = 0.0621$ respectively) overall advertising performance for advertisers who ran one more experiment prior to the outcome period (see Table 3, last column). As mentioned in Section 3.2, DDA-incremental conversions can be expected to be unbiased and have lower variance, making it easier to uncover significant associations. This property explains why results under last-click and DDA-incremental conversions as outcome measure are equivalent but slightly more significant under the DDA specification.

When considering incremental conversions per 1,000 USD measured by experiments directly as outcome measure, there is no significant positive effect. Note that this specification is estimated

only on advertisers who ran at least one experiment as that is a necessary condition for observation of the outcome measure, explaining the smaller sample sizes in result columns two and five in Tables 3 and 4. We will return to this finding in Section 5.4 in the discussion.

To assess potential non-linear associations, we also estimate each model shown in Table 3 using a RF regressor (Ho 1995; Breiman 2001). The second-to-last row of Table 3 shows the R-squared of these estimations: While higher than with linear regression, a purely linear specification is able to capture a lot of the explainable variance for the DDA-based performance outcome (0.1348 compared to 0.1510 with a flexible RF regressor, during-outcome year specification; 0.1329 compared to 0.1579 for the prior-to-outcome year specification). For last-click-based performance, it captures less compared to a flexible non-linear regressor. This picture is in line with the mentioned variance reduction effect of the DDA model (see above and Section 3.2). Overall, the DDA specification serves as a robustness check, corroborating results obtained under the last-click specification.

To visualize associations identified by the RF regressor, Figure 3 and 4 show partial dependency plots of the three log-transformed measures of advertising performance on experimentation use variables. The plots confirm insights from linear regression: A higher number of experiments during or prior to the outcome period are correlated with better advertising performance as measured by last-click and DDA-incremental conversions per 1,000 USD. A linear fit seems to work particularly well for a during-outcome year specification (left-most top and bottom panel in Figure 3). For purely experiment-based performance (that is estimated only on advertisers who ran at least one experiment), the association is less linear and non-monotonic with decreasing performance for a higher number of experiments, possibly suggesting decreasing marginal returns to the degree of exploration.

[INSERT FIGURE 3 AND 4 ABOUT HERE]

4.3 Associations between experimentation adoption and advertising performance

To investigate if earlier adoption of exploration associates with better exploitation performance in the environment, we devise an indicator of experimentation adoption and study its association with advertising performance of the firm. The indicator of experimentation adoption is one for advertisers who were the first to adopt purchase conversion experimentation on the platform and zero for advertisers who have not run any purchase-related holdout experiments. Coefficient estimates in Table 3 can hence be understood as 53% / 73% higher advertising performance ($p = 0.2551$ and $p = 0.0525$) among advertisers who adopted experimentation immediately versus not at all (last-click and DDA-based performance measure respectively, during-outcome year specification). The same estimate is lower and less statistically significant (34%, $p = 0.5323$, and 58%, $p = 0.1831$, respectively) for the prior-to-outcome-year specification shown in the three last columns in Table 3.

The difference in estimates can be rationalized by the different underlying samples: While 131 out of 776 advertisers had run experiments prior to the outcome year, an additional 85 advertisers run their first experiment during the outcome year (see last row of Table 3). Overall, these findings suggest a positive association between (earlier) adoption of experimentation and advertising performance on the platform. Again, this relationship holds for overall advertising, but not for within-experiment advertising outcomes (see negative coefficient for experimentation adoption in the second-to-last column in Table 3 and the middle row of panels in Figure 3 and 4) – we will come back to this in the discussion in Section 5. A RF regressor corroborates results from linear

regression as can be seen in the two-dimensional partial dependency plots in Figure 3 and 4. Highest performance in terms of last-click- and DDA-based measures can be found towards the upper right (light yellow areas), i.e., for advertisers who ran a higher number of experiments and adopted experimentation earlier.

[INSERT FIGURE 5 ABOUT HERE]

4.4 Variable importance and interaction effects

While experiment use associates with advertising performance, meta variables achieve higher variable importance on average as shown in Figure 5. This result is not surprising as experiment use-related variables are zero for all advertisers who did not run an experiment which is the clear majority in the sample. Nonetheless, it seems worthwhile to understand if and how experiment use interacts with important meta variables. Particularly, it would seem intuitive if more experience with advertising on the platform correlates with advertising (exploitation) performance in conjunction with and/or independently of experimentation (exploration). Three variables measure experience with advertising on the platform: Years active, years managed, and overall advertising spend prior to the outcome period. We assess interaction effects with these variables by including interaction terms in the linear regression specifications, and overall advertising spend surfaces as the most significant interacting variable. Table 4 shows regression results when an interaction effect between the number of experiments and overall advertising spend on the platform prior to the outcome period is included. The positive main effect of the number of experiments on advertising performance remains significant. The interaction effect between experimentation use and advertising spend is negative possibly indicating that it becomes harder to drive incremental

performance by means of experimentation the more scaled the advertising operation is. In light of commonly purported diminishing returns to scale for the case of advertising (Johansson 1979; Lewis and Rao 2015), this negative interaction effect seems sensible.

4.5 Robustness checks

To check the robustness of the presented associations across verticals, the authors run the linear regression specification with the overall advertising spend interaction term (see Table 4) in the other nine largest verticals from Table 2. For four, and hence a total of five, we find the same pattern of a statistically significant positive main and negative interaction effect with revenue for the number of experiments run by an advertiser. In another four verticals, we find the same pattern (positive main effect, negative interaction), but the effects are not statistically significant. And in one vertical, effect patterns are different and inconsistent across performance measures, but also not statistically significant. In summary, these patterns corroborate a positive association of experimentation use and advertising performance across a large sample of advertisers operating in a multitude of different verticals.

Further, as there is non-trivial collinearity between predictor variables, we want to ensure that estimates are robust to different specifications of independent variables, especially to the exclusion of non-significant predictors. We hence run regressions while iteratively excluding the least significant independent variable, rendering the coefficient estimates slightly larger and more significant. The model with the complete set of independent variables hence produces conservative estimates of the association between experimentation and advertising performance, all reported estimates hence were obtained under this specification.

[INSERT TABLE 4 ABOUT HERE]

5 DISCUSSION

Speaking to literature on organizational learning and innovation more widely (March 1991; Thomke, Von Hippel and Franke 1998; Kane and Alavi 2007; Thomke 2020) and on experimentation in advertising and product development more specifically (Kohavi et al. 2009, 2013; Kalyanam et al. 2018; Gordon et al. 2019; Koning, Hasan and Chatterji 2019; Rao and Simonov 2019; Kolsarici, Vakratsas and Naik 2020), the present paper presents evidence that firms engaging in exploration by means of experimentation experience higher concurrent and subsequent performance in exploiting the (advertising) environment. Drawing on literature that shows how reinforcement learning enables firms to exploit given environments more effectively (Rana and Oliveira 2015; Schwartz, Fader and Bradlow 2017; García-Galicia, Carsteanu and Clempner 2019), the authors conjecture that exploring advertisers use a reinforcement learning-based process (see Figure 1; for details see Section 2.2) that allocates more spend to advertising policies that have been found to be more performant through experimentation (Schwartz, Fader and Bradlow 2017; Rao and Simonov 2019; Kolsarici, Vakratsas and Naik 2020). Empirical analysis is based on detailed data from one of the largest online advertising platforms capturing the advertising (step 1 in Figure 1) and experimentation practice (step 3 in Figure 1) of thousands of large advertisers across 15 industries. Findings suggest that experimenting firms indeed use experimentation to reinforce their advertising practice on the platform as they experience higher performance than advertising firms who do not use experimentation. In the next sections, we will address research questions from Section 2.4 in more detail.

5.1 Do firms use experiment-based exploration in online advertising?

In this section, we want to address the first set of research questions outlined in Section 2.4: What share of firms runs an experiment and hence chooses a learning rate (epsilon, see Section 2.1) greater than zero? Is this share similar or different across different industries? What share of their overall budget on the platform do experimenting firms allocate to exploration?

Based on the presented results, it appears surprising that a large part of advertising firms does not yet use exploration by means of experimentation, i.e., sets an epsilon of zero. As shown in Table 2, up to 98.6% of 418 sampled firms in a specific industry ("Vertical 13" in Table 2) do not yet engage in experiment-based exploration in a given recent year. The share of experimenting firms is further quite different across industries. Across all verticals, at least 70% of advertisers do not explore by means of experimentation in a recent year and, in line with expectations, experience lower performance in the environment. Finally, speaking to the last question, it appears that firms that use experimentation allocate around 10% of their overall spend to experiment-based exploration. In future research, it could be interesting to see if there is evidence for an optimal share of spend on experimentation. Kolsarici, Vakratsas and Naik (2020) indicate that this share may be dependent on the uncertainty given in the environment. Our findings suggest that there may be further industry-specific factors beyond uncertainty.

Overall, results speak to literature studying how firms solve the exploration-exploitation dilemma (March 1991; Mc Namara and Baden-Fuller 2007; Kane and Alavi 2007) and provide unique evidence from online advertising: The largest part of firms in this environment still seems to solve the dilemma by focusing purely on exploitation (March 1991). The subset of firms who choose to experiment (and hence explore and set an epsilon greater than zero per the framing outlined in Section 2.1) do better than the firms who do not. Speaking to the second set of research questions

outlined in Section 2.4, the next section discusses why there may be a performance differential for experimenting firms.

5.2 Do experimenting firms experience higher performance?

In this section, the authors wish to address the second set of research questions outlined in Section 2.4: Do experimenting firms experience higher performance than their counterparts who do not use experimentation? How big is the performance differential between experimenters and non-experimenters? Speaking to these questions, findings show that firms in the e-commerce vertical, on average, see two percent higher performance per experiment run in the same year, and three percent higher performance per experiment run in the year prior (see Section 4.2). Further, firms which adopted experimentation sooner, experience higher performance than firms who adopted it later. These associations are equivalent in the other ten largest verticals at the exception of one (see Section 4.5). They further persist when an interaction between experiment use and experience in the environment (total advertising spend on the platform before the outcome year) is included. Interestingly, this interaction is negative, indicating that advertisers who have spent more on the platform see less additional performance per experiment run compared to advertisers who spend less. In light of widely assumed negative returns to scale (Johansson 1979; Lewis and Rao 2015), this finding seems sensible and further corroborates that firms using experiments do see higher performance when including numerous control variables.

The authors propose that experimenting firms are able to realize this better performance through use of a reinforcement learning-based expert system as shown in Figure 1. Each period *t*, a fully automated expert system or a human decision-maker in complementarity with a decision support system (Turban and Watkins 1986) devises new advertising policies, tests them against existing

policies in an experiment and then adopts the best performing policy in the next period $t+1$. Use of such a system is in line with existing research on firms' advertising practices (Eastlack and Rao 1989; Fischer et al. 2011; Schwartz, Fader and Bradlow 2017; Kolsarici, Vakratsas and Naik 2020) and accounts of reinforcement learning systems used by firms to explore and exploit other environments (Rana and Oliveira 2015; García-Galicia, Carsteanu and Clempner 2019).

In the authors' framing, the expert system can be both executed by a machine or a human. In most real-world scenarios, it is likely executed by both jointly (Basu and Batra 1988; Kane and Alavi 2007; Fischer et al. 2011; Schwartz, Fader and Bradlow 2017; Zia and Rao 2019). It will be interesting for future research to study how firms actually implement such a system; e.g., what share of firms uses a fully automated expert system versus a managerial decision support system versus almost completely relies on discretionary human action. It will further be relevant to understand how often firms update their advertising policies using experimentation, i.e., how they set the t in Figure 1. Kolsarici, Vakratsas and Naik (2020) provide first evidence in this direction.

A complementary explanation for the better performance among experimenting firms is that they may be better at data-driven decision making and advertising operations more generally (Brynjolfsson and McElheran 2016; Brynjolfsson, Hitt and Kim 2011). More advanced modeling techniques will be required to shed light on this possible extension of the current study. Further, it has been reported that advertisers like to rely on longer standing, non-experimental decision support mechanism (Lavrakas 2010) that have been shown to be less precise than experiments (Gordon et al. 2019). Along these lines, firms who do not experiment, may still explore the environment by comparing different advertising policies, but do so relying on imprecise methods, contributing to an explanation of their lower performance. Also, in this account though, firms likely rely on an expert system similar to the one shown in Figure 1.

5.3 Managerial implications

On the level of practical implications, regardless if firms' lower concurrent and subsequent performance results from a lack of exploration or from exploration by means of flawed methods, firms who do not yet use experimentation in advertising seem well advised to adopt it swiftly. To benefit from the adoption of experimentation, they further will need to employ a process similar to the one discussed above and shown in Figure 1 that allocates spend to advertising policies that have been found to be higher performing in experiment-based exploration. Such a process is akin to a reinforcement learning-based expert system as studied in Rana and Oliveira (2015) for the case of pricing, in Schwartz, Fader and Bradlow (2017) for the case of advertising, and in García-Galicia, Carsteanu and Clempner (2019) for the case of portfolio management. In a stylized form the related expert system can be represented as: Compare available advertising approaches in an experiment, identify most performant advertising approach, deploy advertising budget against this approach, repeat. Schwartz, Fader and Bradlow (2017) show how such a system can be implemented in an automated manner.

A related question of high managerial relevance is how *much* to invest in experiment-based exploration and how to best solve the exploration-exploitation trade-off (Sutton and Barto 2018). While beyond the scope of the current study, it would be highly relevant for future research to investigate optimal levels of exploration for (online) advertising. A recent white paper by the Boston Consulting Group (Mank et al. 2019) provides guidance based on anecdotal evidence: "Some leaders earmark a portion of their budget for experiments or testing new methods ([…] 10% [for] new experiments)." In line with this finding, advertisers across verticals who do engage in experiment-based exploration spend on average about 10% of their overall advertising budget

on experimental advertising campaigns. A practical approach to get experiment-based exploration and a related expert system "off the ground" could hence be to dedicate 10% of advertising spend to experiment-based exploration and then regularly evaluate if ongoing advertising campaigns should be divested or increased based on experiment results. Earmarking a portion of the advertising budget for experimentation in such a way can be the first step in catalyzing the organization towards a "culture of experimentation" that also requires executive endorsement and alignment on how to measure performance (Mank et al. 2019).

In any case, to succeed with experimentation, firms will need to adopt it sustainably and experiment on an ongoing basis. Experimentation will need to be built into a firm's strategy and not be regarded as a one-off or marginally important tactical function (Kohavi et al. 2009, 2013; Ghosh, Thomke and Pourkhalkhali 2020; Thomke 2020; Kolsarici, Vakratsas and Naik 2020).

5.4 Limitations and future research

While our analysis controls for numerous platform and advertiser variables and for non-linearities, it is important to bear in mind that it remains correlational. Findings can hence only be interpreted in this and not in a causal way. Firms who use experiment-based learning might have better analytical capabilities and a more data-savvy culture more generally that catalyze the benefits derived from experimentation (Brynjolfsson, Hitt and Kim 2011). Firms who adopt experimentation might further see better advertising performance independent of their use of experiment-based exploration for other reasons. Future analysis and more advanced modeling approaches will be required to address these limitations. It could also be interesting to complement and extend purchase conversion-based performance measures with outcomes that address monetary value directly.

Despite these limiting considerations, it should be noted that result patterns are consistent with a perspective where performance increases deriving from experimental learning do not materialize in experiments themselves (costly exploration), but rather in wider advertising outside experimental campaigns (profitable exploitation – March 1991; Sutton and Barto 2018). Put differently, experiment-based learning is not supposed to be and is not profitable in and of itself, but the insights it provides enable increased performance in the environment.

It should further be noted that our analysis approach does not allow us to observe the exact exploration and/or reinforcement policy that firms use. It could be very interesting to investigate through further analysis if certain exploration and reinforcement policies associate with or lead to better performance. Speaking to this inquiry, Kolsarici, Vakratsas and Naik (2020) suggest that advertising firms should experiment more in environments with higher uncertainty. In essence, this attempt has also been made by numerous researchers studying how firms allocate advertising budget and how they should go about this allocation problem (Basu and Batra 1988; Fischer et al. 2011; Schwartz, Fader and Bradlow 2017; Zia and Rao 2019), but, to the authors' best knowledge, reinforcement learning frames are rarely used for such analysis to date.

6 Conclusion

The present study establishes that, while some large and economically successful advertisers use exploration by means of experimentation – across a multitude of industries and thousands of the largest firms, the majority has not adopted experimentation. Experimentation adopters' measured performance is further better than that of advertisers who do not use experiment-based exploration, suggesting that experimenting firms indeed successfully use a reinforcement learning-based expert system that allocates spend to advertising policies that have been found to be more performant

through experiments (see Figure 1; Schwartz, Fader and Bradlow 2017; Sutton and Barto 2018). While these results confirm the widely held assumptions on the benefits of experimentation, the presented survey of firm behavior shows that most firms have not been able to implement firm-level learning via experimentation (Rao and Simonov 2019).

Using a reinforcement learning frame, concrete recommendations how firms can adopt robust experimentation and learning practices for their advertising operations can be derived: (1) *Devise an exploration policy*: Earmark a portion of the advertising budget for experimentation; 10% seems to be commonly advised, both in a recent survey among top managers at leading firms (Mank et al. 2019) and based on the data presented in this study. (2) *Define a reward that is to be maximized through learning*: In collaboration with other company functions, e.g., finance, identify a key performance indicator that advertising is to drive. Examples are return on marketing investment (ROMI) or a composite of brand and direct response metrics. (3) *Continuously innovate and develop new candidate policies*: Devise new advertising approaches that are expected to impact the defined reward. (4) *Rigorously reinforce*: Adopt better performing advertising approaches, abandon poorly performing ones. The authors hope that this concise frame can provide guidance to firms how to adopt robust experimentation practices and improve their advertising performance.

REFERENCES

M. M. Abraham and L. M. Lodish (1990). Getting the most out of advertising and promotion. Harvard Business Review, 68(3), 50-1.

A. K. Basu and R. Batra (1988). ADSPLIT: a multi-brand advertising budget allocation model. Journal of Advertising, 17(2), 44-51.

R. Berman, L. Pekelis, A. Scott and C. Van den Bulte (2018). p-Hacking and False Discovery in A/B Testing. SSRN Working Paper.

T. Blake, C. Nosko and S. Tadelis (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. Econometrica, 83(1), 155-174.

L. Breiman (2001). Random forests. Machine learning, 45(1), 5-32.

E. Brynjolfsson, L. M. Hitt and H. H. Kim (2011). Strength in numbers: How does data-driven decision-making affect firm performance? Available at SSRN 1819486.

E. Brynjolfsson and K. McElheran (2016). The rapid adoption of data-driven decision-making. American Economic Review, 106(5), 133-39.

R. Du, Y. Zhong, H. Nair, B. Cui and R. Shou (2019). Causally Driven Incremental Multi Touch Attribution Using a Recurrent Neural Network. arXiv preprint arXiv:1902.00215.

J. O. Eastlack and A. G. Rao (1989). Advertising experiments at the Campbell soup company. Marketing Science, 8(1), 57-71.

Facebook (2007). Facebook Unveils Facebook Ads. Facebook Newsroom, accessed online at https://newsroom.fb.com/news/2007/11/facebook-unveils-facebook-ads/ on September 13 2019.

Facebook (2019). About Facebook Conversion Lift Tests. Facebook Business, accessed online at https://www.facebook.com/business/help/688346554927374 on September 27 2019.

Facebook (2019a). Facebook Data-driven Attribution. Facebook Business, accessed online at https://www.facebook.com/business/m/one-sheeters/data-driven-attribution on March 16 2020.

Facebook (2019b). 4 Keys to Using Machine Learning for Campaign Measurement. Facebook IQ, accessed online at https://www.facebook.com/business/news/insights/4-keys-to-using-machine-learning-for-campaign-measurement on March 14 2020.

Facebook (2019c). About the data-driven attribution model. Facebook Business, accessed online at https://www.facebook.com/business/help/224795638078128 on March 17 2020.

M. Fischer, S. Albers, N. Wagner and M. Frie (2011). Practice prize winner—dynamic marketing budget allocation across countries, products, and marketing activities. Marketing Science, 30(4), 568-585.

M. García-Galicia, A. A. Carsteanu and J. B. Clempner (2019). Continuous-time reinforcement learning approach for portfolio management with time penalization. Expert Systems with Applications, 129, 27-36.

S. Ghosh, S. Thomke and H. Pourkhalkhali (2020). The Effects of Hierarchy on Learning and Performance in Business Experimentation. Harvard Business School Working Paper.

B. R. Gordon, F. Zettelmeyer, N. Bhargava and D. Chapsky (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Marketing Science, 38(2), 193-225.

M. T. Hannan and J. Freeman (1984). Structural inertia and organizational change. American Sociological Review, 149-164.

T. K. Ho (1995). Random decision forests. In Proceedings of 3rd IEEE International Conference on Document Analysis and Recognition (Vol. 1, pp. 278-282).

J. Jankowski, P. Kazienko, J. Wątróbski, A. Lewandowska, P. Ziemba and M. Zioło (2016). Fuzzy multi-objective modeling of effectiveness and user experience in online advertising. Expert Systems with Applications, 65, 315-331.

J. K. Johansson (1979). Advertising and the S-curve: A new approach. Journal of Marketing Research, 16(3), 346-354.

G. A. Johnson, R. A. Lewis and E. I. Nubbemeyer (2017). Ghost ads: Improving the economics of measuring online ad effectiveness. Journal of Marketing Research, 54(6), 867-884.

K. Kalyanam, J. McAteer, J. Marek, J. Hodges and L. Lin (2018). Cross channel effects of search engine advertising on brick & mortar retail sales: Meta analysis of large scale field experiments on Google.com. Quantitative Marketing and Economics, 16(1), 1-42.

G. C. Kane and M. Alavi (2007). Information technology and organizational learning: An investigation of exploration and exploitation processes. Organization Science, 18(5), 796-812.

C. Kim, K. Kwon and W. Chang (2011). How to measure the effectiveness of online advertising in online marketplaces. Expert Systems with Applications, 38(4), 4234-4243.

R. Kohavi, R. Longbotham, D. Sommerfield and R. M. Henne (2009). Controlled experiments on the web: survey and practical guide. Data mining and knowledge discovery, 18(1), 140-181.

R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu and N. Pohlmann (2013). Online controlled experiments at large scale. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1168-1176). ACM.

C. Kolsarici, D. Vakratsas, and P. A. Naik (2020). The Anatomy of the Advertising Budget Decision: How Analytics and Heuristics Drive Sales Performance. Journal of Marketing Research, 1-21.

R. Koning, S. Hasan and A. Chatterji (2019). Experimentation and Startup Performance: Evidence from A/B Testing. Harvard Business School Working Paper.

P. J. Lavrakas (2010). An evaluation of methods used to assess the effectiveness of advertising on the internet. Interactive Advertising Bureau Research Papers.

R. A. Lewis and J. M. Rao (2015). The unfavorable economics of measuring the returns to advertising. The Quarterly Journal of Economics, 130(4), 1941-1973.

R. A. Lewis and J. Wong (2018). Incrementality bidding & attribution. Available at SSRN 3129350.

H. Li and P. K. Kannan (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. Journal of Marketing Research, 51(1), 40-56.

H. Li, P. K. Kannan, S. Viswanathan and A. Pani (2016). Attribution strategies and return on keyword investment in paid search advertising. Marketing Science, 35(6), 831-848.

S. H. Liao (2005). Expert system methodologies and applications—a decade review from 1995 to 2004. Expert Systems with Applications, 28(1), 93-103.

X. Lin, H. S. Nair, N. S. Sahni and C. Waisman (2019). Parallel Experimentation in a Competitive Advertising Marketplace. arXiv preprint arXiv:1903.11198.

L. M. Lodish, M. Abraham, S. Kalmenson, J. Livelsberger, B. Lubetkin, B. Richardson and M. E. Stevens (1995). How TV advertising works: A meta-analysis of 389 real world split cable TV advertising experiments. Journal of Marketing Research, 32(2), 125-139.

P. Manchanda, P. E. Rossi and P. K. Chintagunta (2004). Response modeling with nonrandom marketing-mix variables. Journal of Marketing Research, 41(4), 467-478.

T. Mank, N. Rich, C. Bona, N. De Bellefonds and T. Recchione (2019). Marketing Measurement Done Right. Boston Consulting Group White Paper, accessed online at https://www.bcg.com/publications/2019/marketing-measurement-done-right.aspx on March 18 2020.

J. G. March (1991). Exploration and exploitation in organizational learning. Organization Science, 2(1), 71-87.

E. Mason (2019). Building and Validating Media Mix Models. Thirdlove Tech Blog, accessed online at https://github.com/mecommerce/ThirdLove-Tech-Blog/blob/master/Media_Mix_Model/ThirdLove_MMM_Whitepaper.pdf on September 7 2019.

P. Mc Namara and C. Baden-Fuller (2007). Shareholder returns and the exploration–exploitation dilemma: R&D announcements by biotechnology firms. Research Policy, 36(4), 548-565.

R. Rana and F. S. Oliveira (2015). Dynamic pricing policies for interdependent perishable products or services using reinforcement learning. Expert Systems with Applications, 42(1), 426-436.

J. M. Rao and A. Simonov (2019). Firms' reactions to public information on business practices: The case of search advertising. Quantitative Marketing and Economics, *17*(2), 105-134.

E. M. Schwartz, E. T. Bradlow and P. S. Fader (2017). Customer acquisition via display advertising using multi-armed bandit experiments. Marketing Science, 36(4), 500-522.

Statista (2020). Number of monthly active Facebook users worldwide as of 4th quarter 2019, accessed online at https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ on March 17 2020.

R. S. Sutton and A. G. Barto (2018). Reinforcement learning: An introduction. MIT press.

S. Thomke, E. Von Hippel and R. Franke (1998). Modes of experimentation: an innovation process—and competitive—variable. Research Policy, 27(3), 315-332.

S. H. Thomke (2020). Experimentation Works: The Surprising Power of Business Experiments. Boston, MA: Harvard Business Press.

E. Turban and P. R. Watkins (1986). Integrating expert systems and decision support systems. MIS Quarterly, 121-136.

D. Zantedeschi, E. M. Feit and E. T. Bradlow (2016). Measuring multichannel advertising response. Management Science, 63(8), 2706-2728.

M. Zia and R. C. Rao (2019). Search advertising: Budget allocation across search engines. Marketing Science, 38(6), 1023-1037.

FIGURES

FIGURE 1: A generalized expert and decision support system used by firms for learning in advertising.

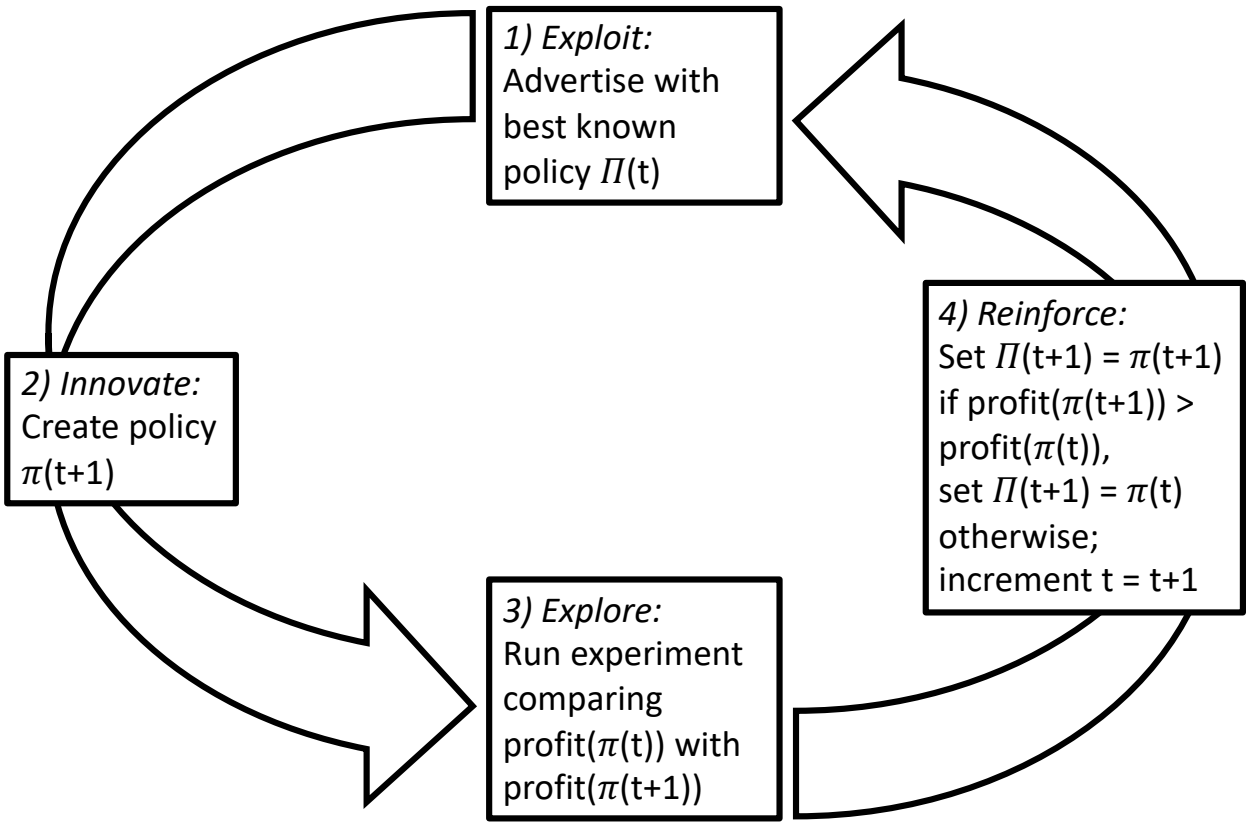


FIGURE 2: Specification of data sampling approach. Firm eligibility for inclusion is based on activity in the sampling year, historical advertising and experimentation behavior is observed from prior periods and performance from the subsequent ("outcome") year.

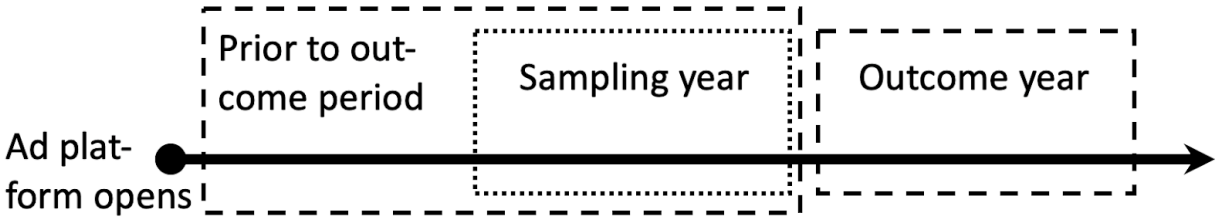


FIGURE 3: Partial dependency plots (via a random forest regressor) to isolate effect of experiment use by firms during the outcome year. Column one considers the number of experiments carried out by the firm. Column two considers both the number of experiments as well as how early the firm adopted experimentation compared to its competitors on the platform (1 = earliest adopters, 0 = non-adopters). Bright yellow indicates highest, dark blue lowest comparative performance. Axes have been removed in accordance with data policies of the online advertising platform.

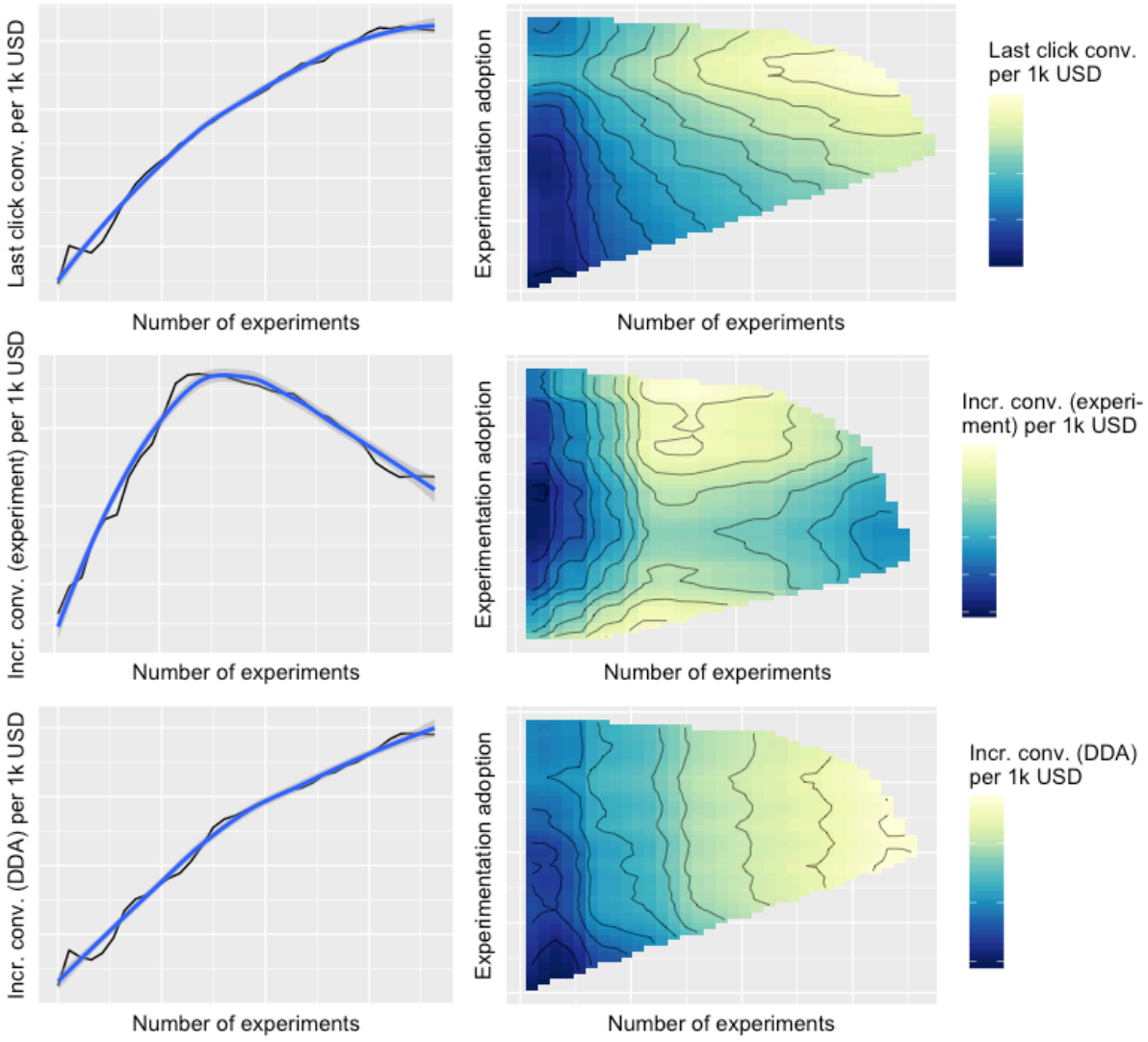


FIGURE 4: Partial dependency plots (via a random forest regressor) to isolate effect of experiment use by firms prior to the outcome year. Column one considers the number of experiments carried out by the firm. Column two considers both the number of experiments as well as how early the firm adopted experimentation compared to its competitors on the platform (1 = earliest adopters, 0 = non-adopters). Bright yellow indicates highest, dark blue lowest comparative performance. Axes have been removed in accordance with data policies of the online advertising platform.

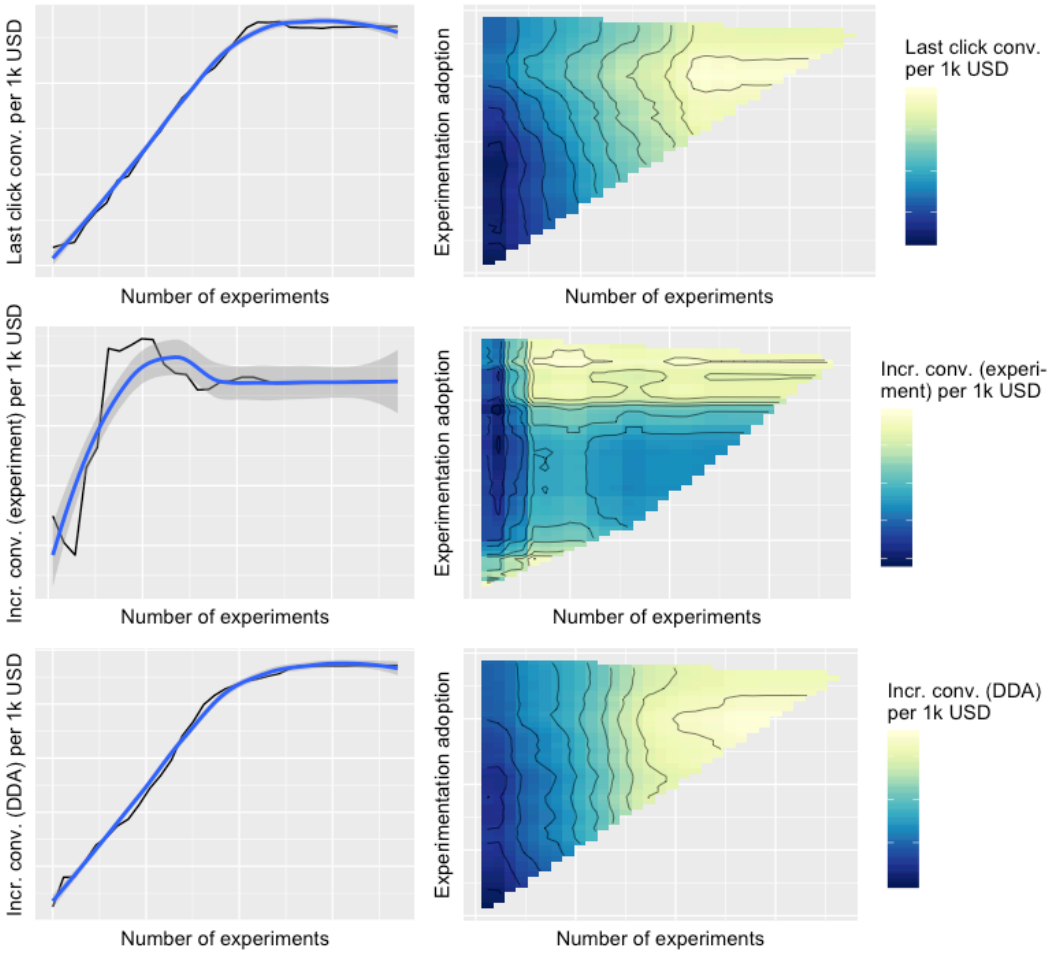
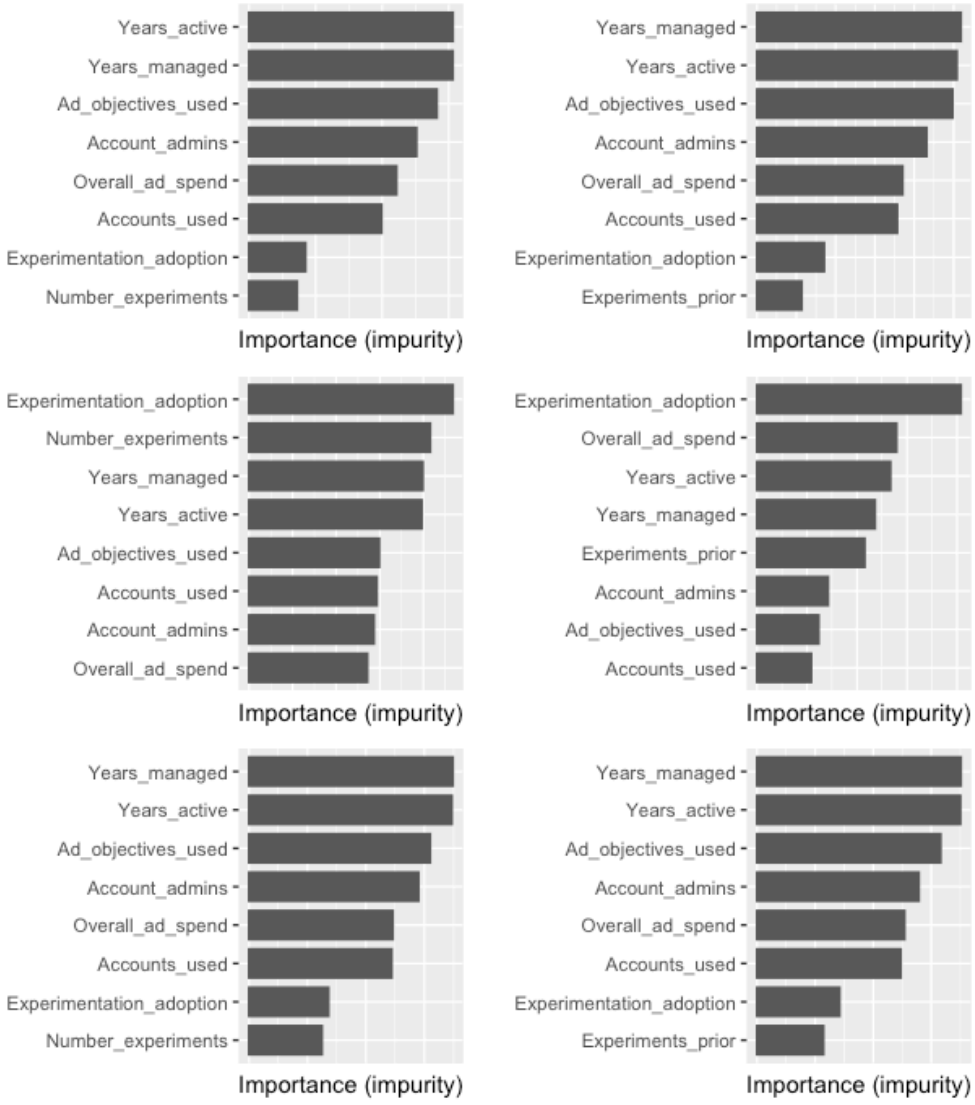


FIGURE 5: Variable importance as obtained from a random forest regressor using decision tree node impurity to assess importance. Column one is based on the model fit to outcome year while column two is for the prior year specification. Rows one to three utilize last-click per $1k advertising spend, incremental conversions per $1k advertising spend during experiments, and incremental conversions per $1k advertising spend from an experimentally calibrated attribution model, respectively. The axis values for node impurity have been removed in accordance with data policies of the online advertising platform.



TABLES

TABLE 1: Variables used in analysis. Meta variables include potential conflating variables. Experiment use variables detail if and how firms are implementing experiments. Three separate types of outcomes are considered. Each of the outcomes are commonly used as success metrics within online advertising.

Data type	Variables
Meta	Observed prior to the outcome period:
	Vertical
	Years active on platform
	Years managed by dedicated Facebook partner
	Overall advertising spend prior to outcome period
	Number of accounts used
	Number of registered account administrators
	Number of different advertising objectives used
Experiment use	Observed prior to the outcome period:
	Experimentation adoption: Years since first experiment completed$_i$ / max(years since first experiment) for each advertiser *i* (0 for advertisers who never experimented, 1 for advertisers who were the first to use experimentation)
	Observed both prior to and during the outcome period:
	Experiments with purchase conversion objective
	Spend on experiments
	Share of spend on experiments in overall spend (exploration ratio)
	Spend on significant experiments ($p < .05$)
	Average p-value achieved (spend-weighted)
	Average holdout group size
	Average experiment duration in days
	Average significant purchase conversion lift (spend-weighted)
	Minimum significant purchase conversion lift
	Maximum significant purchase conversion lift
Outcomes	Observed during the outcome period:
	Last-click purchase conversions per 1,000 USD
	Incremental (experiment) purchase conversions per 1,000 USD
	Incremental (DDA) purchase conversions per 1,000 USD

TABLE 2: Summary of key variables by industry. Samples sizes by industry along with mean advertising and experimentation behavior. Note that industries have been blinded at the request of the online advertising platform, aside from the economically most important (e-commerce) to be used as a point of reference and for development of the analysis.

Measurement frame	***Prior to outcome year***		***During outcome year***					***During outcome year, conditional on experimenting***			
Vertical (sorted by overall adv. spend)	**Number of firms**	**Average years on platform**	**Avg. adv. spend** (as % of e-comm. vertical)	**Avg. distinct ad campaigns**	**Avg. distinct ads**	**Share of spend on purchase objectives**	**Share of firms running experiments**	**Avg. number of purchase conversion experiments**	**Spend on purchase experiments over total adv. spend**	**Avg. holdout group size**	**Avg. experiment duration in days**
E-commerce	971	3.7	100.0%	10,158	25,281	66.0%	22.2%	14.8	9.4%	15.9%	24.1
Industry 2	684	3.2	72.7%	3,368	9,499	22.1%	6.7%	7.3	2.8%	17.3%	25.4
Industry 3	534	4.1	90.7%	5,352	16,470	34.5%	27.9%	11.3	7.8%	20.6%	25.6
Industry 4	1,070	1.7	42.4%	4,589	15,139	32.0%	8.9%	16.1	6.5%	13.5%	29.8
Industry 5	498	3.5	71.3%	10,433	33,934	33.8%	10.4%	22.6	11.1%	15.4%	29.4
Industry 6	618	3.6	41.5%	3,632	14,986	36.2%	7.6%	6.6	7.6%	11.4%	47.9
Industry 7	313	3.4	79.1%	52,398	83,342	38.5%	16.9%	8.3	12.6%	20.9%	35.1
Industry 8	289	2.9	65.6%	5,137	16,377	40.8%	4.2%	8.0	3.5%	6.1%	54.5
Industry 9	316	4.3	58.4%	3,752	17,542	41.4%	29.4%	47.2	9.8%	22.1%	34.4
Industry 10	128	4.7	83.4%	6,427	19,808	33.3%	26.6%	9.8	9.2%	17.7%	30.5
Industry 11	245	3.3	38.9%	3,368	15,388	45.0%	8.6%	6.7	7.2%	11.1%	32.8
Industry 12	154	3.9	44.0%	5,531	16,462	52.9%	10.4%	5.1	6.3%	8.8%	38.9
Industry 13	418	2.7	10.2%	920	3,638	35.9%	1.4%	23.0	18.2%	22.1%	26.0
Industry 14	291	2.9	13.3%	2,286	5,909	41.8%	2.7%	7.0	8.3%	21.4%	13.1

TABLE 3: Results of linear regression for the three performance outcomes and the two model specifications in the e-commerce vertical; p-values in brackets, * significant at 10%-, ** at 5%-, *** at 1%-level. For confidentiality reasons only qualitative results are shown for meta variables. R-squared of random forest regressor is included to assess increase in explanatory ability when allowing non-linearities and additional model complexity.

		During outcome year (spec. 1)		Prior to outcome year (spec. 2)		
Outcome measure / dep. var.	Last-click conv.	Exper. incr. conv.	DDA incr. conv.	Last-click conv.	Exper. incr. conv.	DDA incr. conv.
Years active on platform	+ ** (.0166)	- (.3973)	+ *** (.0002)	+ ** (.0134)	- (.5615)	+ *** (.0002)
Years managed	+ *** (.0000)	+ * (.0581)	+ *** (.0000)	+ *** (.0000)	+ ** (.0233)	+ *** (.0000)
All-time advertising spend	- (.8528)	+ (.8890)	- (.8681)	- (.8342)	+ (.8815)	- (.8454)
Accounts used	+ (.8797)	- (.8546)	+ (.5600)	+ (.8225)	- (.6669)	+ (.5138)
Account admins	- (.6250)	- (.8804)	- (.5209)	- (.5213)	- (.9707)	- (.4299)
Advertising objectives used	+ (.2967)	+ (.5759)	+ (.3706)	+ (.3160)	+ (.6226)	+ (.3950)
Experimentation adoption	.5289 (.2551)	.1746 (.7551)	.7296 * (.0525)	.3361 (.5323)	-.9384 (.3152)	.5802 (.1831)
Number of experiments	.019 * (.0737)	.0067 (.4308)	.02 ** (.0224)	.033 (.1080)	.0217 (.2015)	.031 * (.0621)
Intercept	+ *** (.0000)	+ *** (.0000)	+ *** (.0003)	+ *** (.0000)	+ *** (.0002)	+ *** (.0003)
R-squared	.0929	.0549	.1348	.0921	.0652	.1329
R-sq. (RF)	.1797	.0687	.1510	.1713	.1181	.1579
N	776	216	776	776	131	776

TABLE 4: Results of linear regression for the three performance outcomes and the two model specifications in the e-commerce vertical, including an interaction effect of experiment use and total spend on the platform prior to the outcome year; p-values in brackets, * significant at 10%-, ** at 5%-, *** at 1%-level. For confidentiality reasons only qualitative results are shown for meta variables.

				During outcome year (spec. 1)	Prior to outcome year (spec. 2)	
Outcome measure / dep. var.	Last-click conv.	Exper. incr. conv.	DDA incr. conv.	Last-click conv.	Exper. incr. conv.	DDA incr. conv.
All-time ad spend	+ (.5835)	+ (.6747)	+ (.2286)	+ (.3988)	+ (.8775)	+ (.1983)
Number of experiments	.032*** (.0083)	.0102 (.3420)	.039*** (.0000)	.064*** (.0016)	.0186 (.3616)	.07*** (.0000)
Interaction	- (.2750)	- (.4216)	- ** (.0324)	- * (.0566)	- (.8168)	- *** (.0098)
Other meta variables	Qualitative results equivalent to the ones in Table 3					
Intercept	+ *** (.0000)	+ *** (.0003)	+ *** (.0003)	+ *** (.0000)	+ *** (.0003)	+ *** (.0001)
R-squared	.0928	.0574	.1357	.096	.0578	.1384
N	776	216	776	776	131	776