Web Appendix

This Web Appendix has six main sections. First, we outline our main estimation procedure in more detail. Second, we report results from running our analysis on our entire sample, not just the campaigns that hit the recommended weekly minimum conversions in expectation. Third, we describe how we generated our experiment-level treatment effect estimates (and the necessary assumptions on the data). Fourth, we conduct a randomization check on our experimental design. Fifth, we describe an illustrative case study to provide additional intuition behind our experiment. Sixth, we provide results by the three main verticals in our sample (E-commerce, CPG, and Retail).

A Methodology

First, we describe our estimation procedure in more depth. We leverage the approach of Efron (2014, 2016); this section borrows heavily from those sources.

Conditional on our observed distributions of treatment effects and standard errors, our goal is to estimate the true, latent distribution of effects across experiments. Further, we want to do so in as flexible a way as possible, especially given the nonstandard distributions we observe. We start by articulating the general problem at hand and then show how recently developed deconvolution methods can allow us to flexibly estimate our quantities of interest under minimal assumptions.

Suppose from a set of N experiments we observe a set of treatment effects X_1, X_2, ..., X_N. Each X_i is a noisy measure of experiment i's true, unobserved treatment effect Θ_i, and we assume that the Θ_i are distributed according to an unobserved distribution $g(\theta)$. We are interested in making inferences on g based on our realized X_i's. Next, assume that the unobserved distribution of treatment effects are drawn iid from g. That is, $\Theta_i \overset{\text{ind}}{\sim} g(\theta)$ for all $i \in \{1, 2, ..., N\}$. Further, assume that each X_i is drawn independently from Θ_i according to a known distribution p_i: $X_i \overset{\text{ind}}{\sim} p_i(X_i \mid \Theta_i)$. Note that this specifies a hierarchical distribution: first the set of Θ_i are drawn from g, and then conditional on the Θ_i's, we draw our realized X_i's.

Given this set up, if we assume a broad exponential family of models for g, not only can we attain flexible functional forms, but the optimization problem also becomes tractable.[35]

[35] See Kline et al. (2022), Walters (2022) for other recent applications of this methodology. We note Efron (2016) also explores incorporating regularization into the deconvolution estimate; we chose the non-regularized version to

We now lay out that approach and introduce further notation.

Specifically, assume the support of g is a finite discrete set $\mathcal{T} = \{\theta_1, ..., \theta_m\}$. (This assumption is not strictly necessary, but it eases the analysis.) This makes the prior $g(\theta)$ an m-vector $g = (g_1, ..., g_m)$ that specifies the probability g_j on θ_j. We assume:

$$g(\alpha) = exp\{Q\alpha - \phi(\alpha)\} \tag{1}$$

where α is a p-dimensional parameter vector and Q is a known $m \times p$ structure matrix. Denoting by Q_j^T the jth row of Q, we have that the jth component of $g(\alpha)$ is

$$g_j(\alpha) = exp\{Q_j^T\alpha - \phi(\alpha)\} \quad \text{for } j = 1,2,...,m \tag{2}$$

where $\phi(\alpha)$ normalizes $g(\alpha)$ to make it a probability distribution:

$$\phi(\alpha) = \log \sum_{j=1}^{m} exp(Q_j^T\alpha) \tag{3}$$

In our estimation, we follow Narasimhan and Efron (2020) in letting Q be a basis matrix for natural cubic splines over $\mathcal{T}$ with degrees of freedom p. Past applications have analyzed relatively small-scale data compared to ours (e.g., Kline et al. (2022)) and assumed that fixing $p = 5$ granted the model enough underlying flexibility to adequately describe the data. In the event that the true prior is not contained in such a low-dimensional exponential family, the estimated model would contain "definitional bias" in the words of Efron (2016). Empirically, we find that models that assume $p = 5$ fail to capture our skewed distributions well, and so we estimate our models over wide ranges of p and then perform model selection over them. We discuss how we perform model selection later, but empirically our results are robust to several different selection procedures.

The discussion above has focused on the 'higher' level distribution g from which the unobserved Θ_i are drawn. We now turn to the 'lower' level distributions that map the unobserved Θ_i to the observed treatment effects, X_i. Given that context, let

$$p_{ij} = p_i(X_i \,|\, \Theta_i = \theta_j) \tag{4}$$

denote the probability that X_i is realized if $\Theta_i = \theta_j$, and let P_i be the m-vector of probabilities for X_i across all possible values of θ_j: $P_i = (p_{i1}, ..., p_{im})^T$. Importantly, note that the i subscripts on $p_i(X_i \,|\, \Theta_i)$ mean that each experiment can have a different conditional probability distribution over the observed treatment effects. This is important in our setting

avoid any bias in our estimates.

because we not only observe X_i, but we also observe an estimate of the variance, $\hat{\sigma_i}^2$, that varies by i.

For our analysis, we assume that this lower level distribution is normal with known variance from the empirical treatment effect estimate: $p_i(X_i \,|\, \Theta_i) \sim N(\Theta_i\,,\, \hat{\sigma_i}^2)$. This assumption is an appeal to the Central Limit Theorem; intuitively, this puts no restriction on our higher level distribution of interest, g, but imposes an assumption that our experiments are sufficiently large that our variance estimates have converged and the treatment effects are normally distributed around the unobserved Θ_i.[36]

Given this set up, the marginal probability for X_i becomes:

$$f_i(\alpha) = \sum_{j=1}^{m} p_{ij} g_j(\alpha) = P_i' g(\alpha) \tag{5}$$

and hence the log likelihood function is $l_i(\alpha) = \log P_i' g(\alpha)$. We use maximum likelihood to generate an estimate $\hat{\alpha}$ for α, which then pins down our distribution of interest, g.[37] Note that in this spline-based setup from Efron, performing maximum likelihood over the space of distributions for g ultimately reduces to maximizing over the weights α on the spline bases. This substantially reduces the state space and makes the problem tractable while maintaining a high degree of flexibility.

A.1 Implementation

First, we define a discrete state space $\mathcal{T}$ for the support of $g(\theta)$. Given the wide range of treatment effect estimates, discretizing the entire range from the smallest to largest treatment effect estimate often proved computationally challenging. Hence, we define a grid of bin size 0.001 over the range between the 1st and 99th percentiles of the treatment effect estimates. Given the extreme spread outside the 1st and 99th percentiles of treatment effects, we drop those observations.[38]

[36] For small experiments, this approximation may be less valid. To check robustness along this dimension, we reran our analysis restricting to experiments with more than 50 converters in the experiments themselves; this number has been used as a heuristic for a sample size that is large enough for the Central Limit Theorem to hold (e.g., Angrist and Pischke (2009) discusses some of the evidence around this). This restriction did not change our results substantively, both for our main results and those for the small and large scale advertiser comparison. We also note that this known variance assumption for the lower level distribution is common in meta-analyses, even under different methodologies (e.g., DellaVigna and Linos (2022)).

[37] This approach is sometimes called g-modeling to emphasize that we are interested in the shape of the prior; most empirical Bayes methods are f-modelling in that they are interested in the marginals.

[38] We re-estimated the main results including both the full sample and singleton points for the support of g that coincided with the treatment effects outside of the 1st and 99th percentiles. In these analyses, the median point estimates were unchanged, though the mean increased slightly, as one would expect given the direction of the skew in the underlying distribution. We prefer the results in the main text since we think the extreme observations are likely spurious.

As mentioned previously, we choose Q to be a basis matrix for natural cubic splines with degrees of freedom equal to p. To minimize the risk of rounding errors, we standardize the basis matrix so the columns have mean zero and sum of squares equal to one. We then run models over $p \in \{10, 20, 30, ..., 200\}$. Intuitively, as we vary p, we change the number of knots, thereby adding increased flexibility at the risk of overfitting. We then must choose between models.

A.2 Model selection

Our preferred method of model selection across the possible degrees of freedom relies on cross-validation. However, we note that under several different model selection procedures (e.g., Bayesian Information Criterion or Akaike Information Criterion) our results do not change substantively. We now describe how we performed our cross-validation procedure.

For a given degrees of freedom, p, we partition the set of experiments into k subsets. We use $k-1$ folds to estimate a prior g and then evaluate the likelihood of observing the treatment effects in the k^{th} fold under that prior. We repeat this process k times for each degree of freedom, so each fold is used $k-1$ times for training and once for evaluation. For each p, this process generates an average value of out of sample performance across folds; we can then select the p that performs best. For each meta-analysis, we then re-estimate the model on the full sample with the selected degrees of freedom. Our main results (Table 7 and Table 8) are from models selected with ten-fold cross-validation; due to computational constraints our remaining results are based on three-fold cross-validation. We note that this means for every distribution we estimate in the main text, we fit 61 different models (20 values of p $\times$ three-fold cross-validation for each p plus 1 final model estimation from selected p value).

A.3 Defining advertisers and campaigns

For the purposes of our experiments, we treat each pixel as a single advertiser, and we treat all ads that are optimized against that pixel as a single campaign. Each experiment is defined at the campaign level, meaning the final number of advertisers, experiments, campaigns, and pixels are equal.

This definition is usually straightforward to implement. However, a wrinkle is that multiple ad accounts can run ads optimizing against the same pixel. This is often done when, for example, a company runs ads from different departments or works with external agencies. In framing the total number of 'advertisers' in our study, we thus did not use the number of advertising accounts, but rather the number of pixels as that seemed a better, if more

conservative, estimate.

Some of our analyses use advertiser-level characteristics. Because there are instances where multiple accounts were used for the same pixel but differed in one of the characteristics (e.g., country), we map account characteristics to advertisers by taking the modal demographic across the accounts within each experiment, weighted by delivered ad impressions. For example, if an experiment involved three accounts and the plurality of impressions came from accounts in the US, we would label the experiment as from the US.

In practice, 85% of experiments have one account, meaning the aforementioned wrinkles affected a relatively small fraction of our sample. Further, rerunning our analyses restricting to accounts that have this one-to-one mapping with experiments yields very similar results.

B Results from the Full Sample

In the main text we restrict our analyses to campaigns that hit the recommended number of conversions per week in expectation; here, we re-run our results on our entire sample. This consists of 150,757 experiments versus the 70,909 in the main text.[39] Figure A1 displays both the distribution of baseline ad effectiveness and the distribution of the change in effectiveness when optimizing for clicks instead of purchases.

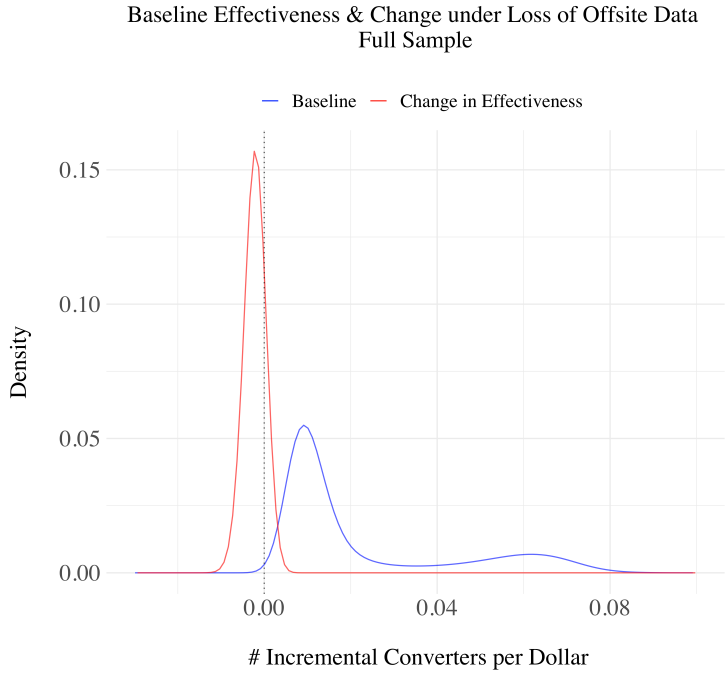
Table A1: Quantiles and means from the estimated baseline distribution of advertising effectiveness.

	10th	25th	50th	75th	90th	Mean
# Incremental Converters per $1,000	5.1	8.1	13.5	53.1	206.6	48.4
Cost per Incremental Converter	$195.03	$123.04	$74.13	$18.82	$4.84	$20.65
# Fewer Incremental Converters per $1,000	-6.0	-4.3	-2.6	-0.9	-0.3	-2.3

Intuitively, given that the optimization algorithm does not perform as well for the additional campaigns that we now retain in the sample, we would expect the cost per incremental converter to increase compared to the main text. We confirm that this is the case – if we recompute the median change at the median effectiveness, we get that cost per incremental

[39] Of the initial 187,922, we drop experiments outside the 1st and 99th percentiles of the treatment effects. We also drop 33,416 (18%) experiments that have no recorded purchases in either holdout or exposed; these treatment effects have zero variance, which is incompatible with our maximum likelihood estimation procedure. We note if you assume all these experiments generate no incremental converters per dollar and are not affected by losing offsite data, and recompute the percentiles of each distribution, the median cost per incremental converter is $105.00 that under the median change in effectiveness would increase to $124.14, an 18% increase.

Figure A1: Estimated distributions from our full sample of the baseline number of incremental converters per dollar and the within-campaign change.



Notes: The blue line traces the baseline ad effectiveness distribution. The red line corresponds to the distribution of the within-campaign change in ad effectiveness. The dashed black lines represent the 95% confidence intervals.

converter increases from \$74.13 to \$92.23, a 24% increase. This is less than the 31% we estimated for our main sample, but still a substantial increase. Notably, the baseline distribution appears bimodal; we suspect this is due to an integer-related point because many of the advertisers who did not meet the minimum threshold were spending small amounts, and so their experiments had fewer users in them. With fewer people in the denominator, across experiments the cost per incremental converter is not uniformly distributed across the range but rather clusters at certain fractions (e.g., 1/30, 1/20).

C Generation of Experiment-level estimates

As input into our meta-analysis, we first need to generate estimates of the treatment effects and standard errors for each individual experiment. In this section, we walk through how we went from the individual-level data to experiment-level results.

First, without loss of generality, we focus on the baseline case (the derivation for the click optimized arm is identical). In the baseline case, define:

$$
\begin{aligned}
N_H &:= \text{the number of users in the Holdout group} \\
N_E &:= \text{the number of users in the Exposed group} \\
N_H^C &:= \text{the number of converters in the Holdout group} \\
N_E^C &:= \text{the number of converters in the Exposed group} \\
d &:= \text{the ad spend in dollars in the Exposed group}
\end{aligned}
$$

Given those definitions, we constructed our main outcome variable, the number of incremental converters per dollar, as:

$$\tilde{N}_I := \frac{1}{d}\left(N_E^C - N_H^C\left(\frac{N_E}{N_H}\right)\right)$$

This expression takes the number of converters in each of the Holdout and Exposed groups, scales them by the number of users in each, and then divides by the spend.

In terms of variance, we first define the variance for the number of converters in the Holdout group (defined analogously for the number in the Exposed group):

$$\sigma_H^2 := N_H\left(\frac{N_H^C}{N_H} - \left(\frac{N_H^C}{N_H}\right)^2\right)$$

The above is derived from the formula for the variance of a binomial distribution, $np(1-p)$, with N_H independent trials and probability of success N_H^C/N_H. The key assumption is that each user represents an independent trial.

Letting σ_E^2 denote the corresponding quantity for the Exposed group, we get that the variance of the number of incremental converters per dollar is:

$$\tilde{\sigma}^2 := \frac{1}{d^2}\left(\sigma_E^2 + \sigma_H^2\left(\frac{N_E}{N_H}\right)^2\right)$$

The above formulas define our treatment effect and standard error for our baseline case. We now turn to those for the change in effectiveness. For this, we need to introduce subscripts referring to the purchase and click optimized conditions.

Let $\tilde{N}_{purchase}$ denote the number of incremental converters per dollar of the purchase optimized arm and $\tilde{N}_{click}$ denote that for the click arm. We then define our main outcome variable of interest as the change in the number of incremental converters per dollar across arms as:

$$\Delta\tilde{N} := \tilde{N}_{click} - \tilde{N}_{purchase}$$

For the variance of this estimate, let $\tilde{\sigma}^2_{purchase}$ denote the variance of the number of incremental converters per dollar in the purchase arm and $\tilde{\sigma}^2_{click}$ denote that for the click arm. We then define the variance of the difference as:

$$\tilde{\sigma}^2_{\Delta} := \tilde{\sigma}^2_{purchase} + \tilde{\sigma}^2_{click} - 2\text{Cov}(\tilde{N}_{click}, \tilde{N}_{purchase})$$

Under the assumption each user is an independent trial, the covariance term drops out, leaving us with only the first two terms.

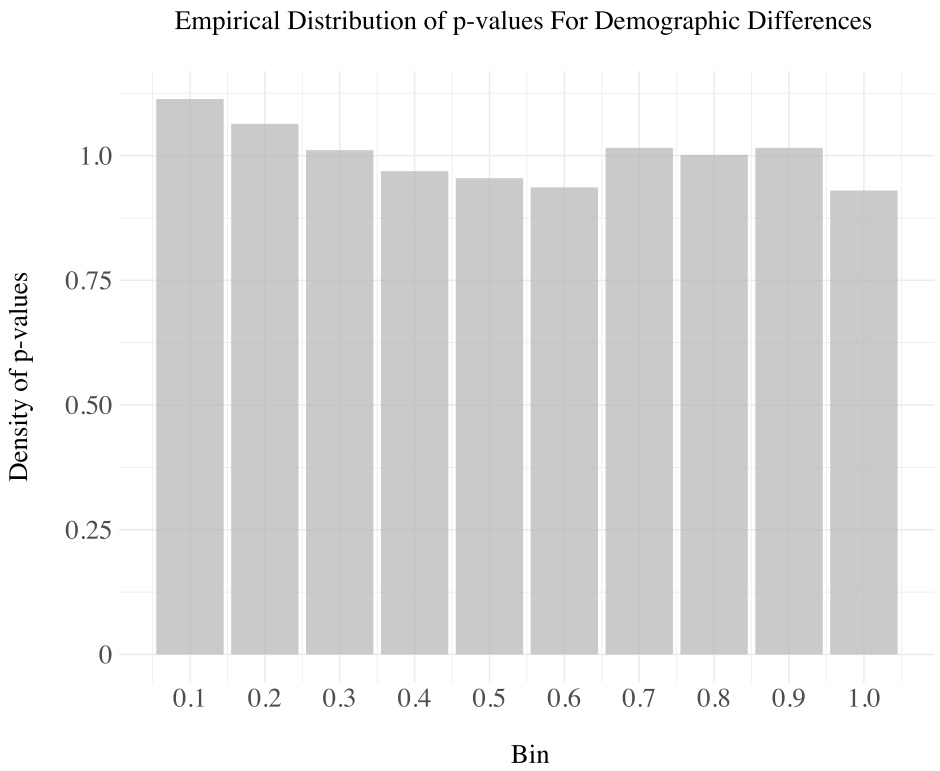
D Randomization Check

As mentioned in the main text, our randomization was done using the same infrastructure that underlies Meta's advertiser-facing lift product. This technology has previously been described in several papers, each of which show that the experiments that were conducted passed randomization checks (Gordon et al. 2022, Athey et al. 2023, Gordon et al. 2019). In this appendix, we report randomization checks on our data as well.

To conduct randomization checks, we collected data on user demographics for a random sample of experiments in our study. We selected demographics that have good coverage in the data and that have standardized values to facilitate analysis: age, gender, account age, number of Facebook friends, iOS user indicator, and web and mobile activity on Facebook in the past month. We restrict to studies with more than 50 users in the control to avoid small sample sizes that could skew our randomization tests. We conducted this check about a year and a half after the experiment was run which means these demographics come from a later point in time. However, while some users may have dropped off or values of some of these demographics may have shifted, any changes seen in the test group we would expect to be mirrored in the control group.

For each of the randomly selected experiments, we conducted t-tests across the test and control groups for each demographic feature. We expect the resulting distribution of p-values will be uniform on [0,1] under correct randomization. Figure A2 plots the density of p-values across 10 evenly spaced bins over [0,1]. Across bins, 6% of p-values are below 0.05, 51% are below 0.50, and 75% are below 0.75, consistent with correct randomization. A chi-squared test comparing our distribution of p-values to what we would expect if they were uniformly distributed leads us to fail to reject the null hypothesis that there are no differences between the distributions, $\chi^2(9, N = 6,918) = 9.94, p = .355$.

Figure A2: Empirical distribution of p-values from randomization checks.



Notes: For each randomly selected experiment, we conduct a t-test for the equality of means of the demographics of users in the treatment and control groups. The distribution reports the resulting p-values, pooled across demographic variables and experiments. Under the null of correct randomization, we would expect to see a uniform distribution of p-values.

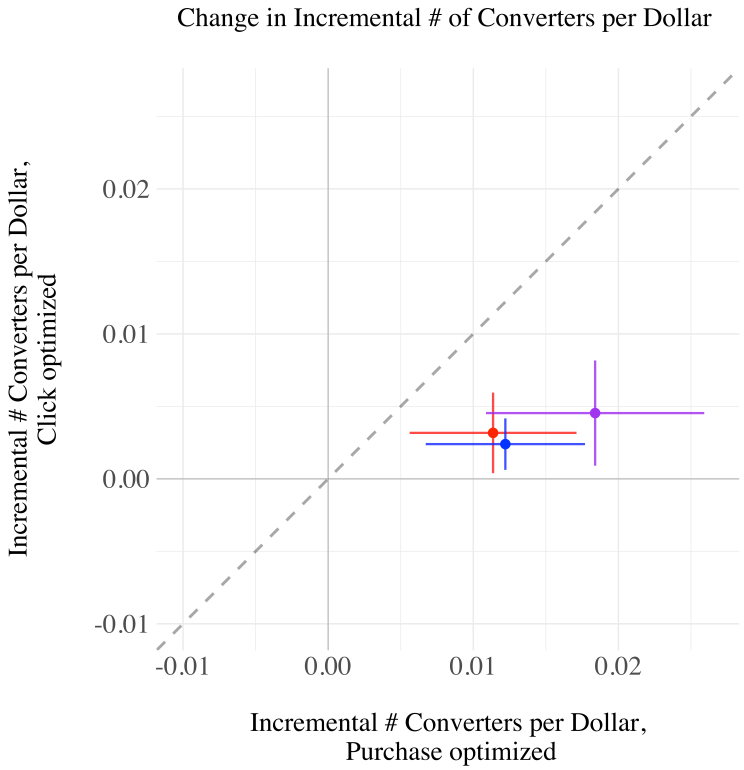
Finally, we end with a caveat to our randomization check. Namely, given these demographics were collected after our experiment was run, if our treatment induced differential attrition on the platform or changed any of our demographics, there could be true imbalances at the time of the experiment that this check fails to detect. Given the small effect sizes we observe in our sample and the fact that several past papers have conducted independent validation of Meta's lift architecture, we are not overly concerned about this caveat but flag it nonetheless.

E Illustrative Case Study and Visualization

To help build intuition behind our experimental design, we first visualize the results from a few select campaigns, and then we visualize the results from the full sample that we use in our main analysis.

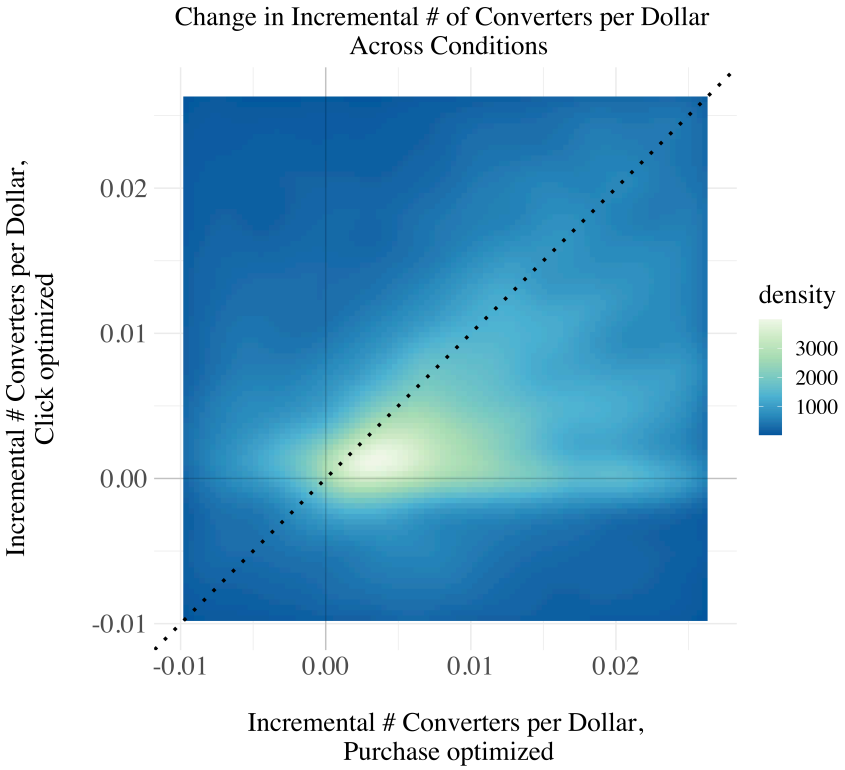
First, in Figure A3, we focus on three example campaigns, and we plot their treatment effects and 95% confidence intervals at baseline (x-axis) and in the counterfactual (y-axis). The red dot represents an apparel company, the blue dot represents a beauty company, and the purple dot represents a jewelry company. The fact that these point estimates lie to the right of the y-axis indicates that, at baseline, each of the campaigns generated a positive number of incremental converters per dollar when the campaigns were optimizing for purchases. Similarly, the fact that they lie below the 45 degree line indicates that they are estimated to earn fewer incremental converters per dollar when they optimize for clicks instead of purchases. For example, we estimate that the jewelry company's (purple dot) cost per incremental converter increased from \$54 to \$154.

Figure A3: Example of treatment effects for three hand-picked advertisers.



In Figure A4, we expand this analysis, showing a heatmap that includes all the treatment effect estimates in our main sample. (This heatmap is similar in spirit to L'Abbé plots that are often used in meta-analyses.) We can clearly see the density to the right of the y-axis and below the 45 degree line: the mass of our estimated treatment effects across all our experiments suggests both a positive baseline effect and a negative effect of losing pixel data.

Figure A4: Heatmap of all 70,909 treatment effects in our main sample.



F Results by Firm Vertical

Here we explore how the results differ by the three largest verticals in our sample: E-commerce, CPG, and Retail. These numbers help provide a sense of heterogeneity across advertisers and may also be useful for practitioners.

Among these three verticals, CPG advertisers have the highest median cost per incremental customer at baseline. As this is a purely descriptive exercise, we cannot say whether that is due to differences in products sold, advertising quality, share of offline sales[40], or other unobservables that differ across verticals. Separately, we also note the change in effectiveness (in an absolute sense) is similar across the verticals.

Table A2: Quantiles and means of estimated distributions by vertical.

	10th	25th	50th	75th	90th	Mean
E-commerce						
# Incremental Converters per $1,000	5.8	12.0	21.5	68.3	94.3	56.4
Cost per Incremental Converter	$172.50	$83.05	$46.52	$14.64	$10.60	$17.74
# Fewer Incremental Converters per $1,000	-15.0	-10.0	-6.4	-3.3	-1.0	-7.7
Consumer Packaged Goods						
# Incremental Converters per $1,000	4.5	9.5	16.7	39.7	95.7	43.0
Cost per Incremental Converter	$222.72	$105.52	$59.89	$25.19	$10.45	$23.25
# Fewer Incremental Converters per $1,000	-12.2	-8.9	-5.9	-3.2	-0.9	-6.5
Retail						
# Incremental Converters per $1,000	6.7	16.9	33.4	132.2	213.2	115.0
Cost per Incremental Converter	$149.16	$59.21	$29.96	$7.57	$4.69	$8.70
# Fewer Incremental Converters per $1,000	-11.6	-8.4	-5.4	-2.5	-2.5	-0.1

[40] We note that a large majority of CPG and Retail sales still occur offline (e.g., Boston Consulting Group estimates that 95% of US CPG sales occur offline (Novacek et al. 2016); eMarketer similarly estimates 81% of global retail sales occur offline Cramer-Flood (2023)). Our estimates of ad effectiveness are biased downward to the extent that the online ads in our study drove offline sales, suggesting there may be more scope for bias in CPG and Retail. At the same time, we note that the vertical with nearly all of its sales online, E-commerce, experiences the largest magnitude change in effectiveness, albeit by a small margin. We leave a deeper exploration of these points to future work.