Periodic Measurement of Advertising Effectiveness Using Multiple-Test-Period Geo Experiments

Jon Vaver, Jim Koehler

Google Inc.

Abstract

In a previous paper [6] we described the application of geo experiments to the measurement of advertising effectiveness. One reason this method of measurement is attractive is that it provides the rigor of a randomized experiment. However, related decisions, such as where and how to spend advertising budget, are not static. To address this issue, we extend this methodology to provide periodic (ongoing) measurement of ad effectiveness. In this approach, the test and control assignments of each geographic region rotate across multiple test periods, and these rotations provide the opportunity to generate a sequence of measurements of campaign effectiveness. The data across test periods can also be pooled to create a single aggregate measurement of campaign effectiveness. These sequential and pooled measurements have smaller confidence intervals than measurements from a series of geo experiments with a single test period. Alternatively, the same confidence interval can be achieved with a reduced magnitude and/or duration of ad spend change, thereby lowering the cost of measurement. The net result is a better method for periodic *and* isolated measurement of ad effectiveness.

Keywords: ad effectiveness, advertising experiment, periodic measurement, experimental design

1 Introduction

Advertisers benefit from the ability to measure the effectiveness of their campaigns. This knowledge is fundamental to strategic decision making and operational efficiency and improvement. However, advertising is dynamic. Competitors come and go, product lines evolve, and consumer behavior changes. Consequently, measuring ad effectiveness is not a one-time exercise. Advertisers with search campaigns need to know if their bidding strategy, keyword sets, and ad creatives are having a consistently compelling impact on consumer behavior. Since an assessment of ad effectiveness is relevant for a limited amount of time, the need for measurement is ongoing. Methods of measurement need to be adapted to accommodate this persistent need for measurement.

There are several key capabilities that a periodic geographically based measurement method should provide. Most importantly, the method needs to provide the ability to generate a sequence of ad effectiveness measurements across time. Additionally, the experimental units should rotate between the test and control groups. This rotation ensures that, over time, all geographic regions, or "geos", experience an equivalent set of campaign conditions, which balances the ad spend opportunity across geos. The capability to evaluate the design of an experiment is also important. Understanding how the measurement uncertainty is impacted by characteristics such as experiment length, test fraction, and magnitude of ad spend change is critical to

executing an effective and efficient experiment.

The application of geo experiments to the measurement of advertising effectiveness was described in a previous paper [6]. These experiments measure ad effectiveness across a single test period. A series of single test period geo experiments meets the requirements above. However, the time required to execute these experiments is less than optimal. Each experiment requires a separate pretest period, which significantly limits measurement frequency. This restriction is particularly undesirable since ongoing measurement is the primary goal. The alternative approach described here avoids this problem by combining the test period of one measurement with the pretest period of the next measurement. This coupling of the pretest and test periods not only avoids the inefficiency of isolated pretest periods, it also uses ad spend more efficiently to reduce the confidence interval of the ad effectiveness measurements.

2 Description of Multiple-Test-Period Geo Experiments

Our objective is to measure the impact of advertising on consumer behavior. Examples of this behavior include clicks, online and offline sales, newsletter sign-ups, and software downloads. We refer to the selected behavior as the response metric. Results of the analysis are in the form of return on ad spend (ROAS), which is the incremental impact that a change in ad spend has on the response metric divided by the change in ad spend.

In this paper, we describe how ad effectiveness can be measured periodically using a multiple test-period geo experiment, which is a generalization of the single test period geo experiment. Consequently, many of the considerations and steps for performing *periodic measurement* are the same, or similar, to those discussed previously for generating an *isolated measurement* of ad effectiveness.

The first step is to partition the geographic region of interest, (e.g. a country), into a set of geos. It must be possible to target ad serving to these geos, and track ad spend and the response metric at this same geo level. The location and size of the geos can be used to mitigate potential ad serving inconsistency due to finite ad serving accuracy and consumer travel across geo boundaries. A process that uses optimization to generate geos will be described in a future paper. In the United States, one possible set of geos is the 210 DMAs (Designated Market Areas) defined by Nielson Media, which is broadly used as a geo-targeting unit by many advertising platforms.

The next step is to randomly assign each of the N geos to a geo-group. Randomization is an important component of a successful experiment as it guards against potential hidden biases. That is, there could be fundamental, yet unknown, differences between the geos and how they respond to the treatment. Randomization helps to keep these potential differences equally distributed across the geo-groups. In an experiment that contains multiple test periods, we rotate the assignment of the test condition between geo groups.. If there are M geo-groups, then the test fraction is $1/M$. That is, only N/M geos are assigned to the test condition at any given point in time. It also may be helpful to use a randomized block design [3] in order to better balance the geo-groups across one or more characteristics or demographic variables. We have found that grouping the geos by size prior to assignment can reduce the confidence interval of the ROAS measurement by 10% or more.

Each experiment contains a series of distinct time periods: one pretest period and one or more test periods (see Figure 1). During the pretest period there are no differences in campaign structure across geos (e.g. bidding strategy, keyword set, and ad creatives). All geos operate at the same baseline level and there are no incremental differences between the test and control geos in the ad spend and response metric.

In each test period, the campaigns of the geos in one geo-group are modified so that they differ

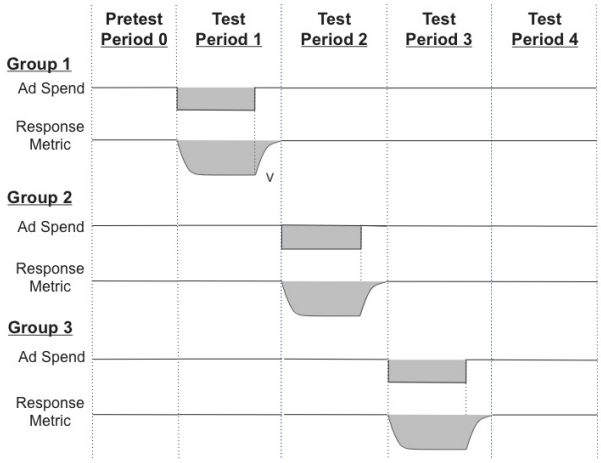

Figure 1: Diagram of a periodic geo experiment with four test periods. Ad spend is modified in a different geo-group during each of the first three test periods, and it returns to the baseline state in the final test period. There may be some delay before the corresponding change in a response metric is fully realized.

from the baseline condition. This modification generates a nonzero difference in the ad spend for these geos relative to the others. That is, the ad spend differs from what it would have been if these campaigns had not been modified. This difference will be negative if the campaign change causes the ad spend to decrease (e.g. campaigns turned off), and positive if the change causes an increase in ad spend (e.g. bids increased and/or keywords added).

The ad spend difference will, hopefully, generate a non zero difference in the response metric, perhaps with some time delay, ν. Each test period extends beyond the end of the ad spend change by ν to fully capture this incremental change in the response metric. Total clicks (paid plus organic) is an example of a response metric that is likely to have $\nu = 0$. Offline sales is an example of a response metric that is likely to have $\nu > 0$, since it takes time for consumers to complete their research, make a decision, and then visit a store to make their purchase.

Each test period provides another opportunity to measure ROAS. So, the length of the test period determines the frequency with which advertising effectiveness can be assessed. In addition to this monitoring capability, the adjacency of these test period transitions also reduces the confidence interval of these ROAS estimates. For example, the transition from Test Period 1 to Test Period 2 provides the opportunity to observe the impact of reducing the ad spend on the response metric in geo-group 2. It also provides the opportunity to observe the impact of restoring the ad spend to the baseline level in geo-group 1. The use of adjacent test periods *effectively doubles the difference in ad spend* for each test period level measurement of ROAS, except for the first measurement (transition from Pretest Period 0 to Test Period 1), and the last (transition from Test Period 3 to Test Period 4, in the Figure 1 example). This effective doubling of the ad spend reduces the confidence interval of the ROAS measurement by increasing the leverage in fitting the linear model described below [1]. So, in Test Period 1 only the geos in geo-group 1 have a test condition with reduced ad spend. In Test Period 2 the geos in geo-groups 1 and 2 have a test condition with increased and reduced ad spend, respectively, and in Test Period 3 the geos in geo-groups 2 and 3 have a test condition with increased and reduced ad spend, respectively.

3 Linear Model

After an experiment is executed, the ROAS for test period j is generated by fitting the following linear model:

$$y_{i,j} = \beta_{0j} + \beta_{1j} y_{i,j-1} + \beta_{2j} \delta_{i,j} + \epsilon_{i,j} \qquad (1)$$

where $i = 1, ..., N$, $y_{i,j}$ is the aggregate of the response metric during test period j for geo i, $\delta_{i,j}$ is the difference between the actual ad spend in geo i and the ad spend that would have occurred without the campaign change associated with the transition to test period j, and $\epsilon_{i,j}$ is the error term. We fit this model using weights

[1] This reduction can be characterized analytically, as demonstrated in Appendix A.

$w_i = 1/y_{i,0}$ in order to control for heteroscedasticity caused by heterogeneity in geo size.

The first two parameters in the model, β_{0j} and β_{1j}, are used to account for seasonal differences in the response metric across periods j and $j-1$. The parameter of primary interest is β_{2j}, which is the return on ad spend (ROAS) of the response metric for test period j. A sequence of ROAS measurements can be calculated by fitting this model separately for each test period in the experiment. Note that when $j = 1$, Equation 1 matches the linear model for the single test period geo experiment described in [6].

More generally, the ROAS can be estimated by pooling the data from a set of test periods, J. In this situation, the model becomes

$$y_{i,j} = \beta_{0j} + \beta_{1j} y_{i,j-1} + \beta_{2J} \delta_{i,j} + \epsilon_{i,j}. \qquad (2)$$

This model has the same form as Equation 1, except here j ranges over *all* of the values of J. Each combination of geo and test period provides another observation. So, instead of fitting the model with N observations, it is fit with $N|J|$ observations. The set J can include any number of test periods, and there is no need for these periods to be consecutive, although typically they will be. J may also include all of the test periods, in which case all of the experiment data are pooled to generate a single ROAS estimate.

The values of $y_{i,j}$ (e.g. offline sales) are generated by the advertiser's reporting system. The geo level ad spend is available through the ad platform reporting system (e.g. AdWords). The process for finding the ad spend counterfactual for each test period, $\delta_{i,j}$, is analogous to the process described in [6]. If there is no ad spend during period $j-1$ then the ad spend difference in test period j, $\delta_{i,j}$, is simply the ad spend during test period j. However, if the ad spend is positive during period $j-1$ and it is either increased or decreased, as depicted in Figure 1, then the ad spend difference is found by fitting a second linear model:

$$s_{i,j} = \gamma_{0j} + \gamma_{1j} s_{i,j-1} + \mu_{i,j} \qquad (3)$$

Here, $s_{i,j}$ is the ad spend in geo i during test period j and $\mu_{i,j}$ is the error term [2]. Assuming $s_{i,0} > 0$, this model is fit with weights $w_i = 1/s_{i,0}$ using only the set of control geos (C).

This ad spend model characterizes the impact of seasonality on ad spend across the transitions between test periods, and it is used as a counterfactual [3] to calculate the ad spend difference. The ad spend differences in the control and test geos (T) of each test period transition are:

$$\delta_{i,j} = \begin{cases} s_{i,j} - (\gamma_{0j} + \gamma_{1j} s_{i,j-1}) & \text{for i} \in T \\ 0 & \text{for i} \in C \end{cases} \qquad (4)$$

The zero ad spend difference in the control geos reflects the fact that these geos continue to operate at the baseline level across the test period transition. Note that, with the exception of the first and last test periods, all test periods will include $\delta_{i,j}$ that are positive and negative, since ad spend increases across the test period boundary for some geos while is decreases for others, as described in Section 2.

4 Example Results

We employed a geo experiment in [6] to evaluate the potential cannibalization of cost-free organic clicks by paid search clicks for an advertiser. We did so because the advertiser was concerned that consumers were clicking on paid search links when they would have clicked on free organic search links had there not been paid links present. The goal of the experiment was to measure the cost per incremental click (CPIC). That is, the cost for clicks that would not have occurred without the search campaign. Here we show the results of a similar geo experiment that was run to monitor the effectiveness of an existing national search advertising campaign across time.

[2] The error term in Equation 3 is scaled by the ROAS as it propagates through to Equation 1 or 2 in an additive way through the application of Equation 4. However, this error term is often smaller than the error term in Equations 1 and 2 by an order of magnitude, or more.

[3] The *counterfactual* is the ad spend that would have occurred in the absence of the campaign change across each test period transition.

The measurement period lasted approximately 12 weeks and included 11 test periods. During this experiment, the advertiser's search ads were not shown in 1/6 of the geos during each of the first 10 test periods. Each of the six geogroups took turns going "dark" across the experiment, which ensured that no geo stopped showing search ads for more than about a week at a time. Search ads were shown in all of the geos again starting with the 11th test period.

Figure 2 shows individual test period level results generated using Equation 1. There is one ROAS measurement for each test period. These measurements are not quite independent of one another because the test period associated with one measurement is the pretest period for the subsequent test period. However, there is no overlap in the data used to generate the measurements in non-adjacent test periods. The ROAS ranges from 1.3 to 2.9 clicks per dollar (CPIC ranges from \$0.34 to \$0.77 per incremental click). Note that the width of the confidence interval remains roughly the same across test periods 2 through 10. It is higher in test periods 1 and 11 where the experiment does not benefit from the effective doubling of the ad spend difference, as described in Section 2.

Figure 3 shows results generated by pooling data across test periods using Equation 2. Each ROAS measurement is generated by pooling the data from each test period with the data from all of the previous test periods. So, the ROAS generated for the first period has $J = \{1\}$, for the second $J = \{1, 2\}$, and for the 11^{th} $J = \{1, 2, 3, ..., 11\}$. The final ROAS estimate is 1.9 clicks per dollar (CPIC = \$0.53 per incremental click). Note that the confidence interval decreases monotonically across the length of the experiment as additional data are added to the model.

In contrast to the individual test period results, this scenario provides a long-term estimate of the response metric. For example, the impact of shorter term factors such as the weather, or a competitor's promotions, in a single week might be smoothed across an entire quarter. It is also

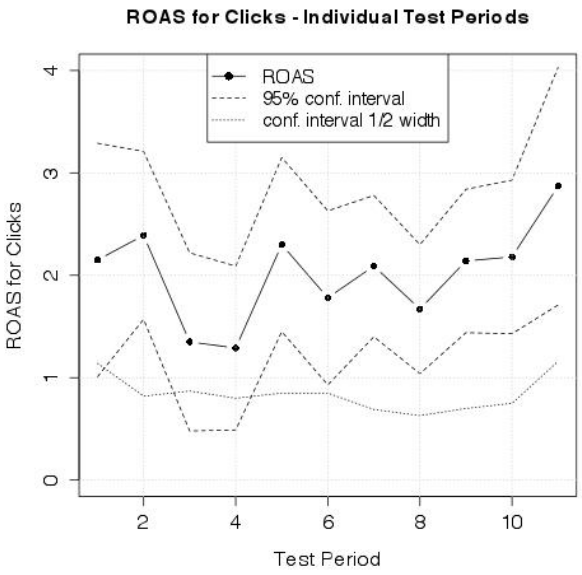

Figure 2: Measurement of return on ad spend for clicks as a function of test period. There is one ROAS measurement for each test period.

possible to balance these alternative views of the data by pooling across a subset of test periods. For example, the ROAS can be calculated by pooling data across consecutive pairs of test periods by letting $J = \{1, 2\}$, $J = \{2, 3\}$, $J = \{3, 4\}$, and so on.

One potential application for the information presented above is the generation of combined Shewhart-CUSUM quality control charts [5]. These charts are used in detection monitoring. They can identify both a sudden change (Shewhart) and a gradual change (CUSUM) in the response metric.

5 Experimental Design

As always, design is an important step for running an efficient and effective experiment. The design considerations include the length of the experiment, the test fraction, the magnitude of the ad spend difference, and the length of the test periods (switching frequency). Although there are more design considerations with a multiple-test-period geo experiment than with a single

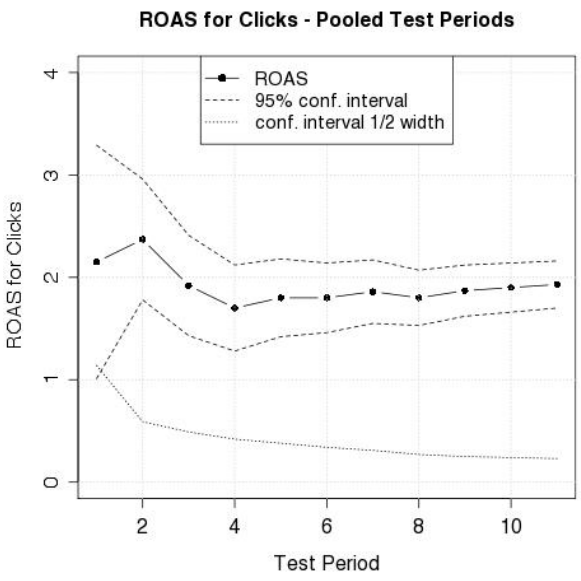

Figure 3: Measurement of return on ad spend for clicks as a function of test period length. Each ROAS measurement is generated by pooling the data from all of the previous test periods.

test period geo experiment, the same approach that was employed in [6] is still applicable.

Consider the matrix form of Equation 2,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5}$$

where $\mathbf{X} = (\mathbf{X_{j_1}}\mathbf{X_{j_2}}...\mathbf{X_{j_{|J|}}}\boldsymbol{\Delta})$ is the concatenation of $|J|$ matrices with dimension $[N|J|$ x $2]$ and one matrix with dimension $[N|J|$ x $1]$;

$$\mathbf{X_{j_1}} = \begin{bmatrix} 1 & y_{1,j_1-1} \\ . & . \\ . & . \\ 1 & y_{N,j_1-1} \\ 0 & 0 \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ 0 & 0 \end{bmatrix}, \mathbf{X_{j_2}} = \begin{bmatrix} 0 & 0 \\ . & . \\ . & . \\ 0 & 0 \\ 1 & y_{1,j_2-1} \\ . & . \\ . & . \\ 1 & y_{N,j_2-1} \\ 0 & 0 \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ 0 & 0 \end{bmatrix}$$

$$\mathbf{X_{j_{|J|}}} = \begin{bmatrix} 0 & 0 \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ . & . \\ 0 & 0 \\ 1 & y_{1,j_{|J|}-1} \\ . & . \\ . & . \\ 1 & y_{N,j_{|J|}-1} \end{bmatrix}, \boldsymbol{\Delta} = \begin{bmatrix} \delta_{1,j_1} \\ . \\ . \\ \delta_{N,j_1} \\ \delta_{1,j_2} \\ . \\ . \\ \delta_{N,j_2} \\ . \\ . \\ . \\ \delta_{1,j_{|J|}} \\ . \\ . \\ \delta_{N,j_{|J|}} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} y_{1,j_1} \\ . \\ . \\ y_{N,j_1} \\ y_{1,j_2} \\ . \\ . \\ y_{N,j_2} \\ . \\ . \\ . \\ y_{1,j_{|J|}} \\ . \\ . \\ y_{N,j_{|J|}} \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{1,j_1} \\ . \\ . \\ \epsilon_{N,j_1} \\ \epsilon_{1,j_2} \\ . \\ . \\ \epsilon_{N,j_2} \\ . \\ . \\ . \\ \epsilon_{1,j_{|J|}} \\ . \\ . \\ \epsilon_{N,j_{|J|}} \end{bmatrix}.$$

The coefficient vector, $\boldsymbol{\beta}$, is a vector with length $2|J| + 1$,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{0,j_1} \\ \beta_{1,j_1} \\ \beta_{0,j_2} \\ \beta_{1,j_2} \\ . \\ . \\ \beta_{0,j_{|J|}} \\ \beta_{1,j_{|J|}} \\ \beta_2 \end{bmatrix}.$$

With the model in this form, the variance-covariance matrix of the weighted least squares estimated regression coefficients is:

$$var(\hat{\boldsymbol{\beta}}) = \sigma_\epsilon^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} \tag{6}$$

(see [4]), where $\mathbf{W}$ is an [N $|J|$ x N $|J|$] diagonal matrix containing the weights w_i;

$$\mathbf{W} = \begin{bmatrix} \hat{W} & 0 & . & . & 0 \\ 0 & \hat{W} & & & . \\ . & & . & & . \\ . & & & . & 0 \\ 0 & . & . & . & \hat{W} \end{bmatrix}$$

with

$$\hat{\mathbf{W}} = \begin{bmatrix} w_1 & 0 & . & . & 0 \\ 0 & w_2 & & & . \\ . & & . & & . \\ . & & & . & 0 \\ 0 & . & . & . & w_N \end{bmatrix}.$$

The lower right component of the matrix in Equation 6 is the variance of $\hat{\beta}_2$. So,

$$\text{var}(\hat{\beta}_2) = \sigma_\epsilon^2 \frac{\text{adj}(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})_{N|J|,N|J|}}{\det(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})} \quad (7)$$

where $\text{adj}(A)_{n,n}$ is the n, n cofactor of the matrix A and $\det(A)$ is the determinant.

Using a set of geo-level pretest data in the response variable, it is possible to use Equation 7 to estimate the width of the ROAS confidence interval for a specified design scenario. This process is analogous to the process described in [6].

The first step is to select a consecutive set of days from the pretest data to create pseudo pretest and test periods. The lengths of the pseudo pretest and test periods should match the lengths of the corresponding periods in the hypothesized experiment. For example, an experiment with a 14 day pretest period and three 14 day test periods should have pseudo pretest and test periods with the same lengths using 56 days of data. These data are used to estimate W and all but the last column of $\boldsymbol{X}$ in Equation 7.

The next step is to randomly assign each geo to a geo group. If blocking is used, as suggested in Section 2, then this random assignment should be similarly constrained. It may be possible to directly estimate the value of $\delta_{i,j}$ at the geo level. For example, if the ad spend will be turned off in the test geos, then δ_i is just the average daily ad spend for test geo i times the number of days in test period j. Otherwise, an aggregate ad spend difference Δ_j can be hypothesized for each test period and the geo-level ad spend difference can be estimated using

$$\delta_{i,j} = \begin{cases} \Delta_j (y_{i,0} / \sum_i y_{i,0}) & \text{for i} \in T \\ 0 & \text{for i} \in C \end{cases} \quad (8)$$

The last value to estimate in Equation 7 is σ_ϵ. This estimate is generated by considering the reduced linear model;

$$\mathbf{Y} = \hat{\mathbf{X}} \boldsymbol{\beta} + \hat{\boldsymbol{\epsilon}} \quad (9)$$

where $\hat{\mathbf{X}} = (\mathbf{X_{j_1}} \mathbf{X_{j_2}} ... \mathbf{X_{j_{|J|}}})$. This model has the same form as Equation 6 except the column of ad spend difference terms, Δ, has been dropped. Fitting this model using the pseudo pretest and test period data results in a residual variance of $\sigma_{\hat{\epsilon}}$, which is used to approximate σ_ϵ.

To avoid any peculiarities associated with a particular random assignment, Equation 7 is evaluated for many random control/test assignments. In addition, different partitions of the pretest data are used to create the pseudo pretest and test periods by circularly shifting the data in time by a randomly selected offset. The half width estimate for the ROAS confidence interval is $2\sqrt{\overline{\text{var}(\hat{\beta}_2)}}$, where $\overline{\text{var}(\hat{\beta}_2)}$ is the average variance of $\hat{\beta}_2$ across all of the random assignments. This process can be repeated across a number of different scenarios to evaluate and compare designs.

Figure 4 shows the confidence interval prediction as a function of test period for the example shown in Figure 3. The dashed line corresponds to the predicted confidence interval half width and the solid line corresponds to results from the experiment. For this comparison, the ad spend difference calculated using Equation 4 was used as input to the prediction. That is, we assumed that the ad spend difference was known with certainty, as it will be when the pretest period spend is zero. The relatively good match between these two curves demonstrates that the absolute size of the confidence interval can be

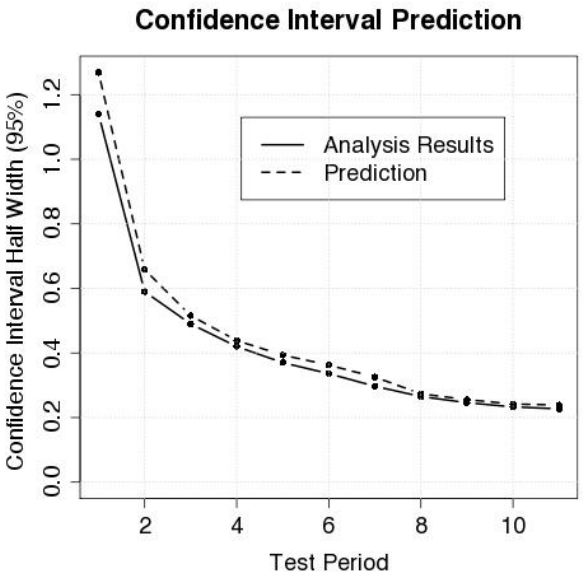

Figure 4: ROAS confidence interval half-width prediction across the length of the experiment.

predicted quite well, at least as long as the ad spend difference can be accurately predicted. In practice, the accuracy of this prediction is impacted by the dynamic environment of the live auctions and uncertainty in the relationship between changes in campaign settings, such as bids and keyword sets, and resulting changes in ad spend. The bid simulator tool [1] and the traffic estimator tool [2] can help with ad spend prediction, and closely monitoring ad spend and adjusting campaign changes in the test group during the early stages of the experiment can also help realize the targeted ad spend difference.

6 Multiple vs. Single-Test-Period Geo Experiments

In this section, we compare the application of multiple and single-test-period geo experiments to scenarios in which the objectives are periodic and isolated measurement. Multiple-test-period experiments are a better choice for periodic measurement, and the same is true for isolated measurement.

To make the comparisons more clear, all of the scenarios considered below have the same number of geo groups and, for all but one scenario, the same test fraction, $q = 1/3$. Also, when it is nonzero, the ad spend intensity (i.e. the ad spend per geo per unit time) is constant across scenarios. There is no delay in the impact of advertising on user behavior (i.e. $\nu = 0$) for the scenarios described in Sections 6.1 through 6.3. The experiment budget is the cost per measurement and each scenario has an absolute experiment budget (i.e. aggregate magnitude of ad spend difference) of either B or $2B$, depending on whether the ad spend is pulsed or continuous. Computational results were generated using the geos and response data from the example shown in Section 4. Analytical results were generated using variants of the analysis described in Appendix A.

6.1 Periodic Measurement - Pulsed Spend

The measurement objective in the first set of comparisons is to monitor ad effectiveness over time. The most obvious approach for extending single period geo experiments to this situation is to run a series of consecutive experiments. In this case, each experiment has the ad spend profile depicted in Figure 8 in Appendix B. The test group has an aggregate ad spend difference of B in every 9 day test period. The following 9 days are reserved for the pretest period associated with the next test period, which results in a pattern of pulsed ad spend. This scenario corresponds to the first row in Table 1.

The analog for this test in the multiple-test-period paradigm corresponds to the second row in Table 1 (also see Figure 9). The spend profile is the same as the first scenario. Most notably, the budget is still B. The primary difference is that the 9 day period subsequent to test period i is not only used as a pretest period for test period $i + 1$, it is also used to reduce the confidence interval of the ROAS estimate associated with measurement i. That is, the information provided by *increasing* the ad spend in geo

group 1 in transitioning to test period i is combined with the information provided by *decreasing* the ad spend in geo group 1 in transitioning to test period $i+1$. Pooling information in this way reduces the confidence interval by a factor of $1/\sqrt{2}$. [4]. Alternatively, the same confidence interval can be achieved using only one half of the ad spend difference. Along with this lower cost, the ad effectiveness measurement is more relevant to the current level of ad spend. [5]

scen.	# Test Periods	Test Period Length	Ad Spend Difference, Leverage	C. I. Half Width
1	1	9	$B,\ B$	1.59
2	2	9	$B, 2B$	1.18

Table 1: Periodic measurement scenarios with pulsed ad spend. See Figures 8 and 9 in Appendix B for the spend profiles associated with these scenarios.

6.2 Periodic Measurement - Continuous Spend

Both of the scenarios in Section 6.1 have a budget of B. Now consider the situation in which the measurement interval continues to be 18 days, but we allow for a continuous change in ad spend across time. The budget for these scenarios is $2B$. It is not possible to apply a single period geo experiment to a situation in which there is a continuous change in ad spend across time. So, a budget-equivalent comparison is made using Scenario 3, which is identical to Scenario 1 except that the test fraction has been doubled to achieve a budget of $2B$ (see Figure 10).

scen.	# Test Periods	Test Period Length	Ad Spend Difference, Leverage	C. I. Half Width
3	1	9	$2B, 2B$	0.79
4	1	18	$2B, 4B$	0.64
5	2	9	$2B, 4B$	0.68
6	3	6	$2B, 4B$	0.66
7	6	3	$2B, 4B$	0.67
8	9	2	$2B, 4B$	0.66
9	18	1	$2B, 4B$	0.64

Table 2: Sequence of multiple-test-period scenarios. The leverage from the ad spend is twice as large as the actual ad spend when switching is used. See Figures 10 through 14 in Appendix B for the spend profiles associated with Scenarios 3-7.

The base scenario in the multiple-test-period paradigm is Scenario 4 in Table 2 with the corresponding spend profile in Figure 11. In this scenario, the test period spans the entire 18 day measurement interval. While the budget is $2B$, the corresponding leverage is $4B$ because at each test period transition there is an increase in the spend for one geo group, and a decrease in spend for another. This scenario has a confidence interval that is smaller than Scenario 1 by a factor of $(1/2)\sqrt{1-q}$, and smaller than Scenario 3 by a factor of $\sqrt{1-q}$.

Additional test group rotations are included in the 18 day measurement period in scenarios 5 through 9 (also see the spend profiles in Figures 12 through 14). The ROAS measurements are generated by pooling data across these shorter test periods. In all these cases, the ad spend difference is $2B$ and the ad spend leverage is $4B$. The ad spend leverage remains constant because the more frequent switching is offset by the shorter length of the test periods. As a result, the confidence interval remains constant across these scenarios (see Figure 5). So, once a measurement period is established, there is no benefit, or harm, in rotating the test condition more frequently with regard to the width of the confi-

[4] In Table 1 the confidence interval improvement for Scenario 2 over Scenario 1 is slightly less than expected (1.18 versus 1.12). This discrepancy is caused by the use of an 18 day pretest period, which was chosen for its compatibility with subsequent scenarios. It disappears if a 9 day pretest period is used to match the length of the 9 day test periods in Scenarios 1 and 2.

[5] Generally speaking, the effectiveness of judiciously applied advertising spend decreases as ad spend volume increases. So, using a smaller ad spend difference provides a more precise measure of the marginal value of the ad spend.

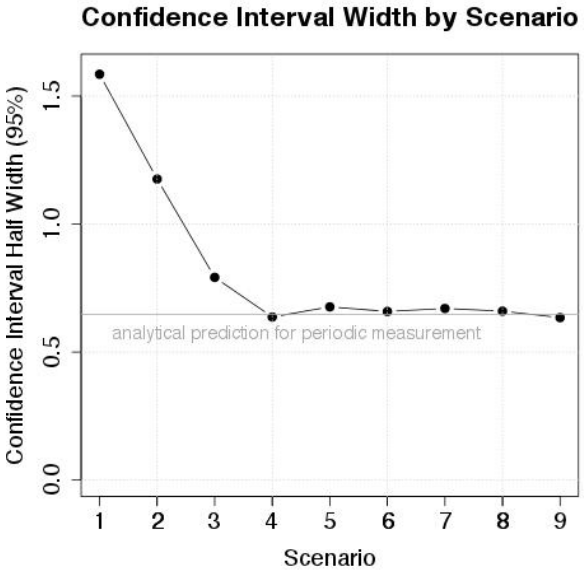

Figure 5: ROAS confidence interval prediction for the periodic measurement scenarios in Tables 1 and 2.

dence interval [6]. Switching more frequently provides the option of reducing the measurement period during the analysis phase of the experiment, although it comes with the additional logistics associated with rotating the test condition more frequently.

The results demonstrate that multiple-test-period experiments use budget and/or time more efficiently than consecutive single test period experiments.

6.3 Isolated Measurement

Now we consider the objective of generating a single measurement of ad effectiveness. The first scenario considered follows the isolated measurement approach described in [6], which corresponds to the first row in Table 3. The spend profile for this scenario is depicted in Figure 15 in Appendix B. The ad spend difference across the 18 day test period is $2B$, as is the ad spend leverage.

In Scenarios 2-6, the 18 day measurement pe-

[6]See Section 6.4 for the exception to this rule.

scen.	# Test Periods	Test Period Length	Ad Spend Difference, Leverage	C. I. Half Width
1	1	18	$2B, 2.00\ B$	1.10
2	2	9	$2B, 3.00\ B$	0.84
3	3	6	$2B, 3.33\ B$	0.75
4	6	3	$2B, 3.67\ B$	0.70
5	9	2	$2B, 3.78\ B$	0.68
6	18	1	$2B, 3.89\ B$	0.64

Table 3: Sequence of isolated measurement scenarios. The ad spend leverage approaches twice the value of the ad spend difference as switching frequency is increased. See Figures 15 through 18 in Appendix B for the spend profiles associated with Scenarios 1-4.

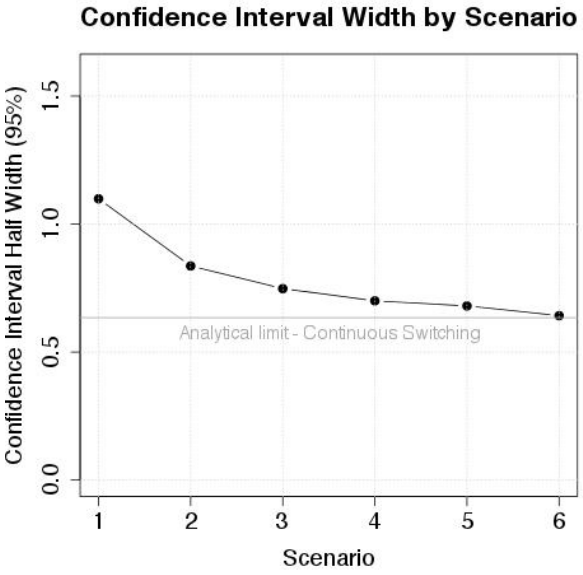

Figure 6: ROAS confidence interval prediction for the isolated measurement scenarios in Table 3.

riod is partitioned by a set of test group rotations. ROAS measurements are generated by pooling data across these shorter test periods. In all of these scenarios, the ad spend difference is $2B$, and the ad spend leverage approaches $4B$ as the switching frequency increases. The ad spend leverage is less than $4B$ because the first switch does not take place until the start of the second test period. As the switching frequency increases, the impact of not having a switch in the first test period decreases (see Figure 6).

As the switching frequency increases, the confidence interval decreases. In the limit, it becomes smaller than the confidence interval of Scenario 1 by a factor of $(1/\sqrt{2})\sqrt{1-q}$ [7]. These results indicate that, even when the goal is isolated measurement, a multiple-test-period experiment uses the ad spend difference more efficiently than a single test period experiment.

6.4 Implications of Delayed Ad Impact

In the fourth set of comparisons the objective is the same as in Section 6.1: monitor ad effectiveness over time. However, in this case we consider the implications of a non-zero delay for the impact of the advertising on the response metric (i.e. $\nu > 0$). In this situation, more frequent rotation of the geos through the test condition results in a reduction in the ad spend difference and leverage, and a correspondingly larger confidence interval.

The scenarios in Table 4 are the same as the first six scenarios in Tables 1 and 2, except here $\nu = 3$ days. Consequently, the ad spend change is truncated 3 days prior to the end of each test period to allow the full impact of the advertising to be realized within each test period, which is similar to the situation depicted in Figure 1. As a result, the ad spend difference and the ad spend leverage are less than the analogous values in Tables 1 and 2.

The ad spend difference in Scenario 1 is reduced by a factor of 2/3 because of the impact delay. This ad spend reduction increases the confidence interval by a factor of $1/(2/3) = 1.5$. The same logic extends to scenarios 2-6. For a measurement period of length L, the ad spend is reduced by a factor of $(L-m\nu)/L$, where m is the number of test periods during the measurement period [8]. The confidence intervals for Scenarios 4-6 in Table 4 are larger than the confidence intervals of the corresponding scenarios in Table 2 by about a factor of $L/(L-m\nu)$, where $L = 18$, $\nu = 3$, and m=1, 2, 3, respectively.

The confidence intervals for Scenarios 1-6 are plotted in Figure 7. Even with a delay in ad impact, the multiple-test-period alternatives to Scenarios 1 and 3 (i.e. Scenario 2 for pulsed ad spend difference, and Scenario 4 - for continuous ad spend difference) will always have a lower confidence interval. However, the larger confidence intervals of Scenarios 5 and 6 demonstrate that partitioning the measurement interval with additional switching is not always harmless. The

[7] This reduction in the confidence interval is the same reduction that would have occurred if an additional 18 day "observation" period were added to the analysis. This additional period would be used to observe the impact on the response variable of returning the ad spend to the baseline level in Group 1, similar to Scenario 2 in Table 1. So, one interpretation of the benefit of switching is that it allows the length of the analysis period to be cut in half without impacting the confidence interval.

[8] Note that $m\nu$ must be less than L to avoid a situation in which some of the impact of the ad spend is shared across adjacent test periods.

sc.	# Test Periods	Test Period Length	Ad Spend Difference, Leverage	C. I. Half Width
1	1	9	$0.67B, 0.67B$	2.37
2	2	9	$0.67B, 1.33B$	1.77
3	2	9	$1.33B, 1.33B$	1.17
4	2	9	$1.67B, 3.33B$	0.77
5	3	6	$1.33B, 2.67B$	1.02
6	6	3	$1.00B, 2.00B$	1.32

Table 4: Sequence of multiple-test-period scenarios in which the impact of the advertising on the response metric lasts up to three days.

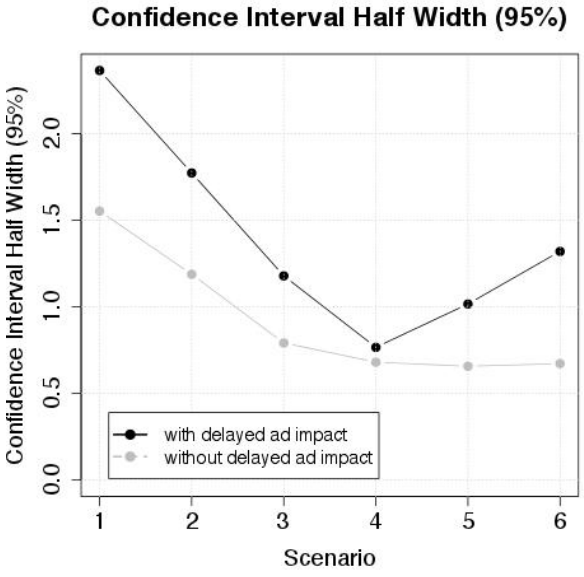

Figure 7: ROAS confidence interval prediction for the delayed ad impact scenarios in Table 4.

presence of a non-zero delay in ad impact makes it necessary to trade-off confidence interval size with measurement frequency.

7 Additional Design Notes

The approach described in Appendix A can be used to analyze a variety of design choices. This section includes the results of several of these analyses.

7.1 Impact of Modifying Test Fraction and Ad Spend Difference

Most advertisers prefer to measure ad effectiveness with as little impact as possible to their existing campaigns. Smaller test fractions have a smaller impact on existing campaigns. However, they also generate ROAS measurements with larger confidence intervals. Modifying the test fraction by a factor of f_t changes the expected confidence interval by a factor of $\sqrt{1/f_t}$. So, reducing the test fraction from 1/4 to 1/8 will increase the confidence interval by a factor of $\sqrt{2}$.

Note that when the test fraction was scaled by f_t, the magnitude of the aggregate ad spend difference was also scaled by f_t. This additional scaling keeps the average geo level ad spend difference, i.e. the *intensity* of ad spend difference, constant. However, scaling this intensity is another way to impact the confidence interval. Smaller intensities correspond to smaller changes in the existing campaigns. Increasing the ad spend difference by a factor of f_δ modifies the expected confidence interval by a factor of $1/f_\delta$. This means that halving the ad spend difference will double the confidence interval.

These results indicate that changing the number of test geos has less impact on the confidence interval than changing the leverage of the linear model by modifying the ad spend difference. However, increasing the magnitude of the ad spend difference has other implications. The efficiency of ad spend typically decreases as ad spend increases. So, the ROAS associated with a large ad spend difference may be smaller than the ROAS associated with a smaller one. Unfortunately, the exact relationship between ROAS and the volume of ad spend is usually unknown. Using an ad spend difference that is too large may lower the ROAS and measure ROAS at a level of ad spend that is not relevant to the advertiser.

7.2 Trade-off: Test Fraction and Test Length

Some advertisers may want to limit the impact of running an experiment on their existing campaigns by using a smaller test fraction, but they may prefer not to do so at the expense of measurement precision. An alternative is to offset the use of a smaller test fraction by increasing the length of the measurement period. If the test fraction is scaled by f_t, then the confidence interval can be kept constant by scaling the length of the measurement period by $f_t^{(-2/3)}$. So, if the test fraction is cut in half, then the length of the measurement needs to increase by a factor

of $(1/2)^{(-2/3)} \approx 1.6$ to keep the same confidence interval.

7.3 Impact of Geo Expansion and Geo Splitting

One potential method for reducing the confidence interval of ROAS measurement is to add more geos to the experiment [9]. Scaling the number of geos by f_g changes the expected confidence interval by a factor of $\sqrt{1/f_g}$. This means that doubling the number of geos will decrease the confidence interval by a factor of $1/\sqrt{2}$.

An alternative to expanding the geographic coverage of an experiment is to re-partition the same geographic area into a larger number of geos. Once again, scaling the number of geos by f_g changes the expected confidence interval by a factor of $\sqrt{1/f_g}$. So, this approach has the same impact as adding new geos, but it does so without increasing the aggregate ad spend difference. The down side of increasing the number of geos via geographic re-paritioning is that smaller geos are more likely to suffer from control/test contamination. Finite geo location accuracy may inconsistently label consumers who live near boundaries between control and test geos. Consumers are also more likely to travel across these boundaries during the course of their daily activities, including commuting to work.

8 Concluding Remarks

Our previous work demonstrated that geo experiments deserve consideration in many decision-making situations that require the measurement of ad effectiveness. They provide the rigor of a randomized experiment, and they can be applied to a variety of user behavior while avoiding privacy concerns that may be associated with alternative approaches. Here we have demonstrated that these experiments can also be used to track ad effectiveness over time. This additional step expands the applicability of geo experiments to the common situation in which one time measurement is not sufficient to meet the needs of advertisers. As an added benefit of generalizing the application of geo experiments, we also identified a better framework for both periodic and isolated measurement of ad effectiveness.

[9] Decreasing the number of geos included in the experiment is another way to reduce the impact of the experiment on existing campaigns.

Acknowledgments

We thank Tony Fagan for reviewing this work and providing valuable feedback and many helpful suggestions.

References

[1] AdWords Help. "What is the bid simulator, and how does it work?" March 20, 2012. http://support.google.com/adwords/bin/answer.py?hl=en&answer=138148

[2] AdWords Help. "Traffic Estimator" March 20, 2012. http://support.google.com/adword/bin/answer.py?hl=en&answer=6329

[3] Box, G., et al. Statistics for Experimenters. New York: John Wiley & Sons, Inc, 1978.

[4] Kutner, M.H. et al. Applied Linear Statistical Models. New York: McGraw-Hill/Irwin, 2005.

[5] Lucas, J.M. 1982. "Combined Shewhart-CUSUM quality control schemes." *Journal of Quality Technology 14, 51-59.*

[6] Vaver, J., Koehler, J. "Measuring Ad Effectiveness Using Geo Experiments." *http://googleresearch.blogspot.com/2011/12/measuring-ad-effectiveness-using-geo.html*, 2011.

9 Appendix A

The adjacent test periods in a multiple-test-period geo experiment allow information to be used more efficiently than in a single-test-period experiment. Using Equation 7, along with several reasonable assumptions, this efficiency can be characterized analytically.

For this comparison, we consider the confidence interval associated with an ROAS measurement from a single-test-period geo experiment and its multiple-test-period analog; a single test period transition from test period $j-1$ to test period j, where $j > 1$. The length of the pretest period in the single-test-period geo experiment is the same as the length of test period $j-1$. The lengths of the single test period and test period j from the multiple-test-period experiment are also the same, as is the associated ad spend difference.

Now, let p and q be the fraction of geos in the test group for the single and multiple-test-period experiments, respectively. Assume that all groups of geos (i.e. the control and test groups in the single-test-period experiment and the geo-groups in the multiple-test-period experiment) are statistically identical. For example, the distribution of the response metric volume is the same for all groups of geos. Similarly, the mean ad spend difference in each group of test geos is the same as the ad spend difference that would have occurred in each group of control geos, if they had been assigned to the test condition. Let $\bar{\delta}$ be the, realized or unrealized, ad spend difference for each group of geos.

Furthermore, assume that the ad spend difference of each geo in a test group is proportional to the response metric in the pretest period. So, for the single-test-period experiment

$$|\delta_i| = \alpha\ y_{i,0} \qquad i \in \{1, ..., N\} \tag{10}$$

and for the continuous experiment

$$|\delta_{i,j}| = \alpha\ y_{i,0} \qquad i \in \{1, ..., N\}. \tag{11}$$

This assumption is reasonable since we expect geos with larger ad spend to have a larger volume in the response metric, and we expect campaign changes to have a larger absolute impact on ad spend in the larger geos. Going one step further, assume that the impact of the ad spend change is small relative to the differences in the response metric volume across geos so that

$$|\delta_{i,j}| = \alpha\ y_{i,0} \approx \alpha\ y_{i,j} \qquad i \in \{1, ..., N\}. \tag{12}$$

In this analysis, the linear model is modified by ignoring the less important β_{0j} terms. So, for the standard experiment

$$\mathbf{X} = \begin{bmatrix} y_{1,0} & 0 \\ . & . \\ . & . \\ y_{(1-p)N,0} & 0 \\ y_{(1-p)N+1,0} & \delta_{(1-p)N+1} \\ . & . \\ . & . \\ y_{N,0} & \delta_N \end{bmatrix}$$

and

$$\mathbf{X^T W X} = \begin{bmatrix} \sum_{i=1}^{N} w_i y_{i,0}^2 & \sum_{i=1}^{N} w_i y_{i,0} \delta_i \\ \sum_{i=1}^{N} w_i y_{i,0} \delta_i & \sum_{i=1}^{N} w_i \delta_i^2 \end{bmatrix}.$$

Since $w_i = 1/y_{i,0}$ and $|\delta_i| = \alpha\ y_{i,0}$,

$$\sum_{i=1}^{N} w_i y_{i,0}^2 = \frac{1}{\alpha} \sum_{i=1}^{N} |\delta_i| = \frac{1}{\alpha} N|\bar{\delta}|$$

$$\sum_{i=1}^{N} w_i y_{i,0} \delta_i = \sum_{i=(1-p)N+1}^{N} w_i y_{i,0} \delta_i = pN\bar{\delta}$$

$$\sum_{i=1}^{N} w_i (\delta_i)^2 = \alpha \sum_{i=(1-p)N+1}^{N} |\delta_i| = \alpha p N |\bar{\delta}|.$$

Then from Equation 7

$$\begin{aligned} \text{var}(\hat{\beta}_2) &= \sigma_{\hat{\epsilon}}^2 \frac{\frac{1}{\alpha} N|\bar{\delta}|}{[\frac{1}{\alpha} N|\bar{\delta}|][\alpha p N|\bar{\delta}|] - [pN\bar{\delta}]^2} \\ &= \frac{\sigma_{\hat{\epsilon}}^2}{\alpha\ N|\bar{\delta}|\ p(1-p)}. \end{aligned} \tag{13}$$

For the multiple-test-period experiment

$$\mathbf{X} = \begin{bmatrix} y_{1,j-1} & 0 \\ . & . \\ . & . \\ y_{(1-2q)N,j-1} & 0 \\ y_{(1-2q)N+1,j-1} & \delta_{(1-2q)N+1,j} \\ . & . \\ . & . \\ y_{(1-q)N,j-1} & \delta_{(1-q)N,j} \\ y_{(1-q)N+1,j-1} & \delta_{(1-q)N+1,j} \\ . & . \\ . & . \\ y_{N,j-1} & \delta_{N,j} \end{bmatrix}$$

and

$$\mathbf{X^T W X} = \begin{bmatrix} \sum_{i=1}^{N} w_i y_{i,j-1}^2 & \sum_{i=1}^{N} w_i y_{i,j-1} \delta_i \\ \sum_{i=1}^{N} w_i y_{i,j-1} \delta_i & \sum_{i=1}^{N} w_i \delta_i^2 \end{bmatrix}.$$

Since $w_i = 1/y_{i,0}$ and $|\delta_{i,j}| = \alpha\ y_{i,0} \approx \alpha\ y_{i,j-1}$,

$$\sum_{i=1}^{N} w_i y_{i,j-1}^2 \approx \frac{1}{\alpha} \sum_{i=1}^{N} |\delta_i| = \frac{1}{\alpha} N|\bar{\delta}|$$

$$\sum_{i=1}^{N} w_i y_{i,j-1} \delta_i \approx \sum_{i=(1-2q)N+1}^{N} \delta_i = 0$$

$$\sum_{i=1}^{N} w_i (\delta_i)^2 \approx \alpha \sum_{i=(1-2q)N+1}^{N} |\delta_i| = \alpha N|\bar{\delta}|2q.$$

The off diagonal terms in $\mathbf{X^T W X}$ are zero because in $\mathbf{X}$ the ad spend difference terms $\delta_{(1-2q)N+1,j}, ..., \delta_{(1-q)N,j}$ have the same magnitude as the terms $\delta_{(1-q)N,j}, ..., \delta_{N,j}$, but with the opposite sign. Then from Equation 7,

$$\begin{aligned} \mathrm{var}(\hat{\beta}_2) &\approx \sigma_{\hat{\epsilon}}'^2 \frac{\frac{1}{\alpha} N|\bar{\delta}|}{[\frac{1}{\alpha} N|\bar{\delta}|][\alpha 2qN\bar{\delta}]} \\ &= \frac{\sigma_{\hat{\epsilon}}'^2}{\alpha\ N|\bar{\delta}|\ 2q}. \end{aligned} \tag{14}$$

With the assumption that $\sigma_{\hat{\epsilon}}^2 = \sigma_{\hat{\epsilon}}'^2$, the ratio of Equations 13 and 14 gives the ratio of $\mathrm{var}(\hat{\beta}_2)$ from the single-test-period experiment and $\mathrm{var}(\hat{\beta}_{2_j})$ from the multiple-test-period experiment,

$$\frac{\mathrm{var}(\hat{\beta}_2)}{\mathrm{var}(\hat{\beta}_{2_j})} \approx \frac{2q}{p(1-p)}. \tag{15}$$

If $p = q = 1/3$, then the confidence interval of the ROAS in the single-test-period experiment will be $\sqrt{3} \sim 1.73$ times greater than it is in the multiple-test-period experiment. Alternatively, if we assume that both of the confidence intervals are the same, $\mathrm{var}(\hat{\beta}_2) = \mathrm{var}(\hat{\beta}_{2_j})$, then $q \approx p(1-p)/2$. So, for a case in which $p = 1/2$ we have $q = 1/8$. The multiple-test-period experiment delivers the same confidence interval as the single-test-period geo experiment with a test fraction that is only 1/4 as large.

10 Appendix B

This appendix contains ad spend profiles for scenarios from Tables 1 through 3.

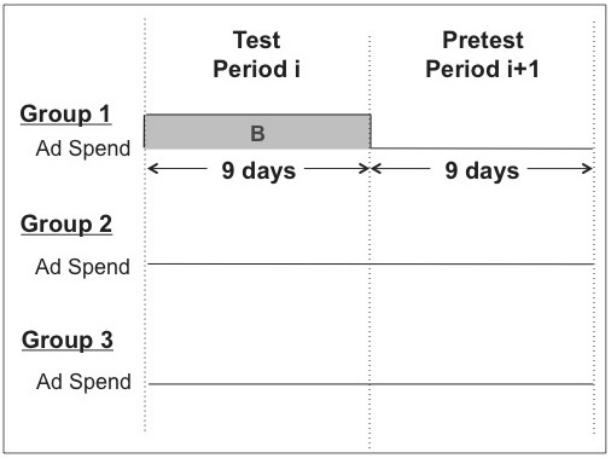

Figure 8: Scenario 1 from Table 1.

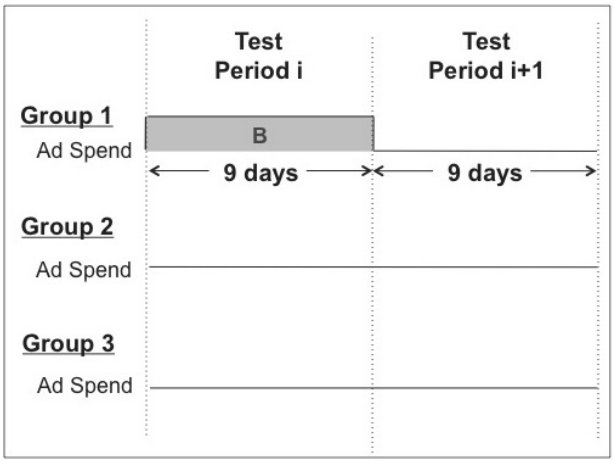

Figure 9: Scenario 2 from Table 1.

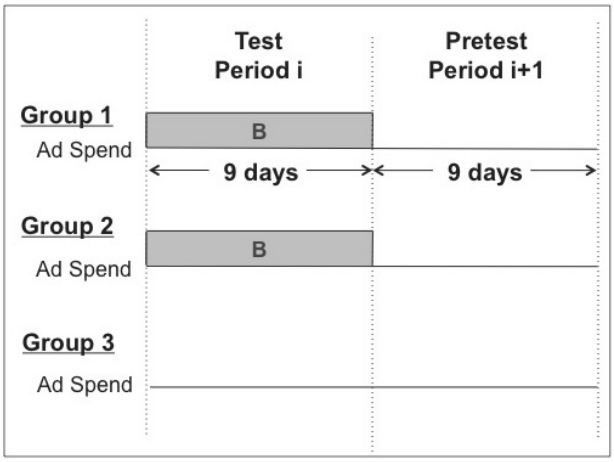

Figure 10: Scenario 3 from Table 2.

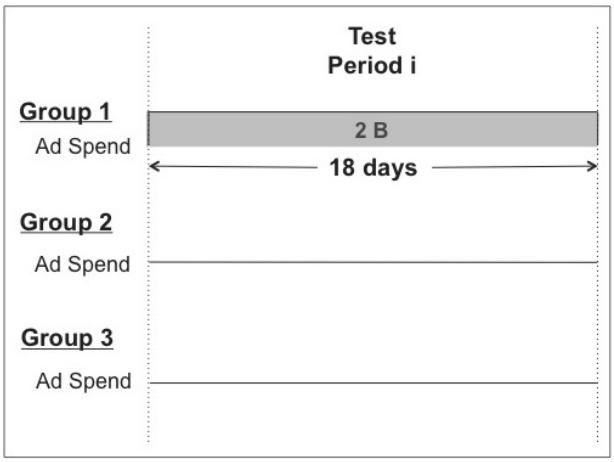

Figure 11: Scenario 4 from Table 2.

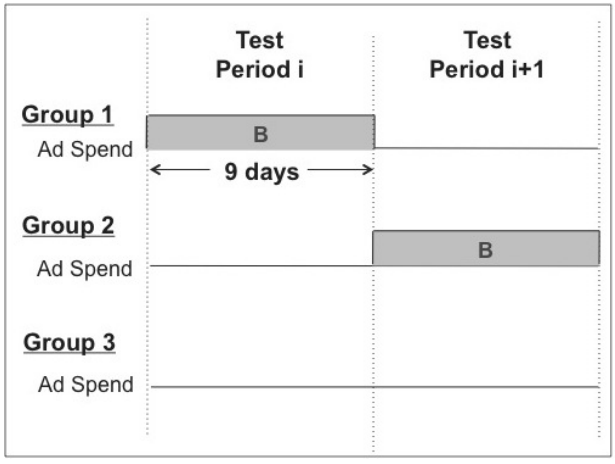

Figure 12: Scenario 5 from Table 2.

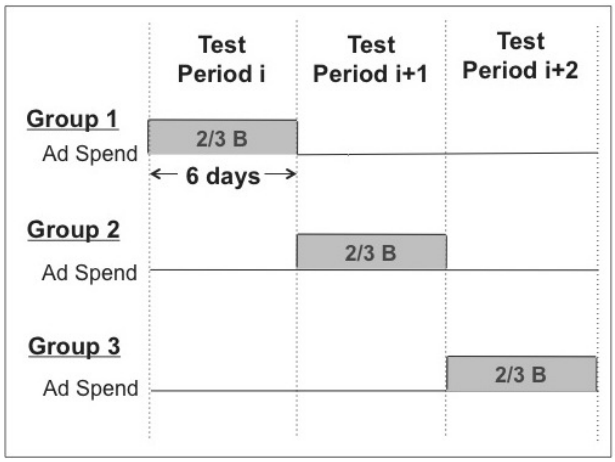

Figure 13: Scenario 6 from Table 2.

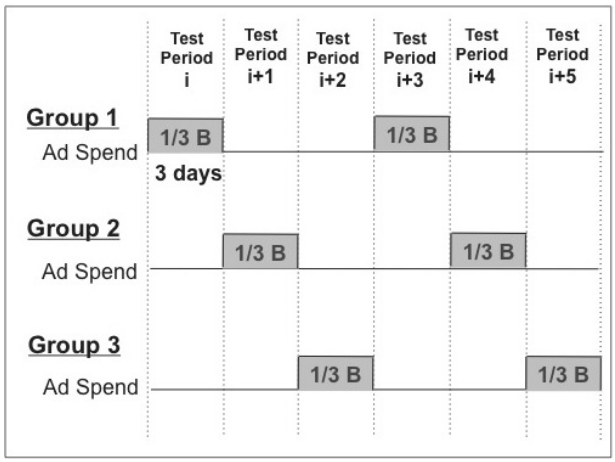

Figure 14: Scenario 7 from Table 2.

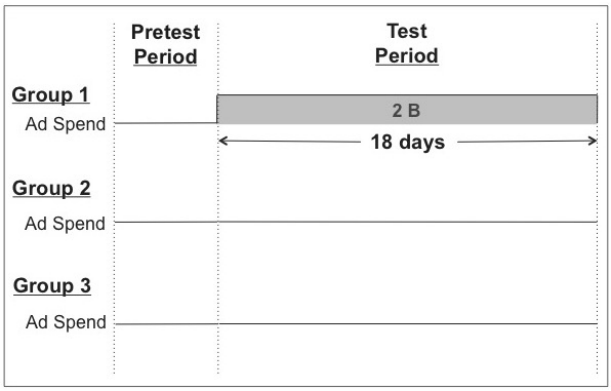

Figure 15: Scenario 1 from Table 3.

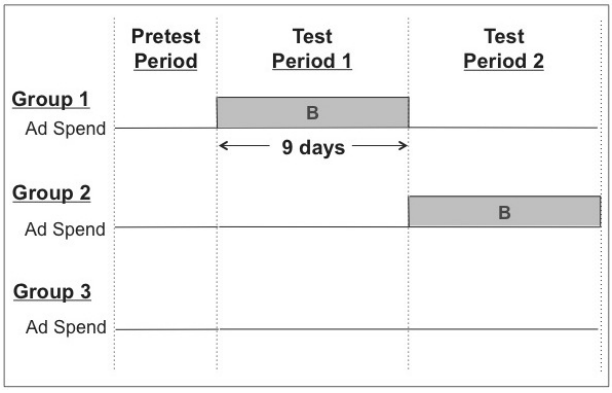

Figure 16: Scenario 2 from Table 3.

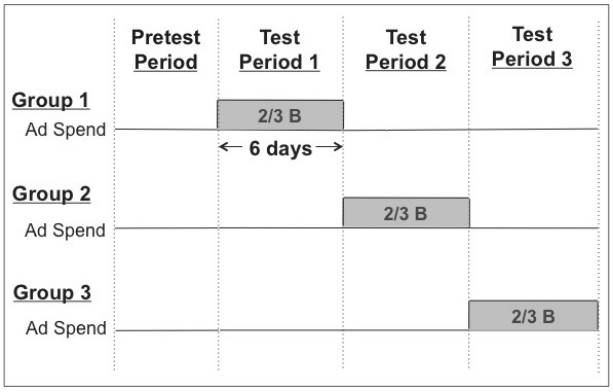

Figure 17: Scenario 3 from Table 3.

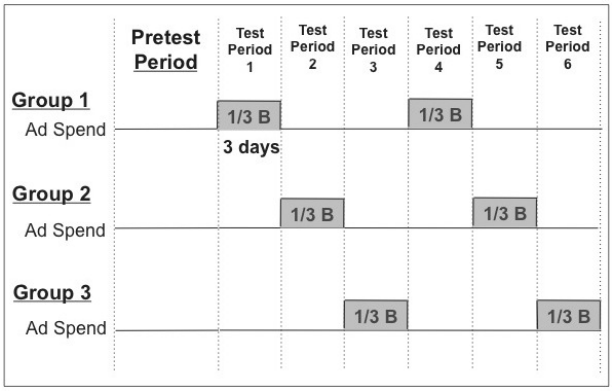

Figure 18: Scenario 4 from Table 3.