BEESWAX WHITE PAPER:

Measuring Incrementality in Digital Media

> Half the money I spend on advertising is wasted; the trouble is I don't know which half.
>
> John Wanamaker (1838–1922)

Digital advertising has long used techniques and tools to attribute the results of ad exposure to commercial outcomes. "Last click", "multi-touch" and "weighted exposure" are some of these techniques, each of which has historically had their own proponents and vendors. Over the past several years, however, marketers have moved beyond the problem of attribution — or, to be less generous, correlation — of ads with outcomes, towards the goal of measuring incrementality, meaning the causation of ads on behaviors.

Incrementality promises to provide the numerator of the return-on-ad-spend ("ROAS") equation. We know what we spent, but what exactly did we return? Finally an answer to Wanamaker's query.

Why is incrementality measurement difficult?

Inevitably you are trying to answer a negative question: If we had not spent this money on advertising, would we have had the same results? Like in a medical context, the gold standard for measurement is a scientifically rigorous double-blind randomized trial. But when you attempt to execute such a rigorous trial in advertising all sorts of complexities arise:

- How do you set up a valid test and control study?
- How do you keep test and control groups consistent over the course of the trial?
- How do you measure the actual outcomes, without falling back on simplistic attribution models?
- How do you pay for and budget the control exposures?
- How do you measure which part of your media techniques are causing the incrementality?
- How do you assure the control group isn't exposed through the myriad of other media channels (TV, etc)?

This is a complex and difficult topic. In this whitepaper we will lay out the challenges and potential approaches to solving each of these problems, with a focus primarily on digital programmatic advertising.

Test/Control: The Gold Standard

If you want to run a randomized trial of your ad campaigns, the first step is to separate your audience into a test group that is exposed to the ads, and a control group that is shown PSAs or another advertiser's creatives. Most ad servers and DSPs offer this functionality as a built-in feature of their platforms. But the underlying assumption, which is no longer valid with the deprecation of cookies and rise of privacy controls, is that the audience can be consistently identified to be properly and consistently segmented.

In order to randomly assign the audience into buckets, platforms will generally take whatever pseudo-random numeric identifier the platform assigns to the user and run a simple modulus function to create a distribution of IDs that can be used to create targeting groups. See the simplistic example below:

How digital platforms create test/control groups

Cookie ID/MAID	Apply Modulus	Normalized Result (0-999)
Sj347sdj10	%1000	561 (test)
22D3A-47	%1000	23.1 (control)

In this example there is a 5% control group so any normalized result <=49 goes into the control group

This methodology is fairly easy to understand, but has significant, and fairly obvious, drawbacks. Both the test and control groups can only be created on users with an ID. And when that ID changes, the user can suddenly and unexpectedly switch groups. *It's as though we're running a pharmaceutical trial and the patients can change or remove their medical bracelets whenever they want.*

The quality of the test/control methodology of a platform is directly proportional to that platform's ability to consistently identify the users. Let's summarize some of the issues of test/control methodology along with potential solutions.

Issue	Description of Problem	Potential Solutions
No ID Users	Users without an ID (cookieless users for example) cannot be persistently placed into a group. These users are not randomly uniform (Apple users have higher incomes, for example) so you cannot assume they will behave the same as other groups.	As discussed in our identity whitepaper [link] there are many emerging solutions to the identity problem, none of which are perfect. Using IP addresses to segment users is a coarse fix, as is potentially using first-party IDs from publishers.
Cross-Device Exposure	The same individual accesses digital content through many devices and channels. Yet the simplistic test/control model is based on a device-identifier.	Platforms can use cross-device graphs to create and maintain user segmentation. However, coverage of these graphs is likely also subject to statistical bias, and deterministic graphs lack scale.
Over-Use of Control Groups	If every customer on a large platform uses the same control group it is possible that those users may become biased since they are exposed to significantly fewer ads than the test group, in aggregate.	Control groups should be randomly distributed, not reused for every experiment.

Take Away: Creating a test/control study can be difficult and has many "gotchas". Be aware of the limitations to your test before assuming your results meet a scientific standard.

So How *Can* Incrementality Be Measured?

To keep with our analogy of clinical trials in medicine, it is vital that at the start of the test we understand how we plan on measuring efficacy. Are we looking to reduce mortality or symptoms? Over what time frame? And with what level of confidence?

In incrementality research this aspect of the study design is often neglected. It is assumed that efficacy will be measured *in the normal way* by the ad platform after properly segmenting the audience. It is important that the measurement method be aligned with the incrementality test design to avoid biased measurement outcomes.

Example: CPG Advertiser Correlating Offline Sales to Online Viewability

Consider a CPG advertiser that wants to measure the correlation between viewability and offline sales. Ads with a third-party viewability vendor tag are shown to all users. Viewable impressions and non-viewable impressions are correlated to offline sales and there appears to be a big uplift from viewability. However, this analysis could be seriously flawed if there are differences between the users who the vendor was able to measure vs those who were difficult to measure (for example, those with ad blockers), as well as the types of sites each of those two groups may frequent.

Relying on the typical last-click and view through attribution within a test/control study can also lead to incorrect or messy interpretations. If the control group is created using a cross-device graph (see discussion above), then the attribution had better use the same graph! If not, results could be biased significantly towards more- or less-measurable conversions. If most of the conversions are view-through based, and users are based on a graph that includes platforms where view-through is less measurable (e.g. in-app), then the study might not show strong results, even when they are present.

Finally, there are special considerations when measuring digital incrementality against an offline sale or activity. This measurement requires a further bridging of the online exposure data to the offline data set, and that bridge may itself have incremental bias, in similar ways to the cross-device example above. If the ID linkages, for example, are exclusively third-party cookie-based but the test/control was built using a more robust graph, then the matches to offline behavior will be biased towards the more measurable users (those with cookies) regardless of which consumers were impacted by the advertising.

One way to avoid these biases is to run the same attribution methodology on the raw auction logs from both the test and control groups instead of using an intermediate set of data like clicks or impressions. This may seem counterintuitive since the test group's impressions (and clicks) provide meaningful signals, but since those signals may be distorted or missing from the control group, looking at the broadest possible correlation may provide more accurate results.

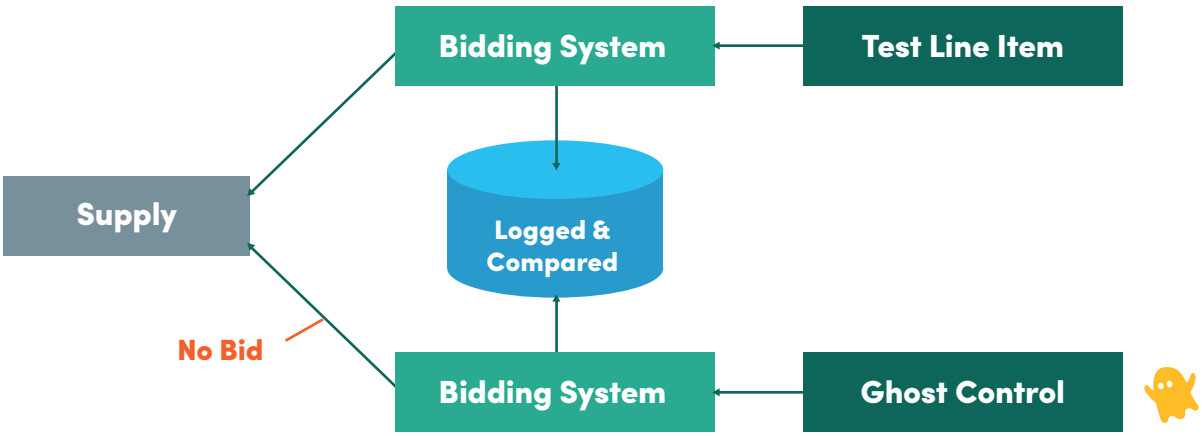
Take Away: Make sure your success measurement is aligned with the test/control group selection and that there are no biases that may be reasonably introduced through measurement. Consider analyzing auction logs instead of impression logs in your attribution model.

Control Ad Wastage: Ghost Ads

The practical problem with incrementality measurement is that it is time consuming and expensive. Specifically, the requirement to serve real, paid-for, control ads can waste up to 8 or 10% of your media budget!

Well some clever folks over the years have developed a deceptively simple sounding technique for overcoming this cost — Ghost Ads. The concept is essentially to architect the ad delivery platform to identify the opportunity to show the control ad, but then simply not serve that ad. These missing ad impressions — the ghost ads — are logged and analyzed as if they were shown, and then the results from these ads are compared to the exposed ads to determine if there was incremental lift.

Highly Simplified View of "Ghost Bidding"



This sounds awesome. But, as is the pattern with all of these incrementality measurement techniques, there are some significant challenges to making this work. Primarily, the questions arise on how to simulate that the ghost exposures are unbiased vs the exposed impressions, since one is a simulation and the other is reality.

Test Ads
(Really Served)

- Actual win rates
- Real auction dynamics
- Frequency caps
- Publisher-enforced creative restrictions
- Ad blockers
- Budgets and caps

Ghost Ads
(Simulated)

- Simulated win rates
- Simulated auction dynamics
- Frequency caps?
- Creative restrictions?
- Ad blockers?
- Budgets and caps?

Issue	Challenges	Best Practice
Pricing and Win Rates	You can't compare ghost ad exposures unless you also simulate which potential exposures would have been won and at what prices. The raw log files of potential ghost exposures are not nearly enough data to measure incrementality since they contain 10x or 20x as many exposures as were available to the test group, which actually had to win auctions to be counted. Further, if pricing is not properly accounted for significant bias can creep into the model as the ghost ads may have been simulated to have been served to lower quality sites with lower clearing prices.	One option is for both test ads and ghost ads to be simulated using the same algorithm based on potential exposures instead of actual exposures.
Auction and Publisher Dynamics	After the platform determines that a ghost exposure would have bid and won a given auction, there are still downstream impacts that would affect the delivery of the test group but are difficult to simulate. For example, DSPs that run an "internal auction" across many customers may result in a lower win rate for the test group than would be estimated by the ghost ad methodology. Also, the impact of viewability vendors that may block exposures after the auction or other publisher-specific blocking technology may under-deliver test vs modeled ghost exposures.	Comparisons should be made between test and ghost ads to see if there are non-representative patterns in the data, such as distribution differences between groups.
Budgets and Frequency Caps	If the test exposures are limited by frequency caps or daily budgets, those limits need to be reflected in the ghost ad simulation. This adds to the difficulty of proper modeling.	Assure these factors are taken into account by the ghost ad model.
Attribution	The problems with measurement described in the section above also apply to ghost ads, and in some cases can be difficult to model. If your primary attribution model is last-click, for example, you would need to simulate which ghost ads were clicked in addition to exposure. Using viewability as a metric has a similar challenge. You may wish to run an entirely distinct attribution model across test and ghost ad logs instead of using your existing attribution model.	Last click attribution should be avoided. A broad correlation between auction data of test vs ghost ads is a good approach, but a more traditional view-through model could also make sense.

Take Away: Ghost ads are a great technique for more affordably running incrementality experiments, but there are significant differences and drawbacks versus a true test/control randomized experiment. It is worth spending some time understanding how your chosen platform executes their methodology.

Digging-In: What Caused the Incrementality?

Suppose you have an incrementality study to test whether certain creative messages caused uplift and the results show positive incrementality. Does this prove the hypothesis that the creative works? Not necessarily. Suppose that early in a test a small difference emerges between the test and control group. Now the bidding algorithm starts to improve bids based on these results, making the test group perform better and better, while the control group doesn't have that feedback loop. In the end you conclude there is a massive difference caused by the creative, when in fact it may have been actually quite small.

The question to ask is “what exactly am I testing”?

- Creative
- Algorithms
- Other factors
- Targeting
- Weighting and capping

Take Away: Know and articulate what you are testing. Try to eliminate factors out of your control in the test unless you want your measurement to cover those elements.

The Challenge of Cross-Channel Incrementality

Rarely does a brand’s entire marketing expenditure take place within the confines of a single platform within the confines of a test. Even with the best designed clinical trial you still need to allow the patients to go home and do all kinds of uncontrolled things. How do you measure incrementality in digital if your test subjects are eating a rich diet of non-tested media outside of your test?

There isn’t a single answer for this question, but there are some approaches:

Approach	Description	Pros/Cons
Pricing and Win Rates	Shut off all marketing activities on certain channels while tests take place on other channels.	Depending on the business, this can be very expensive and potentially hard to execute. Seasonality needs to be considered to avoid biasing towards or against the tested channel. The blackout test length needs to be considered in light of the customer’s purchase consideration length.
Blackouts by geo	Shut off marketing activities in certain zip or postal costs while tests take place.	Much less complex and costly than channel-based blackouts and relatively easy to coordinate across multiple media channels. Need to be careful to make sure the geographic scope is large enough to cover commuting and travel behaviors.

Take Away: No incrementality technique will get you “always on” results across media channels, but there are one-time studies you can arrange to give you useful information.

Conclusion

Digital marketing and advertising has always promised accountability and measurement. Marketers and agencies have long relied on flawed correlation metrics to measure ad effectiveness, but now new techniques and approaches promise to help measure *causation and incrementality* instead.

Measuring causation is inherently difficult as the decision to purchase a product lies solely within the grey matter of the consumer’s brain. We can use scientific methods to tease out the data using randomized test/control methods and ghost bidding. But the devil is in the details. Running unbiased and scientifically valid studies can be quite hard and the default capabilities of many platforms may not be as robust as desired when used to make multi-million dollar media decisions. Smart marketers should ask the tough questions and really understand the benefits and limitations to the experimental methodologies they are utilizing and make informed decisions about what the data says, and what it doesn’t.