Causal inference under limited outcome observability: A case study with Pinterest Conversion Lift

Min Kyoung Kang
Pinterest, Inc.
1099 Stewart St, Seattle
WA, USA
mkang@pinterest.com

ABSTRACT

This paper compares the performance of several established causal inference estimators in measuring conversion related metrics for advertising measurement applications. Conversion lift measurement in advertising industry presents unique challenges due to complex data collection process, potential data losses, and complex customer behaviors leading up to conversion. Case studies with both simulated and real-world data demonstrated that doubly robust estimators outperform regression adjustment estimators in variance reduction for ad measurement use-cases. To further understand the results, we examine the impact of data loss on variance reduction by the estimators and find that the relationship between data loss and variance reduction performance varies by the estimators. Doubly robust estimators could effectively manage complex relationships introduced by data loss, maintaining superior performance over the difference-in-means and regression adjustment estimator in terms of precision under various circumstances. We provide computational cost perspectives as practical considerations for implementing doubly robust estimators in advertising measurement business solutions.

1 Introduction

Accurately measuring the impact of placing advertisement in online platform industry, often referred to as advertising measurement, is a challenging but one of the most important tasks for customer behavior understanding. In particular, measurement reporting for conversion count and volume metrics provides critical guidance on business decisions for optimizing marketing directions based on user understanding. Such measurement process encompasses multiple phases of execution including holdout experiments, data collection, and inferential analysis. Each of these stages has potential to introduce unwanted uncertainty into the data, thereby introducing challenges for the inference [12; 2]. Holdout experiments can suffer from compliance issues meaning that customers may not actually impress the ads even if they are assigned to treatment group. Customer behaviors related to conversions are inherently complex and sparse since in general customers convert occasionally after a number of non-linear interactions. For privacy concerns and data sharing agreement limitations, data loss can occur limiting the observability of outcome metrics. These factors increase the complexity of treatment effect estimation and can potentially decrease the power of experiments. Extending duration of the holdout experiments to account for power incurs opportunity cost for advertisers, as holdout experiments withhold advertisements to control users, limiting the reach to high-potential customers within campaign period.

This work aligns with a number of literature and researches on ways to improve the power and sensitivity of controlled randomized experiments [10; 3; 5; 9; 15]. Such effort is commonly referred as variance reduction techniques, which leverage covariates that explain the variability unrelated to treatment from outcomes to reduce variance of the outcome metrics of interest. This is the area of active investigation and interests as it can decrease the costs of running experiments and extract insights from otherwise inconclusive experimental results. However, there is limited research on variance reduction techniques under the context of advertising measurement with causal inference estimators. Literature in advertising measurement mainly discusses causal inference approach to measure impact of advertising with observational data [4; 7; 13]. We apply causal estimators to holdout experimental settings to evaluate their performance in terms of precision within the unique context of the ad measurement industry, which includes data loss. This work extends existing variance reduction causal inference literature into advertising measurement domain by exploring the opportunities at the intersection of variance reduction in ad measurement research. It aims to inform practitioners in advertising industry who want to improve the sensitivity of their measurement reporting with practical considerations.

The remainder of the paper are organized as follows. Section 2 introduces four established estimators for inferential analysis on holdout experiments in ad measurement use-cases. In section 3, we compare the performance of the estimators with case studies and performs empirical analysis with both real-world and simulated data. We conclude the paper with the discussion on production adoption and business impact considerations while implementing the method in practice.

2 Estimators for advertising measurement

This section introduces the formal definitions of estimators that will be compared in case studies to empirically analyze their performance in the context of conversion advertising measurement. A randomized controlled trial (RCT), also referred to as online randomized experiment, is an industry-standard measure for measuring incrementality of advertis-

ing without confoundings. To formally denote the experimental settings, we denote a treatment as T, outcome as Y, covariate as X, and one instance of experimental data points with i. Historical data we observe provide scientists with a set of data for each instance $(T_i,\ Y_i,\ X_i)$, where X is a set of covariate vectors that are presumed to be closely related with Y according to empirical evidence. Under controlled randomized experimentation, $(Y_i(0),\ Y_i(1)) \perp T_i$ holds. Sometimes experiment experiences opportunistic imbalance due to experiment quality issues, but we can assume $(Y_i(0),\ Y_i(1)) \perp T_i|X_i$.

2.1 Difference-in-Means estimator

The difference-in-means estimator (DIM) is a traditional statistical approach that calculate delta between average outcome of treatment and control groups. This provides information on the significant differences between the teams of two populations.

$$\text{ATE}_{dim} = \overline{Y}_{tr} - \overline{Y}_{ctl} \tag{1}$$

, where $\overline{Y}_{group}$ is for users that belong to a particular group (treatment or control) and defined as $\frac{1}{n_{group}} \sum_{j \in group} y_j$ for outcome of interest y for individuals from control and treatment groups.

2.2 Regression adjustment estimator

Leveraging unit-level covariates in post-experiment inferential stage is known to enhance the precision and efficiency of causal estimates. A traditional method for adjusting treatment effect with covariates is ordinary least square (OLS) regression adjustment (RA). This approach has been studied in various literatures and known for its ability to asymptotically improve the precision of estimators [15; 11; 9; 14; 17]. This work examines two model specifications for regression adjustment estimators and the only difference between the two models is the inclusion of interaction terms between the treatment and covariates.

$$Y_i = \alpha + \tau T_i + \beta X_i + \epsilon_i \tag{2}$$

$$Y_i = \alpha + \tau T_i + \beta X_i + \gamma(T_i \cdot X_i) + \epsilon_i \tag{3}$$

The literature demonstrated that (3) yields better statistical properties [15]. For both of model specifications, the average treatment effect is estimated by solving the regular OLS optimization process to determine model parameter τ.

2.3 Doubly Robust Estimator

Doubly robust (DR) estimator is a causal inference approach to estimate the treatment effect in a doubly robust manner [6; 1]. The robustness comes from incorporating two modeling approaches in calculating estimators, which are propensity score model and outcome model. DR estimators have desirable statistical properties of consistency and efficiency. As long as either of the propensity score model or the outcome model is correctly specified, the doubly robust estimator yields a consistent estimator. Thus, subtle misspecification of either of the outcome and propensity model does not affect the treatment effect estimation. Such double (or dual) robustness characteristics increase the chances that the estimator has minimal population risk in practical applications. When both models are correctly specified, the doubly robust estimator can account for available covariates' into the model and achieves the resulting lowest possible variance. As the methodology can accommodate various machine learning models with regularization, it can handle high-dimensional covariates with both nonparametric and parametric functions to account for complex relationships among various factors. To formally introduce DR estimator, the potential outcome under treatment assignment of t for unit i is formulated with outcome model $f_t(x) = E[Y(T = t)|X = x]$ and propensity score model $p(t, x) = E[T = t|X = x]$ as

$$\widehat{Y}_i(t) = f_t(x) + \frac{(Y_i - f_t(x))}{p(t, x)} \mathbf{1}(T = t) \tag{4}$$

With the provided potential outcome, treatment effect function ϕ can be estimated with the following minimization process, which can be customized to account for effect heterogeneity as needed.

$$argmin_\phi \sum_i (\widehat{Y}_i(1) - \widehat{Y}_i(0) - \phi(X))^2 \tag{5}$$

To obtain ϕ through this optimization process, nuisance parameters such as f (outcome model) and p (propensity score model) are estimated from the data using machine learning techniques. For the analysis of empirical results, we selected random forest models for both f and p, which is tuned using 4-fold cross validation processes on the entire dataset to effectively capture sparse conversion data. As advertisers are most interested in average impact of the advertisement, we estimate average treatment effect assuming the effect homogeneity across population. Thus, for the analysis for the following case studies, $\phi(X)$ is reduced to a simple estimation with $\alpha + \tau T_i$ that measures homogeneous treatment effects in this analysis.

To maintain the consistency of the inference process across various estimators, we construct the analytical confidence interval for all estimators with OLS parameter τ, coefficient of treatment assignment variable T_i. Based on the asymptotic normality of τ, we construct Wald 95% confidence interval, which is a traditional approach for regression analysis. More detailed explanation of calculating confidence interval from OLS optimization can be found in [16].

3 Empirical results with case studies

This section utilizes causal estimators introduced in previous section under the context of controlled randomized experimentation to understand the variance reduction performance of the estimators introduced. For this analysis, both real-user data from pinterest conversion lift studies and simulated data are used to understand the impact of various environmental factors in variance reduction performance under advertising measurement context. This analysis adopts three evaluation metrics: the first assesses the level of variance reduction, the second measures the coverage probability of the confidence interval when ground truth value is available, and the third quantifies the deviation between true and estimated effects at point-estimate level. The second and third metrics are calculated exclusively when ground truth treatment effects are available, as in the case of simulated data. The level of variance reduction is measured against the relative confidence interval ratio of 3 estimators against of difference in means estimator, which is baseline measurement solutions in industry. All confidence interval

is calculated at a 95% of significance level.

$$\text{VR}_{est} = CI_{est}/CI_{DIM} \tag{6}$$

The coverage percentage metrics indicate the proportion of confidence intervals, each obtained from iteration of simulated data, that includes the treatment effect within their lower and upper bound.

$$\text{Cov\%}_{\text{est}} = \frac{\sum_{j=1}^{N} \mathbf{1}(\text{lower bound}_j \leq \tau_j \leq \text{upper bound}_j)}{N} \tag{7}$$

The mean squared error (MSE) metric represents the averaged of squared error, which is the delta between ground truth and calculated treatment effect from the estimators.

$$\text{MSE}_{\text{est}} = \sum_{j=1}^{N} (\hat{\tau}_j - \tau_j)^2 \tag{8}$$

We assess the performance of four estimators ATE_{DIM} (difference in mean), $ATE_{RA(3)}$ (OLS adjustment without interaction), $ATE_{RA(4)}$ (OLS adjustment with interaction), ATE_{DR} (doubly robust estimator) utilizing three performance metrics in the remainder of sections.

3.1 Pinterest conversion lift (PCL) case studies

Pinterest conversion lift (PCL) study quantifies the value of placing advertisements on the Pinterest platform for various lower-funnel performance advertisements. This quantification includes a wide range of customer conversion-related engagement following exposure of performance ads in Pinterest, allowing advertisers to measure and optimize campaign performance. One of the goals of PCL study is to provide advertisers with sufficient statistical power, so that the study delivers maximum value to advertisers with accurate and actionable insights within their campaign budget constraints. This aims to minimize the number of studies yielding inconclusive results attributing from the high variability of outcome metrics. This case study compares the estimators aiming to decrease variability of outcome metrics unrelated to the treatment of the holdout experiment to increase statistical power of the experiment.

Table 1: Quantitative summary of the performance of the estimators for outcome metrics from PCL studies. ('Change in stat sig rate' column displays the difference in statistical significance rate between the DIM estimator and the selected estimator of interest.)

Outcome	Avg %			
	Reduction in Variance (DR)	Reduction in Variance (RA(3))	Reduction in Variance (RA(4))	Change in stat sig rate
Metric 1	7.6%	0.06%	0.06%	3.0%
Metric 2	9.8%	0.19%	0.21%	6.9%
Metric 3	13.1%	0.05%	0.06%	11.2%

To evaluate the variance reduction performance, we compared the confidence interval using variance reduction metrics with difference-in-means estimator as the baseline model; PCL currently utilizes difference in means estimators following the holdout experiment data collection. Treatment effect estimates yielded from various estimators based on selected historical PCL's experimental data. These experiments are randomly selected from delivered studies to advertisers dating back as early as 2023 July. The advertisers selected for this study represent a diverse set of industries, including but not limited to online retail, telecom services, finance, etc. Experiment data is collected from various regions, including North and South America, as well as the EU.

For each of experiment, three types of outcome metrics are considered, which are denoted as metric 1, 2, 3, anonymized to maintain the confidentiality of sensitive internal data. Due to the unkown ground truth treatment effect, we solely focused on comparing variance reduction metrics and did not calculate coverage probability or mean squared error (MSE) metrics for the performance measurement. However, it is important to note that the estimated treatment effects across different estimators are generally consistent, exhibiting minor variations and largely overlapping confidence intervals. The performance metrics indicate relative measures to compare the level of variance reduction relative to difference-in-means estimator. Experimental outcome metric names are anonymized and absolute values of performance metrics are not shown to maintain the confidentiality of sensitive internal data. For the estimation of the treatment effect and confidence interval, we utilized holdout experiment data from historical PCL studies and a number of covariates are collected by summarizing various information of individual user activities on Pinterest platform prior to the experiment start date.

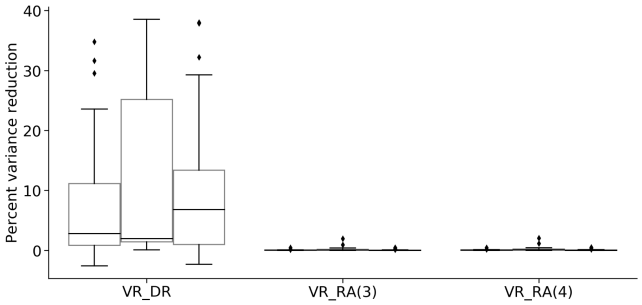

Figure 1: Box plots for variance reduction performance comparisons. This plot compares the percentage of variance reduction across metrics 1, 2, and 3 for the estimators introduced in section 2. Each group of box plots represents the performance of each estimator; within each group, three boxes correspond to metrics 1, 2, and 3, respectively. The vertical axis displays the percentage of reduction in variance, providing a visual representation of the distribution of reduction across selected PCL studies.

For almost all experimental cases selected, DR estimator consistently outperformed regression adjustment estimators in Fig.1 and TABLE 1, demonstrating its effectiveness in accounting for variability orthogonal to the experimental treatment under complex data structures.

For the doubly robust estimator, the percentage reduction in variance is consistently higher across all three metrics. This indicates a reliable performance in reducing variance under various circumstances, leading to the greatest increase in the statistical significance rate. Subtle differences in the level of percentage reduction are explained by metric types; metrics involving volumes tend to have higher variance due to the larger values associated with data. In contrast, metrics such as counts, which involve smaller integer values, typically exhibit lower variance. The doubly robust estimator is more

effective at reducing variability in metrics with larger volumes and higher inherent variability compared to those with lower activity and smaller numerical values. Interestingly, regression adjustment was ineffective in reducing variance and increasing the statistical significance rate. These estimators utilize linear models, which may have limitations in capturing the complex nature of user behaviors and advertising conversion data, leading to lower statistical significance in measuring treatment effects. We deep dive into such dynamics with simulated data in the next subsection.

3.2 Simulated data case studies

To assess the performance of the estimators for the purpose of advertising measurement, we generated simulated data with constant and homogeneous treatment effect across units of observation. Treatment assignment is randomized without confounding, while outcomes are influenced by both treatment effect based on assignment status and other randomly generated covariates. For the simulated data, we adopt a data generating process (DGP) in [8] to account for complex non-linear relationship between outcome and various factors as in real experimental set-up. This DGP exhibits varying fluctuation and smoothness across multi-dimensional variables included in the formulation. Each iteration of simulation randomly generates data of size $N = 100,000$ units and this process is repeated $5,000$ times to assess their average performance. To account for unique data collection process in ad measurement scenario, we randomly mask outcome data under the scenario of varying level of data loss. The following information describes the variables used to generate simulated data and to estimate treatment effects:

Table 2: Coverage %, Variance Ratio, and Mean Squared Error for different estimators

	DR	**RA(3)**	**RA(4)**	**DIM**
Cov %	94.72%	94.64%	94.45%	94.52%
VR	15.0	11.5	11.6	1
MSE	0.024	0.025	0.025	0.028

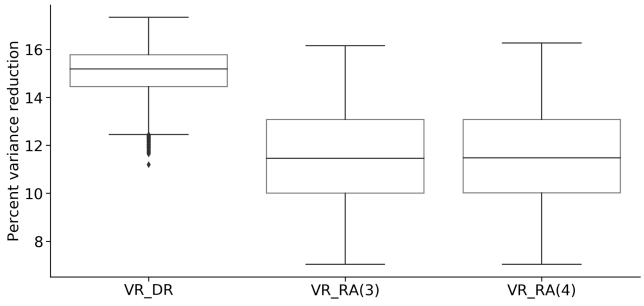

Figure 2: Comparison of percentage variance reduction for various estimators versus difference in means estimator

- Y_i: denotes the outcome variable for the i-th instance. Outcome, which is affected by both four covariates and treatment.
- T_i: denotes the treatment variable for the i-th instance. Under holdout experimental set-up, treatment assignment mechanism is randomized. Thus, treatment is simulated with Bernoulli distribution as $T_i \sim Bernoulli(1/2)$
- X_{ij}: denotes the j-th covariates for the i-th instance. All covariates follow independent and identically distributed normal distributions with mean zero and variance one. We generated 100 covariates and only 4 of them are used for DGP process, while rest of 96 covariates are unrelated to outcome. All of 100 covariates are included in the treatment effect estimation process.
- τ_i: represents the injected treatment effect which we adjust to account for random data loss when calculating the ground truth effect that we aim to estimate in this case study. For each of simulation data generation, $\tau_i \sim \text{Uniform}(0.5, 1.5)$ to validate the model performance under various conditions.
- m_i: represents the observability rate for the outcome Y_i. Given the data loss is common in collecting conversion labels in ad measurement industry, we randomly mask some portion of data at a rate of $1 - m_i$, where $m_i \sim \text{Uniform}(0.7, 0.95)$
- I_i: represents the indicator variable for each unit to decide its outcome loss status. $I_i \sim Bernoulli(m_i)$ and I_i is used to mask outcome Y_i to be zero. $I_i = 0$ induces the scenario of outcome data loss for i-th instance.
- ϵ : represents normally distributed random noise with mean zero and variance one $\sim \mathcal{N}(0, 1)$

With the notations defined, the outcome of interest is formulated to reflect complex relationship adopted from [8]:

$$\begin{aligned} Y_i * I_i \;\; = \;\; & \tau T_i + \\ & \exp\left(\sin\left(0.9 \cdot (X_{i1} + 0.48)^{10}\right)\right) + \\ & (X_{i2} \cdot X_{i3}) + X_{i4} + \epsilon \end{aligned} \tag{9}$$

Fig.2 compares the percentage of variance reduction for various estimators compared to difference-in-means estimator. The vertical axis displays the percentage of reduction in variance, informing the distribution of reduction achieved by each estimator for simulated data under various data loss. The simulation study confirmed again that DR estimator consistently outperforms regression adjustment estimator for variance reduction. Quantitative comparisons of the performance of estimators are in TABLE 2. DR estimator also enhances the precision of metrics decreasing mean squared errors.

We further deep dive into the variance reduction performance of each estimator by the observability rate, which mimic various data loss scenario unique to ad measurement data. The purpose is to understand the relation between data loss and variance reduction performance to identify more appropriate estimators under ad measurement use-cases. The simulation analysis suggests that the regression adjustment models' variance reduction performance decreases as data loss increases. This results explain the PCL case study where we found that regression adjustment analysis was not able to further reduce the variance compared to DIM estimator.

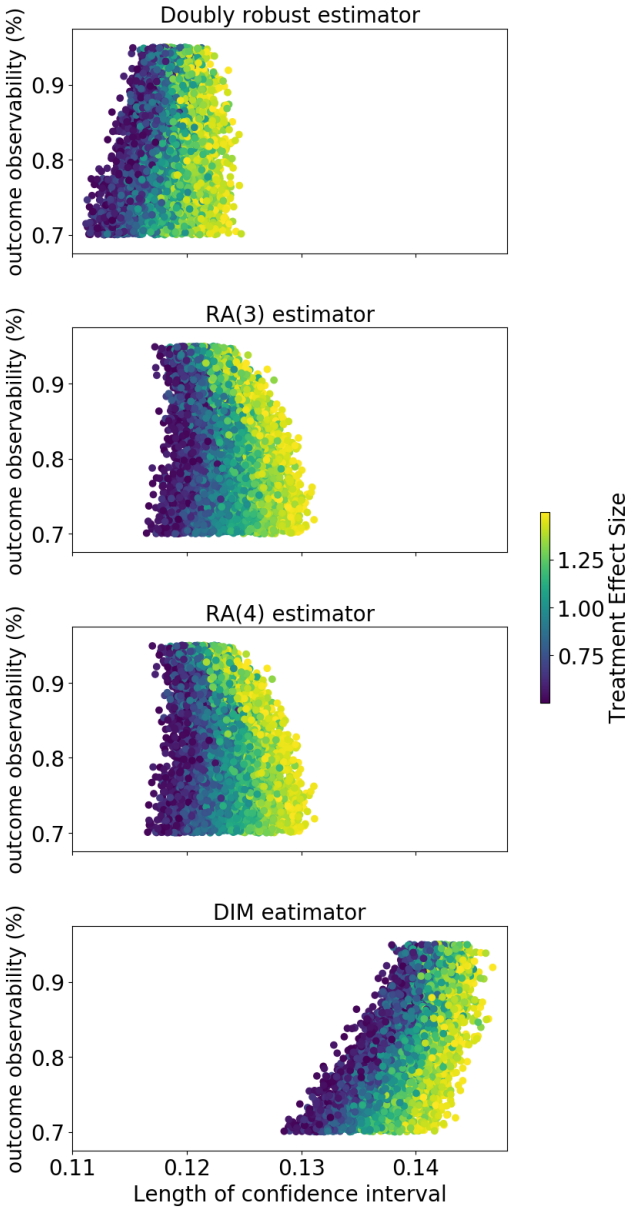

Figure 3: Scatter plots that compare the confidence intervals across estimators by outcome observability and treatment effect size

Fig.3 compares the confidence interval calculated from four different estimators by the various level of outcome observability and treatment effect size. Scatter plots from four estimators show that as outcome observability grows, regression adjustment based estimator's confidence interval starting to overlap with DIM estimator's, while doubly robust estimator maintains its relative advantage.

4 Conclusions

In this paper, we compared several established estimators in the causal inference literature, particularly focusing on their performance for advertising measurement purposes. Advertising measurement presents unique challenges due to the complexities involved in data collection, subsequent data losses, and the intricate relationships across various factors within the data. Our case studies, which included both simulated and real-world data, demonstrated that doubly robust estimators outperform regression adjustment estimators in variance reduction while maintaining coverage probability and reducing the error in point estimates. We also identified the impact of data loss on variance reduction through simulation studies. By examining the relationship between the proportion of data loss and relative variance reduction performance, we observed that as the data loss percentage increases, the regression adjustment model decrease its variance reduction performance trending toward the level of difference in means estimator, while the doubly robust estimator maintains its superiority in variance reduction. Machine learning models used in doubly robust estimators could better handle unexpected intricate relationships introduced into the data due to data loss, compared to regression adjustment-based estimators. Doubly robust estimators are also well-known for their ability to provide robust estimates under mis-specifications of either outcome or propensity score models. In complex non-linear datasets with stochastic factors, the properties of doubly robust estimator improves the precision of inference results.

The inference process of online experiments often involves massive data volumes and training machine learning models on sizeable data can pose challenges due to computational costs. While cross-fitting could further reduce bias, it may introduce challenges due to massive data size, operational costs, and data sparsity. By adopting cross-validation process in model training, we prevent over-fitting of ML models utilized for outcomes and propensity score models, thereby minimizing potential bias issues due to over-fitting. To further mitigate cost related challenges, one can also consider a two-tiered services to leverage high variance reduction solutions when necessary. With this approach, we offer the practical perspectives of the measurement solution, balancing variance reduction performance against costs and scalability considerations.

5 REFERENCES

[1] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[2] J. Barajas, N. Bhamidipati, and J. G. Shanahan. Online advertising incrementality testing: practical lessons and emerging challenges. In *Proceedings of the 30th ACM*

International Conference on Information & Knowledge Management, pages 4838–4841, 2021.

[3] S. Baweja, N. Pokharna, A. Ustimenko, and O. Jeunen. Variance reduction in ratio metrics for efficient online experiments. In *European Conference on Information Retrieval*, pages 292–297. Springer, 2024.

[4] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–16, 2010.

[5] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 123–132, 2013.

[6] M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

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

[8] R. B. Gramacy and H. K. Lee. Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51(2):130–145, 2009.

[9] K. Guo and G. Basse. The generalized oaxaca-blinder estimator. *Journal of the American Statistical Association*, 118(541):524–536, 2023.

[10] Y. Guo, D. Coey, M. Konutgan, W. Li, C. Schoener, and M. Goldman. Machine learning for variance reduction in online experiments. *Advances in Neural Information Processing Systems*, 34:8637–8648, 2021.

[11] G. W. Imbens and J. M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

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

[13] R. A. Lewis, J. M. Rao, and D. H. Reiley. Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th international conference on World wide web*, pages 157–166, 2011.

[14] X. Li and P. Ding. Rerandomization and regression adjustment. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):241–268, 2020.

[15] W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. 2013.

[16] J. Neter, M. H. Kutner, C. J. Nachtsheim, W. Wasserman, et al. Applied linear statistical models. 1996.

[17] A. A. Tsiatis, M. Davidian, M. Zhang, and X. Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in medicine*, 27(23):4658–4677, 2008.