Correcting the Winner's Curse in Large-Scale A/B Testing: A Simulation-Based Evaluation

Yinghao Liu[1], Chenyang Ren[1], Yuchen Chen[1], Hong Ru Chan[1], and Henry Su[1]

[1] Department of Statistics, Columbia University

Team Members and Contributions

- **Yinghao Liu (yl5645):** Analysis, Methodology Review
- **Chenyang Ren (cr3417):** Analysis, Methodology Review
- **Yuchen Chen (yc4502):** Context, Executive Summary, Methodology Review, Conclusion
- **Hong Ru Chan (hc3545):** Context, Executive Summary, Methodology Review, Conclusion
- **Henry Su (hs3539):** Context, Executive Summary, Methodology Review, Conclusion

1 Context

1.1 Problem Description

In this project, we address a fundamental statistical bias that arises in large-scale online controlled experiments, commonly known as A/B testing. Tech companies like Airbnb, Google, and Meta often run hundreds of concurrent experiments to test small changes to websites or apps, such as new layouts, pricing strategies, or recommendation algorithms. Typically, only experiments that demonstrate statistically significant and positive results are selected for launch to all users.

This selection process introduces a hidden pitfall known as the Winner's Curse. Even when no true improvement exists, random chance can cause some experiments to appear more successful than they actually are. Adding up the observed effects of only these "winning" experiments leads to inflated estimates, making product changes seem more effective than they truly are.

The key problem we aim to solve is: How can we accurately measure the true total impact of features selected from many A/B tests, without being misled by selection bias?

To solve this problem, we:

- Quantify the magnitude of bias introduced by selecting only statistically significant ("winning") experiments
- Develop a correction formula to adjust the estimated total effect and remove this bias
- Use bootstrap simulations to construct confidence intervals and quantify the uncertainty around the corrected estimate
- Validate the approach through simulation experiments that mimic real-world A/B testing conditions, following the methodology from the original research paper

In simpler terms:

When companies test many ideas and only keep the ones that seem to "work," the total gains often look better than they truly are. Our goal is to correct this statistical illusion and measure the real, unbiased success of the launched features.

1.2 Why this matters

This problem has serious implications for real-world data-driven product development, especially at companies like Airbnb, Google, or Meta, where hundreds of A/B tests are run simultaneously to evaluate new features, pricing strategies, and design changes. These experiments guide high-stakes decisions such as whether to launch a feature, shift investment, or redesign parts of the user experience.

When teams rely only on tests that show strong, statistically significant results, they unknowingly introduce selection bias. This can cause them to overestimate progress or attribute success to features that performed well by chance. If results are skewed due to the Winner's Curse, decisions may ultimately be based on inflated metrics rather than true performance.

Correcting this bias is essential because it:

- Provides a more honest and realistic picture of business impact, improving return-on-investment (ROI) estimates
- Reduces overconfidence in experimental results, lowering the risk of flawed product or investment decisions
- Prevents false attributions of success driven by random noise rather than real improvements
- Preserves statistical integrity when many experiments are being run and interpreted at once

By applying a correction formula and constructing confidence intervals around the estimated effect, companies can:

- Make better-informed, data-backed product decisions
- Improve experimentation platforms, such as Airbnb's ERF, by automating bias correction
- Scale experimentation to support rapid product development without sacrificing measurement accuracy

In short, solving this problem helps companies stay innovative while remaining statistically rigorous, ensuring that decisions are based on reliable evidence, not just promising-looking results.

1.3 Challenges

The difficulty in addressing this problem lies not in how experiments are conducted, but in how results are selected and interpreted. In A/B testing, teams tend to focus on experiments that yield statistically significant outcomes (e.g., $p < 0.05$), treating these as "successful" and prioritizing them for reporting or implementation.

While this practice is widespread, it introduces structural bias. Even when there is no true effect, random variation can make some results appear favorable. When only those results are aggregated, the average estimated effect becomes misleadingly high.

What makes this bias especially challenging is that it often goes undetected. Because it stems from a common selection practice rather than experimental error, it can easily be overlooked unless explicitly measured and corrected — even in otherwise well-designed testing systems.

1.4 Illustrative Example for a General Audience

To illustrate the issue more concretely, imagine a company runs 100 A/B tests. Only 10 appear to improve outcomes and are selected for launch. The company adds up the results from those 10

and reports a 10% overall improvement. However, many of those "successful" results may have been due to random chance. Ignoring the remaining 90 experiments leads to an overly optimistic estimate of total impact.

Our work addresses this problem by:

- Quantifying the bias introduced by selecting only statistically significant results
- Applying a correction formula to more accurately estimate the true overall effect
- Running simulations to demonstrate the magnitude of the correction and validate its effectiveness

2 Executive Summary

In this project, we identified and corrected the "Winner's Curse" bias that occurs when only statistically significant A/B tests are selected to estimate the total impact of product changes. By implementing a bias-corrected estimator and validating it through simulation, we showed that naive aggregation overestimates the true effect, while our correction method provides a more accurate and fair estimate. This method improves the reliability of data-driven decision-making in large-scale experimentation platforms.

3 Methodology Review

To address the selection bias introduced by launching only statistically significant A/B tests, the authors propose a debiasing method that adjusts the total estimated effect to better reflect the true underlying impact. The central idea is to quantify how much of the observed effect is likely due to random chance and subtract that from the naive total.

3.1 Problem Setup

In large-scale experimentation, teams typically launch features based on statistically significant results. Summing the effects of only these selected experiments results in an overestimate of the true aggregate impact. The goal of the method is to correct this overstatement by estimating and removing the average "lucky uplift" associated with selection.

3.2 Bias Correction Procedure

The debiasing approach involves computing a correction term for each experiment and subtracting the total estimated bias from the naive sum. The process works as follows:

1. **Collect inputs:** For each experiment, obtain the estimated treatment effect (e.g., lift) and its standard deviation (i.e., variability).
2. **Identify selected experiments:** Mark those that exceed the significance threshold (e.g., p < 0.05).
3. **Calculate naive total:** Sum the observed treatment effects from selected experiments — this is the naive total, denoted as S_A.
4. **Estimate bias:** For each experiment, estimate how much of its observed effect might be due to random chance. This bias depends on:
    - The experiment's standard deviation
    - How far the observed effect lies above the selection threshold
5. **Debias the total:** Subtract the sum of all estimated biases from the naive total to obtain the corrected estimate. In other words: $\hat{T}_A = S_A - \hat{\beta}$, where $\hat{\beta}$ is the total estimated bias across all experiments.

This adjustment helps correct for the fact that selected experiments may appear stronger than they actually are due to random variation, not true underlying effects.

3.3 Confidence Interval Construction

To assess the uncertainty of the corrected estimate, the paper uses a bootstrap-based procedure:

1. Resample the original experiment results with replacement to create simulated datasets.
2. For each resample, compute the naive total, bias estimate, and corrected total.
3. Repeat this process many times (e.g., 1,000 iterations) to form an empirical distribution of corrected estimates.
4. Use the spread of those results to create upper and lower bounds of percentile-based confidence intervals.

This allows for robust inference even in the presence of selection effects and variability across experiments.

3.4 Why This Method Is Effective (Compared to Naive Aggregation)

- Corrects for selection bias introduced by focusing only on statistically significant results
- Adjusts for the random "lucky uplift" that inflates observed effects
- Produces more accurate and realistic estimates of total impact
- Uses only effect sizes and standard errors — simple, scalable, and easy to implement
- Successfully deployed in Airbnb's internal experimentation platform, demonstrating real-world value

Together, these improvements make the method both statistically rigorous and practical for use in high-volume experimentation systems.

4 Analysis

4.1 Simulation Parameters

To replicate the simulation experiments from Section 4 of the original paper, we used the following setup:

- **Number of experiments (n):** 30
- **True effects (a_i):** Drawn from a truncated normal distribution with mean 0.2, standard deviation 0.7, bounds $[-1.5,\ 2.0]$
- **Variances (σ_i^2):** Sampled from an inverse gamma distribution with shape 3 and scale 1
- **Observed outcomes (y_i):** Simulated as $y_i \sim \mathcal{N}(a_i,\ \sigma_i^2)$
- **Significance threshold:** $p < 0.05$ (i.e., $y_i > 1.96 \cdot \sigma_i$)
- **Bootstrap iterations:** $B = 1000$
- **Number of simulation runs:** 1000

This configuration follows the structure of the original paper's Section 4, with a modified bootstrap size to balance accuracy and runtime.

All simulations, calculations, and visualizations presented in this section — including **Tables 1 and 2** and **Figures 3, 4, and 5** — were implemented in Python. The full source code and logic are provided in the accompanying Jupyter notebook.

4.2 Simulation Summary

In this simulation, the `simulate()` function is executed 1,000 times to generate repeated experimental outcomes. Each iteration produces six key metrics:

- $|A|$: the number of significant experiments
- S_A: the naive total effect estimate (before bias correction)
- $T_{A,\text{cond}}$: the conditionally debiased total effect estimate
- $\hat{T}_A$: the fully debiased total effect estimate
- T_A: the true total effect
- $\hat{\beta}$: the estimated bias introduced by selecting only significant experiments

These results are stored in a Pandas DataFrame and summarized using descriptive statistics, including the minimum, first quartile (1Q), median (2Q), mean, third quartile (3Q), and maximum values. All values are rounded to two decimal places for interpretability. The resulting summary, presented in **Table 1**, offers a clear and comprehensive overview of the simulation outcomes.

Table 1
Simulation Summary

Statistic	$\lvert A\rvert$	S_A	$T_{A,\text{cond}}$	$\hat{T}_A$	T_A	$\hat{\beta}$
Min	0.00	0.00	0.00	-10.83	-0.25	-0.60
1Q	3.00	4.56	4.27	2.31	2.76	1.35
2Q	4.00	6.60	6.08	5.76	4.05	2.34
Mean	4.29	6.84	6.30	5.80	4.24	2.60
3Q	5.00	8.87	8.18	9.28	5.54	3.62
Max	12.00	23.49	19.85	25.61	11.82	11.72

For example, at the median (2Q) across the 1,000 simulations:

- There are approximately 4 significant experiments per run
- The naive total effect estimate (S_A) is 6.6
- The true total effect (T_A) is 4.05
- The estimated bias ($\hat{\beta}$) is approximately 2.34

This comparison illustrates how the naive estimate tends to overstate the actual impact. In contrast, the debiased estimators—particularly the fully debiased estimate ($\hat{T}_A$)—successfully adjust for this upward bias, producing results that more accurately reflect the underlying true effect.

4.3 Confidence Interval Evaluation

Table 2
Confidence Interval Coverage

Target Coverage	Naive	Bootstrap	Debiased Bootstrap
0.70	0.1462	0.0554	0.0595
0.80	0.2188	0.0675	0.0685
0.90	0.3226	0.0867	0.0806
0.95	0.4365	0.1442	0.1119

In this section, the `simulate_with_CI()` function is executed with 1,000 bootstrap samples ($B = 1000$) and 1,000 simulation runs ($n\_sims = 1000$) to evaluate the performance of different methods for constructing confidence intervals. The results, summarized in **Table 2**, report the empirical coverage probabilities at four target confidence levels: 70%, 80%, 90%, and 95%.

Three types of confidence intervals are compared:

- **Naive:** constructed using standard errors without any correction for selection bias
- **Bootstrap:** based on basic resampling, but without applying debiasing
- **Debiased Bootstrap:** constructed after applying the bias correction method before resampling

Across all confidence levels, the Debiased Bootstrap method consistently outperforms both the Naive and standard Bootstrap approaches in terms of empirical coverage. While it does not fully achieve the nominal coverage rates, its performance is noticeably closer to the target values.

These findings underscore two key insights:

- Naive intervals are substantially biased, often failing to contain the true total effect due to the influence of selection bias
- Applying bias correction significantly improves coverage reliability, even though some minor undercoverage may persist

Overall, these results reinforce the necessity of correcting for selection bias when reporting inference metrics from A/B tests in large-scale experimentation settings.

4.4 Naive Estimate vs True Effect

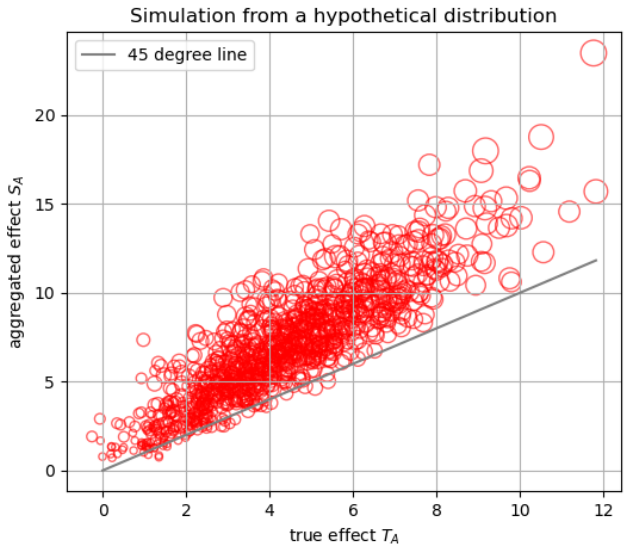

Figure 3. Naive Estimate (S_A) vs True Effect (T_A)

Figure 3 displays a scatter plot comparing the naive aggregated effect estimates (S_A) with the true total effects (T_A) across 1,000 simulation runs. Each point represents the outcome of a single simulation, and the size of each point is proportional to the number of significant experiments identified ($|A|$) in that run.

A 45-degree reference line is included to indicate the ideal scenario in which the estimated effect perfectly matches the true effect. However, most points lie above this line, demonstrating that

the naive estimate (S_A) consistently overestimates the actual total effect (T_A).

This visualization illustrates the core issue of the Winner's Curse: by summing only statistically significant results, the estimate becomes systematically biased, tending to overstate the true total effect. The extent of overestimation is reflected in how far each point lies above the diagonal line — the farther the distance, the more the estimate exaggerates the true underlying effect.

4.5 Conditional Debiased Estimate vs True Effect

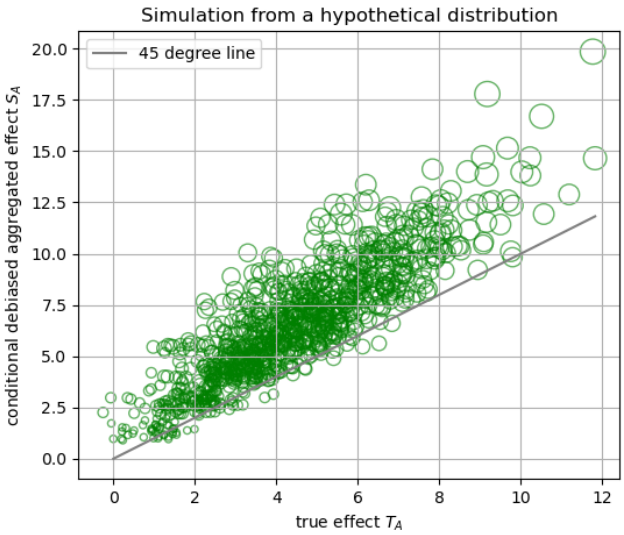

Figure 4. Conditional Debiased Estimate ($T_{A,\text{cond}}$) vs True Effect (T_A)

This plot compares the conditionally debiased aggregated effect estimates ($T_{A,\text{cond}}$) to the true total effects (T_A) across 1,000 simulation runs. Each point in the plot represents one simulation, and the size of the point corresponds to the number of significant experiments ($|A|$) found in that run. The 45-degree reference line represents the ideal scenario, where the estimated effect matches the true effect exactly.

Compared to **Figure 3**, the points in **Figure 4** lie noticeably closer to the 45-degree reference line, indicating that the conditional debiasing method effectively reduces the upward bias introduced by the Winner's Curse. However, some bias remains, particularly in simulations where the estimated effects are large. This demonstrates that while conditional debiasing improves accuracy, it does not fully eliminate the bias, especially in more extreme cases.

4.6 Fully Debiased Estimate vs True Effect

Figure 5 displays a scatter plot comparing the fully debiased aggregated effect estimates ($\hat{T}_A$) to the true total effects (T_A) across 1,000 simulation runs. Each point represents the outcome of one simulation, and the size of the point corresponds to the number of significant experiments ($|A|$) in that run. The 45-degree line on the plot indicates perfect estimation — where the estimated effect exactly equals the true effect.

Compared to **Figures 3 and 4**, the points in **Figure 5** are more tightly clustered around the diagonal line, indicating that the full

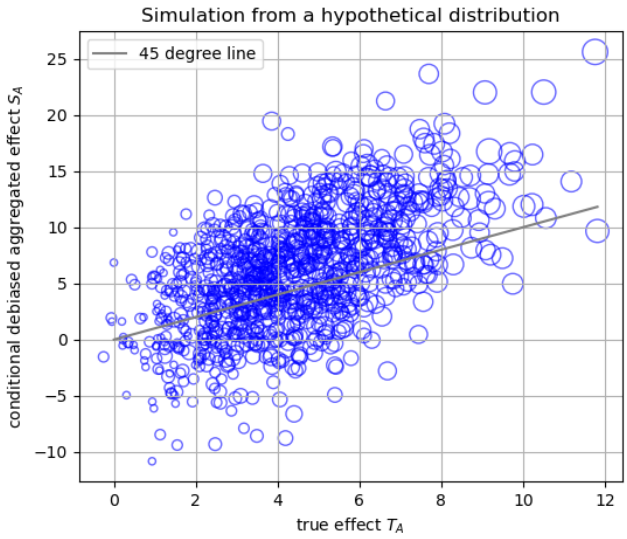

Figure 5. Fully Debiased Estimate ($\hat{T}_A$) vs True Effect (T_A)

debiasing method yields the most accurate estimates of the true total effect.

This visualization demonstrates that the proposed bias correction method significantly reduces the Winner's Curse bias, thereby enhancing the reliability and credibility of reported outcomes for decision-making in large-scale A/B testing contexts.

4.7 Figures 3, 4, 5 – Visual Comparison Summary

- **Figure 3:** Naive estimate (S_A) vs. true effect (T_A) — consistently overstates impact
- **Figure 4:** Conditional debiased estimate ($T_{A,\text{cond}}$) — improved, but still some bias
- **Figure 5:** Fully debiased estimate ($\hat{T}_A$) — closely aligned with T_A, demonstrating effectiveness

5 Conclusion

This project explored a common yet often overlooked issue in large-scale A/B testing: the Winner's Curse — a systematic bias that arises when only statistically significant experiments are selected for reporting or decision-making. This selective aggregation inflates the estimated overall impact of product changes, leading to overly optimistic conclusions. To address this, we implemented and evaluated a bias correction method proposed in the original research paper, combining theoretical insights with extensive simulation experiments to assess how different estimators perform in capturing the true underlying effects.

Our simulations confirmed that the naive approach, which simply sums the treatment effects of significant experiments (S_A), consistently overestimated the true total effect (T_A), as shown in **Figure 3**. Applying a conditional bias adjustment ($T_{A,\text{cond}}$) reduced this inflation to some extent (**Figure 4**), but visible discrepancies remained. In contrast, the fully debiased estimator ($\hat{T}_A$) demonstrated the closest alignment with the true effect (**Figure 5**), providing the most accurate and reliable estimates among the methods tested.

Descriptive statistics from 1,000 simulation runs (**Table 1**)

showed that the naive estimate (S_A) exceeded the true effect (T_A) by an average of approximately 2.6 units — a non-trivial bias. Our evaluation of confidence interval methods (**Table 2**) further revealed that both naive and standard bootstrap intervals often failed to reach their target coverage. The debiased bootstrap method significantly improved empirical coverage, though it still fell slightly short of ideal targets.

These findings carry important practical implications. In data-driven organizations where product and business decisions are often guided by A/B testing outcomes, failing to correct for selection bias can lead to exaggerated expectations and suboptimal choices. By applying the debiasing approach examined in this project, companies can make more trustworthy inferences, reduce the risk of misleading product evaluations, and strengthen the integrity of their experimentation platforms.

In short, bias correction is not merely a statistical refinement — it is a strategic necessity for ensuring accurate, evidence-based decision-making in large-scale online experimentation, ultimately supporting more effective product development and innovation.