Generalized Experiment Conclusion Framework for Extremely Volatile Data

Satyam Anand
Games24x7 Pvt. Ltd
Bengaluru, IN
satyam.anand@games24x7.com

Deepanshi Seth
Games24x7
Bengaluru, IN
deepanshi.seth@games24x7.com

Sanjay Agrawal
Games24x7
Bengaluru, Karnataka, India
sanjay.agrawal@games24x7.com

Debanjan Sadhukhan
Games24x7 Pvt Ltd
Bengaluru, IN
debanjan.sadhukhan@games24x7.com

Tridib Mukherjee
Games24x7
Bengaluru, IN
tridib.mukherjee@games24x7.com

Abstract

We have developed a specialized framework for evaluating online controlled experiments in the highly volatile and rapidly growing skill gaming industry. Addressing challenges such as a limited user base, the outsized influence of top-percentile players, and constantly evolving user behavior, the framework streamlines the experimentation process to deliver timely and accurate insights. Key innovations include the *Strata-Smart A/B Experiment Analysis Algorithm (SEAA)*, which integrates a novel variance reduction technique and a method for distinguishing between outliers and genuine high-value players, alongside a User-Friendly Interface that simplifies analysis for non-technical stakeholders. Additionally, the framework mitigates pre-existing bias by identifying historically similar populations, ensuring robust and reliable conclusions. This system enhances data-driven decision-making, providing consistency and integrity in a fast-paced, competitive environment.

ACM Reference Format:
Satyam Anand, Deepanshi Seth, Sanjay Agrawal, Debanjan Sadhukhan, and Tridib Mukherjee. 2024. Generalized Experiment Conclusion Framework for Extremely Volatile Data. In *8th International Conference on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD) (CODS-COMAD Dec '24), December 18–21, 2024, Jodhpur, India.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3703323.3704276

1 Introduction

Online controlled experiments, or A/B tests, are the gold standard in the technology industry to determine the true impact of new treatments on key business metrics [11]. In online skill-gaming platforms, where users engage for self-esteem, relaxation, and social gratification [20], optimizing the user experience is paramount. At Games24x7, we embrace a culture of testing ideas rapidly and at scale- following the principle of "more, better, faster"[23]. This approach is vital for accurately estimating the impact on key metrics, enabling us to scale successful treatments and terminate ineffective ones, thereby optimizing both costs and user experience.

In the skill gaming industry, the limited user base, combined with the highly volatile and outcome driven metrics [18], and the emphasis on fast experimentation makes it challenging to design adequately powered A/B experiments. Metrics are often driven by a small subset of the users, who account for most of the business, amplifying volatility. Even without any treatment, significant impacts can be observed across a wide range of sample sizes (Figure 3). While industry-standard methods like metric transformation [5], control variates, stratified sampling [6] reduce variance, they often fail to address the pre-existing biases common in our domain. To tackle this, we modified the stratified sampling, *"Delta-Adjusted Stratified Sampling" (DASS)* to ensure that the samples from control and treatment groups are representative and similar (2.1.1). Additionally, on our platform, many treatments primarily impact the top 2-3% of users, making it difficult to distinguish between statistical outliers and genuine high-value users, increasing the likelihood of false positives. Most prevalent methods like imputation or exclusion of high algebraic values often result in removal of important users from the analysis, making the insights inadequate. We propose a novel technique called *"Rareness-in-commonality"* (2.1.2) which accurately identifies outlier-ish behavior in metrics, ensuring that high value users remain in the analysis.

In this paper, we present the comprehensive framework featuring the ***Strata-Smart A/B Experiment Analysis Algorithm (SEAA)*** to conclude AB tests with high power and a user-friendly interface, empowering the business teams with limited technical expertise to perform analysis independently. To mitigate the risk of misinterpreting p-values in statistical tests, we offer a confidence or support percentage—indicating the likelihood of a metric being positive—making the results accessible and easier to interpret. By streamlining the experimentation process across multiple teams, this framework ensures timely and accurate insights, enabling more informed decisions critical for maintaining a competitive edge.

1.1 Related Work

A/B testing began gaining widespread use in the late 1990s with the rise of the Internet, though the underlying statistical theory has been long established [16]. These experiments provide a data-driven way to compare treatments, with statistical significance as the prevailing mechanism for evaluating impact. Tests such as the t-test [10], Kolmogorov–Smirnov test [22] and rank-based tests like Mann–Whitney test [21] are frequently used. The t-test assumes normality or, at the very least, minimal skewness, which is violated

CODS-COMAD Dec '24, December 18–21, 2024, Jodhpur, India
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1124-4/24/12
https://doi.org/10.1145/3703323.3704276

by our gamma-like metrics, making it unsuitable. While transformations can normalize data, they distort the real-world business interpretations [12]. The KS test, though effective in comparing entire distributions, can detect differences that may not align with central tendencies or practical business significance [24]. Rank-based tests handle non-normality better but rely on data ranking rather than actual values, potentially overlooking the influence of top-percentile players—critical in the skill-gaming. Furthermore, these tests are highly dependent on p-values, which are prone to misinterpretation and misuse, increasing the risk of false positives [1] [15]. This underscores the need for business-friendly, actionable metrics that can guide decision-making.

Apart from this, efforts have been made to increase the sensitivity of A/B tests. One approach involves transforming the metric into a lower variance metric. Examples include capping skewed metrics [14], binarizing count metrics, or creating weighted linear combinations [5] [13]. While useful, these methods come with trade-offs—capping may exclude key business-contributing users, and binarization or linear combinations complicate the estimation of impacts on key metrics in the transformed space. Another popular approach is the use of control variates [2] [3], CUPED [6] [4], which have become industry standards but work effectively only when prior information about users is available [11] [7]. Despite advancements such as non-linear adjustments [19] or variance-weighted estimators [17], they often fail to adequately address pre-existing biases common in skill-gaming environments [20].

Our work builds on these methods by modifying traditional stratified sampling. While CUPED approximates stratified sampling when using categorical control variates [6], we ensure historical similarity between samples, mitigating pre-experiment bias and reducing variance effectively. In line with the practices of leading companies with strong experimentation cultures, such as Microsoft [9], LinkedIn [8], Uber [25], we developed our own in-house platform for experiment analysis and conclusion. Tailored to meet our needs, it offers the control and flexibility required to navigate our unique business challenges.

1.2 Contributions

Towards building the robust experiment conclusion framework, we have made several significant contributions that enhance the experimentation capability in the organization - **(i)** We introduced ***SEAA*** which combines techniques for variance reduction, outlier handling and statistical testing to effectively conclude A/B tests. **(ii)** We developed an innovative technique, ***DASS*** to reduce variance in key metrics. While similar in intent to CUPED to increase sensitivity, our approach also addresses pre-existing bias by identifying historically similar populations, mitigating the risk of chance findings. **(iii)** We implemented a novel method for distinguishing between outliers and genuine high-value players. This improves the accuracy of impact measurement by ensuring that our analysis focuses on meaningful data rather than anomalies. **(iv)** We developed a comprehensive end-to-end framework for the business teams, simplifying the process for non-technical users by requiring minimal input to generate detailed analysis reports. Additionally, we translated complex statistical results into clear, actionable insights, enhancing communication and confidence in data-driven decision making.

2 Framework Overview

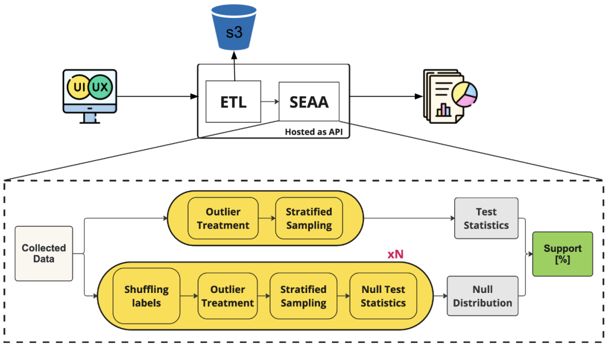


Figure 1: Framework Overview

The framework has been hosted as a tool within a Kubernetes environment. Each analysis triggers a job that runs in the tool, ensuring scalability and reliability with minimal operational costs. Figure 1 illustrates a detailed overview of the entire workflow.

- **User Interface:** Stakeholders interact with a user-friendly UI where they provide experiment details, such as the experiment ID, start date, the metrics to analyze and historical metrics to identify look-alike samples. The information entered in the UI generates a payload that triggers the API of the tool.
- **ETL (Extract Transform Load) Pipeline & Data Storage:** The first component of the tool is an ETL pipeline, which extracts the required data based on the provided details. The collected data is stored in Amazon s3 for deep-dive analysis if needed.
- **Strata-Smart A/B Experiment Analysis Algorithm (SEAA):** This module processes the data according to the configurations provided by the stakeholder and generates the experiment analysis reports, which are compiled into PDF documents and emailed to the stakeholder.
- **Data Accessibility, Logging & Monitoring:** The entire process is logged in a database, enabling monitoring and easy debugging. Stakeholders can track the real-time status of the jobs, ensuring transparency and reliability. Also, the location of the data, the user set considered (post-sampling and outlier removal), and the percentile levels of users based on the specified metric are provided to them for any in-depth analysis.

The core component of this workflow is SEAA, which has been explained next.

2.1 Strata-Smart A/B Experiment Analysis Algorithm (SEAA)

SEAA is an innovative approach for analyzing A/B experiments. It primarily evaluates *support*[%], the percentage of instances in the null distribution falling below the observed lift (*test statistics*). A higher *support*[%] indicates greater confidence or likelihood of a positive impact. The overall SEAA flow is described in Figure 1, with the detailed steps in Algorithm 1. It integrates components for variance reduction, outlier treatment, and statistical testing, each of which is thoroughly described in the subsequent sub-sections.

Algorithm 1 SEAA

```
Clean the collected/observed data for analysis.
Apply outlier treatment if required.
Perform stratified sampling on the treated data till $K$ representative samples are obtained, with all having prebias < $\delta$. If K samples cannot be collected, results are deemed inconclusive.
Compute the test statistics from the $K$ representative samples.
Repeat the following steps till you get $N$ null test statistics to create a null distribution:
    (i) Shuffle the label assignments of the observed data while applying "split-ratio adaptation", i.e., maintain a consistent split ratio (the fraction of actual test size to control size) in both groups formed after shuffling the labels.
    (ii) Apply outlier treatment if required.
    (iii) Perform stratified sampling on the shuffled data, and consider it representative for calculating null test statistics only if prebias < $\delta$.
Combine the test statistics with the null distribution to determine whether there is any positive impact.
```

2.1.1 ***Variance Reduction:*** In this work, we introduce a modified version of conventional stratified sampling, termed ***"Delta-Adjusted Stratified Sampling" (DASS)***, to improve variance reduction. This method divides the population into distinct subgroups or strata based on pre-experiment metrics and includes only those samples with a prebias[1] within a specified δ range when calculating the test statistic.

Equation 1 highlights how the variances of the estimator based on stratified ($var(S_tS)$) and simple random sampling ($var(SRS)$) differ, with the former yielding a lower variance when strata are appropriately chosen [6]

$$var(SRS) = var(S_tS) + \sum_{h=1}^{H} \left(\frac{N_h}{N} \frac{(\overline{y_h} - \overline{y})^2}{n} \right) \tag{1}$$

where $var(SRS) = \frac{S^2}{n}$, $var(S_tS) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \frac{{S_h}^2}{n} \right)$ and $\frac{N_h}{N} = \frac{n_h}{n}$. Here $n, n_h, \overline{y_h}, \overline{y}$, N, N_h, H, S_h^2, and S^2 respectively denote sample size, sample size of stratum h, population mean of stratum h, population mean, population size, population size of stratum h, number of stratum, population variance within stratum h, and population mean. Our goal is to minimize $var(S_tS)$ which can be achieved by reducing S_h^2 (indicating greater homogeneity within strata) and increasing $(\overline{y_h} - \overline{y})^2$ (indicating greater heterogeneity across strata). As a result, careful strata formation is crucial for effective stratified sampling.

DASS ensures strata remain distinct and internally consistent throughout the experiment phase, effectively preventing any treatment effects from altering them. However, this approach necessitates maintaining the relevance and correlation of the strata over time. Hence, in the skill-gaming domain, where key metrics such as engagement, transaction sizes are highly skewed, we transform them to approximate a normal distribution. This enhances their correlation over time significantly and contributes to more balanced and evenly distributed strata formation. Furthermore, DASS addresses challenges related to small sample sizes. In the gaming industry, where the cost of campaigns is high and traffic for hypothesis testing is limited, efficient sample use is crucial. By reducing variance, it enables drawing conclusions with similar impact, even with relatively smaller sample sizes, as illustrated in Figure 3.

2.1.2 ***Outlier Treatment:*** We propose a novel outlier treatment method called ***"Rareness-in-Commonality"***. This approach views outliers not just as high algebraic values, but as behaviors that deviate from the norm. Entities are removed only if their conduct significantly differs from others within the same stratum (based on pre-experiment metrics). The definition of "rare" behavior can vary depending on domain-specific datasets. Here, we suggest measuring rareness by infrequent occurrences, as illustrated in Figure 2. Addressing outliers also significantly reduces variance, not achieved through *"imputation"* methods where outliers are substituted with aggregate metrics derived from domain knowledge. Another common approach—removing the *"top x%ile"* of data points before analysis—does reduce variance but also excludes valuable, highly engaged populations. As demonstrated by Figure 4, our proposed method balances variance reduction with the retention of key contributors, enhancing the reliability of the results.

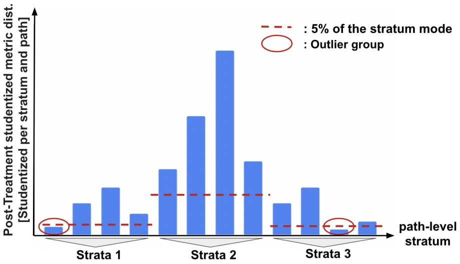


Figure 2: Rareness-in-Commonality

2.1.3 ***Statistical Test:*** We employed a Permutation test to facilitate direct comparisons of metrics without relying on normality assumptions or being sensitive to rank changes, effectively addressing the limitations of both the t-test and the Mann-Whitney test. This approach compares the observed test statistic with a distribution of null-test statistics generated by randomly permuting the labels or group assignments. To ensure the groups being compared are valid and consistent under the null hypothesis, we employed *"split-ratio adaptation"*, which helps produce a robust null distribution for any metric, allowing for meaningful interpretation.

3 Results and Discussion

In this section, we present the findings and results of our explorations. To facilitate understanding, the following key terms are used throughout the discussion - **(i)** $Lift\%$: Calculated as $[(\frac{mean(success\ metric)\ of\ Test}{mean(success\ metric)\ of\ Control}) - 1] * 100$, representing the change in the success metric of the test relative to the control. **(ii)**[2] $value\_contribution\_ratio$: % of the total engagement contributed by the users in the evaluation relative to all targeted users in the A/B test. **(iii)** $liquidity\_ratio$: % of users considered in the evaluation (count). **(iv)** $stimulus[\%]$: Quantum of intervention applied in

[1] prebias = $|[(\frac{mean(past\ metric)\ of\ Test\ group}{mean(past\ metric)\ of\ Control\ group}) - 1]| * 100$

[2] (ii) and (iii) specifies the information loss resulting from excluding entities during prebias adjustment and outlier treatment if any.

the success metric of control group to create a synthetic test group, simulating lift relative to the control. **(v)** *top*[%]: % of users in the test group who are modified, specifically focusing on the top x% of users ranked by success metric contribution.

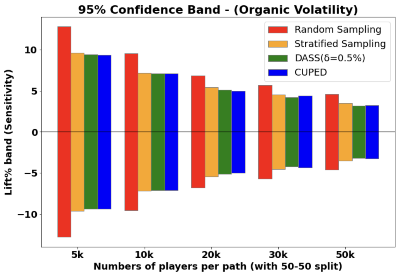
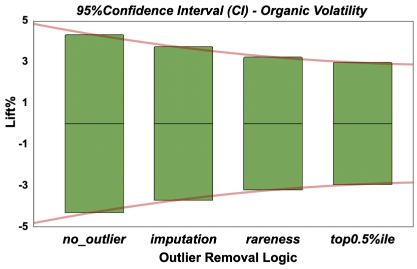


Figure 3: Variance Reduction

Figure 3 illustrates the sensitivity of various variance reduction techniques in A/B testing. The findings indicate that variance decreases as group size increases, indicating that larger sample sizes yield reliable impact estimates. Notably, DASS significantly outperforms random sampling in reducing variance and shows marked improvements over stratified sampling while performing similar to CUPED, the industry standard, for equivalent sample sizes. This enhancement in variance reduction boosts test sensitivity. For example, the variance associated with DASS at 5k is lower than that of random sampling at 10k, allowing for effective experimentation with smaller sample sizes while achieving comparable or even better confidence levels in the results.



Outlier Removal Logic	no_outlier	imputation	rareness	top0.5%ile
value_contribution_ratio	99.9%	91%	89%	81%
liquidity_ratio	99.9%	99.9%	97.8%	99.4%

Figure 4: Outlier Impact

Figure 4 examines the performance of our proposed method of *"Rareness-in-Commonality"*, using a total sample size of 60k, equally divided between two untreated groups over multiple seed values with DASS. The result shows that, while outlier treatment typically reduces volatility, our method significantly narrows the volatility range, without a considerable drop in *value_contribution_ratio*. Although the *"top0.5%ile"* approach further reduces volatility but results in a substantial drop in *value_contribution_ratio*, limiting its usefulness. Conversely, the *"imputation"* approach—where top 0.5%ile users were tagged as outliers and replaced with the maximum of the remaining values across the path—retains key engagement contributors but achieves a lesser reduction in variance.

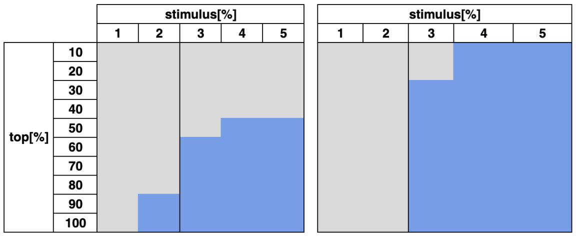


Figure 5: Sensitivity comparison of Statistical tests

Figure 5 compares the sensitivity of two statistical tests: the Mann-Whitney test (left) and the Permutation test (right). To create this plot, multiple synthetic test scenarios were generated by applying various levels of *stimulus*[%] to the top x% of success metric contributors within an untreated group of 30k users. For instance, a patch where *top*[%]=20 and *stimulus*[%]=4, indicates that the top 20% of success metric contributing users from the control group received a 4% boost to generate a synthetic test group and simulate a lift. For this demonstration, a *support* cutoff of 85% was used to annotate the blue patch in both plots, representing regions where *support* exceeded 85% at the simulated lifts. The blue patch indicates the region where the test successfully detected a positive impact, while the gray patch indicates areas where test failed to detect any impact. The results clearly demonstrate that the Permutation test is more sensitive than the Mann-Whitney test, particularly in the regions where the top success-metric contributors are pre-dominantly impacted.

4 Demonstration

URL of the demo video on YouTube: https://youtu.be/JvfjxHmxxg4

In this video, we demonstrate the functionality of the AB tool, where a stakeholder enters the details of the experiment through a UI hosted on the organization's intranet. The tool is triggered via an API call, fetching data from the database and invoking the SEAA module to generate analysis reports, which are then emailed to the relevant stakeholders.

5 Conclusion & Future Work

This work streamlines the experimentation process at Games24x7. By addressing challenges related to data volatility and evolving user behavior, it provides accurate insights and empowers non-technical stakeholders with clear, actionable results. Notably, its applicability extends beyond the skill gaming domain. With minimal modifications, it can be adapted for A/B experiments in various fields such as E-Commerce, Healthcare, FinTech, Investment, OTT platforms handling the issues of skewed metrics and volatile data. Future work will focus on accommodating the new users on the platform. This group exhibits distinct challenges due to a more pronounced data skewness and a lack of historical context, necessitating further development in variance reduction and outlier handling techniques.

References

[1] Ron Berman and Christophe Van den Bulte. 2022. False Discovery in A/B Testing. *Management Science* 68, 9 (2022), 6762–6782. https://doi.org/10.1287/mnsc.2021.4207

[2] Matteo Courthoud. 2022. Understanding CUPED. https://towardsdatascience.com/understanding-cuped-a822523641af. Accessed: 2024-09-07.

[3] Craig. 2022. CUPED on Statsig. https://blog.statsig.com/cuped-on-statsig-d57f23122d0e. Accessed: 2024-09-07.

[4] Alex Deng, Luke Hagar, Nathaniel Stevens, Tatiana Xifara, Lo-Hua Yuan, and Amit Gandhi. 2023. From Augmentation to Decomposition: A New Look at CUPED in 2023. arXiv:2312.02935 [stat.AP] https://arxiv.org/abs/2312.02935

[5] Alex Deng and Xiaolin Shi. 2016. Data-Driven Metric Development for Online Controlled Experiments: Seven Lessons Learned. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 77–86. https://doi.org/10.1145/2939672.2939700

[6] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. 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*. 123–132.

[7] Alexey Drutsa, Gleb Gusev, and Pavel Serdyukov. 2015. Future User Engagement Prediction and Its Application to Improve the Sensitivity of Online Experiments. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 256–266. https://doi.org/10.1145/2736277.2741116

[8] LinkedIn Engineering. 2020. Our Evolution Towards T-Rex: The Prehistory of Experimentation. https://www.linkedin.com/blog/engineering/ab-testing-experimentation/our-evolution-towards-t-rex-the-prehistory-of-experimentation-i. Accessed: 2024-09-07.

[9] Microsoft Experimentation Platform (ExP). 2024. A/B Testing Infrastructure Changes at Microsoft (ExP). https://www.microsoft.com/en-us/research/group/experimentation-platform-exp/articles/a-b-testing-infrastructure-changes-at-microsoft-exp/. Accessed: 2024-09-07.

[10] Shubham Gupta and Sneha Chokshi. 2020. Digital marketing effectiveness using incrementality. In *Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4*. Springer, 66–75.

[11] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, Mike Curtis, Alex Deng, Weitao Duan, Peter Forbes, Brian Frasca, Tommy Guy, Guido W. Imbens, Guillaume Saint Jacques, Pranav Kantawala, Ilya Katsev, Moshe Katzwer, Mikael Konutgan, Elena Kunakova, Minyong Lee, MJ Lee, Joseph Liu, James McQueen, Amir Najmi, Brent Smith, Vivek Trehan, Lukas Vermeer, Toby Walker, Jeffrey Wong, and Igor Yashkov. 2019. Top Challenges from the first Practical Online Controlled Experiments Summit. *SIGKDD Explor. Newsl.* 21, 1 (may 2019), 20–35. https://doi.org/10.1145/3331651.3331655

[12] Khairul Islam and Tanweer J Shapla. 2020. A New Transformed t-test for Skewed Data: A Goodness-of-fit Approach. *International Journal of Statistics and Probability* 9, 5 (2020), 1–30.

[13] Eugene Kharitonov, Alexey Drutsa, and Pavel Serdyukov. 2017. Learning Sensitive Combinations of A/B Test Metrics. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) *(WSDM '17)*. Association for Computing Machinery, New York, NY, USA, 651–659. https://doi.org/10.1145/3018661.3018708

[14] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, New York, USA) *(KDD '14)*. Association for Computing Machinery, New York, NY, USA, 1857–1866. https://doi.org/10.1145/2623330.2623341

[15] Ron Kohavi, Alex Deng, and Lukas Vermeer. 2022. A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 3168–3177. https://doi.org/10.1145/3534678.3539160

[16] Ron Kohavi and Roger Longbotham. 2017. *Online Controlled Experiments and A/B Testing*. Springer US, Boston, MA, 922–929. https://doi.org/10.1007/978-1-4899-7687-1_891

[17] Kevin Liou and Sean J. Taylor. 2020. Variance-Weighted Estimators to Improve Sensitivity in Online Experiments. In *Proceedings of the 21st ACM Conference on Economics and Computation* (Virtual Event, Hungary) *(EC '20)*. Association for Computing Machinery, New York, NY, USA, 837–850. https://doi.org/10.1145/3391403.3399542

[18] Koyel Mukherjee, Deepanshi Seth, Prachi Mittal, Nuthi S Gowtham, Tridib Mukherjee, Dattatreya Biswas, and Sanjay Agrawal. 2022. ComParE: A User Behavior Centric Framework for Personalized Recommendations in Skill Gaming Platforms. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*. 186–194.

[19] Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. 2016. Boosted Decision Tree Regression Adjustment for Variance Reduction in Online Controlled Experiments. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 235–244. https://doi.org/10.1145/2939672.2939688

[20] Debanjan Sadhukhan, Sachin Kumar, Swarit Sankule, and Tridib Mukherjee. 2023. t-RELOAD: A REinforcement Learning-based Recommendation for Outcome-driven Application. In *Proceedings of the Third International Conference on AI-ML Systems*. 1–7.

[21] Suhrid Satyal, Ingo Weber, Hye-young Paik, Claudio Di Ciccio, and Jan Mendling. 2019. Business process improvement with the AB-BPM methodology. *Information Systems* 84 (2019), 283–298.

[22] Shahriar Shariat, Burkay Orten, and Ali Dasdan. 2017. Online evaluation of bid prediction models in a large-scale computational advertising platform: decision making and insights. *Knowledge and Information Systems* 51, 1 (2017), 37–60.

[23] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping experiment infrastructure: more, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC, USA) *(KDD '10)*. Association for Computing Machinery, New York, NY, USA, 17–26. https://doi.org/10.1145/1835804.1835810

[24] Ramesh S.V. Teegavarapu. 2019. Chapter 1 - Methods for Analysis of Trends and Changes in Hydroclimatological Time-Series. In *Trends and Changes in Hydroclimatic Variables*, Ramesh Teegavarapu (Ed.). Elsevier. https://www.sciencedirect.com/topics/earth-and-planetary-sciences/kolmogorov-smirnov-test#chapters-articles

[25] Uber. 2022. Supercharging A/B Testing at Uber. https://www.uber.com/en-IN/blog/supercharging-a-b-testing-at-uber/. Accessed: 2024-09-07.