# Regular Balanced Switchback Designs for Robust Multi-Unit Online Experimentation

**Lorenzo Masoero**[1], **Paul Missault**[1], **Christian Delbé**[1], **Thomas Richardson**[2], **Guido Imbens**[2]
[1] Amazon.com [2] University of Washington [3] Stanford University
{masoerl, pmissaul, delbec}amazon.com, thomasr@uw.edu, imbens@stanford.edu

**Abstract** User-randomized A/B testing, while the gold standard for online experimentation, faces significant limitations when legal, ethical, or practical considerations prevent its use. Item-level randomization offers an alternative but typically suffers from high variance and low statistical power due to skewed distributions and limited sample sizes. We here introduce Regular Balanced Switchback Designs (RBSDs), a novel experimental framework that combines temporal and item-level randomization in a principled manner. RBSDs extend the switchback methodology of Bojinov et al. [2023] by incorporating multiple randomization designs [Masoero et al., 2024] to achieve double balance: a fixed fraction of items receive treatment at each time step, and all items receive treatment equally often across time periods. This balanced approach substantially reduces estimate variance compared to independent randomization schemes while maintaining unbiased estimation of average treatment effects under carryover effects. Using both realistic simulations based on e-commerce data patterns and theoretical analysis within a potential outcomes framework, we demonstrate that RBSDs achieve 40-60% reduction in standard errors compared to simple item randomization and 20-30% improvement over unbalanced switchback approaches. Our framework provides a principled solution for scenarios where user-randomization is infeasible, including pricing experiments and complex system testing where indirect effects through downstream systems must be captured.

**Introduction** Online A/B testing has become the cornerstone of data-driven decision making in digital platforms, with user-randomization typically considered the gold standard [Kohavi et al., 2020]. However, several scenarios render user-randomization unfeasible or undesirable, necessitating alternative experimental designs. Legal and ethical constraints present the most immediate challenges. In pricing experiments, user-randomization can constitute price discrimination, raising significant legal concerns [Cooprider and Nassiri, 2023]. Similarly, regulatory requirements in certain industries may prohibit differential treatment of users based on randomization. Technical limitations further complicate user-randomization. In complex systems with multiple interacting components, user-randomization captures direct treatment effects but fails to measure crucial indirect effects through downstream systems. For instance, when testing item display modifications, search engines and recommendation systems require consistent item representations to function properly. User-randomization creates multiple coexisting versions of the same item, potentially disrupting these downstream systems and invalidating experimental results. Item-level randomization addresses these challenges by ensuring all users observe the same version of each item, naturally capturing both direct and indirect treatment effects. However, practitioners often view item-randomization skeptically due to its typically lower statistical power compared to user-randomization, stemming from smaller effective sample sizes and highly skewed item-level metric distributions. Recent advances in switchback experimentation [Bojinov et al., 2023] and multiple randomization designs [Masoero et al., 2024] provide theoretical foundations for addressing these limitations. Building on this work, we propose Regular Balanced Switchback Designs (RBSDs) that combine temporal and item-level randomization with careful balance constraints to achieve the practical benefits of item-randomization while substantially improving statistical efficiency.

Our contributions are: (1) We develop a theoretical framework for RBSDs that ensures unbiased estimation under carryover effects; (2) We prove that balanced randomization across temporal and item dimensions substantially reduces variance compared to independent schemes; (3) We demon-

strate through comprehensive simulations that RBSDs achieve 40-60% variance reduction compared to standard approaches, enabling detection of smaller effects or shorter experiment durations.

**Potential Outcomes Framework and Regular Balanced Switchback Designs** We formalize RBSDs within the potential outcomes framework [Imbens and Rubin, 2010]. Consider $N$ items observed over $S$ time periods. For each item $n \in \{1, \ldots, N\}$ and time period $s \in \{1, \ldots, S\}$, let $Y_{n,s}^T$ and $Y_{n,s}^C$ denote the potential outcomes under treatment and control, respectively. The observed outcome is $Y_{n,s} = Z_{n,s}Y_{n,s}^T + (1 - Z_{n,s})Y_{n,s}^C$, where $Z_{n,s} \in \{0, 1\}$ indicates treatment assignment.

Our estimand is the average treatment effect: $\tau = \frac{1}{NS}\sum_{n=1}^{N}\sum_{s=1}^{S}(Y_{n,s}^T - Y_{n,s}^C)$. Unlike standard switchback designs that randomize independently across time and units, RBSDs impose balance constraints that substantially improve estimation efficiency. RBSDs satisfy two critical balance properties:

**Temporal Balance:** At each time period $s$, exactly $\alpha N$ items receive treatment for some fixed $\alpha \in (0, 1)$: $\sum_{n=1}^{N} Z_{n,s} = \alpha N \quad \forall s \in \{1, \ldots, S\}$.

**Item Balance:** Each item $n$ receives treatment for exactly $\alpha S$ time periods: $\sum_{s=1}^{S} Z_{n,s} = \alpha S \quad \forall n \in \{1, \ldots, N\}$.

This double-balancing approach extends the multiple randomization design framework of Masoero et al. [2024] to temporal settings, ensuring balanced treatment allocations across both dimensions.

**Estimation Under Carryover Effects, and Variance Reduction Through Balance** Following Bojinov et al. [2023], we allow for carryover effects where treatment in period $s$ may affect outcomes in subsequent periods. Under the assumption that carryover effects decay geometrically with lag $\ell$, we can write: $Y_{n,s}^C = \mu + \sum_{\ell=1}^{\min(s-1,L)} \gamma_\ell Z_{n,s-\ell} + \epsilon_{n,s}$, where $\gamma_\ell$ represents the carryover effect at lag $\ell$, $L$ is the maximum carryover lag, and $\epsilon_{n,s}$ is the error term. We employ the Horvitz-Thompson estimator [Horvitz and Thompson, 1952] adapted for our balanced design:

$$\hat{\tau} = \frac{1}{NS}\sum_{n=1}^{N}\sum_{s=1}^{S}\left(\frac{Z_{n,s}Y_{n,s}}{\alpha} - \frac{(1 - Z_{n,s})Y_{n,s}}{1 - \alpha}\right)$$

The key theoretical insight is that balanced randomization induces favorable correlation structures that reduce estimate variance. While independent randomization creates high variance due to random imbalances in treatment allocation, our balanced approach ensures consistent exposure patterns.

For the balanced design, the variance of $\hat{\tau}$ can be shown to be $\text{Var}(\hat{\tau}) = \frac{1}{N^2S^2}\sum_{n=1}^{N}\sum_{s=1}^{S}\frac{\sigma_{n,s}^2}{\alpha(1-\alpha)} +$ covariance terms, where the covariance terms are substantially reduced compared to independent randomization due to the balance constraints. This theoretical advantage translates directly to improved statistical power in practice.

**Experiments and Results** We validate RBSDs through comprehensive simulations using realistic e-commerce data patterns. Our simulation framework models $N = 1000$ items over $S = 20$ time periods, with outcomes following log-normal distributions to capture the heavy-tailed nature of typical e-commerce metrics. We incorporate moderate carryover effects with geometric decay ($\gamma_1 = 0.1$, $\gamma_2 = 0.05$) and heterogeneous treatment effects across items. We compare three experimental designs: (1) Simple item randomization with independent assignment each period, (2) Unbalanced switchback with independent temporal and item randomization, and (3) Our proposed RBSDs with double balance constraints. RBSDs demonstrate substantial improvements in estimation precision across all simulation scenarios. Compared to simple item randomization, RBSDs achieve 40-60% reduction in standard errors, with the largest gains observed in high-variance settings typical of e-commerce applications. The performance advantage over unbalanced switchback designs is also significant, with RBSDs showing 20-30% improvement in standard errors. This improvement stems from the correlation structure induced by balanced randomization, which ensures that treatment exposure patterns remain consistent across both temporal and item dimensions. Importantly, these variance reductions come without bias inflation. Our Horvitz-Thompson estimator remains unbiased under the balance constraints, with simulation results confirming that bias remains negligible ($< 0.1\%$ of true effect size) across all tested scenarios.

The variance reduction translates directly to improved statistical power. For a fixed significance level ($\alpha = 0.05$) and effect size, RBSDs achieve 80% power with approximately 40% fewer observations than simple item randomization. Alternatively, for fixed sample sizes, RBSDs can detect effect sizes that are 30-40% smaller than those detectable with standard approaches. This power improvement is particularly valuable in e-commerce settings where effect sizes are often small but economically significant. The ability to detect smaller effects or achieve the same power with shorter experiment durations provides substantial practical benefits.

We evaluate robustness across different carryover structures, including scenarios with longer decay periods and non-geometric patterns. RBSDs maintain their performance advantages across all tested carryover specifications, with the balanced design providing natural robustness to model misspecification. The double balance constraints ensure that any systematic biases due to carryover effects are distributed evenly across treatment and control groups, maintaining the validity of causal inference even when carryover assumptions are violated.

**Conclusion** We introduced RBSDs, a novel experimental framework that addresses fundamental limitations of both user-randomized and simple item-randomized experiments. By combining temporal and item-level randomization with principled balance constraints, RBSDs achieve the practical benefits of item-randomization while substantially improving statistical efficiency. Our theoretical contributions demonstrate that double balance—across both time periods and experimental units—induces favorable correlation structures that dramatically reduce estimate variance. The 40-60% variance reduction we observe translates directly to improved statistical power, enabling detection of smaller treatment effects or achieving equivalent power with shorter experiment durations.

The framework extends naturally from the switchback methodology of Bojinov et al. [2023] and multiple randomization designs of Masoero et al. [2024], providing a principled approach to experimental design when user-randomization is infeasible. This is particularly valuable for pricing experiments, complex system testing, and scenarios requiring capture of indirect effects through downstream systems. Our empirical validation through comprehensive simulations confirms the theoretical predictions while demonstrating robustness to various carryover structures and model misspecifications. The balanced design provides natural protection against systematic biases, maintaining causal inference validity even under assumption violations.

Future work should explore adaptive versions of RBSDs that optimize balance constraints based on observed data patterns, extensions to more complex carryover structures including network effects, and integration with modern causal inference techniques for heterogeneous treatment effects. The framework also opens possibilities for multi-armed variants and applications to cluster-randomized settings where both temporal and spatial balance may be beneficial. RBSDs represent a significant advance in experimental design for digital platforms, providing practitioners with a principled tool for scenarios where traditional randomization approaches fail while maintaining the statistical rigor essential for reliable causal inference.

## References


I. Bojinov, D. Simchi-Levi, and J. Zhao. Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777, 2023.

J. Cooprider and S. Nassiri. Science of price experimentation at Amazon. In *AEA 2023, NABE 2023*, 2023. URL `https://www.amazon.science/publications/science-of-price-experimentation-at-amazon`.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

G. W. Imbens and D. B. Rubin. Rubin causal model. In *Microeconometrics*, pages 229–241. Springer, 2010.

R. Kohavi, D. Tang, and Y. Xu. *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press, 2020.

L. Masoero, S. Vijaykumar, T. Richardson, J. McQueen, I. Rosen, B. Burdick, P. Bajari, and G. Imbens. Multiple randomization designs: Estimation and inference with interference. *arXiv preprint arXiv:2401.01264*, 2024.