Sequential-Halving Switchback

Paul Missault, Blake Mason, Dhivya Eswaran, Hancao Li
Amazon
{pmissaul, bjmason, deswaran, lihancao}@amazon.com

ABSTRACT

Adaptive designs redistribute experiment traffic throughout the experiment, and are used in practice to allow experimentation on a large amount of arms in an A/B test. Our work notes that the requirements on the estimators for such designs can be less strict than the typical unbiasedness that is required in a standard A/B test. We leverage this to introduce daily re-randomization to further increase the discriminating power of adaptive designs. Our proposed Sequential-halving switchback (SHS) independently splits each day's traffic evenly across the arms still active, and the worst half is eliminated at the end of each round. We show that pooling observations across rounds in this setup is often desirable, contrary to common belief. To guard against mistakes during exploration, we end with a conventional A/B on the winning arm(s). Finally, the efficacy of our proposal is shown in simulation and in a proof-of-concept study over nine arms on an experimentation platform , where we validate the results of the two finalists against a conventional A/B.

1 INTRODUCTION

Testing *many* candidate treatments in a standard A/B design forces a choice between calendar time and power: either the experiment runs for long enough for each arm to collect sufficient observations to be powered, or we divide traffic across more arms, maintaining the time cost but hurting experimental power. Adaptive experimentation offers a tempting alternative by shifting traffic away from *underperforming* arms while the experiment runs, so the remaining traffic goes where it is most informative. For this problem of "best-arm identification", once there are $K \geq 3$ arms, adaptive designs exist that dominate any non-adaptive design [4].

We propose one such design, a combination of two familiar ideas. *Sequential halving* [5] is a near-optimal, parameter-free algorithm for best-arm identification under a fixed budget: for K initial arms the exploration horizon is divided into $R = \lceil \log_2 K \rceil$ rounds, the per-day budget is split evenly across the arms still active, and at the end of each round the worst half is eliminated on sample means from that round's observations. *Daily re-randomization* is a "switchback" design on multiple units [2, 6] that re-draws every unit's assignment each day. The unit of observation becomes the unit-day, since a unit in control today may be in treatment tomorrow, and the variance of the treatment effect estimator falls accordingly.

Combining the two is attractive but raises concerns about the bias that re-randomizing daily may introduce. Our design, *sequential-halving switchback* (SHS), therefore ends with a *validation* A/B test on the surviving treatments. We introduce the design, derive what data to use for selection, validate our choice in simulation and report a proof-of-concept run.

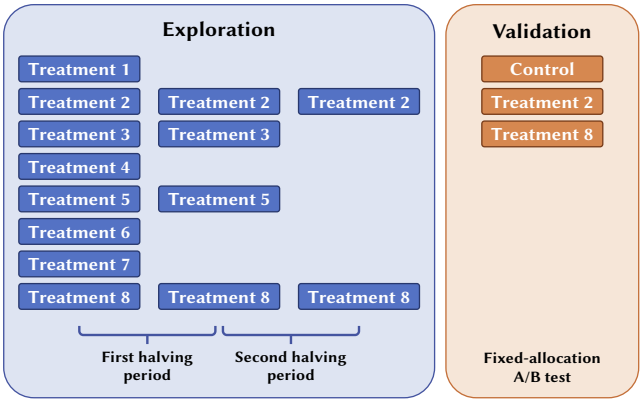


Figure 1: Two-phase experimental design. *Exploration*: starting from many candidate treatments, and re-randomize daily. The set is narrowed by halving: after each halving period the worst performing half is dropped and traffic is re-allocated towards the best performing treatments on a single primary metric. *Validation*: the surviving treatments are compared against control in a standard A/B test

2 THE DESIGN

SHS runs in two phases. In the *exploration phase* sequential halving and daily re-randomization operate jointly: each day the active arms are re-randomized across units, the day's traffic is split evenly among them, and at the end of each round the bottom half is eliminated. Once the target number of finalists is reached, the design enters a *validation phase* in which the survivors are compared against control in a conventional, single-randomization A/B.

For each arm a, let τ_a be its true average effect on the exploration metric relative to control, and $\hat{\tau}_a$ the difference-in-means estimate from the exploration data. Estimators exist for the re-randomized setting that recover each τ_a without bias, but they need strict assumptions [2, 3, 6]. *Exploration doesn't need unbiased estimates of the true effect.* Write $\mathbb{E}[\hat{\tau}_a] = \tau_a + s_a$, where s_a is a shock or bias introduced in the estimator for the true effect of arm a. Exploration works even for non-zero s_a as long as the best arm survives selection, for which it is sufficient that $\mathbb{E}[\hat{\tau}_{\text{best}}] > \mathbb{E}[\hat{\tau}_a]$ for every arm a. In particular: only a shock unique to an arm matters: an arm whose effect lingers into the next day biases every arm equally under uniform re-randomization, so the contrast, and hence selection, is unaffected.

Efficiency of the re-randomization. Re-randomization can improve exploration by changing the unit of analysis and randomization from unit to unit-day. Visitors to a website are a common example: some users visit regularly and trigger daily, but only a fraction of the population triggers on any given day. We model this with two rates. On a total population of N users, let b be the

base rate, the probability that a unit triggers on a day having never triggered before, and r the *retrigger rate*, the per-day probability that a unit which has already triggered does so again.

A single randomization analyses and randomizes on units, and so each unit's "re-triggers" carry the same treatment and data is aggregated to that unit. Its effective sample is the count of *distinct* units triggered so far, $n_{\mathrm{AB}}(T)/N = 1-(1-b)^T$, which is concave and *saturates*. Daily re-randomization instead re-draws assignments every day and unit-days scale more favorably:

$$\frac{n_{\mathrm{re}}(T)}{N} \;=\; \big(1-(1-b)^T\big) \;+\; r\Big(T-\frac{1-(1-b)^T}{b}\Big)$$

which is asymptotically *linear* in the horizon T with slope r.

Since the same unit is re-randomized, in practice these extra samples aren't independent and thus true error shrinks by less than $1/\sqrt{rT}$. The study of Section 5 illustrates the shortfall: there, a 2.6× increase in the observed sample bought only a ≈ 1.4× reduction in standard error, short of the $\sqrt{2.6} \approx 1.6$ independence would give. Approaches exist to estimate the true variance, which is left out of scope for this abstract [9].

3 POOLING ACROSS ROUNDS

Sequential halving [5, 8] eliminates arms after each round by using that round's data. It is tempting to retain or *pool* data across rounds since more observations will reduces variance further. However, this is discouraged[1, 7] since it introduces a selection bias . In our setting, we can have a closed form expression for the variance gain under assumptions, and thus can discuss this bias-variance tradoff in more detail.

Let round r run for d_r days with K_r active arms, each with equal variance σ. With N triggers per day split evenly across arms, each arm receives $n_r = Nd_r/K_r$ observations. By round R, pooling uses $n_{1:R} = \sum_{r\le R} n_r$ observations per arm, whereas per-round estimation uses only n_R.

Precision gain. Pooling reduces the standard error by

$$G = \sqrt{\frac{n_{1:R}}{n_R}} = \sqrt{\frac{K_R}{d_R}\sum_{r\le R}\frac{d_r}{K_r}},$$

which under exact halving and equal round lengths is $G = \sqrt{2-2^{-(R-1)}} < \sqrt{2}$. Note that this is bounded, and even on an arbitrarily long horizon, pooling improves precision by less than a factor $\sqrt{2}$.

Bias from a localized shock. Consider a shock of size s affecting one arm k during rounds $J \subseteq \{1,\dots,R-1\}$. It therefore affects $n_J = \sum_{j\in J} n_j$ observations, and pooling retains it with weight $w_J = n_J/n_{1:R}$, shifting the pooled estimate of arm k by sw_J. We call this *localized* since only the differential shifts matter: a common shift cancels in the treatment contrast, so we read s as a shock specific only to one arm k.

Assume equal-length rounds, exact halving and normally distributed outcomes, and let $\Delta = \tau_{\mathrm{best}} - \tau_k > 0$ be the "true gap" between the best arm and the shocked arm. If the shock occurs only before the final round, pooling replaces Δ by $\Delta - sw_J$. Per-round estimation uses only the final, unshocked round, so it retains the full gap Δ but has a standard error larger by the factor G.

We first define the gap in pooled standard-error units:

$$\lambda \equiv \frac{\Delta}{\sigma\sqrt{2/n_{1:R}}} = \frac{\Delta}{\sigma}\sqrt{\frac{n_{1:R}}{2}}. \tag{1}$$

The effective signal-to-noise ratio for the shocked arm is then $\lambda_k = \lambda\,(1 - sw_J/\Delta)$ under pooling, against $\lambda_k = \lambda/G$ under per-round estimation. Comparing the two arms on the final round, each scheme misranks with probability $\Phi(-\lambda_k)$ at its own λ_k. Equating the two cancels λ, so pooling is the worse estimator exactly when the shock it retains exceeds its own fractional precision gain, $s\,w_J > \Delta\,(1 - 1/G)$. In the worst case, a shock covers every round but the last, $w_J = 1 - 1/G^2$ in which case the crossing sits at $s = \Delta\,G/(G+1)$. For a 2-round exploration this is $0.55\,\Delta$, rising to $0.59\,\Delta$ as $R \to \infty$. So as a rule of thumb, *pooling is worse only if the bias on an arm exceeds half of its gap to the best arm.*

4 POOLING VALIDITY IN SIMULATION

To empirically validate the findings, we use an additive model $Y_a = \tau_a + \varepsilon$ for each arm a, with Δ defined on the τ_a as in Section 3. We simulate eight arms and $R = \lceil \log_2 8 \rceil = 3$ rounds of equal length. One arm is the true best, a second trails it by a gap Δ, and the remaining six trail by 3Δ so they rarely interfere. Outcomes are normal with common standard deviation σ on ε. A shock adds s to the second arm's outcomes on every round but the last, the worst case of Section 3.

The closed form from Section 3 compares two finalists, so conditional on that, we emperically validated our closed form in Table 2. To validate the selection itself, we sweep the shock and the standard deviation, both expressed against the gap: the shock as $s/\Delta \in [0, 1]$, and the standard deviation as λ of Equation (1), taking $\lambda \in \{0.5, 1, 2, 3\}$. A lower λ means a higher σ. Table 1 reports the probability each scheme's final pick is the true best arm over 600,000 replications per cell.

At $\lambda \ge 2$, the noise reduces and thus both the shock and true effect dominate the estimators. In this case, the tradeoff sits at the predicted $s/\Delta = 0.570$, measured at 0.576 and 0.570. Importantly, we that as λ decreases, the crossing moves to 0.75 at $\lambda = 1$ and for $\lambda = 0.5$ where pooling remains dominant even at a shock the full size of the gap.

5 A VALIDATION STUDY

Setup. As a proof of concept, we ran an SHS exploration over nine arms (a control plus eight treatments) on an experimentation platform, with pooling of data and narrowing to two finalists plus control over two halving rounds, then ran a three-week validation A/B on those finalists. Metric identities are withheld, so we provide magnitude in units of the standard error σ for interpretability. Taking the validation on the winning treatments as the reference, we compare the two designs on 7 metrics × 2 treatments = 14 comparisons. For each we ask whether the estimates agree in *sign*, differ significantly from each other, and agree on *significance* against control at $p < 0.05$. The windows between exploration and validation are disjoint, so the estimates are independent. Note that we do not explicitly validate the selection here, and assume that if the final pooled estimates of exploration meet those 3 criteria, it serves as an empirical indication that arm selection was correct since the

Table 1: Probability the top-ranked arm is the true best, by shock ratio s/Δ and gap-to-noise λ (Eq. (1)), over 600,000 replications per cell. The shock hits every round but the last; the middle band is the predicted break-even. Bold marks the higher probability in each pair, unless the two are within three Monte Carlo standard errors.

	$\lambda = 0.5$		$\lambda = 1$		$\lambda = 2$		$\lambda = 3$	
s/Δ	pool	per-rd.	pool	per-rd.	pool	per-rd.	pool	per-rd.
0	**.501**	.470	**.776**	.736	**.974**	.935	**.999**	.989
0.10	**.496**	.465	**.766**	.729	**.969**	.934	**.998**	.988
0.25	**.488**	.460	**.748**	.719	**.960**	.933	**.996**	.988
0.50	**.474**	.450	**.721**	.703	**.939**	.932	**.991**	.988
0.570	**.471**	.447	**.713**	.700	.932	.931	.988	.988
0.75	**.460**	.443	.690	.690	.910	**.931**	.979	**.988**
1.00	**.447**	.432	.657	**.679**	.870	**.931**	.957	**.989**

Table 2: The closed form under pooling, conditional on the best arm and the shocked arm being the two finalists. *Sim.* is the simulated probability of then picking the true best arm over 400,000 replications, against the predicted $\Phi(\lambda(1-sw_J/\Delta))$. The largest gap is 0.019, at $\lambda = 0.5$ under no shock. Per-round estimation matches its own $\Phi(\lambda/G)$ to 0.005 in every cell.

	$s/\Delta = 0$		$s/\Delta = 0.570$		$s/\Delta = 1$	
λ	Sim.	Φ	Sim.	Φ	Sim.	Φ
0.5	.673	.692	.647	.647	.625	.613
1	.826	.841	.774	.775	.725	.716
2	.976	.977	.935	.935	.875	.874
3	.999	.999	.988	.988	.957	.957

metrics were estimated in line with our criteria. Validating arm selection directly would require a multi-way non-adaptive experiment across all pairs of the 9 arms which is underpowered in any realistic experimentation.

Results. The designs agree in sign on all but one comparison, and that exception is a metric neither design found significant. No exploration estimate differs significantly from its validation counterpart. The significance verdicts moreover *nest*: every metric significant under validation is also significant under exploration, and never the reverse, so exploration is the more sensitive of the two and contradicts validation nowhere. The median absolute disagreement between point estimates is $0.18\,\sigma$, and the largest discrepancy is $1.63\,\sigma$ ($p = 0.10$). On a metric significant in both designs (Figure 2), the interval shrinks steadily as daily re-randomized data accumulate, leading to tighter intervals than the A/B validation which saturates.

What pooling bought. Pooling cuts the mean exploration standard error by a factor of 1.4, improving precision. We evaluated both schemes (pooled and non pooled) on the collected data and note that both eliminate the same four arms in round 1 and return the same two finalists after round 2.

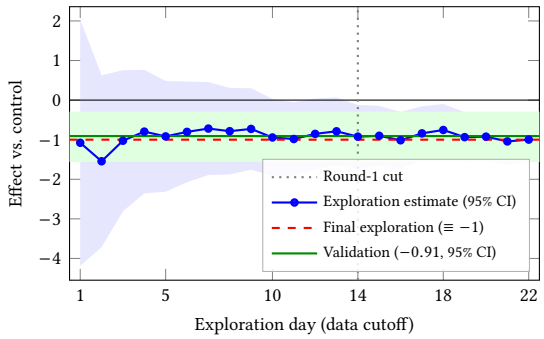


Figure 2: Cumulative pooled exploration estimate (blue, 95% CI band) by exploration day, for a representative treatment on a metric that moved. The axis is effect versus control in multiples of the final exploration estimate, so that estimate is −1 by construction. Green is the 95% CI of the independent three-week validation A/B, at −0.91 on the same scale; the dotted line marks the round-1 elimination. Exploration ends tighter than validation.

6 CONCLUSION

We proposed sequential-halving switchback, which exploits the fact that best-arm selection needs correct rankings rather than unbiased estimates to run exploration on daily re-randomized traffic. Pooling across rounds is preferable unless an arm-specific shock exceeds roughly half the gap to the best arm, and a proof-of-concept run over nine arms agreed with a confirmatory A/B in sign on all but one comparison.

REFERENCES

[1] Peter Bauer and Karl Köhne. Evaluation of experiments with adaptive interim analyses. *Biometrics*, pages 1029–1041, 1994.

[2] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777, 2023. doi: 10.1287/mnsc.2022.4583.

[3] Yuchen Hu and Stefan Wager. Switchback experiments under geometric mixing. *arXiv preprint arXiv:2209.00197*, 2022.

[4] Guido Imbens, Chao Qin, and Stefan Wager. Admissibility of completely randomized trials: A large-deviation approach. *arXiv preprint arXiv:2506.05329*, 2025.

[5] Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*, pages 1238–1246. PMLR, 2013.

[6] Paul Missault, Lorenzo Masoero, Christian Delbé, Thomas Richardson, and Guido Imbens. Robust and efficient multiple-unit switchback experimentation. *arXiv preprint arXiv:2506.12654*, 2025.

[7] Philip Pallmann, Alun W Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V Hampson, Jane Holmes, Adrian P Mander, Lang'o Odondi, Matthew R Sydes, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC medicine*, 16(1):29, 2018.

[8] Qining Zhang, Tanner Fiez, Yi Liu, and Wenyang Liu. Multi-metric adaptive experimental design under a fixed budget with validation. *arXiv preprint arXiv:2506.03062*, 2025.

[9] Jing Zhou, Jiannan Lu, and Anas Shallah. All about sample-size calculations for a/b testing: Novel extensions & practical guide. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3574–3583, 2023.