When and How to Pilot: Design Rules for Two-Wave Experiments

Juan C. Yamin*
Brown University

Nearly one in ten of the 10,905 trials registered in the AEA RCT Registry over the past decade reports running a pilot. Digital experiments often share this two-wave structure, a small first wave of traffic, the ramp-up, preceding the full experiment [Kohavi et al., 2020]. Experimenters run these first waves for many reasons, but pilot data can also inform a consequential design choice, the treatment assignment probability, in platform terms the traffic split. The assignment probability determines how many observations each arm receives, and with them the precision of the average treatment effect estimator and the smallest effect the experiment can reliably detect from its traffic. This paper asks how much authority a small pilot should have over the main-wave design, and answers with a closed-form assignment rule that moves from balance only as far as the pilot's evidence has earned, together with a finite-sample guarantee on what the move can cost.

If the potential-outcome variances under treatment and control were known, the variance-minimizing rule would be the Neyman allocation [Neyman, 1934]. Assigning a main-wave share π of the sample to treatment yields a variance proportional to

$$V(\pi, \sigma_1^2, \sigma_0^2) = \frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}, \qquad \text{minimized at} \qquad \pi^*(\sigma_1^2, \sigma_0^2) = \frac{\sigma_1}{\sigma_1 + \sigma_0},$$

so the optimal rule tilts traffic toward the noisier arm, attaining the benchmark $V^*(\sigma_1^2, \sigma_0^2) = (\sigma_1 + \sigma_0)^2$. These variances are unknown when the main wave is designed. A natural rule estimates them from the pilot and substitutes the sample variances into the Neyman formula, yielding the feasible Neyman allocation (FNA). When the pilot is large this plug-in logic is well founded, and pilot-based variance estimates deliver asymptotically efficient assignments [Hahn et al., 2011, Armstrong, 2026].

When the pilot is small, however, the premise of the plug-in fails, and the experimenter must decide how much the design should respond to variance estimates that are not necessarily trustworthy. Keeping the main wave evenly split caps the possible precision loss whatever the true variances, at the cost of forgoing the gains the pilot could deliver. Following the pilot estimates as if they were true removes this protection, and nothing limits how far noisy estimates can tilt the assignment. In the registry, the median reported pilot has about 300 observations and roughly a tenth have fewer than 30. At such sizes, FNA can lose to balance even in large main waves [Cai and Rafi, 2024].

We formalize this choice as a finite-sample statistical decision problem, on the premise that a design rule should be evaluated with the information actually available at the moment of design, not in an asymptotic world where pilot estimates are treated as known. In a two-wave experiment with bounded outcomes and no parametric restrictions, the experimenter observes a pilot of total size M, then chooses the main-wave assignment, and is judged by regret, the excess variance over the infeasible Neyman allocation. Unlike fully sequential rules that reassign as outcomes arrive [Dai et al., 2023, Zhao, 2024], the design is committed in one shot, as when traffic shares are locked after a ramp-up, experiments run in batches, or the primary metric arrives with delay. The two canonical rules mark the endpoints of this problem. Balance, the no-data rule, is minimax-risk optimal and uniquely minimax-regret optimal without a pilot, but it gives the realized pilot no authority over the assignment. FNA gives the pilot's point estimates full authority, and its worst-case regret is

*Brown University, Department of Economics. Email: juan_yamin_silva@brown.edu. The full paper is available at arXiv:2607.16982.

unbounded at every finite pilot size. In short, balance refuses to move, while feasible Neyman can move too far. The finite-pilot question is how much authority the realized evidence has earned.

Exact minimax regret, the natural ex ante criterion, shows that adaptation is worth it. With as few as two pilot observations per arm its value falls strictly below the no-pilot level, so any optimal rule must move for some realizations. Yet the criterion is not computable, since it ranges over full outcome distributions and evaluates complete pilot-to-assignment mappings.

The Conditional Minimax Regret (CMR) rule answers the authority question by acting on what the pilot has not ruled out rather than on a point estimate [Manski, 2021, Chernozhukov et al., 2025]. It builds a confidence rectangle $\mathcal{C}_M$ for the treatment and control variances, the configurations still consistent with the realized pilot, combining distribution-free bounds on each arm's variance valid at any pilot size [Maurer and Pontil, 2009]. It then chooses the assignment that minimizes worst-case regret over that set,

$$p_{\mathrm{CMR}} \in \arg\min_{\pi} \sup_{(\sigma_1^2,\sigma_0^2)\in\mathcal{C}_M} \left[\, V(\pi,\sigma_1^2,\sigma_0^2) - V^*(\sigma_1^2,\sigma_0^2) \,\right].$$

The solution is a formula rather than an optimization. With $[\underline{\sigma}_d, \overline{\sigma}_d]$ the standard-deviation interval of arm d, $p_{\mathrm{CMR}} = (\underline{\sigma}_1 + \overline{\sigma}_1)/(\underline{\sigma}_1 + \overline{\sigma}_1 + \underline{\sigma}_0 + \overline{\sigma}_0)$, the Neyman formula applied to the midpoint of each arm's interval. When the rectangle is wide, CMR stays close to balance. As it contracts around the true variances, CMR moves toward the Neyman allocation. Alongside the assignment, the rule reports a certificate, a bound on the regret the chosen assignment can incur [Andrews and Chen, 2025]. The rule comes with finite-sample and large-pilot guarantees.

Theorem (informal). *(i) With probability at least $1-\alpha$, the realized regret of p_{CMR} is bounded by its certificate, which never exceeds the no-pilot guarantee of $1/4$ for outcomes normalized to $[0,1]$. (ii) At interior variance pairs, p_{CMR} converges to the Neyman allocation at rate $M^{-1/2}$, with realized regret and certificate of order M^{-1}. (iii) Uniformly over bounded outcome distributions, expected regret is $O(M^{-1/2})$, matching the minimax-regret lower bound up to constants.*

The construction extends to shared-control A/B/n tests and to stratified designs that split traffic across user segments or markets, where CMR selects an allocation vector and again moves from the no-information allocation toward Neyman as the pilot informs. Binary outcomes admit exact finite-sample confidence sets, a bounded-kurtosis condition replaces the known outcome bound, and the rule applies when the ramp-up observes a fast proxy while the primary outcome arrives later.

We also ask how large the pilot itself should be. Pilot observations could have gone to the main wave, so a useful pilot must be informative enough to repay its cost. With a total budget of B observations, a balanced pilot of roughly $B^{2/3}$ observations, followed by CMR, attains worst-case performance within a constant factor of the best any two-wave design can achieve, and any design that comes close must use a pilot of the same order. A budget of 1,000 thus pilots with about 100, and one of 100,000 with about 2,000, larger pilots consuming smaller shares.

We stress-test the rules in two large digital experiments whose contrast maps the two regimes of the design problem (Figure 1). Each calibration takes the outcome rates the experiment realized in its randomized arms as the truth. We repeatedly draw a balanced ramp-up, apply each rule, and evaluate the implied main-wave variance at those rates. The first is a typical Upworthy test [Matias et al., 2021]. Upworthy randomly assigned each reader one of four images for the same story headline, and the outcome is whether the reader clicked, with click rates below two percent. At a 100-impression ramp-up, feasible Neyman assigns some arm zero traffic with probability 0.78, since an arm with no clicks is estimated to have zero variance, and clipping the assignment, the standard repair, still needs three times the traffic. With only one percent to gain, any movement mostly chases sampling noise, and CMR pays a small price for adapting, at most two and a half percentage points over balance at any ramp-up size. The second is a stratified design calibrated to

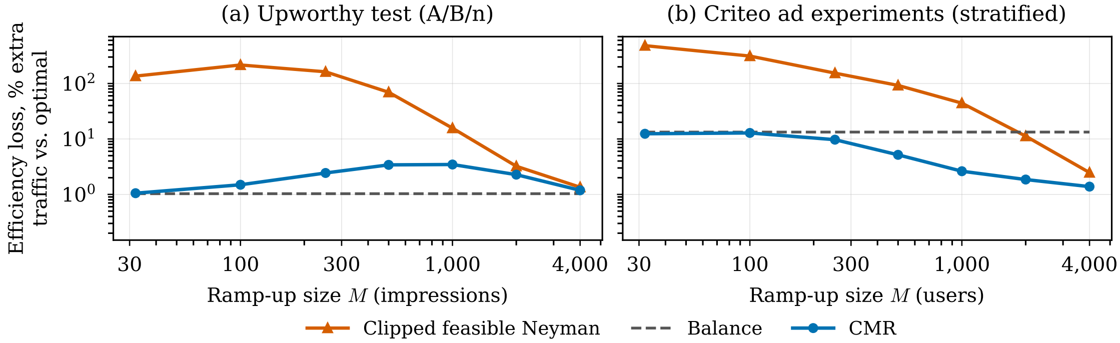

Figure 1: Efficiency loss by ramp-up size, the percent additional main-wave traffic a rule needs to match the precision of the optimal, infeasible Neyman design, both axes in logs. Each panel reports the mean loss over ramp-up draws for one calibrated design. Panel (a) is the archive's median-gain four-arm test among those with every arm above 5,000 impressions and 30 clicks, testing three images against the first-created one as shared control and minimizing the summed variance of the contrasts. Panel (b) is the four-segment Criteo calibration. Balance is each design's no-information allocation, shared-control balance in (a) and representative sampling with even splits in (b). Unclipped feasible Neyman is omitted because it assigns a positive-variance arm zero traffic in up to 83 percent (a) and 100 percent (b) of ramp-ups below 1,000, making its expected loss infinite. CMR uses the exact binary rectangle at $\alpha = 0.05$. Clipped FNA floors arm shares at 5 percent (a) and cell shares at a tenth of their representative share (b), renormalizing.

Criteo's 13.9-million-user collection of randomized advertising experiments [Diemert et al., 2018], where treatment is eligibility for an ad campaign and the outcome is a visit to the advertiser's site. Visit rates across four user segments, quartiles of a baseline covariate, range from 1.2 to 8.5 percent, so the optimal design oversamples the noisiest segment and representative balance needs 13 percent more traffic to match it. CMR captures 61 percent of that attainable gain at a 500-user ramp-up and about 90 percent by $M = 4{,}000$. Across both designs, CMR avoids feasible Neyman's severe small-pilot losses while capturing most of its large-pilot gains.

References

Isaiah Andrews and Jiafeng Chen. Certified decisions, 2025. URL https://arxiv.org/abs/2502.17830.

Timothy B. Armstrong. Asymptotic efficiency bounds for a class of experimental designs, 2026. URL https://arxiv.org/abs/2205.02726.

Yong Cai and Ahnaf Rafi. On the performance of the Neyman allocation with small pilots. *Journal of Econometrics*, 242(1):105793, 2024.

Victor Chernozhukov, Sokbae Lee, Adam M Rosen, and Liyang Sun. Policy learning with confidence. *arXiv preprint arXiv:2502.10653*, 2025.

Jessica Dai, Paula Gradu, and Christopher Harshaw. Clip-OGD: An experimental design for adaptive Neyman allocation in sequential experiments. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2305.17187.

Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD*, London, 2018.

Jinyong Hahn, Keisuke Hirano, and Dean Karlan. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108, 2011.

Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, Cambridge, 2020.

Charles F Manski. Econometrics for decision making: Building foundations sketched by Haavelmo and Wald. *Econometrica*, 89(6):2827–2853, 2021.

J. Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. The upworthy research archive, a time series of 32,487 experiments in U.S. media. *Scientific Data*, 8(195), 2021.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934.

Jinglong Zhao. Adaptive Neyman allocation. In *Proceedings of the 25th ACM Conference on Economics and Computation*, 2024. arXiv:2309.08808.