Bayesian sequential testing for optimal experimental design

Mark Robinson
STATS 209 - Final project

1 Abstract

Experimental programs in the technology industry are typically designed and carried out using null hypothesis statistical testing. This paper develops an alternative Bayesian framework that employs sequential testing for optimal experimental program design under limited resources. The framework generates 1.9-3.8× higher returns than a static testing baseline, with larger gains in more challenging testing environments. Critically, the framework also endogenously determines all testing parameters that remain discretionary in existing frequentist and Bayesian approaches, including testing duration, implementation rules, and resource allocation. The sequential Bayesian approach strictly weakly sequential frequentist testing under any frequentist design, and exhibits simple implementation rules providing compelling justification for adoption in practice.

2 Introduction

Technology companies repeatedly test product ideas through randomized experiments as part of experimentation programs. Standard frequentist approaches based on null hypothesis testing prioritize statistical significance of individual tests. This does not reflect the true goal of these programs, which is typically optimization of some return (e.g. user satisfaction, or click-through rate). Sudijono et al. (2025) show that in a Bayesian framework, experimentation programs can be designed both based on the relevant business metrics, and accounting for resource constraints, leading to optimized returns. However, existing frameworks for Bayesian experimental program design have been static, requiring commitment to fixed testing durations before observing results, missing the value of sequential decision-making.

This paper extends the Bayesian experimental design framework to incorporate sequential testing with endogenous stopping decisions. Similar to the approach first employed by Wald (1945) for general hypothesis testing, I show that organizations will optimally shelve unpromising ideas, ship promising ones, or continue testing, based on accumulating evidence over time. I find this Bayesian sequential approach generates 1.9-3.8× higher returns than static baselines by: (i) rapidly eliminating bad ideas to free resources, (ii) quickly deploying good ideas to accelerate returns, and (iii) concentrating extended testing on marginal cases where information is most valuable.

Critically, the sequential Bayesian framework with resource constraints endogenously determines all testing parameters that remain discretionary under frequentist approaches. Organizations need only estimate prior distributions from historical data; the framework then optimally determines testing duration, stopping thresholds, and resource allocation. This contrasts with frequentist sequential testing where practitioners must specify significance levels, power allocations across stages and other design parameters. As importantly, due to these parameters being determined optimally, the Bayesian approach weakly dominates sequential frequentist testing under any frequentist testing design.

2.1 Related literature and contributions

Work on frequentist sequential testing has become increasingly of interest from the perspective of experimentation teams. Common approaches include traditional 'peeking' (e.g. O'Brien and Fleming (1979)), where practitioners commit to fixed tests but check results at intermediate stages, and more recently anytime-valid confidence intervals (Maharaj et al., 2023), which permit valid inference at any stopping time, overcoming the pre-registration constraints of 'peeking' approaches. While anytime-valid methods provide strong theoretical guarantees, their

conservativeness and restrictive assumptions have somewhat limited practical adoption. Bayesian sequential testing has been also been explored, for example by Stucchio (2015) and Deng et al. (2016). However, these works optimize individual experiments without considering overall design with program-wide resource constraints.

This paper embeds sequential testing within an experimental program framework where organizations face resource constraints and must allocate testing capacity across a portfolio of ideas. This setting resolves parameter selection issues that make alternative approaches subjective and/or difficult to implement optimally: frequentist 'peeking' requires discretionary choices of significance levels and power allocations across stages; and existing Bayesian methods require arbitrary tolerance parameters governing testing duration (Stucchio, 2015). By contrast, the program-level framework with optimally determined stopping thresholds endogenously determines all program testing parameters from the prior distribution alone.

Three key findings strengthen the case for Bayesian adoption. First, the sequential Bayesian framework weakly dominates sequential frequentist testing under any frequentist design (Corollary 1.1), extending beyond Sudijono et al. (2025)'s similar result for static Bayesian and frequentist approaches. Second, optimal implementation at each period is made straightforward by a simple straightforward threshold rule (Proposition 1). Finally, all testing parameters are optimally determined by the framework given only the prior distribution, eliminating discretionary choices that remain in both frequentist sequential testing and existing Bayesian implementations. These properties make the approach particularly suitable for practical implementation.

3 Model

3.1 Framework

I develop a Bayesian framework with sequential testing for optimal experimental design for organizations that repeatedly test and deploy new product ideas. Organizations possess fixed testing resources with which to experiment and assess the value of new ideas. Ideas are evaluated through controlled randomized experiments such A/B testing, and based on accumulated evidence, the organization decides whether to ship the idea (deploy to all units), shelve it (discard), or continue testing to gather more information, with a view to shipping the idea at a later date.

Following Azevedo et al. (2020), I formalize this as an A/B testing problem as follows. Consider an experimentation program with a pool of ideas indexed by $i = 1, \ldots, I_t$ at discrete time t. Each period, k new ideas are generated, where k is sufficiently large to ensure interior solutions.[1] The organization maximizes expected returns over an infinite horizon with discount factor $\gamma \in (0, 1)$.

Each idea i has an associated treatment effect Δ_i representing the return from implementation of an idea, measured in a common metric across all ideas.[2] Following empirical evidence from past experiments (Sudijono et al., 2025), I assume treatment effects are drawn independently from a normal prior distribution: $\Delta_i \sim N(\mu, \tau^2)$.

Each period, the organization allocates its testing resources by choosing $(n_{i,t})_{i=1}^{I_t}$, where $n_{i,t} \geq 0$ denotes the testing resources assigned to test idea i at time t. I use n for notation as these resources typically denote one user to be tested on, but they equally might denote groups of users (e.g. 000s of users) or testing spend. The allocation is constrained by the total resource pool size: $\sum_{i=1}^{I_t} n_{i,t} \leq N_t$, and each testing resource can only be assigned to one test per period. For ideas allocated some testing resources ($n_{i,t} > 0$), the organization observes a noisy signal:

$$\hat{\Delta}_i \sim N\left(\Delta_i, \frac{\sigma^2}{n_{i,t}}\right).$$

The organization's objective is to select allocations $(n_{i,t})_{i=1}^{I_t}$ and shipping decisions $S_t \subseteq [I_t]$ at each period to maximize total expected discounted utility:

$$\max \ \mathbb{E}\left[\sum_{t=1}^{\infty} \sum_{i \in S_t} \gamma^{t-1} u(\Delta_i)\right],$$

where $u(\cdot)$ is an increasing utility function.

[1] For indexing convenience, I write $I_t = k \cdot t$, though in practice ideas would exit the pool upon deployment. This simplification does not affect the analysis.

[2] This is presented as a one-off return without loss of generality. Δ_i could equivalently be the present discounted value of implementing the idea for all future periods.

3.2 Optimal decision rules

In a static, single-period setting, the shipping decision is straightforward: implement all ideas with positive expected utility given their posterior distribution. Formally, $i \in S \Leftrightarrow \mathbb{E}[u(\Delta_i)|\hat{\Delta}_i; n_i] \geq 0$. Without future testing opportunities, the organization simply maximizes:

$$\sum_{i=1}^{I} \mathbb{E}\left[\max\{0, \mathbb{E}[u(\Delta_i)|\hat{\Delta}_i; n_i]\}\right]$$

However, in the sequential testing framework, this policy is suboptimal. Ideas not shipped at time t remain available for additional testing at $t+1$, creating a value for deferring decisions. The optimal policy must balance three considerations: (i) immediate implementation of promising ideas, (ii) continued testing of uncertain ideas to gather more information, and (iii) efficient abandonment of unpromising ideas to free resources for new tests.

To characterize the optimal solution, I impose several simplifying assumptions. First, I assume constant resources across time: $N_t = N$ for all t. Second, that the number of ideas j selected for testing each period is time-invariant.[3] Third, that resources are allocated equally among tested ideas: $n_{i,t} \in \{0, N/j\}$ given j. This follows Sudijono et al. (2025) who show this allocation is optimal in the static case given a sufficient number of ideas and resources.

Under these assumptions, we can characterize our knowledge of each idea's value by the number of completed tests $x_{i,t}$ and the cumulative posterior mean $\hat{\Delta}^c_{i,t}$, which are sufficient for the posterior distribution given the normal-normal conjugate structure. The organization's decision problem reduces to: (i) selecting the optimal j, (ii) choosing which j ideas to test each period, and (iii) determining which ideas to ship.

The following proposition establishes that optimal decisions for (ii) and (iii) above follow a simple threshold structure, allocating ideas to shelve, continue testing, and ship:

Proposition 1. *At the optimal solution, there exist threshold functions $\alpha(x)$ and $\beta(x)$ such that for an idea with x completed tests and posterior mean $\hat{\Delta}^c$:*

- *If $\hat{\Delta}^c < \alpha(x)$: shelve the idea (discontinue testing)*
- *If $\hat{\Delta}^c \in [\alpha(x), \beta(x)]$: continue testing (re-test the idea next period)*
- *If $\hat{\Delta}^c > \beta(x)$: ship the idea (implement immediately)*

The proof is provided in the appendix. Proposition 1 simplifies the organization's problem to identifying the optimal threshold functions $(\alpha(x), \beta(x))$ and the optimal number of ideas j to test per period. Given optimal thresholds, the organization maximizes:

$$\max_j \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \sum_{i=1}^{j} \mathbf{1}\{\hat{\Delta}^c_{i,t} > \beta(x_{i,t})\} \cdot \mathbb{E}[u(\Delta_i)|\hat{\Delta}^c_{i,t}; x_{i,t}]\right]$$

Then with j, $\alpha(x)$ and $\beta(x)$ optimally determined, Proposition 1 implies program testing proceeds simply as follows:

- *First period* $(t = 1)$:
  - **Selection**: Select and test j ideas at random, given all are i.i.d. distributed with the same prior.
  - **Implementation**: Ship ideas i with $\hat{\Delta}^c_{i,1} > \beta(x_{i,1})$
- *Future periods* $(t > 1)$:
  - **Selection**: Select and test the j'_{t-1} ideas i with $\hat{\Delta}^c_{i,t-1} \in [\alpha(x_{i,t-1}), \beta(x_{i,t-1})]$ from time period $t-1$; and $j - j'_{t-1}$ new ideas selected at random.
  - **Implementation**: Ship ideas i with $\hat{\Delta}^c_{i,t} > \beta(x_{i,t})$

The following corollary compares the Bayesian and frequentist sequential testing approaches.

[3] I note this likely to be approximately optimal in any case as competing effects (information accumulation versus resource constraints) balance out across the distribution of ideas

Corollary 1.1. *The sequential Bayesian policy achieves weakly higher expected returns than sequential frequentist testing for any frequentist design (choice of significance level and size allocations), with strict inequality except in degenerate cases where frequentist parameters are reverse-engineered to replicate the Bayesian solution exactly.*

This extends the equivalence result in Sudijono et al. (2025), who shows that static Bayesian and frequentist approaches are equivalent if frequentist parameters are reverse-engineered to replicate the Bayesian solution, and the Bayesian approach dominates otherwise. The result provides theoretical justification for adopting Bayesian sequential testing in experimental program design.

The following proposition establishes convergence of the decision thresholds.

Proposition 2. *As $x \to \infty$, the continuation region $[\alpha(x), \beta(x)]$ collapses to a point: $\lim_{x\to\infty}[\beta(x) - \alpha(x)] = 0$. Moreover, there exists finite $x^* < \infty$ such that for all $x \geq x^*$, no continuation region exists.*

The proof appears in the appendix. Proposition 2 gives the desirable guarantee that no ideas will be tested indefinitely: beyond some threshold x^*, all ideas are either shipped (if posterior mean is positive) or shelved (otherwise). Previous approaches such as that of Stucchio (2015) provide similar guarantees, but only via specification of arbitrary stopping thresholds. A strength of this framework is that these thresholds are determined optimally given only the priors over idea values: testing is stopped precisely because the opportunity cost of further testing exceeds the potential gain.

3.3 Numerical solution

I solve the model numerically using dynamic programming with value function iteration. The solution proceeds in two stages:

Stage 1: Computing decision thresholds. I solve for the value function $V(\hat{\Delta}^c, x)$ representing the expected continuation value for an idea with posterior mean $\hat{\Delta}^c$ and x completed tests. This requires finding the restart cost c (the opportunity cost of continuing to test an existing idea versus starting a new one) through fixed-point iteration. Given a candidate c, I solve the infinite-horizon Bellman equation via value iteration. The consistency condition $c = \gamma\mathbb{E}[V(\hat{\Delta}_1^c, 1)]$ pins down the equilibrium restart cost, where $\hat{\Delta}_1^c$ is the posterior mean after the first test. Once $V(\cdot)$ converges, I extract thresholds $\alpha(x)$ and $\beta(x)$ for each testing stage $x = 1, 2, \ldots$.

Stage 2: Optimizing the number of tested ideas. Given the optimal thresholds from Stage 1, I compute the expected utility for each candidate number of tested ideas $j \in \{1, 2, \ldots\}$, accounting for the resource allocation $n_i = N/j$ per idea. The optimal j^* maximizes total expected discounted utility.

Further details of the dynamic programming algorithm are provided in the appendix.

3.4 Baseline comparison and steady-state approach

To quantify the gains from Bayesian sequential testing, I compare my framework to a static-style baseline that commits to a fixed testing duration before making decisions. In this baseline, the organization chooses both the number of ideas j to test and a fixed testing duration T, accumulating signals over T periods without intermediate decision points. At time T, all ideas are simultaneously evaluated and each is either shipped (if expected utility is positive) or shelved.

For fair comparison across frameworks, I evaluate both approaches at a steady state rather than from a forward-looking perspective. The baseline optimization problem maximizes steady-state return:

$$\max_{j,T} \frac{j}{T} \cdot \mathbb{E}\left[\max\{0, \mathbb{E}[u(\Delta_i)|\hat{\Delta}_{i,T}^c; x_{i,T}]\}\right]$$

where j/T represents the steady-state number of ideas resolved per period and the expectation term is the expected utility per idea. Each idea receives $n_i = N/j$ units per period, accumulating $x_{i,t} = T$ observations with variance $\sigma^2 j/(NT)$ before a decision is taken.

Similarly, the sequential testing framework is evaluated at steady state by computing the expected lifetime of an idea $L = \sum_{t=1}^{\infty} \varphi_t$ where φ_t is the survival probability at t (the probability that an idea is tested at stage t, i.e. that it was in the 'continue testing' region at all $t-1$). The number of ideas assessed per period on average is is then j/L, each with expected utility $\mathbb{E}[V(\hat{\Delta}_1^c, 1)]$ computed under the optimal sequential policy.

This steady-state normalization accounts for differences in the timing of decisions and provides a common metric for evaluating performance across approaches.

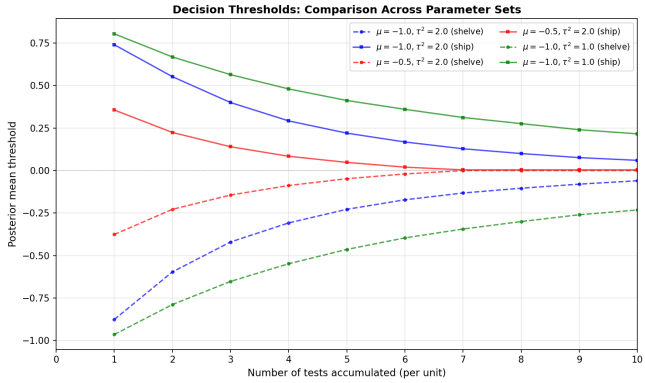

Figure 1: Optimal decision thresholds

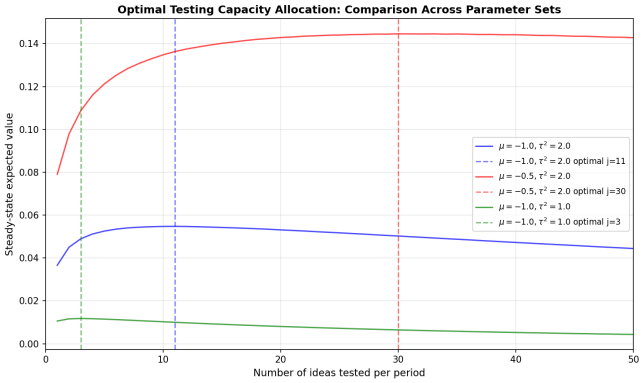

Figure 2: Optimal # ideas tested

4 Results

I demonstrate the benefits of the sequential testing framework through numerical analysis across three parameter configurations, fixing $\sigma^2 = 8.0$, $\gamma = 0.99$, $N = 1$[4] and varying: (i) $\mu = -1.0, \tau^2 = 2.0$ (standard), (ii) $\mu = -0.5, \tau^2 = 2.0$ (high mean), and (iii) $\mu = -1.0, \tau^2 = 1.0$ (low variance). These cases span environments from highly selective (most ideas are bad) to moderately selective, and from high to moderate prior uncertainty. All analyses assume linear utility $u(\Delta_i) = \Delta_i$.

The results demonstrate steady-state performance gains over the baseline ranging from 1.9× to 3.8×, with larger improvements in more challenging environments (lower μ or higher τ^2). The framework's superiority stems from its ability to dynamically reallocate resources by terminating clearly bad or good ideas early.

To explore this performance gain, I present the following: (i) optimal decision thresholds converge monotonically toward zero as testing progresses, in line with Proposition 2, with convergence rates determined by prior uncertainty; (ii) action probabilities reveal efficient filtering with rapid shelving of unpromising ideas and continuation only for marginal cases; (iii) ideas in the continuation region are the source of efficiency gains relative to naive single-period testing; and (iv) the optimal number of per-period tests balances testing quantity and quality given limited resources. Finally, I provide results for an extension to a case with a cost of idea generation.

4.1 Optimal decision thresholds

Figure 1 displays optimal decision thresholds $\alpha(x)$ (shelve/continue) and $\beta(x)$ (continue/ship) as functions of accumulated tests x across the three parameter configurations. The threshold structure validates Propositions 1 and 2. I discuss key observations below

Decisions: For all cases, $\alpha(x) < 0 < \beta(x)$, exhibiting a continuation region. The shelve threshold lies above the prior mean, as negative signals close to the mean merely confirm the prior and justify rapid abandonment. The ship threshold exceeds zero, imposing a higher standard for implementation that accounts for the value of further testing under uncertainty.

Convergence: Both thresholds satisfy $\lim_{x\to\infty}[\beta(x) - \alpha(x)] = 0$ as predicted by Proposition 2, with convergence rates inversely proportional to prior variance τ^2. For $\tau^2 = 1.0$, the continuation region remains substantial even at $x = 10$, while for $\tau^2 = 2.0$ it nearly vanishes by $x = 10$. This reflects that higher prior uncertainty requires more observations to achieve comparable posterior precision.

Comparative statics: Across parameter sets, threshold separation $\beta(x) - \alpha(x)$ is wider for lower μ and higher τ^2. The width of the continuation region quantifies decision-relevant uncertainty: when priors are more pessimistic or uncertain, greater evidence is required before commitment to either shipping or shelving.

[4]Recall N can be thought of as a generic measure of testing resources. I use $N = 1$ as it leads to neat visualization of results, without loss of generality.

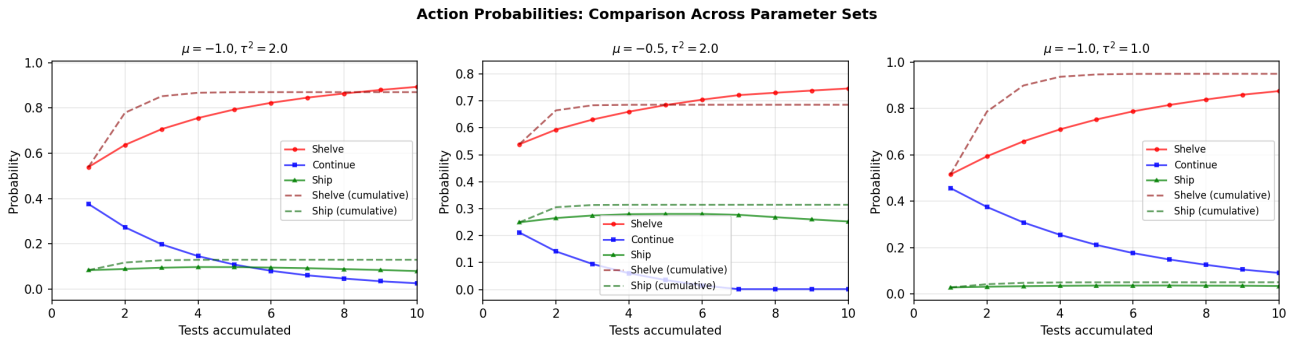

Figure 3: Optimal action probabilities

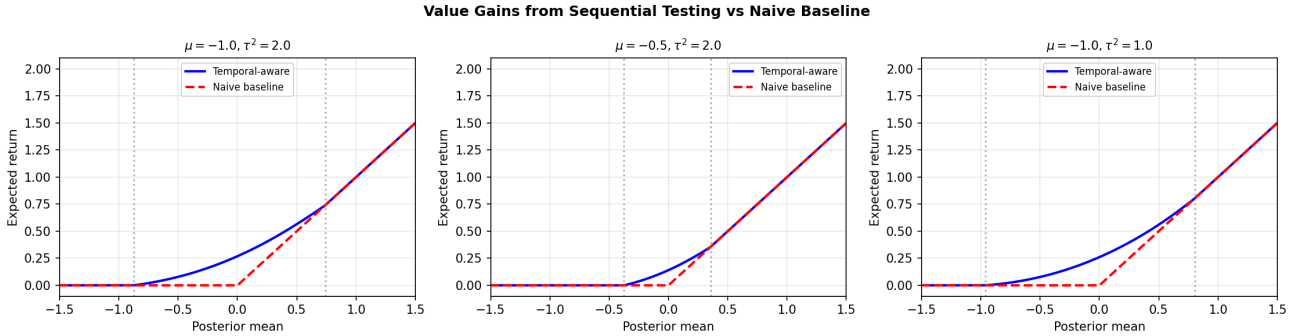

Figure 4: Implied gain compared to naive implementation after one test

4.2 Action probabilities and testing dynamics

Figure 3 decomposes action probabilities $P(\text{shelve}|x)$, $P(\text{continue}|x)$, and $P(\text{ship}|x)$ at each stage x, computed by integrating threshold policies over the prior distribution $N(\mu, \tau^2)$. Three patterns emerge:

Early filtering: For $\mu = -1.0, \tau^2 = 2.0$, approximately 50% of ideas are shelved after the first test, rising cumulatively to 85% by test 10. This reflects efficient early termination when signals confirm the negative prior.

Rapid resolution: Continuation probability decays exponentially. The probability decreases with μ (less selective environments resolve faster) and increases with τ^2 (higher uncertainty prolongs testing). By $x = 6$, continuation probability falls below 20% for all cases, indicating few ideas warrant extended testing.

Selective shipping: Cumulative shipping probability ranges from 5-30% across parameter sets, with higher probability for $\mu = -0.5$. The ratio of shipped to shelved ideas reflects the prior: more optimistic priors lead to higher shipping rates.

Figure 4 uses these thresholds to compare the sequential testing policy with a naive single-period policy that ships all ideas with positive posterior means and shelves the rest after one test. These policies are compared based on the expected value per idea, given the posterior mean after one test. The sequential testing framework generates strictly higher value for all posterior means in the continuation region $[\alpha(1), \beta(1)]$ where single-period decisions have highest error rates. In this region, sequential testing achieves both lower false positive rates (stricter shipping standards) and lower false negative rates (avoiding premature shelving) by allowing evidence to accumulate before committing to irreversible decisions

For posterior means outside the continuation region, the policies coincide: ideas with strongly negative signals are shelved, and ideas with strongly positive signals are shipped by both policies.

4.3 Optimal testing capacity allocation

Figure 2 illustrates the trade-off determining the optimal number of ideas j^* to test per period. This choice balances two competing effects:

Testing quantity effect: Increasing j allows the organization to evaluate more ideas per period, potentially identifying more good ideas to ship. The relative size of this effect depends on the proportion of good ideas in the population.

Testing quality effect: Increasing j spreads fixed testing resources N more thinly, reducing resource allocation per idea (as $n_i = N/j$). Smaller sample sizes per test lead to noisier signals, requiring more testing rounds to achieve comparable posterior precision. This degrades both decision quality and increases the time taken for each decision.

The optimal j^* balances these effects. Comparison between my parameter cases shows the specific optimum depends on the prior distribution: environments with more negative mean ideas (lower μ) favor lower j^* as screening ideas is more important, such that the testing quality effect dominates.

4.4 Performance comparison

Table 1 summarizes steady-state performance across the three parameter cases. The sequential Bayesian framework achieves 1.9-3.8× higher per-period returns than the baseline, with higher gain for more challenging parameter cases: those where it is harder to identify good ideas.

Table 1: Steady-state performance: Baseline vs Sequential Bayesian Testing

Parameters	**Baseline EV**	**Sequential EV**	**Improvement**	**Lifetime L Ratio**	**# Test F Ratio**
$\mu = -1.0, \tau^2 = 2.0$	2.22	5.48	2.46×	92.26×	17.22×
$\mu = -0.5, \tau^2 = 2.0$	7.75	14.46	1.87×	104.54×	20.91×
$\mu = -1.0, \tau^2 = 1.0$	0.31	1.18	3.75×	14.12×	14.07×

Mechanism: Performance gains arise from optimized resource reallocation. Expected lifetime L measures the average number of testing periods before resolution under the sequential policy. Short lifetimes ($L \approx 1.3 - 1.7$) indicate rapid decision-making, with most ideas resolved after 1-2 tests. By contrast, the baseline commits to fixed durations $T \in [25, 150]$ periods, leading to far higher 'testing lifetimes' under the baseline[5]. This gain enables higher a number of tests to be processed per period under Sequential Bayesian testing: $F = j/L$, compared to only j/T ideas per period under the baseline.

Comparative statics: Improvement factors are highest for challenging environments: (i) $\mu = -1.0, \tau^2 = 1.0$ yields 3.75× gains because the negative prior combined with moderate uncertainty makes early shelving highly valuable, but the baseline over-tests due to the restricted set-up; (ii) $\mu = -0.5, \tau^2 = 2.0$ yields only 1.87× gains because the less negative prior reduces the value of early filtering. The sequential approach dominates uniformly but provides more advantage when screening is critical.

4.5 Extension: cost of idea generation

The framework can be flexibly extended to account for richer and more realistic settings. Here, I extend to incorporate costs of generating ideas. Introducing a cost $z > 0$ for each new idea affects the effective opportunity cost of continuing to test an existing idea versus generating and testing a new one: $c' = \gamma \mathbb{E}[V(m_1)] - z$. The subtraction of z makes continuation relatively more attractive, as the alternative (starting fresh) now entails a cost of generating a new idea.

To illustrate the impact, I set $z = 0.002$ using the standard parameters ($\mu = -1.0, \tau^2 = 2.0, \sigma^2 = 8.0, \gamma = 0.99, N = 1$). Figures 5 and 6 display the results. Introducing idea generation costs produces three effects:

Longer testing duration: Both $\alpha(x)$ and $\beta(x)$ shift outward, expanding the continuation region. The expected lifetime L increases as a result ideas undergo more testing rounds before resolution. Organizations become more cautious about both shelving and shipping, as continuing to test an existing idea is now more valuable relative to the costly alternative of generating a new idea.

Reduced testing capacity: The optimal number of tests per period j^* decreases. Organizations shift towards fewer tests of higher quality. This reflects the trade-off between testing quantity and quality: when idea generation

[5]Technically the Lifetime Ratio in the table should be inverted (lifetimes are smaller for Bayesian testing), but it is presented like this to more clearly show the gain from the Bayesian approach

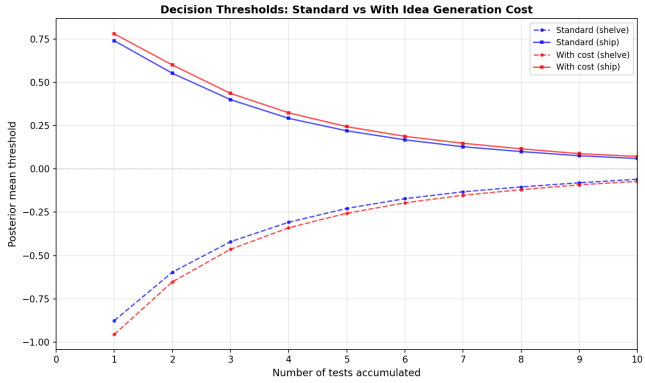

Figure 5: Optimal decision thresholds - with costs

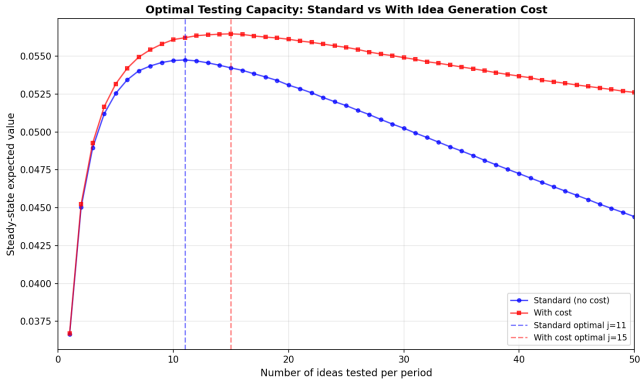

Figure 6: Optimal # ideas tested - with costs

is expensive, it becomes optimal to test fewer ideas more thoroughly rather than cycling quickly through many ideas.

5 Conclusion

This paper develops a framework that combines Bayesian inference and sequential testing for experimental program design under limited resources. Numerical results demonstrate 1.9-3.8× performance gains over a static baseline, with larger improvements in more challenging environments where screening is most valuable.

Three features make this approach particularly suitable for practical implementation. First, as established in Corollary 1.1, the sequential Bayesian framework weakly dominates sequential frequentist testing under any significance level specification. This extends prior results showing equivalence between static Bayesian and frequentist approaches (Sudijono et al., 2025).

Second, as established in Proposition 1, optimal decision-making to shelve, continue testing, or ship ideas each period follows a simple threshold structure based on the average observed signal and the number of tests conducted.

Third, the framework endogenously determines all testing parameters from the prior distribution alone. Organizations need only estimate priors from historical data; the framework then optimally determines stopping thresholds, testing duration, and resource allocation. This eliminates discretionary parameter choices that remain in frequentist sequential testing (significance levels, power allocations between tests) and existing Bayesian implementations (tolerance parameters). The combination of performance superiority and straightforward optimal implementation provides a compelling case for adoption in experimental program design.

Limitations and extensions. My analysis maintains several simplifying assumptions that merit relaxation in future work. First, I assume equal resource allocation across tested ideas ($n_i = N/j$), though heterogeneous allocation might improve efficiency when ideas exhibit different levels of uncertainty. Second, I fix the number of tests per period j in advance, while dynamic adjustment of j in response to the evolving pool of ideas could yield additional gains. Additional extensions could model for richer scenarios such as risk aversion, which may amplify the continuation region by making decision-makers more cautious before shipping, similar to increasing prior variance. Modeling for implementation costs could be incorporated by adjusting the prior mean (for implementation cost c_2: shift μ to $\mu - c_2$).

More generally, it would be valuable to perform empirical calibration using real experimentation data from technology firms to validate the quantitative predictions and guide practical implementation. Finally, I highlight wider applications of this framework beyond A/B testing in the technology industry: the basic set-up extends to any setting involving sequential evaluation of uncertain and heterogeneous opportunities under resource constraints, including hiring decisions and consumer choice.

References

Azevedo, E. M., Deng, A., Montiel Olea, J. L., Rao, J. and Weyl, E. G. (2020), 'A/b testing with fat tails', *Journal of Political Economy* **128**(12), 4614–000.

Deng, A., Lu, J. and Chen, S. (2016), Continuous monitoring of a/b tests without pain: Optional stopping in bayesian testing, *in* '2016 IEEE International conference on data science and advanced analytics (DSAA)', IEEE, pp. 243–252.

Maharaj, A., Sinha, R., Arbour, D., Waudby-Smith, I., Liu, S. Z., Sinha, M., Addanki, R., Ramdas, A., Garg, M. and Swaminathan, V. (2023), Anytime-valid confidence sequences in an enterprise a/b testing platform, *in* 'Companion Proceedings of the ACM Web Conference 2023', pp. 396–400.

O'Brien, P. C. and Fleming, T. R. (1979), 'A multiple testing procedure for clinical trials', *Biometrics* pp. 549–556.

Stucchio, C. (2015), 'Bayesian a/b testing at vwo', *Whitepaper, Visual Website Optimizer* .

Sudijono, T., Ejdemyr, S., Lal, A. and Tingley, M. (2025), Optimizing returns from experimentation programs, *in* 'Proceedings of the 26th ACM Conference on Economics and Computation', pp. 869–869.

Wald, A. (1945), 'Sequential tests of statistical hypotheses', *The Annals of Mathematical Statistics* **16**(2), 117–186.

6 Appendix

6.1 Proof of Proposition 1

I prove that the optimal decision rule can be characterized by threshold functions $\alpha(x)$ and $\beta(x)$ that depend on the number of tests x previously conducted on an idea.

Setup: Define $V(\hat{\Delta}^c_{i,t}, x_{i,t})$ as the expected value of continuing to test an idea i that has cumulative signal $\hat{\Delta}^c_{i,t}$ from $x_{i,t}$ previous tests. The organization faces three choices at each decision point:

1. **Shelve:** Discard the idea and receive zero payoff, then test a new idea with value $V(0,0)$ next period.
2. **Ship:** Implement the idea immediately and receive expected utility $\mathbb{E}[u(\Delta_i)|\hat{\Delta}^c_{i,t}; x_{i,t}]$, then test a new idea with value $V(0,0)$ next period.
3. **Continue testing:** Test the idea again, obtaining an updated signal and moving to state $(\hat{\Delta}^c_{i,t+1}, x_{i,t+1})$ next period.

The value function satisfies the Bellman equation:

$$V(\hat{\Delta}^c_{i,t}, x_{i,t}) = \max\{0, \mathbb{E}[u(\Delta_i)|\hat{\Delta}^c_{i,t}; x_{i,t}] + \gamma V(0,0), -c + \gamma\mathbb{E}[V(\hat{\Delta}^c_{i,t+1}, x_{i,t+1})|\hat{\Delta}^c_{i,t}]\}$$

where c is the opportunity cost of allocating resources to continue testing idea i rather than starting a new idea, and $V(0,0) = \mathbb{E}[V(\hat{\Delta}^c_{i,1}, 1)]$ is the expected value of testing a new idea.

Ship threshold: The firm ships idea i if and only if:

$$\mathbb{E}[u(\Delta_i)|\hat{\Delta}^c_{i,t}; x_{i,t}] + \gamma V(0,0) \geq -c + \gamma\mathbb{E}[V(\hat{\Delta}^c_{i,t+1}, x_{i,t+1})|\hat{\Delta}^c_{i,t}]$$

This defines the upper threshold $\beta(x_{i,t})$ such that $i \in S_t \Leftrightarrow \mathbb{E}[u(\Delta_{i,t})|\hat{\Delta}^c_{i,t}; x_{i,t}] \geq \beta(x_{i,t})$, where

$$\beta(x_{i,t}) = \gamma\mathbb{E}[V(\hat{\Delta}^c_{i,t+1}, x_{i,t+1})|\hat{\Delta}^c_{i,t}] - \gamma V(0,0) - c$$

Shelve threshold: Similarly, the firm shelves idea i (choosing zero payoff) over continuing to test if and only if:

$$0 \geq -c + \gamma\mathbb{E}[V(\hat{\Delta}^c_{i,t+1}, x_{i,t+1})|\hat{\Delta}^c_{i,t}]$$

This defines the lower threshold $\alpha(x_{i,t})$ such that the firm shelves if $\hat{\Delta}^c_{i,t} \leq \alpha(x_{i,t})$, where $\alpha(x_{i,t})$ satisfies $V(\alpha(x_{i,t}), x_{i,t}) = V(0,0)$.

Monotonicity: Given the normal-normal conjugate structure, the posterior mean $\mathbb{E}[\Delta_i|\hat{\Delta}^c_{i,t}; x_{i,t}]$ is monotone increasing in $\hat{\Delta}^c_{i,t}$ and the posterior variance decreases with $x_{i,t}$. This ensures that $V(\hat{\Delta}^c_{i,t}, x_{i,t})$ is monotone in $\hat{\Delta}^c_{i,t}$, establishing the existence of thresholds $\alpha(x_{i,t})$ and $\beta(x_{i,t})$ such that:

- For $\hat{\Delta}^c_{i,t} < \alpha(x_{i,t})$: shelve the idea
- For $\hat{\Delta}^c_{i,t} \in [\alpha(x_{i,t}), \beta(x_{i,t})]$: continue testing
- For $\hat{\Delta}^c_{i,t} > \beta(x_{i,t})$: ship the idea □

6.2 Proof of Corollary 1.1

We show that sequential Bayesian testing achieves weakly higher returns than sequential frequentist testing under any significance level allocation, with strict inequality except in reverse-engineered frequentist cases.

In sequential frequentist testing following Wald (1945), the decision maker chooses error probabilities (α', β') at each stage, which determine stopping boundaries. Let $\{S_t^F\}$ denote the set of ideas shipped under the frequentist policy, determined by comparing test statistics to critical values calibrated to achieve target error rates.

The Bayesian policy maximizes expected utility directly:

$$\max \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \sum_{i \in S_t} u(\Delta_i)\right]$$

subject to resource constraints. By Proposition 1, the optimal Bayesian policy is characterized by thresholds $(\alpha^*(x), \beta^*(x))$ derived from the value function.

The frequentist policy instead maximizes utility subject to error rate constraints:

$$\max \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \sum_{i \in S_t^F} u(\Delta_i)\right] \quad \text{s.t.} \quad P(\text{Type I}) \leq \alpha',\ P(\text{Type II}) \leq \beta'$$

Since the Bayesian policy faces strictly fewer constraints (no error rate requirements), it achieves weakly higher returns: the Bayesian solution lies in the feasible set of the frequentist problem if we ignore error constraints. Equality holds only if the frequentist error rates are chosen such that implied thresholds replicate the Bayesian solution exactly (reverse-engineered case). In all other cases, the frequentist test yields strictly suboptimal decisions relative to the unconstrained Bayesian optimum. □

6.3 Proof of Proposition 2

I prove that the continuation region $[\alpha(x), \beta(x)]$ converges to $\{0\}$ as $x \to \infty$, and that convergence occurs in finite time.

Part 1: Convergence. Under the normal-normal conjugate structure, the posterior variance after x observations evolves as:

$$\tau_x = \left(\frac{1}{\tau_0} + \frac{x}{\sigma^2}\right)^{-1}$$

Thus $\lim_{x\to\infty} \tau_x = 0$, implying posterior beliefs converge to a point mass at the true parameter value as $x \to \infty$.

The value of continuation at stage x is:

$$V_{\text{cont}}(m, x) = -c + \gamma \mathbb{E}[V(m', x+1)|m, x]$$

where m' is the posterior mean after one more observation. As $\tau_x \to 0$, the distribution of m' given m concentrates around m (since $\text{Var}(m') = \tau_x - \tau_{x+1} \to 0$). Therefore:

$$\lim_{x\to\infty} V_{\text{cont}}(m, x) = -c + \gamma \max\{0, m\}$$

For the continuation region to be non-empty, we require:

$$-c + \gamma \max\{0, m\} > \max\{0, m\}$$

For $m > 0$: this requires $-c + \gamma m > m$, or $c < -m(1-\gamma)$, which cannot hold for all $m > 0$ since $c > 0$. For $m < 0$: this requires $-c > 0$, which is impossible. For $m = 0$: this requires $-c > 0$, which is impossible.

Thus as $x \to \infty$, the continuation region vanishes: $\lim_{x\to\infty}[\beta(x) - \alpha(x)] = 0$.

Part 2: Finite convergence. Since τ_x decreases monotonically and $c > 0$ is fixed, there exists finite x^* such that for all $x \geq x^*$, the value of information $\gamma\mathbb{E}[V(m', x+1)|m] - \max\{0, m\}$ is less than the cost c for all posterior means m. At this point, no continuation region exists, and all ideas are immediately shipped (if $m > 0$) or shelved (if $m \leq 0$). □

6.4 Dynamic programming approach

The sequential testing framework is solved numerically using dynamic programming with value function iteration. The solution proceeds in two stages: finding the optimal restart cost c through fixed-point iteration, and computing decision thresholds for each testing stage.

6.4.1 Value function iteration

I discretize the state space of posterior means on a grid $\{m_1, m_2, \ldots, m_N\}$ spanning the range $[m_{\min}, m_{\max}]$. For a given restart cost c and posterior variance τ_t after t tests, I solve the infinite-horizon Bellman equation iteratively:

$$V^{(k+1)}(m) = \max\{0, m, -c + \beta\mathbb{E}[V^{(k)}(m')|m]\}$$

where m is the current posterior mean, and m' is the posterior mean after one additional test.

The transition probability $P(m'|m)$ is computed using the normal-normal conjugate updating rules. Given current posterior mean m with variance τ_t, and observing a new signal, the next posterior mean m' follows a normal distribution with mean m and variance $\text{Var}(m') = \tau_t - \tau_{t+1}$, where $\tau_{t+1} = \frac{\tau_t \sigma^2}{\tau_t + \sigma^2}$ is the updated posterior variance.

I construct a transition kernel matrix P where entry $P_{ij} \approx P(m_j|m_i) \cdot dx$ represents the probability of transitioning from grid point m_i to m_j, with dx being the grid spacing. The expected value $\mathbb{E}[V(m')|m_i]$ is then computed as $(P \cdot V)_i$. Iteration continues until convergence: $\max_i |V^{(k+1)}(m_i) - V^{(k)}(m_i)| < \epsilon$.

6.4.2 Fixed-point iteration for restart cost

The restart cost c represents the opportunity cost of testing a previously-tested idea versus starting with a new idea. It must satisfy the consistency condition $c = \gamma\mathbb{E}[V(m_1)]$, where m_1 is the posterior mean after the first test, distributed as $m_1 \sim N(\mu_0, \tau_0 - \tau_1)$.

I find c through fixed-point iteration:

1. Initialize $c^{(0)}$ (e.g., $c^{(0)} = 0.05$)
2. Solve the value function $V^{(k)}$ given $c^{(k)}$ using value iteration
3. Compute $\mathbb{E}[V^{(k)}(m_1)] = \sum_i V^{(k)}(m_i) \cdot P(m_1 = m_i) \cdot dx$
4. Update: $c^{(k+1)} = \lambda c^{(k)} + (1 - \lambda)\gamma\mathbb{E}[V^{(k)}(m_1)]$ (with damping $\lambda \in (0, 1)$)
5. Repeat until $|c^{(k+1)} - c^{(k)}| < \epsilon$

6.4.3 Threshold extraction and optimization

Once the value function converges, I extract decision thresholds α_t and β_t for each testing stage $t = 1, \ldots, T$. For each posterior variance τ_t (after t tests), I solve the Bellman equation and identify:

- $\alpha_t = \max\{m : V(m, \tau_t) = 0\}$ (highest posterior mean where shelving is optimal)
- $\beta_t = \min\{m : V(m, \tau_t) = m\}$ (lowest posterior mean where shipping is optimal)

To optimize the number of ideas j tested per period, I perform a grid search over candidate values. For each j, I compute the effective observation variance $\sigma^2_{\text{eff}} = \sigma^2 \cdot (j/N)$ (reflecting resource allocation N/j per idea), solve the complete dynamic program with this adjusted variance, compute the expected lifetime L under the resulting policy, and evaluate steady-state number of ideas tested $F = j/L$ and expected value $EV = F \cdot \mathbb{E}[V(m_1)]$. The optimal j^* maximizes EV across all candidates.

The baseline comparison model follows a similar approach but solves a simpler optimization over (T, j) where T is the number of periods to accumulate signals before making a single ship/shelve decision, and j is the number of ideas tested. This baseline does not allow for sequential decision-making or early stopping.