Multivariate Empirical Bayes Launch Rule: A Practical Case Study

Bartek Kozakowski, Kenneth Hung, Jingang Miao, and Houssam Nassif

Meta
{bko, kenhung, jingang, houssamn}@meta.com

October 26, 2026

1 Why combine metrics?

Experimentation platforms often have a high-variance primary metric that is important for a decision and several related metrics that are measured more precisely (Li, Nassif, and Luedtke, 2027). Looking only at the high-variance metric can leave many experiments inconclusive. A multivariate Bayesian model offers an established way to combine these readings: it uses historical experiments to estimate cross-metric true-effect correlations, then uses the more precise metrics to sharpen inference about the noisier one. Although the same posterior can also provide denoised effect estimates and help prioritize experiments (Coey and Hung, 2025) by their probability of benefit, this case study evaluates only launch decisions based on posterior-probability thresholds.

There is a long line of prior academic and industry work that covers surrogate and proxy metrics, weighted combinations, experiment splitting, and empirical Bayes methods for one or several outcomes (Athey et al., 2026; Azevedo et al., 2019; Bibaut et al., 2024; Chou, 2026; Coey and Cunningham, 2019; Deng, 2015; Soloff, Guntuboyina, and Sen, 2025; Stephens, 2017; Tripuraneni et al., 2024; Urbut et al., 2019). Our contribution here is a detailed implementation and validation case study. We hope to supplement the literature by offering a practitioner's view—by describing what worked, what failed, and what teams building similar capabilities should check in their own systems.

Our application tracks the relative treatment effects on 5 metrics: 1 high-variance primary metric, used to make launch decisions, and 4 related, lower-variance metrics. The model returns the posterior probability that the true effect on the primary metric is positive. A platform could compare that probability with a posterior-probability threshold for launch, but it should do so only in a range where calibration has been demonstrated.

2 Model and assumptions

For experiment i, let X_i be the observed effects vector and θ_i the true effects vector. We use the measurement model

$$X_i \mid \theta_i \sim \mathcal{N}(\theta_i, \Sigma_i), \qquad \Sigma_i = D_i R D_i,$$

where D_i contains the standard errors reported by the experimentation platform, and R is the cross-metric measurement-error correlation matrix.

We separate the reported standard errors from the cross-metric measurement-error correlation matrix because experimentation systems often report the former for every metric and experiment, but do not record the latter. We estimate R from A/A experiments, where true effects are zero. Using it for A/B experiments assumes that the cross-metric measurement-error correlation matrix is stable across the two settings. Changes in traffic mix, exposure, analysis windows, variance reduction, missingness, or metric implementation can violate that assumption; Section 3 checks it with split-half A/B data.

We model the true effect vector θ_i as drawn from a 3-component multivariate Gaussian mixture. Each component has its own weight, mean, and covariance. We fit the mixture by expectation-maximization (EM) on the marginal likelihood, which accounts for each experiment's measurement error. We initialize from a fixed starting point and stop after 80 iterations:

$$\theta_i \sim \sum_{k=1}^{3} w_k \mathcal{N}(\mu_k, T_k).$$

The 3 components represent small, ordinary, and large true effects, following Deng et al. (2021). We selected this specification using held-out likelihood grouped by experiment. We call the resulting 80-iteration fit the *adopted prior* because it is used for the main results; Section 5 compares it with constrained alternatives. The posterior remains a mixture. It provides 3 quantities used below: $P = \mathbb{P}(\theta_p > 0 \mid X)$, where θ_p denotes the primary metric, the posterior mean $\mathbb{E}[\theta_p \mid X]$, and the posterior probability of belonging to each mixture component. The first quantity is a multivariate analogue (Urbut et al., 2019) of the local false sign rate (Stephens, 2017).

Launch rules and component label. We evaluate three launch rules:

- The **probability-threshold rule** launches when $P \geq t$; the "0.70 rule" uses $t = 0.70$.
- The **posterior-mean rule** launches when $\mathbb{E}[\theta_p \mid X] \geq m$, for a prespecified minimum effect m.
- The **robust rule** launches "likely small-effect" when their $P \geq t$ is satisfied for 9 "perturbed" versions of the small effect component of the prior (Section 5.4).

We call an experiment "likely small-effect" when its posterior probability of belonging to the small-effect component exceeds 0.5.

The model assumes accurate standard errors and cross-metric measurement-error correlations, and assumes historical experiments represent current true effects. For the posterior probabilities to support decisions, they must be calibrated near the posterior-probability thresholds used in practice.

3 Validation design

We divide experiments into 3 groups by time, where the first one is for fitting the prior and the next two cohorts are hold-outs for validation. This mimics what happens in practice, where the distribution of experiments' true effects can change over time (Section 5.3). Both validation cohorts combined together we call the "pooled validation set". With separate A/A readings we estimate the cross-metric measurement-error correlation matrix. Each validation experiment is then randomly divided into two disjoint halves (Coey and Cunningham, 2019) that share the same treatment effect but have independent sampling errors

$$X^{(A)} = \theta + \varepsilon^{(A)}, \qquad X^{(B)} = \theta + \varepsilon^{(B)}.$$

We write $X_p^{(A)}$ and $X_p^{(B)}$ to refer to readings for the primary metric. Additionally, for variance estimation we cluster test-control pairs within experiments because their readings are not independent.

Table 1: Validation checks and the failure modes they diagnose.

Check	**Reference comparison**	**What it diagnoses**
Standardized split difference	Compare $(X_\cdot^{(A)} - X_\cdot^{(B)})/\sqrt{(SE_\cdot^{(A)})^2 + (SE_\cdot^{(B)})^2}$ with mean 0 and variance 1. Separately compare split-half cross-metric measurement-error correlations with A/A estimates.	Reported standard-error scale and cross-metric measurement-error correlation, without using the prior.
Split cross-product	Compare the average $X_p^{(A)} X_p^{(B)}$ with the fitted prior's $\mathbb{E}[\theta_p^2]$.	Spread of true primary-metric effects, without relying on the measurement-error model.
Prediction of $X_p^{(B)}$ from the posterior given $X_p^{(A)}$	Compare the posterior predictive probability that $X_p^{(B)}$ is positive with the observed positive share among experiments above each posterior-probability threshold.	End-to-end calibration of the posterior used by the decision rule.

Three comparisons isolate different failure modes; see Table 1. We fit the prior on historical experiments that excluded both validation cohorts (Malmir, Nassif, et al., 2026; Malmir, Radwan, et al., 2026). For each held-out experiment, the model sees only $X^{(A)}$. Because θ is latent, validation uses its observable implication: the posterior-predictive probability $\mathbb{P}(X_p^{(B)} > 0 \mid X^{(A)})$. We compare this probability with the observed share of positive $X_p^{(B)}$ readings among experiments above each proposed posterior-probability threshold. This jointly checks the prior and measurement-error model and indirectly assesses the launch probability $\mathbb{P}(\theta_p > 0 \mid X^{(A)})$.

4 Validation results and additional findings

Sections 4.1 and 4.2 report the validation checks from Section 3 in the same order, on the pooled validation set. Section 4.3 reports an additional simulation-based finding.

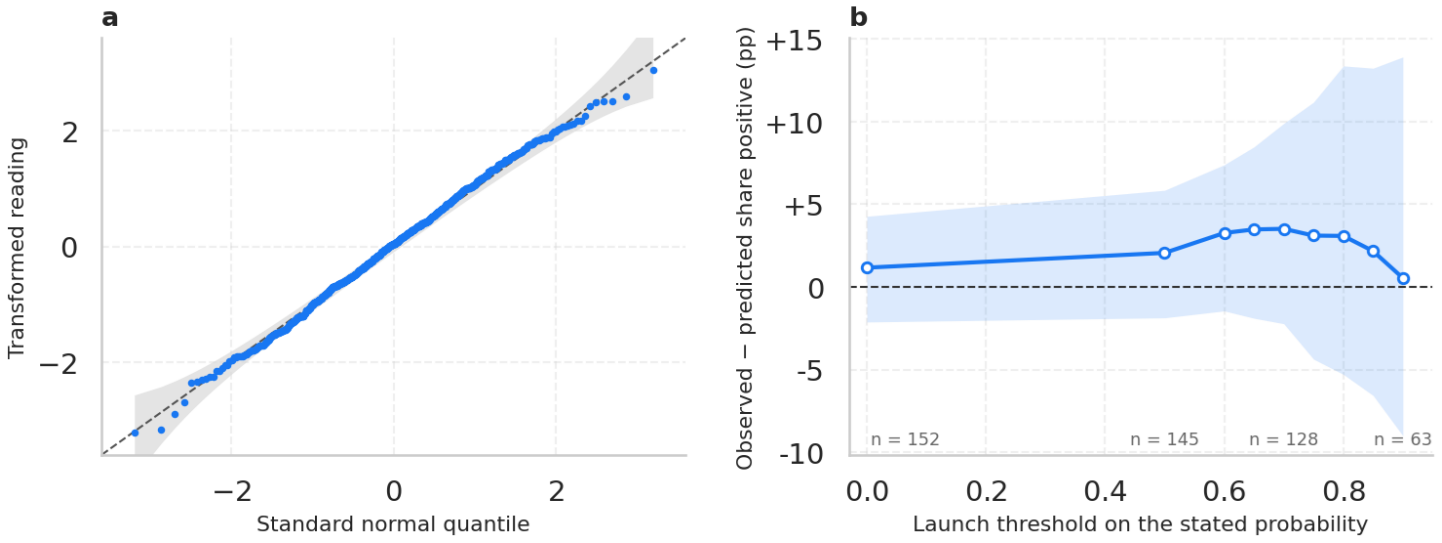

Figure 1: (a) Normal QQ plot of the primary-metric readings, each transformed by its prior predictive CDF. (b) The posterior-calibration check from Section 3, on the pooled validation set. For each posterior-probability threshold, the point uses only experiments above that threshold; the line is the difference between the observed and predicted shares of positive held-out readings, the dashed line marks zero, bands are 90% intervals, and n is the number of experiments.

4.1 Checks: measurement-error model and prior spread

Check	Key result	Interpretation
Scale of reported SEs	Primary metric: variance 1.010 (SE 0.005) against 1, and mean 0.006 (0.004) against 0. Other metrics: variance 0.999 to 1.008.	Reported SEs are accurate to within 1%.
Cross-metric error correlation	A/A against A/B correlations over the 10 metric pairs: RMS difference 0.095. Using the A/B correlation changes P by 0.047 on average.	Some mismatch, with moderate sensitivity here.
Spread of true primary-metric effects	The mean of $X_p^{(A)} X_p^{(B)}$ gives a root mean square of true effects 0.80 times the adopted prior's RMS (90% interval from 0 to 1.16). By this point estimate, the prior is wider than the data indicate, but the interval includes 1	Consistent with the prior, but too imprecise to test it.

4.2 Posterior calibration

The adopted prior's mean absolute calibration gap was 0.0245. Prediction and observation agreed through a posterior-probability threshold of 0.90: the gap stayed between −0.007 and +0.034, and all 90% intervals included zero (Figure 1). At 0.90 the check rests on 64 experiments. Above that threshold, support fell to 36 experiments at 0.95, so this study does not establish calibration in that range.

4.3 Additional findings

In simulations from the adopted prior, true effects of all five metrics were drawn, measurement noise was added, and the model's estimate of the primary metric's true effect was compared with its true value. The squared correlation between the two was 80.9% using only the primary metric, 97.3% using the four auxiliary metrics, and 97.4% using all five. Under the fitted prior, the auxiliary metrics contain almost all of the information about the primary metric.

5 Lessons for practitioners

The validation findings suggest a broader set of implementation lessons. The numbers below describe this case study. They illustrate diagnostics and failure modes; they are not universal performance guarantees.

5.1 Measurement error

Check the noise model before fitting a prior. The two halves of one experiment have the same true effect, so their difference contains only measurement error. If the reported standard errors and error correlations are right, the standardized difference behaves like standard normal noise. The check needs no prior. A failure therefore points to the data path. Section 4.1 found no mismatch in the scale of the standard errors.

Store the error correlations from the experiment itself whenever possible. Platforms report standard errors but rarely the correlation of errors between metrics (Section 2). A/A tests can supply it, but an A/A correlation describes the few A/A experiments that were run, not the A/B experiments being decided. Since the prior is fitted while accounting for the error correlations, a different error correlation will give a different prior. Here the in-experiment estimate differed from the A/A estimate by at most 0.1 on the primary metric's row and by 0.2 on one pair of auxiliary metrics. Used only for scoring, error covariance estimated on individual A/B tests changed probabilities by about 0.05 on average (Section 4.1). Used to fit the prior, it raised the share of readings launched by the 0.70 rule by about a third.

5.2 The prior

Check that the fit has converged in its decisions, not only in its likelihood. A mixture likelihood can be almost flat in some directions. The EM algorithm then takes very small steps, and the likelihood looks converged while some parameters still move. If decisions depend on those parameters, they change with the number of iterations. The adopted prior was stopped after 80 iterations (Section 2). Running it to 2,000 iterations moved only the small-effect component and changed 16% of the decisions under the 0.70 rule.

Prefer simpler priors with fewer parameters. The adopted prior has 62 parameters. Several hundred noisy readings cannot determine all of them, and the result depends on when the fit stops. If the components share one covariance shape, they describe small, ordinary, and large effects with the same pattern across metrics. Such priors had 17 to 24 parameters and reached the same optimum from every starting point. The three-component version with all means at zero was as well calibrated as the adopted prior. It launched 43% fewer experiments as measured on A/A comparisons under the 0.70 rule, and its decisions changed less when it was refitted on resampled data.

Keep decisions away from parameters that the data cannot determine. Put thresholds where decisions are stable, or require each decision to hold over plausible parameter values. The small-effect component is much narrower than the noise of a single reading. A reading therefore hardly moves the posterior of its members. Their probability of a positive effect is set by the component's mean and width, two values that the data determine poorly. Hence we prefer a rule that imposes a minimum posterior mean for stability: among eight priors that passed the convergence check, 27% of decisions under the 0.70 rule differed, compared with 4% under the posterior-mean rule. For edge experiments we use a robust launch rule (Section 5.4).

5.3 Posterior calibration

Validate on a quantity that can be observed, and check sizes as well as signs. The launch probability concerns the true effect, which is never observed. The model also predicts something that is observed: the reading of the held-out half of the same experiment (Section 3). Comparing predicted and observed signs of held-out readings tests the posterior indirectly (Kessler, McQueen, and Rokkanen, 2025). A sign check can pass while the size is wrong, because small and large positive effects have the same sign. For small-effect component launches, the share of positive held-out readings matched the prediction, but their held-out average effect was 3.7 times the predicted average.

Check calibration across kinds of experiments (heterogeneity), not only on average. A pooled check averages over different kinds of experiments. If one kind is overconfident and another is under-confident, the errors cancel, and a cohort with an unusual mix can show a large gap of either sign. We used two cohorts for validation. At 0.90 the first cohort had a negative gap, the new experiments of the second cohort a positive one, and the pooled gap was close to zero. The first cohort's experiments, measured again in the later window, had a negative gap, with an interval that excludes zero. This means the experiments in the first cohort may be different in nature from those in the second cohort.

Compare the full posterior with one that uses the primary metric alone. Section 4.3 showed that the auxiliary metrics carry almost all of the usable signal. A high reading on the noisy primary metric alone is often high by chance; the auxiliary metrics show whether related metrics support it. With the primary metric alone, the share of 0.70 rule launches with a positive held-out reading was 21 points below the model's prediction. With all five metrics, it was within 4 points of the prediction.

5.4 The launch rule and its use

Choose the launch rule on false positives, effect shipped, and stability together. Each criterion alone favors a different threshold. A low threshold ships more effect but also launches more experiments without effect. A high threshold avoids those launches but gives up gains. A decision that changes when the prior is refitted on slightly different data is not reproducible, so stability is the third criterion. One example supplements the 0.70 rule: An experiment labeled "likely small-effect" launches only if it satisfies the 0.70 probability threshold under 9 plausible versions of that component, where its primary-metric mean is set to 0, 0.4, or 0.8 times the component's fitted standard deviation, and its standard deviation is set to 0.4, 1, or 2 times its fitted value, with all other parameters unchanged. This kept 80% of the effect shipped by the plain 0.70 rule, at 20% of its A/A false-positive rate, and its decisions changed less under refits.

References

Athey, Susan et al. (2026). "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely". In: *The Review of Economic Studies* 93.4, pp. 2284–2312.

Azevedo, Eduardo M. et al. (2019). "Empirical Bayes Estimation of Treatment Effects with Many A/B Tests: An Overview". In: *AEA Papers and Proceedings* 109, pp. 43–47.

Bibaut, Aurélien et al. (2024). "Learning the Covariance of Treatment Effects Across Many Weak Experiments". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 153–162.

Chou, Winston (2026). "Blending Proxy Metrics with a North Star". In: *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. arXiv: 2606.21745.

Coey, Dominic and Tom Cunningham (2019). "Improving Treatment Effect Estimators Through Experiment Splitting". In: *The World Wide Web Conference*, pp. 285–295.

Coey, Dominic and Kenneth Hung (2025). "Empirical Bayes Selection for Value Maximization". In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 357–368.

Deng, Alex (2015). "Objective Bayesian Two Sample Hypothesis Testing for Online Controlled Experiments". In: *Proceedings of the 24th International Conference on World Wide Web Companion*, pp. 923–928.

Deng, Alex et al. (2021). "On Post-Selection Inference in A/B Testing". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2743–2752.

Kessler, Ryan, James McQueen, and Miikka Rokkanen (2025). *Evaluating A/B Testing Methodologies via Sample Splitting: Theory and Practice*. arXiv: 2512.03366.

Li, Zhaoqi, Houssam Nassif, and Alex Luedtke (2027). "Estimation of Subsidiary Performance Metrics under Optimal Policies". In: *Statistica Sinica* 37.3.

Malmir, Mohsen, Houssam Nassif, et al. (2026). "Impression Share Prediction: An Offline Evaluation Task for Ranking Systems". In: *Conference on Recommender Systems (RecSys)*.

Malmir, Mohsen, Mohammed A Radwan, et al. (2026). "Beyond IPS: Reliable Counterfactual Evaluation in Multi-Stage Ad Systems without Logged Propensities". In: *Conference on Neural Information Processing Systems (NeurIPS)*.

Soloff, Jake A., Adityanand Guntuboyina, and Bodhisattva Sen (2025). "Multivariate, Heteroscedastic Empirical Bayes via Nonparametric Maximum Likelihood". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 87.1, pp. 1–32.

Stephens, Matthew (2017). "False Discovery Rates: A New Deal". In: *Biostatistics* 18.2, pp. 275–294.

Tripuraneni, Nilesh et al. (2024). "Choosing a Proxy Metric from Past Experiments". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5803–5812.

Urbut, Sarah M. et al. (2019). "Flexible Statistical Methods for Estimating and Testing Effects in Genomic Studies with Multiple Conditions". In: *Nature Genetics* 51.1, pp. 187–195.