

Approximate Sampling for Randomization Tests: A Total-Variation Certification Framework

Mohamed Faisal Sindhi

sindhi@usf.edu

University of South Florida <https://orcid.org/0009-0000-1348-0566>

Hammaz Buksh

University of South Florida

Research Article

Keywords: approximate sampling, randomization tests, total variation distance, covariate-constrained randomization, cluster randomization, quantum optimization

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11322841/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Approximate Sampling for Randomization Tests: A Total-Variation Certification Framework

Mohamed Faisal Sindhi^{1,*}

Hammaz Buksh¹

¹School of Information Systems, University of South Florida

*Corresponding author: `sindhi@usf.edu`

Abstract

Covariate-constrained cluster randomization draws a treatment assignment from a constrained set of balanced allocations, and randomization-based inference relies on that set’s distribution. We study what happens when an approximate sampler replaces the exact target. On one fixed, reproducible benchmark built from the 16-county Dickinson cluster-randomization example, we define an exact accepted set of 1,287 of 12,870 balanced allocations, obtained by exact rational scoring with a canonical index tie-break. We certify sampler outputs by total variation (TV) distance to this target, using full-statevector probabilities where available. On the anchor, ideal statevector results show that a fixed-count amplification control retains raw leakage of 0.00144, while two fixed-depth QAOA headline runs remain far from the target (raw TV 0.833 and 0.948). On four small synthetic instances, the energy-selected XY-QAOA outputs also remain far from the target (raw TV 0.59 to 0.72). The penalty control places 51% of its raw mass on wrong-weight strings. Its randomization-test p-value is undefined without an extension rule; we report an envelope from two fully defined extensions. For any fixed outcome realization and fully defined randomization-test event, the difference between the p-value under a sampler and under the exact target is bounded by TV. In a simulation anchored on the observed covariates with simulated outcomes, the exact reference holds sharp-null rejection near 5%, a biased classical control inflates type-I error, and the two statistics respond differently to the same distortion. The study demonstrates a certification workflow and its inferential consequences on one empirical-anchor benchmark, with a synthetic extension that is pre-specified, hash-recorded, and frozen before the relevant execution; it was not externally preregistered. Primary inference in the extension is limited to equal-size Gaussian outcomes; binary, heterogeneous, and unequal-size analyses are sensitivity-only. The study does not establish quantum advantage, hardware performance, noise robustness, or clinical benefit.

1 Introduction

Cluster randomized trials assign whole units, such as counties, clinics, or schools, to arms. When the number of clusters is small and covariate balance matters, investigators often restrict randomization to a set of allocations that satisfy a balance criterion [21]. The allocation is then drawn uniformly from the accepted set, and the randomization distribution used for inference is that same set. Balance-based rerandomization is due to Morgan and Rubin [15]; its asymptotic theory is developed by Li, Ding and Rubin [12], and covariate adjustment under constrained randomization by Wang and Li [19]. Accelerated rerandomization software, FastRerandomize, handles larger constrained sets [8]. A 2026 preprint proposes a Metropolis–Hastings sampler for rerandomization and claims exact and asymptotic validity for that sampler [13]. These methods sample from balance-defined

sets; neither certifies the output law of an approximate sampler. Whenever the constrained search is performed by an approximate method, the sampler changes the randomization distribution, and so changes the inference that depends on it.

This paper asks one narrow question. If treatment assignments are generated from an exact constrained target P^* , what inferential error occurs when the randomization reference distribution is replaced by an approximate sampler’s distribution Q ? We answer it on a fixed empirical-anchor benchmark, extended by a set of synthetic instances that is pre-specified, hash-recorded, and frozen before the relevant execution. The benchmark uses real pre-randomization county-level covariates from the 16-county Colorado cluster-randomization example reported by Dickinson et al. (2015) [5]. No trial outcomes are analyzed. All outcomes in the inference study are simulated under pre-specified families, and the quantum results are ideal statevector-simulator results.

The contribution has three parts. First, we define an exact accepted-set target and a fixed certification workflow. Samplers are assessed by the exact total variation (TV) distance from P^* , with a close certificate and a far certificate defined by thresholds that are pre-specified, hash-recorded, and frozen before the relevant execution, and by full-statevector probabilities wherever a statevector is available. Second, we apply a standard inequality to show that for any fixed outcome realization and any fully defined randomization-test event, the difference between the sampler’s p-value and the exact p-value is bounded by TV. Third, we measure the downstream consequences on type-I error and power in a simulation anchored on the observed covariates. The contrast between the two statistics shows that the worst-case bound does not predict the size of the error for any particular test. The synthetic extension, pre-specified and hash-recorded as above, tests whether the same workflow behaves consistently across instances and across deployment and reference conditions; it is a robustness check, not a replacement for the anchor.

The paper does not claim quantum advantage. The Grover-style construction is a correctness control with a fixed accepted-set phase oracle, not a scalable reversible implementation or a resource estimate. The QAOA arms use a balance-score phase separator and bounded local optimization. Their role is to stress-test whether a balance-oriented sampler preserves the randomization law, and to show what certification reveals, including a negative result. Appendix A reports a separate, exploratory hardware run, not a certification study.

2 Exact covariate-constrained target

Let Ω_8 denote the set of $\binom{16}{8} = 12,870$ allocations of 8 treated and 8 control counties. Each county has covariates x_i from the frozen Dickinson input: `location`, `inciis`, `uptodateonimmunizations`, `hispanic`, and `incomecat`. Categorical variables are coded with the alphabetically first level as reference. Each covariate column is standardized using its sample mean and standard deviation over the 16 counties, and the ℓ_2 balance score of an allocation z is

$$\text{BL}(z) = \sum_j \left(\sum_{i \in T(z)} Z_{ij} \right)^2,$$

where $T(z)$ is the treated set and Z_{ij} the standardized covariate. This treated-sum score, with minimum, mean and maximum 1.161, 24 and 116.66, equals $16 \sum_j (\bar{Z}_{T,j} - \bar{Z}_{C,j})^2$ for an 8-versus-8 allocation: sixteen times the summed squared standardized treatment–control mean differences, because with column-centred Z each mean difference is $\sum_{i \in T(z)} Z_{ij} / 4$. This is an algebraic identity, not an expression computed by `cvcrand`.

We compute $\text{BL}(z)$ in exact rational arithmetic, using the raw integer covariates and the exact variances, and rank allocations in ascending order. Ties are broken by the canonical allocation index, which is the lexicographic order of the treated sets. The accepted set $\mathcal{A}$ consists of the first $m = \text{round}(12,870 \times 0.1) = 1,287$ allocations in that order. The target is the uniform distribution P^* on $\mathcal{A}$, so $P^*(z) = 1/1,287$ for $z \in \mathcal{A}$ and $P^*(z) = 0$ otherwise.

Exactly two allocations share the cutoff score. They are complementary: their treated sets partition the 16 counties. Because the standardized columns sum to zero, every allocation has the same score as its complement, so the tie is structural. The index rule therefore decides membership at the boundary. The tie-break is a stated definition, not a numerical accident. A floating-point ranking gives the same membership on this instance, but that agreement does not justify the float ranking as the definition. The canonical tie-break breaks complement closure only at this boundary pair, so county treatment probabilities are $643/1287$ or $644/1287$ rather than one half (largest deviation $1/2574 \approx 0.00039$). After the target was locked, we evaluated a complement-closed 1,288-state sensitivity target with 644 complement pairs and exact one-half county marginals. This analysis is post-lock. It is not the reported target, and the 1,287-state target is unchanged. Under it, the reported anchor laws move by at most 3.3×10^{-5} in raw TV (biased descent 0.958042 to 0.958075; XY headline 0.833403 to 0.833404).

We cross-checked the target against the `cvcrand` 0.1.1 package on the same frozen covariates, using its documented settings (metric ℓ_2 , cutoff 0.1, seed 12345). Allocation identifiers, design-matrix values, balance scores (within 2×10^{-14}) and membership agree, so the frozen treated-sum score matches `cvcrand`'s ℓ_2 calculation (no weights) for this fixed instance. This agreement is for this frozen instance only; we do not claim that `cvcrand` implements an explicit exact tie rule.

The covariates are observed pre-randomization county-level measurements from the Dickinson study [5], and they are the only empirical input to the target. The package also distributes a simulated individual-level outcome dataset. We do not use it. Every outcome in the inference study is generated under the models of Section 5, and the benchmark is therefore a controlled simulation anchored on empirical covariates, not an analysis of trial results.

Exact rational arithmetic is deliberate. Floating-point summation can reorder nearly equal scores, and the boundary pair is exactly the case where rounding could in principle reorder the tied allocations. Exact rational scores remove that ambiguity.

Table 1 summarizes the benchmark and each sampler definition.

Table 1: Benchmark target and sampler definitions. All quantities are defined on the single frozen Dickinson benchmark.

Element	Definition
Benchmark	16 Colorado counties from the example reported by Dickinson et al. (2015) [5] (cvcrand 0.1.1 <code>Dickinson_design</code>); 8 treated; 12,870 balanced allocations.
Covariates	location, inciis, uptodateonimmunizations, hispanic, incomecat; l_2 balance score; cutoff 0.1.
Target P^*	Uniform on the 1,287 accepted allocations (exact rational scores; canonical index tie-break); zero elsewhere.
Controls	Direct uniform draws from the accepted set; uncapped rejection sampling (acceptance 1/10).
Swap chain	Symmetric proposal over 64 treated–untreated swaps; accepted-state chain; one-time marginals at steps 0, 10, 41, 170, 1000. Correlated; not an IID sampler.
Biased control	Steepest descent on the balance score from uniform starts; 54 local minima; exact TV 137/143.
Phase 5	Ideal statevector; uniform Dicke initial state; fixed accepted-set phase oracle (correctness control, not a reversible circuit); $t^* = 2$.
Phase 6 XY	Ideal statevector; Dicke initial state; fixed ordered two-matching R_{XX}/R_{YY} mixer; balance-score phase separator; depth $p \leq 3$; bounded Nelder–Mead.
Phase 6 penalty	Ideal statevector; $ +\rangle^{\otimes 16}$ initial state; R_X mixer; quadratic cost $\text{BL}_{\text{all}} + \lambda(x - 8)^2$ with $B_{\text{scale}} = \overline{\text{BL}} = 24$ and $\lambda = 2B_{\text{scale}} = 48$; depth $p \leq 3$.
Outcome cells	Eight cells; 5,000 replicates each; $J = 16$, 8 treated, 50 individuals per cluster; baseline $0.5 \bar{x}$; simulated outcomes only.
Statistics	T1 unadjusted difference in cluster means; T2 OLS coefficient adjusted for standardized inciis; unstudentized; two-sided; ties count as extreme.

Exact targets and controls are exact by construction; all other distributions are numerical or Monte Carlo values, with their precision stated where they appear.

3 Certifying approximate sampling distributions

A sampler is characterized by its output distribution Q on the balanced domain. We assess Q by the total variation distance $\text{TV}(Q, P^*) = \frac{1}{2} \sum_{z \in \Omega_8} |Q(z) - P^*(z)|$. Because P^* has no mass outside $\mathcal{A}$, the sum splits as $\frac{1}{2} [\sum_{z \in \mathcal{A}} |Q(z) - 1/m| + Q(\Omega_8 \setminus \mathcal{A})]$. Aggregating all outside-accepted states into one bin therefore preserves the TV exactly. We report every distance on this 1,288-bin representation (1,287 accepted bins and one outside bin). When a sampler also produces strings of the wrong Hamming weight, we add their mass to the outside bin. The same identity then holds on the extended domain.

We use two certificate levels, both defined before any result was computed. A distribution is *certified close* when its exact TV is at most $\tau_{\text{close}} = 0.02$, and *far* when it is at least $\tau_{\text{far}} = 0.10$; between the thresholds the result is inconclusive. Full-statevector probabilities carry no sampling error and are labelled *exact*, meaning direct evaluation from the ideal statevector to numerical precision. Sampled distributions require a finite-sample bound. We use the distribution-free method-of-types inequality for the L1 deviation of an empirical distribution, as restated by Mardia et al. [14]. For

n IID draws over k bins and any radius ε ,

$$\Pr(\|\hat{P}_n - P\|_1 \geq \varepsilon) \leq (2^k - 2) e^{-n\varepsilon^2/2},$$

where the exponent uses the constant $\phi \geq 2$ of the cited source. The bound is distribution-free for IID sampling. With $k = 1,288$ bins and $\alpha = 0.05$, it gives under IID sampling the conservative confidence set $[\max\{0, \widehat{\text{TV}} - h\}, \min\{1, \widehat{\text{TV}} + h\}]$ for TV, with $h = \varepsilon/2$. The minimum shot budgets are 179,154 for $\varepsilon = 0.10$ and 4,478,847 for $\varepsilon = 0.02$; the calibration used the upward-rounded tiers 179,200 and 4,480,000. Sharper plug-in, bias-corrected, bootstrap, asymptotic and test-inversion intervals are not validated here. For sampled distributions, close means the upper endpoint of the TV confidence set is at most τ_{close} , and far means its lower endpoint is at least τ_{far} .

These budgets make the limits of finite-shot certification concrete. Under the confirmatory budget, the simple worst-case sufficient condition for closeness covers only $\delta_{\text{TV}} = 0$, because the interval half-width is 0.01 and $\delta_{\text{TV}} + 2h \leq \tau_{\text{close}}$ then requires $\delta_{\text{TV}} = 0$. The distribution-free bound proves coverage under IID sampling; the simulations show what the bound does in practice. They do not replace the proof.

The calibration of this bound is an analytic check, not a second proof. We applied the frozen verdict rules to synthetic distributions whose exact TV to P^* is known in closed form. The families are an exact target, leakage outside the accepted set, concentration on one accepted state, a structured within-set bias, a diffuse within-set bias, and a postselection distortion. Each family was evaluated on a grid of distances from 0 to 0.20 under the two shot budgets, with 10,000 replicates per cell.

Across all cells the observed coverage of the interval for TV was at or above the nominal level, and no verdict was reported in the wrong direction beyond Monte Carlo error. The screening budget never produced evidence-close, because its half-width exceeds τ_{close} . It produced evidence-far only at $\delta_{\text{TV}} = 0.20$, not at the far threshold itself. The confirmatory budget produced evidence-close at the exact target with probability 1.000 and at $\delta_{\text{TV}} = 0.005$ with probability 0.999, and never at $\delta_{\text{TV}} \geq 0.01$. These figures describe the power of the frozen rules under these budgets only.

Identity testing [17] is a different task from TV interval estimation. Three distinct tasks are separated here: evidence that a distribution is far, certification of closeness, and event-specific safety bounds. A close verdict certifies that the exact TV is at most τ_{close} at the stated confidence. The procedure does not convert non-rejection into a claim of closeness.

4 Classical and ideal quantum sampler constructions

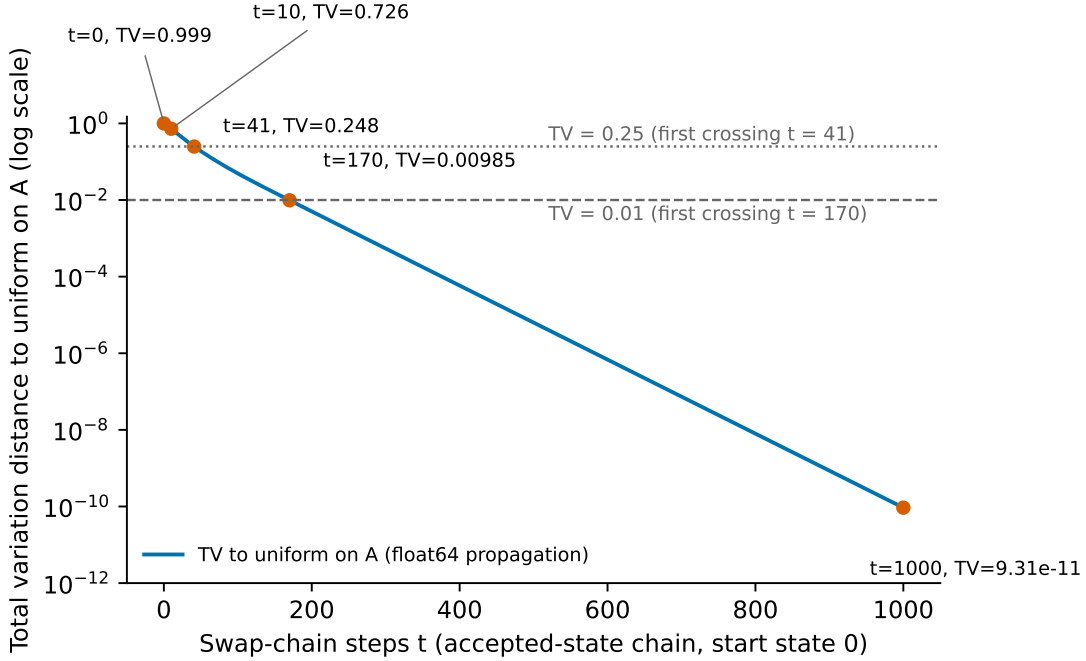
4.1 Classical controls and the swap chain

Direct uniform sampling from $\mathcal{A}$ has exact TV zero by construction and serves as the positive control. Ordinary rejection sampling draws uniform balanced allocations and accepts those in $\mathcal{A}$, so each attempt succeeds with probability $1/10$. Its uncapped output is exactly uniform on $\mathcal{A}$. A capped variant reports failures as a separate outcome. Its raw distribution has exact TV equal to the failure mass, and its conditional distribution after discarding failures is exactly uniform.

A swap chain on the accepted set uses a symmetric proposal. A state is a treated set, and a proposal exchanges one treated county with one untreated county, chosen uniformly among the $8 \times 8 = 64$ ordered pairs. A proposal that leaves $\mathcal{A}$ is rejected and the chain stays in place. The accepted-state graph is connected, forming one component of 1,287 states, and symmetric. Self-loop counts range

from 41 to 57, so the chain is aperiodic. The integer transition matrix is exactly symmetric with row sums 64, so its stationary law is exactly uniform on $\mathcal{A}$. Finite-time distributions are computed by float64 propagation of this matrix. We report them to numerical precision, and we do not treat chain output as IID.

The spectral gap of the chain is 0.022003. From the start state, the TV to uniform first falls below 0.25 at step 41 and below 0.01 at step 170, and reaches the nonzero numerical value 9.3×10^{-11} at step 1,000 (Figure 1). These are float64 values, not exact zeros. In the inference study the chain enters only through its one-time marginals at the fixed steps.



Stationary law exact (integer matrix); finite-time values are numerical float64 propagation.

Figure 1: Swap-chain total variation to uniform on the accepted set, from the exact integer transition matrix. Markers at steps 0, 10, 41, 170, and 1,000 show the reported values. Stationary-law properties are exact. Finite-time values are numerical float64 propagation of that matrix and are not IID samples. Horizontal lines mark $TV = 0.25$ (first crossing at step 41) and $TV = 0.01$ (first crossing at step 170).

A deliberately biased classical control is defined before any result is seen. It starts from a uniform random allocation and performs steepest descent on the balance score over the same 64 swaps, stopping at a local minimum. Its exact output distribution is the basin distribution over all 12,870 starts. It places no mass outside $\mathcal{A}$, but its outputs concentrate on 54 distinct local minima, and its exact TV to P^* is $137/143 \approx 0.958$.

The one-time marginals used in the inference study are fixed-time distributions of a correlated chain, not IID samples.

4.2 Ideal statevector amplification control

The amplification control starts from the uniform superposition over the 12,870 balanced states, which is the uniform Dicke state $D(16, 8)$ in the idealized form. It applies a fixed phase oracle that multiplies the amplitude of each of the 1,287 accepted states by -1 , followed by the reflection about the initial state. This is a fixed-oracle correctness control for this benchmark. It is not a scalable circuit or a resource estimate, and the construction follows amplitude amplification [3].

The Grover angle is $\theta = \arcsin \sqrt{1287/12870} = 0.321751$. The iteration count $t^* = 2$ was fixed before the sweep by the rule in Supplementary Table S2 of the separate supplement, and predicts acceptance $\sin^2(5\theta) = 0.99856$. At $t^* = 2$ the raw distribution has accepted mass 0.998560 and full-statevector TV to P^* of 0.001440, certified close as exact. Its conditional distribution on $\mathcal{A}$ has numerically zero TV. Amplification preserves equal relative amplitudes among accepted states but does not remove the residual raw leakage. Off-count raw TV values (0.90 at $t = 0$, 0.32 at $t = 1$, 0.40 at $t = 3$, 0.94 at $t = 4$) are diagnostic and did not change t^* .

4.3 QAOA arms

Two prospective QAOA arms were fixed in a protocol amendment before optimization. The balance-energy QAOA family studied here is intentionally not optimized to reproduce the uniform accepted-set target. It is a stress test of whether a plausible balance-oriented sampling procedure preserves the required randomization law. The ansatz follows the QAOA framework of Farhi et al. [7] and the alternating-operator and constrained-mixer framework of Hadfield et al. [9]. Both use depth $p \in \{1, 2, 3\}$, a balance-score phase separator, and a mixer. The phase separator is a quadratic pseudo-Boolean cost in the 16 binary variables. It is compiled into single-qubit RZ and two-qubit RZZ rotations, and it is normalized affinely to $[0, 1]$ using constants computed by exact enumeration before any optimization.

The first arm is Hamming-weight-preserving, using the XY mixing construction of Wang et al. [20]. It starts from the uniform Dicke state $D(16, 8)$. Its mixer is a fixed ordered sequence over the 16-cycle: all eight edges of matching E_0 , then all eight edges of matching E_1 , each edge applying $R_{XX}(\beta)$ then $R_{YY}(\beta)$. The sequence is the definition of the mixer. It is not claimed to equal continuous-time evolution under the summed ring Hamiltonian. Weight preservation was verified numerically at predetermined points and at every objective evaluation.

The second arm is a penalty control. It starts from $|+\rangle^{\otimes 16}$ and uses RX mixers. Its cost is the quadratic extension of the balance score to all 2^{16} bitstrings, $\text{BL}_{\text{all}}(x) + \lambda(|x| - 8)^2$, with penalty weight $\lambda = 2B_{\text{scale}}$ fixed in advance (defined below).

Each arm was optimized with Nelder–Mead under box bounds, with $\gamma_l \in [0, \pi]$ and $\beta_l \in [0, \pi/2]$, a 300-evaluation budget per start, and 10 starts per depth from fixed seeds. We selected the run with the lowest final normalized energy, with ties broken by lower depth and seed. Selection never used TV. This is bounded local search within the stated box, not global optimization. Across 120 runs with diagnostic penalty weights excluded from selection, 88 stopped by stagnation and 32 exhausted the budget. Ninety of the 120 runs were flagged by a default-NumPy `isclose` boundary diagnostic, which is a numerical diagnostic and not exact boundary equality; the tolerance arithmetic is in Supplementary Table S3. A strict absolute re-audit disagrees on one run. We report the count descriptively and do not interpret it as optimality or failure. Supplementary Table S1 summarizes the 120 saved runs. We make no claim about other optimizers, a target-aligned ansatz, or QAOA in general; optimizer strategies are studied elsewhere [23].

Table 2 reports the selected runs. The XY-ring headline has raw TV 0.833 and accepted mass 0.168. Its conditional TV is 0.498, which is still far. The penalty headline has raw TV 0.948, accepted mass 0.052, and wrong-weight mass 0.512. Its conditional TV is 0.0024. That conditional value is a postselected diagnostic and does not validate the raw sampler. For the penalty output, the 51% wrong-weight mass means the raw distribution is not a valid 8-versus-8 randomization distribution.

Table 2: Selected quantum-arm results. Raw distributions are the primary sampler outputs over all 1,288 bins. Conditional values are postselected diagnostics (accepted bins renormalized); they do not establish validity of the raw sampler. Quantum probabilities are evaluated from the full ideal statevector to numerical precision. “Certified” means that the raw TV is at most 0.02. Mass decomposition: A is the accepted set; $\Omega_8 \setminus A$ is the set of weight-8 allocations not accepted; $|z| \neq 8$ denotes wrong-weight strings. The three masses sum to one.

Sampler (selected)	Raw TV to P^*	Certified (≤ 0.02)	Conditional TV (postselected)	Mass decomposition
Phase amplification, $t^* = 2$	5 0.001440	True	6.98e-17 (postselected)	$P(A) = 0.998560$; $P(\Omega_8 \setminus A) = 0.001440$; $P(z \neq 8) = 0$
XY-QAOA, $p = 3$, seed 7	0.833403	False	0.497558 (postselected)	$P(A) = 0.168097$; $P(\Omega_8 \setminus A) = 0.831903$; $P(z \neq 8) = 0.000000$
Penalty-QAOA, $p = 3$, seed 3	0.948094	False	0.002420 (postselected)	$P(A) = 0.051906$; $P(\Omega_8 \setminus A) = 0.436401$; $P(z \neq 8) = 0.511693$

The phase separators are normalized affinely so that the cost takes values in $[0, 1]$ over the relevant domain. For the XY arm the normalization uses $\text{BL}_{\min} = 1.161$ and $\text{BL}_{\max} = 116.66$ on the weight-8 subspace. For the penalty arm the normalization uses the minimum and maximum of the extended cost over all 2^{16} bitstrings. The maximum is 3,072, attained at the empty string, and not the shortcut value $\text{BL}_{\max} + 64\lambda \approx 3,189$. The constants and their hashes were frozen before any optimization.

The scale $B_{\text{scale}} = \overline{\text{BL}} = 24$ is the mean balance score over the 12,870 weight-8 strings, defined before any coefficient. The penalty weight $\lambda = 2B_{\text{scale}} = 48$ is the headline value. Two diagnostic weights, $0.5B_{\text{scale}} = 12$ and $8B_{\text{scale}} = 192$, were run with the same procedure. They are fixed multiples of B_{scale} , not functions of any individual $\text{BL}(x)$. They are reported as sensitivity outcomes and never enter selection. For each configuration the minimum cost over wrong-weight strings exceeds the best weight-8 cost, so the penalty does make weight 8 the cost minimum. It does not make the output weight 8, because the amplitudes evolve under the mixer and the cost is not the output distribution.

5 Simulation design for randomization inference

The inference study asks what happens to a randomization test when its reference is replaced by a sampler. Within each replicate we draw the realized allocation Z_{obs} from the exact target P^* . Using the cluster-level model below, we generate outcomes once. We then compute a two-sided randomization p-value using P^* and the same p-value using each approximate reference, on the

same outcomes and the same observed allocation. Sampling Z_{obs} from the approximate reference would mix two sources of error, so we do not do it in the primary validity experiment.

Outcomes are simulated and anchored on the observed covariates. The baseline cluster mean is $\bar{Y}_i(0) = 0.5 \tilde{x}_i + a_i + \bar{e}_i$, where $\tilde{x}_i$ is the standardized observed `inciis` for county i , $a_i \sim N(0, \rho)$ is a cluster effect, and $\bar{e}_i \sim N(0, (1 - \rho)/50)$ is the mean of 50 independent individual errors. The individual-level residual (cluster effect plus individual error) has variance 1 and intra-cluster correlation ρ ; the cluster-mean residual has variance $\rho + (1 - \rho)/50$, which is 0.069 (SD 0.263) at $\rho = 0.05$. The shifts $\tau \in \{0.25, 0.50\}$ are raw cluster-mean values, so $\tau = 0.50$ is a half-SD shift relative to the individual-level SD of 1. Under an alternative, $\bar{Y}_i(1) = \bar{Y}_i(0) + \tau_i$. The study uses eight cells. The base cells use intra-cluster correlation $\rho = 0.05$: a sharp null, constant shifts $\tau = 0.25$ and $\tau = 0.50$, and a heterogeneous shift $\tau_i = 0.50(1 + 0.5 \tilde{x}_i)$. Additional sharp-null and constant-shift cells vary ρ over $\{0.00, 0.20\}$. Each cell has 5,000 replicates, and each cell has its own seed rule.

Two statistics are computed on cluster means and are not studentized. The first, T1, is the unadjusted difference in means between treated and control clusters. The second, T2, is the coefficient on treatment in an ordinary least-squares regression of the cluster means on treatment and $\tilde{x}$. Both are two-sided. A tie counts as extreme, and the decision threshold is $\alpha = 0.05$.

The reference distributions are the exact target; two exact controls, direct uniform and uncapped rejection, which must reproduce the target exactly; swap-chain one-time marginals at steps 0, 10, 41, 170, and 1,000; the biased descent control; the Phase 5 raw output at $t^* = 2$; and the two raw QAOA headline outputs. Conditional distributions appear only as labelled diagnostics.

For the penalty output, wrong-weight mass prevents a single balanced test statistic. The randomization p-value is undefined until an extension rule is chosen for those strings, so we bound it with two fully defined extensions: the lower counts every wrong-weight string as non-extreme, and the upper counts every such string as extreme. The two p-values form an envelope. We report guaranteed and possible rejection rates, the rates at which the envelope lies wholly on one side of α or touches the rejection region. The true rejection rate under any extension lies between these two rates. Neither endpoint is the type-I error or the power of the penalty output.

Monte Carlo uncertainty is reported with Wilson 95% intervals. At 5,000 replicates, a rate near 0.05 has half-width about 0.6 percentage points, and a rate near 0.5 has half-width about 1.4 points.

Each replicate draws one Z_{obs} and one outcome vector; the statistic vectors over all 12,870 allocations are computed once and shared by every reference. T2 uses the residualized representation with respect to the intercept and $\tilde{x}$, which is algebraically equal to the ordinary least-squares coefficient. Its agreement with direct least squares on fixed test outcomes has maximum difference 1.11×10^{-15} in the frozen preconditions record. The p-value under each reference sums its weights over allocations at least as extreme as the observed one, with absolute tolerance 10^{-10} for exact ties.

The exact controls must reproduce P^* exactly. The direct-uniform and uncapped-rejection references were checked for exact equality of p-values with the target on a fixed set of synthetic outcomes, one per cell, using exact rational arithmetic. The Phase 5 and Phase 6 distributions were regenerated from their immutable parameter vectors, and their full-statevector TV values and certified flags agreed with the saved records to within 3.4×10^{-15} . The swap-chain marginals were regenerated from the exact integer matrix and reproduced the first crossings at steps 41 and 170. None occurred.

The cluster-level model is a deliberate simplification. Each cluster mean is Gaussian, with a cluster effect whose variance is ρ and an error whose variance is $(1 - \rho)/50$, so the design effect enters only through the intra-cluster correlation. We do not model unequal cluster sizes, individual-level covariates, or non-Gaussian outcomes. Those features would change the variance of the statistics, but they are not needed to isolate the effect of the reference distribution, which is the question studied here.

6 Results

6.1 Certification of the sampler distributions

Table 2 and Figure 2 show the selected outputs. The amplification control at $t^* = 2$ is certified close on raw leakage. The two QAOA headline outputs are far on the raw distribution. Their conditional distributions are shown in the same figure, labelled as postselected diagnostics. The conditional value of the penalty output, 0.0024, lies below the close threshold. It is a property of the accepted part of the output only, and does not show that the raw penalty sampler is valid.

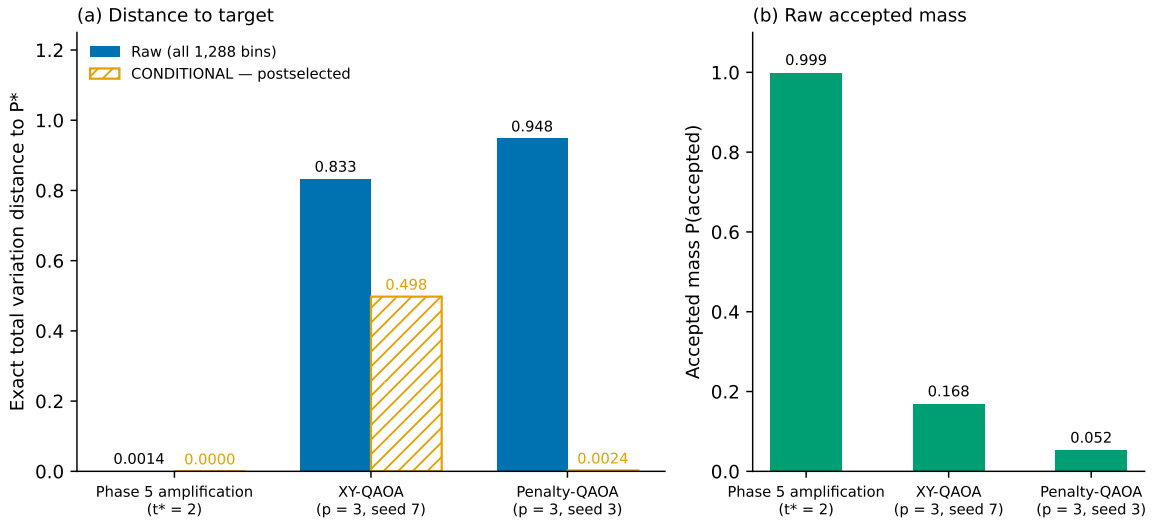


Figure 2: Raw versus conditional outputs for the amplification control at $t^* = 2$ and the two selected QAOA headline runs. Panel (a) shows total variation to P^* on the 1,288-bin representation (exact for classical references; full-statevector values for quantum references). Solid bars are raw distributions, which are the primary sampler outputs. Hatched bars are conditional values, labelled “postselected”: they use only accepted bins and renormalize them. Conditional values are postselected diagnostics and do not establish validity of the corresponding raw sampler. Panel (b) shows the raw accepted mass. The penalty output has wrong-weight mass 0.512, which is not shown in panel (b).

6.2 Downstream inference

Tables 3 and 4 give the downstream results. Under the sharp null in the base cell, the exact target rejects at 5.20% under T1 and 4.40% under T2. The swap-chain marginal at step 170 shows no

disagreement with the target at the displayed Monte Carlo resolution, the step-41 marginal shows small disagreement (1.28% under T1 and 0.64% under T2 at $\tau = 0.50$), and the step-10 marginal, which still carries substantial distance, shows larger disagreement (10.5% under T1 and 5.5% under T2). The biased control inflates type-I error under T1 to 16.68%, with interval 15.67–17.74%. In the cell with $\rho = 0.00$, the T1 rate rises to 29.56%.

Table 3: Downstream results, panel (a): sharp-null type-I rejection rates at $\alpha = 0.05$ in the base cell, for the exact target, the swap-chain marginals, the biased control, and the quantum references. Brackets give Wilson 95% Monte Carlo intervals from 5,000 replicates per cell. Penalty-QAOA entries (D8) are bounds, not point rates: rejection or disagreement under any fully defined extension of the event to wrong-weight strings lies between the two stated values. The quantum-reference distributions are evaluated from full ideal statevectors; the displayed downstream rates and Wilson intervals are Monte Carlo estimates from 5,000 simulated-outcome replicates. Descriptive only; no regression or causal reading.

Reference	Raw TV	Type-I T1 (F0) % [CI]	Type-I T2 (F0) % [CI]
Exact target P^*	0.000000	5.20 [4.62–5.85]	4.40 [3.87–5.00]
Swap, $t = 10$	0.725991	6.22 [5.58–6.92]	7.06 [6.38–7.80]
Swap, $t = 41$	0.248146	5.24 [4.66–5.89]	4.46 [3.92–5.07]
Swap, $t = 170$	0.009850	5.20 [4.62–5.85]	4.40 [3.87–5.00]
Biased descent control	0.958042	16.68 [15.67–17.74]	6.10 [5.47–6.80]
Phase 5 raw, $t^* = 2$	0.001440	5.12 [4.54–5.77]	4.40 [3.87–5.00]
XY-QAOA raw	0.833403	0.18 [0.09–0.34]	4.02 [3.51–4.60]
Penalty-QAOA raw (D8), bounds	0.948094	[0.00, 0.86]	[0.00, 9.16]

Table 4: Downstream results, panel (b): decision disagreement with P^* at $\alpha = 0.05$ for constant shift $\tau = 0.50$ in the base cell. Brackets give Wilson 95% Monte Carlo intervals from 5,000 replicates per cell. Penalty-QAOA entries (D8) are bounds, not point rates: rejection or disagreement under any fully defined extension of the event to wrong-weight strings lies between the two stated values. The quantum-reference distributions are evaluated from full ideal statevectors; the displayed downstream rates and Wilson intervals are Monte Carlo estimates from 5,000 simulated-outcome replicates. The exact target’s disagreement with itself is 0 by construction, so no interval is shown. Descriptive only; no regression or causal reading.

Reference	Raw TV	Disagreement (F1b) % [CI]	T1	Disagreement (F1b) % [CI]	T2
Exact target P^*	0.000000	0 (by construction)		0 (by construction)	
Swap, $t = 10$	0.725991	10.50 [9.68–11.38]		5.50 [4.90–6.17]	
Swap, $t = 41$	0.248146	1.28 [1.00–1.63]		0.64 [0.45–0.90]	
Swap, $t = 170$	0.009850	0.00 [0.00–0.08]		0.00 [0.00–0.08]	
Biased descent control	0.958042	14.14 [13.20–15.13]		8.76 [8.01–9.58]	
Phase 5 raw, $t^* = 2$	0.001440	0.04 [0.01–0.15]		0.00 [0.00–0.08]	
XY-QAOA raw	0.833403	37.12 [35.79–38.47]		2.86 [2.43–3.36]	
Penalty-QAOA raw (D8), bounds	0.948094	[17.70, 68.58]		[0.02, 97.02]	

The XY headline has raw TV 0.833. Under T1 it disagrees with the target on 37.12% of replicates at $\tau = 0.50$, while its sharp-null T1 type-I rate is 0.18%, so it is conservative and distorted at once. Under T2 the same distribution disagrees on 2.86% of replicates and has sharp-null type-I rate 4.02% (Wilson 3.51–4.60%), closer to the exact-target rate (4.40%) than T1 is, but below the nominal 5% level. The two statistics therefore respond very differently to one distribution, even though a fixed TV of 0.833 bounds the error of every event. The Phase 5 control, with raw TV 0.0014, shows a disagreement of 0.04% under T1.

For the penalty output, Tables 3 and 4 report bounds. Under the sharp null in the base cell, the possible rejection rate is 0.86% under T1 and 9.16% under T2, and the definite rejection rate is zero in both cases. At $\tau = 0.50$, guaranteed disagreement is 17.70% under T1 and possible disagreement is 68.58%. Under T2 the bounds are 0.02% and 97.02%. These intervals are wide. They are the honest description of a sampler whose wrong-weight mass prevents a single p-value.

Figure 3 plots raw TV against downstream disagreement and sharp-null type-I behaviour across the references. The figure is descriptive. It shows where the reported references fall. It does not establish a regression, a monotone relationship, or a causal effect of distance on inference.

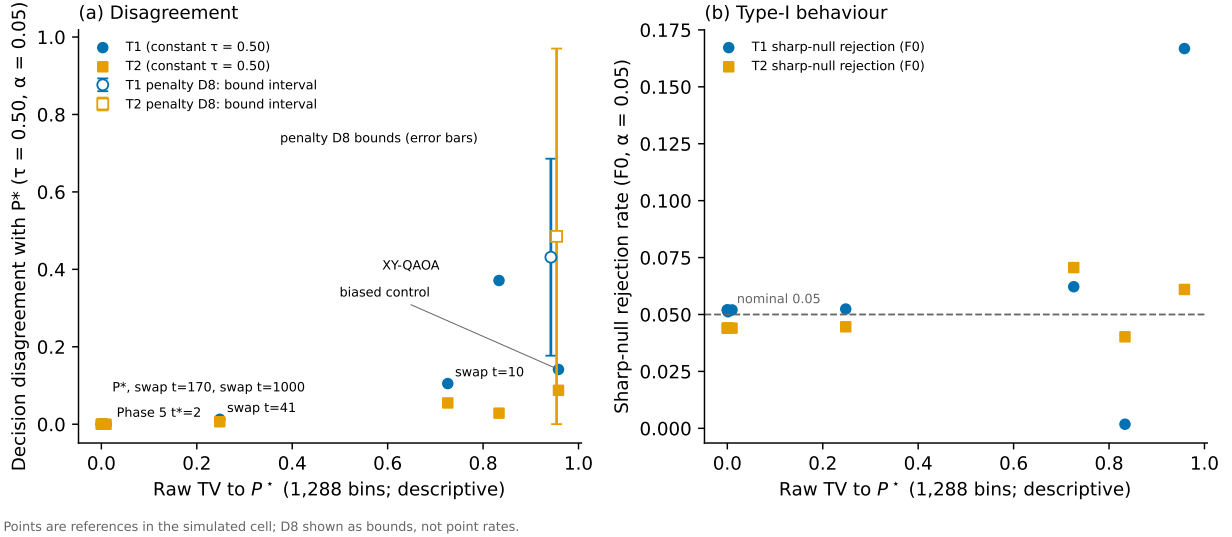


Figure 3: Raw total variation to P^* against downstream behaviour, for the references in the base cell with constant shift $\tau = 0.50$. Panel (a) shows decision disagreement with P^* at $\alpha = 0.05$ for T1 and T2. The penalty output is shown as a bound interval, not a point rate. Panel (b) shows sharp-null rejection rates for the same references, with the nominal level marked. This figure is descriptive only. It shows where the references fall and does not establish a regression, a monotone relationship, or a causal effect of total variation on inference.

The biased descent control inflates the null rate under T1 to 16.7% and under T2 to 6.1% in the base cell, so its distortion is visible under both statistics but differs in size. Neither this nor the XY-headline pattern above generalizes beyond the simulated families and the benchmark.

The swap marginal at step 10 has smaller TV than the XY headline, yet under T2 it shows larger disagreement (5.50% against 2.86%). We therefore make no regression, slope, causal, or general TV-to-inference reading of the figure.

6.3 Synthetic multi-instance extension

The synthetic extension tests whether the workflow behaves consistently across constrained-randomization targets that differ in size, covariate strength, and law. Eight synthetic instances use $J \in \{12, 16\}$ clusters, covariate-strength settings $\kappa_{\text{cov}} \in \{0.2, 0.6\}$, and two seeds per setting. Their covariates mirror the anchor’s structure: three continuous covariates on the anchor’s integer ranges, a balanced binary covariate, and a three-level categorical variable with dummy coding. They do not reproduce the anchor’s joint distribution. Targets are complement-closed, so each cluster has treatment probability one half, and every accepted set has at least 100 allocations.

All eight instances carry the classical laws of Section 4, including Boltzmann laws calibrated to expected normalized costs 0.02, 0.05, and 0.10. On four instances we also ran the XY-QAOA arm, with ten deterministic starts at each depth $p \in \{1, 2, 3\}$ and 300 evaluations per start. This ideal statevector law is far from the target in every case (raw TV 0.59 to 0.72; Table 5), and its postselected conditional TV is diagnostic only. We make no claim of superiority over any classical law, and no claim about speed or hardware.

Table 5: Synthetic multi-instance extension (pre-specified, hash-recorded, and frozen before execution; not externally preregistered; four instances with the ideal-statevector XY-QAOA sampler). Raw TV is the full-statevector raw TV of the sampler’s output law to P^* on the complement-closed target. The certificate status is “far” when raw TV is at least $\tau_{\text{far}} = 0.10$. Only sampler-relevant quantities are shown; exact-target design power is in Table 6.

Instance	κ_{cov}	J	$ \mathcal{A} $	Raw TV	Status
S12-k02-a	0.2	12	100	0.592	far
S12-k06-a	0.6	12	100	0.613	far
S16-k02-a	0.2	16	1,288	0.595	far
S16-k06-a	0.6	16	1,288	0.722	far

Inference uses 10,000 replicates for the final primary cells and 5,000 for screening cells. The primary family is equal-size Gaussian cluster means ($n_i = 50$) at intra-cluster correlation 0.05 or 0.20, under the sharp null and under a constant shift of 0.5. Binary equal-size, heterogeneous Gaussian, and Gaussian unequal-size analyses are sensitivity-only; they are inventoried in Supplementary Table S4 and do not alter sampler selection or primary-cell designation. Power cells use the same replicate counts as their sharp-null counterparts, a rule fixed before outcomes were generated.

Figure 4 plots, for each law and instance, the sharp-null T1 rejection rate at P^* against the raw TV of the law involved. The left panel shows assignment from P^* with a Q reference, and the right panel shows assignment from Q with a P^* reference. The display is descriptive. It does not fit a regression or a threshold, and it does not imply a causal relation. Within each panel, laws with similar raw TV can produce rejection rates on either side of the nominal level, so raw TV alone does not determine the rate.

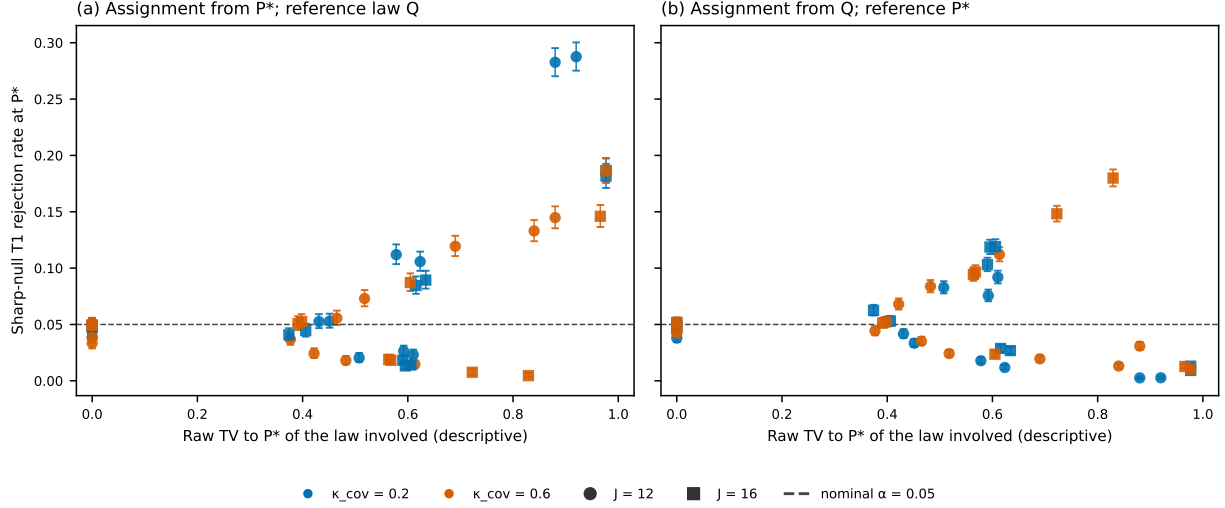


Figure 4: Descriptive display of the synthetic extension. Each point is one eligible law on one synthetic instance, for the equal-size Gaussian primary family at ICC 0.05 under the sharp null. Left: assignment from P^* with reference law Q ; the x -axis is the raw TV of Q to P^* . Right: assignment from Q with reference P^* ; the x -axis is the raw TV of the assignment law to P^* . Error bars show Wilson 95% intervals. The dashed line marks the nominal level $\alpha = 0.05$. The exact controls coincide with the baseline and are omitted. The display is descriptive and is not a regression, a threshold, or a causal relation.

Under the sharp null, the fixed-event bound of Section 7 holds in all 176 primary cells examined, and it is the only bound applied to these results. Table 5 reports the four XY sampler instances. Their exact-target design power is in Table 6, which is not a sampler result. In every row there T2 rejects more often than T1 at the same replicate count, the same pattern as in the anchor.

Table 6: Target-design power under P^* on the four synthetic instances of Table 5. Entries are sharp-alternative rejection counts out of 10,000 for T1 and T2 with assignment and reference both P^* , under a constant shift $\tau = 0.5$, equal-size Gaussian outcomes. These are properties of the exact-target design and do not depend on any sampler; they are not XY-QAOA results.

Instance	J	$ \mathcal{A} $	ICC 0.05		ICC 0.20	
			T1	T2	T1	T2
S12-k02-a	12	100	3,385	6,878	2,029	2,784
S12-k06-a	12	100	4,109	7,090	2,345	2,924
S16-k02-a	16	1,288	6,451	9,249	3,845	4,841
S16-k06-a	16	1,288	6,710	9,265	4,031	4,918

7 Operational bridge to randomization-test events

Fix the observed outcomes Y_{obs} and the observed allocation Z_{obs} . Let T be a statistic, and let $\mathcal{E} = \{z : |T(z; Y_{\text{obs}})| \geq |T(Z_{\text{obs}}; Y_{\text{obs}})| - 10^{-10}\}$ be the two-sided extremeness event. Once the data are fixed, $\mathcal{E}$ is a fully defined subset of the domain, and it does not depend on which reference is

used. For any probability distribution Q on the domain, the randomization p-value is $Q(\mathcal{E})$.

Proposition 1. *For any finite domain, any probability distributions Q and P , and any event $\mathcal{E}$, $|Q(\mathcal{E}) - P(\mathcal{E})| \leq \text{TV}(Q, P)$.*

Proof. Let $D = Q - P$, so $\sum D = 0$. Then $\sum D^+ = \sum D^- = \frac{1}{2} \sum |D| = \text{TV}(Q, P)$. Hence $Q(\mathcal{E}) - P(\mathcal{E}) = \sum_{\mathcal{E}} D \leq \sum_{\mathcal{E}} D^+ \leq \sum D^+ = \text{TV}(Q, P)$. Exchanging Q and P gives the lower bound. $\square$

This inequality is standard. We use it as an operational bridge from distributional distance to event-specific p-value error, not as a new theorem.

Applied to a randomization test, the proposition gives, for every replicate and every reference, $|p_Q - p_{P^*}| \leq \text{TV}(Q, P^*)$. A close certificate therefore bounds the worst-case p-value error over all fully defined events at the certified level. A far certificate does not imply large error for every event. Total variation is a supremum over events, so a particular statistic may show a much smaller error than the bound.

The 1,288-bin representation preserves TV because P^* has no mass outside $\mathcal{A}$, as shown in Section 3. For the penalty output, the extended domain adds the wrong-weight strings. The bound applies separately to each of the two extensions. Let $\mathcal{E}_{\text{lo}}$ denote the event with wrong-weight strings counted as non-extreme, and $\mathcal{E}_{\text{hi}}$ the event with them counted as extreme. Each is a fully defined event on the extended domain, and the target assigns zero probability to the wrong-weight strings. So $|p_{\text{lo}} - p_{P^*}| \leq \text{TV}_{\text{ext}}$ and $|p_{\text{hi}} - p_{P^*}| \leq \text{TV}_{\text{ext}}$, where $\text{TV}_{\text{ext}} = 0.948$ is the extended-domain TV. We checked both inequalities against all saved replicates, for both statistics and all eight cells. The audit found no failures. The global maximum observed difference is 0.998446, attained by the step-0 swap reference under both T1 and T2 in cells F3c and F3d, against a reference TV of 0.999223. Smaller per-reference maxima are 0.388 for the XY-QAOA reference under T1 and 0.543 for the step-10 swap marginal under T2. No difference exceeds its reference TV. The audit confirms the arithmetic of the simulation. It does not improve the bound.

The two extensions give the envelope in the results because no extension is scientifically privileged. Choosing one would make a substantive assumption about invalid assignments that the data cannot settle.

8 Discussion and limitations

The main methodological point is that a sampler’s distance from its target controls the worst case of randomization-based inference, but the realized error depends on the event being tested. A certificate should therefore be reported together with the downstream behaviour of the specific statistics that a trial would use. Disagreement under power and type-I error under the null are different quantities, and each must be reported.

The study has several limits. Its scope is one empirical-covariate anchor with 16 clusters; eight small synthetic instances, four of them ideal-statevector XY-QAOA instances; primary equal-size Gaussian inference, with binary, heterogeneous-effect and unequal-size analyses as sensitivity only; one exploratory raw-hardware campaign; and no computational-effort-matched classical-versus-quantum scaling comparison. Its outcomes are simulated, and the cluster-level model with its variance components is our specification. Its quantum results are ideal statevector simulations. The penalty

output is not a valid 8-versus-8 randomization distribution, and its inference is bounded rather than point-estimated. The selected headline QAOA runs are one run per arm, chosen by energy under bounded local search. The results therefore describe these runs and do not establish generic QAOA performance. Most runs were flagged by the default-NumPy boundary diagnostic described in Section 4; it limits any reading of optimizer quality.

The amplification result is a correctness control. Its phase oracle is a fixed diagonal operator for this benchmark, not a reversible membership circuit, so we report no resource estimate. Dicke-state preparation has a published construction with gate count, depth, and ancilla bounds [1], with short-depth variants [2], but we make no claim about the circuits used here. Statevector results are full-amplitude computations for the ideal model only.

The finite-shot sets are conservative. Under IID sampling and the frozen confirmatory budget, the worst-case sufficient condition for closeness covers only the exact target, although calibration observed evidence-close at $\delta_{TV} = 0.005$. Sharper plug-in, bias-corrected, bootstrap, asymptotic and test-inversion intervals are not validated here, and we give no coverage for them. Far detection, closeness certification and event-specific safety bounds are different tasks. Only the lumped interval of Section 3 supports closeness; the hardware appendix gives far-type event lower bounds and no closeness claim; and the fixed-event bound of Section 7 is a third, event-specific kind. Identity testing against a known target [16] and closeness testing of two unknown distributions [4] are further distinct tasks.

Several claims are outside scope. The paper makes no claim of quantum advantage or speedup, since no scaling study or matched classical comparison was performed, and no claim of noise robustness, scalable oracle cost, clinical benefit or general QAOA performance. Its randomization-based statements apply to the stated simulation families and extension rule. The audit of the interrupted synthetic replicate run supports the saved stream outputs but cannot recover its source hash. Assignment from Q with reference P^* is not a valid design for the constrained target, and internal coherence under the sharp null is not evidence that the design was restored.

Future work would vary the outcome models, the cluster count, and the covariate structure, and would add noise models and a matched classical search comparison. Those extensions are separate studies with their own pre-specification, hash recording and freezing before execution. Quantum-enhanced optimization for patient stratification has been proposed with claimed efficiency and significance gains [6]; that work addresses optimization of covariate stratification and does not certify randomization laws, and its claims are not evaluated here. A different, adaptive preparation model uses mid-circuit measurement, feedback, collective measurements with global rotations, and ancilla resources [22]; it is not a demonstrated replacement for the static, no-ancilla IBM workflow used here.

Three questions apply before an approximate sampler replaces an exact constrained target. What is its exact distance from the target on the output space the randomization test uses? Are invalid outputs, such as wrong-weight strings, excluded, penalized or counted, and under which rule? Which test statistics will the trial use, since the worst-case bound does not say how far a particular statistic moves? The paper answers these for one benchmark.

The results also suggest a reporting standard. Raw distributions should be certified on their own output space, with conditional or postselected quantities reported separately and labelled as such. A close certificate for a conditional distribution does not license the raw sampler, and a far certificate for a raw distribution does not establish the size of the error for any particular test. Both facts must appear in the same report.

9 Reproducibility and availability

All numerical results are generated from frozen scripts, pinned environments, and saved machine-readable artifacts. The protocol, its amendments, and the exact seed rules are recorded in the project files and the prepared release bundle. The protocols and the synthetic extension were internally time-ordered and hash-recorded before execution; they were not externally registered, and no registry identifier exists. A clean regeneration reproduced the 44 synthetic-extension streams that predate an interrupted run bit-for-bit; source hashes were not recorded at that launch, so its code state cannot be proven by hash. Ideal statevector computations use Qiskit [11], and local search uses SciPy [18]. The randomization test follows Imbens and Rubin [10].

Hashed manifests record SHA-256 hashes for the protocol, code, preconditions, aggregate tables, per-replicate results, theorem audit and test record; the manifest families are listed in Supplementary Table S5. The hardware submission record exists in two forms. The original contains account-identifying metadata and is private; the public manifest is a redacted release that records the hash of the original. A redacted release bundle, excluding raw hardware counts, has been prepared and independently verified locally. It is not yet published; a persistent archive link and DOI will be added only after external deposit.

Appendix A: Exploratory hardware run

We ran one exploratory hardware campaign on one IBM device, `ibm_marrakesh`. It used four frozen circuit types: two readout baselines (all-zero and all-one preparations), the Dicke preparation D1 for $c = 2$, and the complete QAOA circuit Q1 with the frozen $p = 1$, seed-0 parameters. Each circuit was run as two separate 2,048-shot jobs, eight jobs in total, and all completed. Raw counts were analysed without mitigation, readout correction or postselection.

The frozen protocol specified the lower bound through a pooled upper Wilson 95% bound on the event proportion. The reported bounds instead use replicate-specific one-sided exact Clopper–Pearson bounds with Bonferroni allocation: four tests (two replicates by two events), familywise level 0.05, and per-test level 0.0125. This replicate-specific analysis was a documented post-protocol exploratory deviation and is not a preregistered analysis. For Q1, the event lower bounds on $\text{TV}(Q, P^*)$ are 0.9794 and 0.9835 for the accepted event, and 0.8572 and 0.8546 for the weight-8 event, for replicates 1 and 2. They use event probabilities only. They are not full-distribution TV estimates and not closeness certificates.

Observed accepted mass for Q1 was 0.0137 and 0.0103, against an ideal value of 0.1155. For D1 the observed weight-8 mass was 0.1416 and 0.1294, pooled 0.1355, against an ideal value of 1. A readout-only diagnostic, which models independent bit flips, predicts a D1 weight-8 mass near 0.70. It is therefore insufficient to explain the observed loss. The evidence is consistent with noisy execution of the deep compiled Dicke preparation on this device. It does not uniquely attribute the loss to state preparation, routing, calibration drift, readout or the QAOA gates. The exact submitted Q1 transpilation could not be reproduced after calibration changes. Table 7 summarises what each layer supports.

Table 7: Implementation-error ladder. Each row states what was measured, what is supported, and what is unsupported.

Layer	Measured	Supported	Unsupported
Mathematical target P^*	Exact rational construction of the 1,287-allocation accepted set and the uniform target	P^* is the intended uniform law on the accepted set	Anything about a physical sampler
Ideal full-statevector sampler	Ideal D1 weight-8 mass; ideal Q1 accepted mass 0.1155	The logical circuits implement the intended ideal laws	Any device claim; any noise attribution
Compiled gate-level circuit	Ideal simulation of transpiled D1 (total variation 6×10^{-14}); current-pipeline Q1 reconstruction (1×10^{-13})	Transpilation preserves the ideal D1 law; decoder and measurement map consistent	Direct re-establishment of the submitted Q1 transpilation
Raw hardware counts	Eight integrity-checked jobs; D1 weight-8 mass about 0.13–0.14; Q1 accepted mass about 0.01–0.014; readout rates	These raw observations for these jobs; readout alone does not explain the D1 weight loss	A device-wide distribution; full-distribution closeness; causal attribution
Downstream inference	Not measured on hardware (no hardware p-values, rejection rates or power)	Nothing new	Any claim that hardware sampling changes type-I error or power

References

- [1] Andreas Bärtshi and Stephan Eidenbenz. Deterministic preparation of Dicke states. In *Fundamentals of Computation Theory (FCT 2019)*, Lecture Notes in Computer Science, pages 126–139. Springer, 2019. doi: 10.1007/978-3-030-25027-0_9.
- [2] Andreas Bärtshi and Stephan Eidenbenz. Short-depth circuits for Dicke state preparation. In *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 87–96, 2022. doi: 10.1109/QCE53715.2022.00027.
- [3] Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. In *Quantum Computation and Quantum Information*, volume 305 of *Contemporary Mathematics*, pages 53–74. 2002.
- [4] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203, 2013. doi: 10.1137/1.9781611973402.88.
- [5] L. Miriam Dickinson, Brenda Beaty, Chet Fox, Wilson Pace, W. Perry Dickinson, Caroline Emsermann, and Allison Kempe. Pragmatic cluster randomized trials using covariate constrained randomization: A method for practice-based research networks (PBRNs). *The Journal of the American Board of Family Medicine*, 28(5):663–672, 2015. doi: 10.3122/jabfm.2015.05.150001.

- [6] Laia Domingo and Christine Johnson. Quantum-enhanced optimization for patient stratification in clinical trials. arXiv preprint arXiv:2601.11413, 2026. Preprint.
- [7] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann. A quantum approximate optimization algorithm, 2014.
- [8] Rebecca Goldstein, Connor T. Jerzak, Aniket Kamat, and Fucheng Warren Zhu. Fastrerandomize: Fast rerandomization using accelerated computing. *SoftwareX*, 33:102508, 2026. doi: 10.1016/j.softx.2026.102508.
- [9] Stuart Hadfield, Zhihui Wang, Bryan O’Gorman, Eleanor G. Rieffel, Davide Venturelli, and Rupak Biswas. From the quantum approximate optimization algorithm to a quantum alternating operator ansatz. *Algorithms*, 12(2):34, 2019. doi: 10.3390/a12020034.
- [10] Guido W. Imbens and Donald B. Rubin. Fisher’s exact p -values for completely randomized experiments. In *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, pages 57–82. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.006.
- [11] Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J. Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D. Nation, Lev S. Bishop, Andrew W. Cross, Blake R. Johnson, and Jay M. Gambetta. Quantum computing with qiskit, 2024.
- [12] Xinran Li, Peng Ding, and Donald B. Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37): 9157–9162, 2018. doi: 10.1073/pnas.1808191115.
- [13] Jiuyao Lu, Tianruo Zhang, and Ke Zhu. Fast rerandomization for balancing covariates in randomized experiments: A metropolis-hastings framework. arXiv preprint arXiv:2602.07613, 2026. Preprint.
- [14] Jay Mardia, Jiantao Jiao, Ervin Tánzos, Robert D. Nowak, and Tsachy Weissman. Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and Inference: A Journal of the IMA*, 9(4):813–850, 2020. doi: 10.1093/imaiai/iaz025. Published online 2019; Lemma 1 restates the method-of-types L_1 deviation bound with $\phi \geq 2$.
- [15] Kari Lock Morgan and Donald B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40, 2012. doi: 10.1214/12-AOS1008.
- [16] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. doi: 10.1109/TIT.2008.928987.
- [17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017. doi: 10.1137/151002526.
- [18] Pauli Virtanen et al. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261, 2020. doi: 10.1038/s41592-019-0686-2.
- [19] Bingkai Wang and Fan Li. On flexible covariate adjustment under covariate-constrained randomization. *Clinical Trials*, 23(3):282–288, 2026. doi: 10.1177/17407745261423479.

- [20] Zihui Wang, Nicholas C. Rubin, Jason M. Dominy, and Eleanor G. Rieffel. XY-mixers: Analytical and numerical results for the quantum alternating operator ansatz. *Physical Review A*, 101:012320, 2020.
- [21] Hengshi Yu, Fan Li, John A. Gallis, and Elizabeth L. Turner. cvcrand: A package for covariate-constrained randomization and the clustered permutation test for cluster randomized trials. *The R Journal*, 11(2):191–204, 2019. doi: 10.32614/RJ-2019-027.
- [22] Jeffery Yu, Sean R. Muleady, Yu-Xin Wang, Nathan Schine, Alexey V. Gorshkov, and Andrew M. Childs. Efficient preparation of Dicke states. *Physical Review Letters*, 136:030601, 2026. doi: 10.1103/9gjk-rgql.
- [23] Leo Zhou, Sheng-Tao Wang, Soonwon Choi, Hannes Pichler, and Mikhail D. Lukin. Quantum approximate optimization algorithm: performance, mechanism, and implementation on near-term devices. *Physical Review X*, 10(2):021067, 2020. doi: 10.1103/PhysRevX.10.021067.