MDPI

Article

Structure-Preserving Randomization Inference for Placement Effects on Exogenous Paths

Aryan Patel

[1] Independent Researcher; ampatel355@gmail.com

Abstract

A finite sequence of decisions is often scored on a realized path that the decisions are assumed not to change. The resulting statistic confounds the path, the realized structure of the decision sequence, and the placement of that structure on the index set. We develop structure-preserving randomization inference for this question: the test holds the realized path and structural profile fixed, draws re-placements from a declared conditional law, and ranks the realized statistic against them. What distinguishes it from a standard permutation test is the re-randomized object: the calendar placement of a template whose ordered durations, gaps, sizes and signs are held fixed. Under the sharp null that the realized placement is exchangeable with draws from that law, the plus-one Monte Carlo p-value is finite-sample super-uniform; no return model or large-sample approximation is required. Because the conditioning law is analyst-declared rather than identified by the data, we report a sensitivity range and an intersection–union test over a declared finite menu. Simulations verify size and monotone power, and a financial validation shows no entry-placement discovery after multiplicity control. The conclusion is conditional: the method tests placement relative to the chosen law and does not assess skill absorbed into the held-fixed structure.

Keywords: conditional randomization inference; permutation test; finite-sample validity; sensitivity analysis; exchangeability; Monte Carlo methods

MSC: 62F03; 62G09; 62G10; 62F40; 62P05

1. Introduction

Across event-time analysis, trading-strategy evaluation, inspection-window scheduling, and other settings with a fixed exogenous path, the same abstract object recurs: a finite sequence of *decisions* is overlaid on a process whose realized trajectory is treated as unchanged by those decisions, and one observes a single realized outcome. The outcome conflates three logically distinct ingredients, the **structure** of the decision sequence (its ordered durations, inter-decision gaps, magnitudes, signs, and a non-overlap constraint), the realized **path** of the process, and the **placement** of the decisions on the index set. The question this paper addresses is narrow: does the placement carry information, separately from the structure and the path, under a declared counterfactual law?

This question is easy to ask and easy to answer badly. The obvious move, compare the realized outcome to a fixed benchmark such as a zero-effect null or an unconditional average, cannot separate an informative placement from a structure that is simply well suited to a favorable path. A sequence can produce a large, "significant" outcome purely because its durations and magnitudes load on a drifting path that *any* structurally matched

Received:
Accepted:
Published:

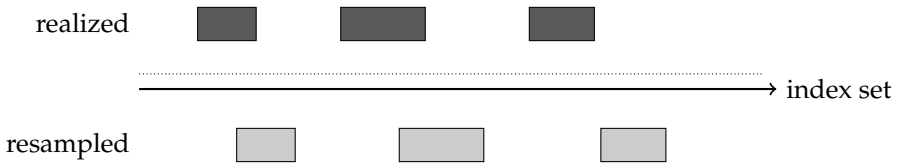
placement would have shared. Tests that reference the realized mean against a fixed benchmark are powerful against exactly this confound, and so reject when there is no genuine timing edge at all. The right comparison is not against a benchmark but against the **counterfactual placements that preserve everything except placement itself**.

We therefore condition on the realized structure and the realized path and re-randomize only the placement margin, drawing from a conditional law over feasible, non-overlapping schedules. This is randomization inference in the classical sense [1,2]: validity flows not from an assumed sampling model but from an exchangeability statement the analyst declares and from a reusable sampler (Algorithm 1) that draws structure-preserving re-placements. The verdict is a finite-sample-valid Monte Carlo p-value under that sharp exchangeability null, not a guarantee attached to arbitrary data. Because "re-randomize the placement" does not name a *single* law, one may additionally hold fixed a warm-up region, or local context; we treat the choice of conditioning measure as model risk and report a sensitivity range across a declared menu rather than a single number. Finance supplies the running example and one validation; the construction is portable only to domains where the path-exogeneity condition is credible.

1.1. Contributions

- **A general structure-preserving null and test.** We formalize the structural profile of a decision sequence and define the no-placement-skill null as exchangeability of the realized placement with structure-preserving re-placements. What distinguishes the construction from a standard permutation or randomization test is the re-randomized object: not a permutation of labels or returns, but the calendar placement of a decision template whose ordered durations, internal-gap multiset, sizes, and signs are held fixed, so the reference set is the orbit of structure-preserving re-placements on a single realized path. We give an explicit, reusable sampler (Algorithm 1).
- **Finite-sample validity, with asymptotics for power.** We prove the Monte Carlo p-value $(1 + k)/(M + 1)$ is super-uniform for every M under exchangeability alone (Theorem 1); under a stated, explicitly-imposed combinatorial central-limit condition (Assumption 2) we characterize a heuristic detection rate, conditional on a stated central-limit condition, namely a minimum detectable edge shrinking as $N^{-1/2}$ (Theorem 3, Corollary 2).
- **Exposure robustness as a conditional implication.** We formalize the finance specialization in which exposure shared by all structure-matched placements leaves the placement law invariant, whereas a mean-against-benchmark test loads on that exposure (Proposition 2).
- **Model risk as a sensitivity range.** We give an admissibility predicate for conditioning measures and report the verdict as a sensitivity range with a valid intersection–union test over a declared finite menu (Theorem 2), clarifying that the resulting region is a declared sensitivity image and not a Manski identified set [3,4].
- **Calibration, competition, and validation.** On known-truth simulations the test holds nominal size and exhibits monotone power; we confirm validity on an exactly enumerable small-N instance and validate on a financial panel, with a self-contained condition-monitoring example (Section 6.5) demonstrating the same algorithm in a non-financial, exogenous-path domain.

Findings preview. The test holds its nominal 0.050 size across four no-skill worlds where mean-based competitors reject a profitable-but-untimed process about two-thirds of the time (Reality Check 0.660, SPA 0.637; the single-strategy specializations of White's Reality Check [5] and Hansen's test for Superior Predictive Ability [6]); power is monotone in signal; and in the financial validation no decision rule shows entry-placement skill under



Fixed: path, ordered durations, weights, signs, gap multiset, non-overlap.
Randomized: placement coordinates under declared law $\mathbb{Q}_\theta$.

Figure 1. Structure-preserving randomization. The realized schedule is ranked against feasible re-placements that preserve the structural profile on the same exogenous path.

the neutral gap-permutation measure after multiplicity control, a finding conditional on the realized duration and gap profile rather than a global no-skill verdict. The method also connects to the conditional randomization and model-X framework of Candès et al. [7], to conditional and generalized Monte Carlo testing [8,9], to exchangeability theory [10,11], and to the data-snooping and superior-predictive-ability literature in the financial application [12]. The validity claim is *conditional*: it holds under H_0^θ, the exchangeability of the realized placement with its $\mathbb{Q}_\theta$-resamples, consistent with the honest framing developed in Section 8.

1.2. Roadmap

Section 2 fixes notation, defines the structural profile, the conditioning set, and the estimand, and states the structure-preserving sampler as Algorithm 1. Section 3 constructs the finite-sample-valid Monte Carlo test, its statistic, and the boundary of what the theorem does not prove. Section 4 develops the admissible measure family, the admissibility conditions, and the sensitivity-range reporting. Section 5 develops the asymptotic behavior: the imposed central-limit condition and its consequences for consistency, local power, and the finance exposure comparison. Section 6 reports the calibration, power, head-to-head competitor, and exact-enumeration studies on known-truth designs. Section 7 presents the financial application as a validation. Section 8 discusses scope and limitations, and Section 9 concludes. Proofs are collected in Appendix A and the asymptotic derivations in Appendix B.

2. The structure-preserving randomization framework

This section sets up the inferential problem in general form. We are given a finite sequence of *decisions* overlaid on an exogenous stochastic process, together with the realized *structure* of that sequence: the ordered holding durations, the inter-decision gaps, the magnitudes (weights) and signs attached to each decision, and a non-overlap constraint. The object under test is the *placement*, that is, where on the index set the decisions were located. The framework holds the realized structure and the realized exogenous path fixed and re-randomizes only the placement, drawing from an admissible conditional law that preserves the structure. Comparing a realized statistic to its distribution over re-placements yields a finite-sample-valid Monte Carlo p-value. The reusable artifact is the sampler of Algorithm 1; everything else in the paper is validation or robustness analysis built on top of it.

Throughout, finance supplies the running example: the decisions are trade entries, the exogenous process is a realized price path, and the statistic is a cumulative return. The notation also covers passive inspection windows, event windows, and other interval-placement problems in which the acted-on path is plausibly exogenous. It does not cover interventions that change the future path without additional causal modeling; Assumption 1 is therefore a scope condition, not a formality.

2.1. Decisions, the exogenous process, and the realized structure

Fix an exogenous realized path $\{P_t\}_{t=0}^{T}$ on the index set $\{0, 1, \ldots, I_{\max}\}$; in the finance instantiation $\{P_t\}$ is the observed open-price calendar and $I_{\max}$ its final index. A *decision sequence* of length N_s is a tuple

$$\mathcal{T}_s = \{(i_j^{\text{in}}, i_j^{\text{out}}, \omega_j, d_j)\}_{j=1}^{N_s}, \tag{1}$$

where i_j^{in} and i_j^{out} are the entry and exit indices of the jth decision on the calendar, ω_j is its weight (the fraction of the resource committed at entry), and d_j is its sign (direction). The realized *holding duration* of decision j is

$$h_j = i_j^{\text{out}} - i_j^{\text{in}}, \qquad h_j > 0, \tag{2}$$

the realized *internal gap* between consecutive decisions is

$$g_j^{\text{int}} = i_{j+1}^{\text{in}} - i_j^{\text{out}}, \qquad j = 1, \ldots, N_s - 1, \tag{3}$$

and the *external slack* aggregates the leading slack before the first decision and the trailing slack after the last,

$$g^{\text{ext}} = i_1^{\text{in}} + \left(I_{\max} - i_{N_s}^{\text{out}}\right). \tag{4}$$

Decisions are required not to overlap: $i_{j+1}^{\text{in}} \geq i_j^{\text{out}}$ for all j, equivalently $g_j^{\text{int}} \geq 0$. We collect the structure into a single object.

Definition 1 (Structural profile)**.** *The* structural profile *of a decision sequence is the tuple*

$$\mathcal{S} = (\mathbf{h}, \mathbf{g}^{\text{int}}, g^{\text{ext}}, \boldsymbol{\omega}, \mathbf{d}),$$

consisting of the realized ordered holding-duration vector $\mathbf{h} = (h_1, \ldots, h_{N_s})$*, the internal-gap multiset* $\mathbf{g}^{\text{int}} = (g_1^{\text{int}}, \ldots, g_{N_s-1}^{\text{int}})$*, the external slack* g^{ext}*, the weights* $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{N_s})$*, and the signs* $\mathbf{d} = (d_1, \ldots, d_{N_s})$*. A* feasible schedule *is any placement of the decision sequence on the index set that preserves* $\mathcal{S}$ *and admits no overlapping decisions. The canonical sampler* $\mathbb{Q}_{\text{gp}}$ *of Algorithm 1 fixes the duration order, leaving* $\mathbf{h}$ *in its realized sequence, while permuting the internal-gap multiset* $\mathbf{g}^{\text{int}}$*.*

The structural profile is exactly the part of the realized sequence we will *not* disturb. What remains free, given $\mathcal{S}$ and the path, is the calendar placement: the leading offset g^{lead} at which the first decision begins, and the order in which the realized internal gaps are consumed. That residual freedom is the placement margin, and it is the only quantity re-randomized.

The entire construction rests on one load-bearing requirement, which we state up front as a named assumption rather than burying it among the limitations.

Assumption 1 (Path exogeneity)**.** *The realized path* $\{P_t\}_{t=0}^{T}$ *is exogenous to the placement: moving the decisions to any feasible re-placement on the index set leaves the path unchanged, so the decisions do not themselves move the exogenous trajectory.*

Assumption 1 is what makes a re-placement a meaningful counterfactual, since the same path is re-scored under each placement. The method does not apply where the path is endogenous, that is, where a decision alters the trajectory it is scored against; in such settings the re-placements no longer share the realized path and the exchangeability that licenses the test fails.

2.2. The conditioning set and the estimand

We condition on everything except placement. The conditioning σ-field for sequence s is

$$\mathcal{C}_s = \sigma\Big(\{P_t\}_{t=0}^{T}, N_s, \mathbf{h}, \mathbf{g}^{\text{int}}, g^{\text{ext}}, \boldsymbol{\omega}, \mathbf{d}, \mathcal{A}_s\Big),$$

where $\mathcal{A}_s$ denotes the calendar and the non-overlap feasibility constraints induced by the realized log. Conditional on $\mathcal{C}_s$ the exogenous path, the count, the durations, the spacing burden, the weights and the signs are all fixed; only the calendar placement of the realized template is left to vary. The test therefore does not ask whether the decision rule as a whole is profitable or effective. It asks whether the realized placement is unusually favorable *after* the realized structure and the realized path have been absorbed into $\mathcal{C}_s$.

Let $\mathcal{F}(\mathcal{C}_s)$ be the feasible schedule space consistent with $\mathcal{C}_s$, let $S_0 \in \mathcal{F}(\mathcal{C}_s)$ be the realized schedule, and for any schedule S let $T(S)$ be the outcome obtained by applying the preserved weights, signs, and durations at the placement S on the fixed path. Let $\mathbb{Q}$ be a law on $\mathcal{F}(\mathcal{C}_s)$, made explicit in Section 2.4. The estimand is the conditional placement advantage

$$\Delta_Q(s) = T(S_0) - \mathbb{E}_{S\sim\mathbb{Q}}[T(S) \mid \mathcal{C}_s],$$

together with the upper-tail probability

$$p_Q(s) = \mathbb{Q}(T(S) \geq T(S_0) \mid \mathcal{C}_s).$$

The reported effect size estimates $\Delta_Q(s)$ by the simulated mean and the Monte Carlo p-value estimates $p_Q(s)$ from finite draws. Both are conditional quantities tied to the chosen $\mathbb{Q}$; neither is an unconditional estimate of total decision quality.

2.3. The structure-preserving null

The placement margin is the natural locus of a sharp null: under it, the realized placement carries no information beyond the structure, so the realized schedule is statistically interchangeable with any structurally identical re-placement on the same path. The following definition fixes the null and the corresponding notion of skill at the level of the realized statistic.

Definition 2 (Structure-preserving null and entry-placement skill)**.** *Let $\{P_t\}_{t=0}^{T}$ be a realized price path. The* structural profile *of a strategy is the tuple $\mathcal{S} = (\mathbf{h}, \mathbf{g}^{\text{int}}, g^{\text{ext}}, \boldsymbol{\omega}, \mathbf{d})$ consisting of the realized ordered holding-duration vector $\mathbf{h}$, the internal gap multiset $\mathbf{g}^{\text{int}}$, the external slack g^{ext}, the capital weights $\boldsymbol{\omega}$, and the direction signs $\mathbf{d}$. A* feasible schedule *is any placement of the trade sequence on the open-price calendar that preserves $\mathcal{S}$ and admits no overlapping positions. Here $\mathbb{Q}_{\text{gp}}$ fixes the duration order while permuting the internal-gap multiset. The structure-preserving null is the distribution $\mathbb{Q}$ over feasible schedules induced by Algorithm 1 below. A strategy is said to exhibit* entry-placement skill *at level α if its realized return CR^{actual} exceeds the $(1-\alpha)$ quantile of CR^{sim} under $\mathbb{Q}$.*

Definition 3 (No-entry-skill null)**.** *Fix the realized price path $\{P_t\}$ and structural profile $\mathcal{S}$. The* no-entry-skill null *H_0 is the hypothesis that the realized schedule S_0 is itself a draw from $\mathbb{Q}$; equivalently, that $S_0, S_1, \ldots, S_M$ are exchangeable conditional on $(\mathcal{S}, \{P_t\})$.*

Under H_0 the realized entry schedule is exchangeable with random timing under $\mathbb{Q}$, conditional on the structural constraints and the observed path; under the alternative the realized schedule produces returns that systematically exceed those of random timing carrying the same structural constraints. Exchangeability here is the substantive modeling assumption, not a fact about the data: it asserts that, given $(\mathcal{S}, \{P_t\})$, the realized placement

is interchangeable with the re-placements the sampler draws. The test is one-sided because the substantive claim is that placement is *better* than chance, not merely different from it. This is randomization inference in the sense of Fisher [1]: a hypothesis stated so that, under it, an observed label is exchangeable with a family of re-labelings, with the reference distribution generated by re-randomization rather than assumed parametrically [2,11]. Here the re-randomized label is placement, the invariant content is the structural profile, and the exchangeable family is the orbit of feasible re-placements of the realized template on the fixed path. Because H_0 holds the exogenous path fixed and conditions on the realized structure, it is a conditional randomization null in the sense of Ernst [8] and the conditional-independence program of Candès et al. [7], with conditioning set $\mathcal{C}_s$.

2.4. The conditional law and its sampler

It remains to specify $\mathbb{Q}$, the admissible conditional law over feasible schedules. The construction perturbs exactly the two free coordinates of the placement margin and nothing else. First, the leading offset is drawn uniformly from its feasible range,

$$g^{\text{lead}} \sim \text{DiscreteUniform}\{0, \dots, g^{\text{ext}}\}. \tag{5}$$

Second, the realized internal gaps are randomly permuted. The randomized entry indices are then generated recursively,

$$\tilde{i}_1^{\text{in}} = g^{\text{lead}}, \qquad \tilde{i}_{j+1}^{\text{in}} = \tilde{i}_j^{\text{in}} + h_j + \tilde{g}_j^{\text{int}}, \tag{6}$$

with randomized exits

$$\tilde{i}_j^{\text{out}} = \tilde{i}_j^{\text{in}} + h_j. \tag{7}$$

This preserves the exact duration profile and the multiset of realized internal gaps, not merely their averages. Because every realized gap is non-negative and the final exit must remain within the calendar, simulated decisions cannot overlap and every simulated schedule is feasible by construction; same-index rotations (zero-gap transitions) survive permutation and so are preserved whenever they occur. Algorithm 1 states the procedure.

Algorithm 1: Structure-preserving randomization of a single trade schedule

Input: Realized durations $\mathbf{h} = (h_1, \dots, h_{N_s})$; realized internal gaps $\mathbf{g}^{\text{int}} = (g_1^{\text{int}}, \dots, g_{N_s-1}^{\text{int}})$; external slack g^{ext}; price calendar of length $I_{\max} + 1$.

Output: Randomized entry indices $(\tilde{i}_1^{\text{in}}, \dots, \tilde{i}_{N_s}^{\text{in}})$ and exit indices $(\tilde{i}_1^{\text{out}}, \dots, \tilde{i}_{N_s}^{\text{out}})$ with no overlapping trades.

Draw $g^{\text{lead}} \sim \text{DiscreteUniform}\{0, \dots, g^{\text{ext}}\}$;
Draw a uniformly random permutation π of $\{1, \dots, N_s - 1\}$ and set $\tilde{g}_j^{\text{int}} = g_{\pi(j)}^{\text{int}}$;
Set $\tilde{i}_1^{\text{in}} \leftarrow g^{\text{lead}}$ and $\tilde{i}_1^{\text{out}} \leftarrow \tilde{i}_1^{\text{in}} + h_1$;
for $j \leftarrow 2$ **to** N_s **do**
    $\tilde{i}_j^{\text{in}} \leftarrow \tilde{i}_{j-1}^{\text{out}} + \tilde{g}_{j-1}^{\text{int}}$;
    $\tilde{i}_j^{\text{out}} \leftarrow \tilde{i}_j^{\text{in}} + h_j$;
end
return $(\tilde{i}^{\text{in}}, \tilde{i}^{\text{out}})$

The algorithm is written in the trade/price-calendar variables of the running example; the generic objects it manipulates are the durations, the gaps, the slack, and the non-overlap constraint. Reading the durations as passive inspection windows, marked event windows, or other intervals that do not alter the future path gives the same mathematical object.

Reading them as interventions that change the path does not: those cases require a different causal or dynamic model.

The sampler draws the leading offset g^{lead} and the gap permutation π as two independent coordinates, and feasibility requires that the resulting schedule fit inside the calendar for *every* such draw. The following lemma establishes that the leading-offset range is the same for every permutation, so the joint support of (g^{lead}, π) is a product set, the realized schedule S_0 lies in it, and S_0 and the resamples are draws from one identical conditional law. This is what makes the exchangeability invoked in Definition 3 and Theorem 1(i) hold, rather than an artifact of any per-permutation truncation of the offset range.

Lemma 1 (Permutation-independent feasibility and a common conditional law)**.** *Fix the structural profile $\mathcal{S}$ and the calendar $\{0, \dots, I_{\max}\}$, and recall that the total external slack $g^{\text{ext}} = i_1^{\text{in}} + (I_{\max} - i_{N_s}^{\text{out}})$ is the sum of leading and trailing slack. Then for every permutation π of the internal gaps and every $g^{\text{lead}} \in \{0, \dots, g^{\text{ext}}\}$, the schedule produced by Algorithm 1 satisfies*

$$\sum_{j=1}^{N_s} h_j \;+\; \sum_{j=1}^{N_s-1} g_j^{\text{int}} \;+\; g^{\text{lead}} \;\le\; I_{\max},$$

and leaves trailing slack $g^{\text{ext}} - g^{\text{lead}} \ge 0$. Consequently the feasible support of the draw is the product set $\{0, \dots, g^{\text{ext}}\} \times \{\text{permutations of } \mathbf{g}^{\text{int}}\}$, independent of π; the realized schedule S_0 is the member with $g^{\text{lead}} = i_1^{\text{in}}$ and $\pi = \text{id}$, so S_0 and the resamples are draws from the identical conditional law $\mathbb{Q}$ given $\mathcal{C}_s$.

Proof. The durations $\sum_j h_j$ and the internal gaps $\sum_j g_j^{\text{int}}$ are sums over the realized multisets and are therefore invariant under any permutation π of the internal gaps; their sum, the total internal span $\Sigma h + \Sigma g^{\text{int}}$, is a constant that does not depend on π. By the definition of the external slack, $\Sigma h + \Sigma g^{\text{int}} + g^{\text{ext}} = I_{\max}$ exactly, since the leading slack, the internal span, and the trailing slack partition the calendar. For any $g^{\text{lead}} \in \{0, \dots, g^{\text{ext}}\}$ the recursion of Algorithm 1 lays the fixed durations and the permuted gaps down starting at index g^{lead}, so the final exit is $\tilde{i}_{N_s}^{\text{out}} = g^{\text{lead}} + \Sigma h + \Sigma g^{\text{int}} = g^{\text{lead}} + (I_{\max} - g^{\text{ext}}) \le I_{\max}$, with equality iff $g^{\text{lead}} = g^{\text{ext}}$. The total span thus satisfies $\Sigma h + \Sigma g^{\text{int}} + g^{\text{lead}} \le I_{\max}$ for every π and every admissible g^{lead}, and the trailing slack is $I_{\max} - \tilde{i}_{N_s}^{\text{out}} = g^{\text{ext}} - g^{\text{lead}} \ge 0$. Because this bound is independent of π, no permutation-dependent truncation of the offset range is ever needed, and the support of (g^{lead}, π) is the product set $\{0, \dots, g^{\text{ext}}\} \times (\text{permutations})$. The realized schedule corresponds to $(g^{\text{lead}}, \pi) = (i_1^{\text{in}}, \text{id})$, an element of this product set, so S_0 and the resamples share one conditional law given $\mathcal{C}_s$, which is the exchangeability premise of Theorem 1(i). □

Sampling distribution.

Algorithm 1 samples from a distribution $\mathbb{Q}$ over feasible schedules whose marginal of g^{lead} is uniform on $\{0, \dots, g^{\text{ext}}\}$ and whose marginal of $\tilde{\mathbf{g}}^{\text{int}}$ is uniform over the permutations of $\mathbf{g}^{\text{int}}$. This $\mathbb{Q}$ is invariant under relabeling of decision indices and preserves the realized spacing multiset exactly; it is not the uniform measure over all non-overlapping placements consistent with $\mathcal{S}$, which is combinatorially expensive to construct when gaps differ in size. The Monte Carlo evaluation draws M schedules $S_1, \dots, S_M$ i.i.d. from $\mathbb{Q}$ and reports the one-sided p-value

$$\hat{p}_M = \frac{1 + \sum_{m=1}^{M} \mathbf{1}\{T(S_m) \ge T(S_0)\}}{M + 1},$$

with the plus-one correction so that the estimate is never exactly zero [13,14]. Throughout we take $M = 5{,}000$ draws per sequence unless stated otherwise.

3. A finite-sample-valid Monte Carlo test

We now build the test that sits at the core of the method. The input is a single realized schedule: a finite sequence of decisions placed on an exogenous index set, together with its realized structure and the realized exogenous path. The output is a Monte Carlo p-value that is valid in finite samples, in the sense that under the null it does not reject more often than its nominal level. The construction asks only that the realized schedule be exchangeable with a set of re-placements that preserve the structure, so the guarantee follows from exchangeability alone and needs no central-limit approximation, no asymptotics in the number of decisions, and no large-sample regularity on the path [1,2,15].

Throughout, $\mathcal{C}$ denotes the conditioning set $\mathcal{C}_s$ of Section 2.2 fixed by the realized structure and path, and $T(S)$ the cumulative-return map evaluated at a schedule S. The single null constructed here is the gap-permutation measure, the minimum element $\mathbb{Q}$ of the admissible family $\mathfrak{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta\}$ developed in Section 4; the validity theorem is stated for an arbitrary member $\mathbb{Q}_\theta$ so that it serves both this section and the sensitivity analysis later.

3.1. The Monte Carlo p-value and its statistic

Run the sampler of Algorithm 1 M times to obtain feasible schedules $S_1, \ldots, S_M$, score each on the fixed path through the cumulative-return map, and rank the realized value CR_s^{actual} against the simulated ones. With $M = 5{,}000$ draws per strategy, each simulated cumulative return is

$$CR_{s,m}^{\text{sim}} = \exp\Big(\textstyle\sum_{j=1}^{N_s} \log(1 + \tilde{r}_{j,m})\Big) - 1, \qquad \tilde{r}_{j,m} = \omega_j \, \tilde{q}_{j,m},$$

and the one-sided p-value uses the plus-one (smoothed) estimator

$$\hat{p}_s = \frac{1 + \sum_{m=1}^{M} \mathbf{1}\{CR_{s,m}^{\text{sim}} \geq CR_s^{\text{actual}}\}}{M + 1}.$$

The plus-one in numerator and denominator counts the realized schedule among the draws; it keeps the p-value strictly positive and is exactly what super-uniformity requires [13,14,16]. We report this quantity as *valid* rather than exact, since on a discrete orbit ties are folded into the upper tail and move the bound toward conservatism.

Alongside the p-value we report a descriptive effect size on the return scale. The *conditional excess return* is

$$\text{RCSI}_s = CR_s^{\text{actual}} - \mu_s^{\text{sim}},$$

the gap between the realized return and the mean return under random placement, and its standardized form is the Monte Carlo z-score

$$\text{RCSI}_{z,s} = \frac{CR_s^{\text{actual}} - \mu_s^{\text{sim}}}{\sigma_s^{\text{sim}}}.$$

RCSI_z is asymptotically pivotal under $\mathbb{Q}$ only in the large-N regime of Assumption 2, and is otherwise reported as a descriptive, non-pivotal standardized summary; when $\sigma_s^{\text{sim}} = 0$ it is set to zero if $CR_s^{\text{actual}} = \mu_s^{\text{sim}}$ and is otherwise left undefined. These two contrasts summarize the magnitude and the standardized position of the realized schedule, but they describe an effect; they do not deliver the verdict. The verdict rests on $\hat{p}_s$, whose level guarantee we state next and which does not invoke any large-sample approximation.

3.2. Finite-sample validity

The guarantee is a super-uniformity statement: under the null the Monte Carlo p-value stochastically dominates a uniform draw, so it controls the level at every attainable

threshold, and for any fixed realized schedule it converges to the population tail probability as M grows. Validity follows from a single invariance: under H_0 the realized placement is exchangeable with its structure-preserving re-placements, so its rank among them is uniform. This is the conditional-Monte-Carlo exchangeability argument of Besag and Clifford [9] and Hemerik and Goeman [15], specialized to the placement orbit, and it rests on exchangeability alone, with no distributional or large-sample assumption on the path.

Theorem 1 (Finite-sample validity and large-M consistency of the Monte Carlo p-value)**.** *Fix any structure-preserving measure $\mathbb{Q}_\theta$ in the family $\mathfrak{Q}$ of Definition 4, and let $S_1, \ldots, S_M$ be i.i.d. draws from $\mathbb{Q}_\theta$ given $\mathcal{C}$. Define*

$$\hat{p}_M = \frac{1 + \sum_{m=1}^{M} \mathbf{1}\{T(S_m) \geq T(S_0)\}}{M+1}.$$

(i) Validity. *Suppose H_0^θ holds in the statistic-level form that $T(S_0)$ is exchangeable with $\{T(S_m)\}_{m=1}^M$ given $\mathcal{C}$; this holds in particular if $S_0 \overset{d}{=} \mathbb{Q}_\theta$ given $\mathcal{C}$. Then $\hat{p}_M$ is super-uniform: for every α on the grid $\{1/(M+1), \ldots, 1\}$, $\Pr(\hat{p}_M \leq \alpha \mid \mathcal{C}) \leq \alpha$, with equality absent ties and strict conservatism when ties are counted into the right tail.* (ii) Consistency in M. *For* any *fixed S_0, whether or not H_0^θ holds, $\hat{p}_M \to q_\theta := \mathbb{Q}_\theta\big(T(S) \geq T(S_0) \mid \mathcal{C}\big)$ almost surely as $M \to \infty$.*

Applied to the gap-permutation measure $\mathbb{Q}$ of Definition 2, with $T(S) = \log(1 + CR(S))$ a fixed measurable map of the schedule on the fixed path, part (i) is exactly the level guarantee for $\hat{p}_s$. The proof is a rank argument: exchangeability makes the rank of $T(S_0)$ among the $M+1$ values uniform, and the strong law gives the large-M limit. We defer it to Appendix A.

Two features of the statement carry the method. Part (i) buys validity with exchangeability alone, so the level holds in finite samples for any number of decisions and any path, with the plus-one correction handling discreteness and ties [2,15]. Part (ii) separates the role of M from the role of the null: more draws sharpen the estimate of the population tail q_θ regardless of whether H_0^θ holds, so M controls Monte Carlo error and not test level. The exchangeability posited in Definition 2 is the substantive assumption; everything downstream is computation [8,10]. We therefore say the p-value is *valid* rather than exact: it attains equality only in the idealized tie-free form, and is conservative once Monte Carlo ties are accounted for.

3.3. What the finite-sample theorem does not prove

The theorem is deliberately narrow. It does not prove that the analyst's chosen law $\mathbb{Q}_\theta$ is the correct counterfactual, that the realized schedule was actually randomized by that law, or that non-rejection means the decision rule has no skill. It proves only that, if the realized placement is exchangeable with $\mathbb{Q}_\theta$-replacements after conditioning on the declared structure and path, then the plus-one rank p-value controls level in finite samples. This distinction is important enough to state before the sensitivity analysis: choosing $\mathbb{Q}_\theta$ is a modeling decision, and the scientific claim is conditional on that decision.

4. Admissible conditioning measures and model-risk sensitivity

The validity argument of Section 3.2 fixes the structural profile $\mathcal{S} = (\mathbf{h}, \mathbf{g}^{\text{int}}, g^{\text{ext}}, \boldsymbol{\omega}, \mathbf{d})$ and the realized path $\{P_t\}$ and re-randomizes only the placement of the trade template on the open-price calendar, in the randomization-inference tradition [1,2,15]. The construction applies to finite decision sequences overlaid on exogenous paths whose realized structure one is willing to hold fixed; domains where decisions change the path require additional modeling and are outside the finite-sample guarantee. But "re-randomize the placement"

does not name a single probability law. Many conditional measures preserve the realized structure while differing in what *else* they hold fixed, and each such measure encodes a different question about where the decisions could admissibly have been put. The choice among them is not pinned down by the data; it is a modeling decision. This section makes that choice explicit. We define the family of measures we are willing to defend, give the admissibility predicate that qualifies a measure for membership, order the family by how much it conditions on, and then specify how the resulting collection of answers is to be reported. Our central methodological commitment is that the output is not a scalar verdict but a *sensitivity range* across a declared menu of conditioning measures, read in the spirit of partial identification rather than as a sharp identified set [3,4].

4.1. The admissible family

Every measure we consider fixes $(\mathcal{S}, \{P_t\})$ and re-randomizes only the placement; the members differ in an additional feature set M_θ of the realized schedule that they additionally hold fixed. The minimal member conditions on the geometry $\mathcal{S}$ alone.

Definition 4 (Admissible menu and sensitivity range)**.** *Let $\mathfrak{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta\}$ be the family of structure-preserving measures that all fix $(\mathcal{S}, \{P_t\})$ and differ only in which additional features M_θ of the realized schedule they hold fixed: the gap-permutation measure (conditioning on the geometry $\mathcal{S}$ alone, with $M_\theta = \emptyset$, the minimum element), the leading-gap restriction (forbidding placement in the universal warm-up region), and the context-matched measure (restricting entries to bars of comparable trailing volatility). Each $\mathbb{Q}_\theta$ defines a null H_0^θ and an estimand $q_\theta = \mathbb{Q}_\theta\big(T(S) \geq T(S_0)\big)$. The* admissible menu $\Theta_0 \subseteq \Theta$ *is the set of measures admissible in the sense of Definition 5. Given Θ_0, the reported quantity is not a scalar "skill" but the* sensitivity range over the declared admissible menu, $\{q_\theta : \theta \in \Theta_0\}$, *summarized by the outer bounds* $\big[\min_{\theta\in\Theta_0} q_\theta,\ \max_{\theta\in\Theta_0} q_\theta\big]$.

The canonical member, and the one used for the headline analysis, is the gap-permutation measure $\mathbb{Q}_{\mathrm{gp}}$ of Algorithm 1: it draws the leading gap uniformly on the feasible slack and permutes the realized internal gaps, thereby conditioning on the ordered durations, the gap multiset, the weights, the directions, and the non-overlap constraint, and on nothing more. The leading-gap restriction forbids placements that begin inside a universal warm-up region common to all strategies, removing a placement degree of freedom that no strategy could have exploited at entry. The context-matched measure restricts each randomized entry to bars whose trailing volatility is comparable to the bar the trade actually occupied, so that what is tested is the *incremental* entry advantage beyond simply matching the local market state. Table 1 lists the features each measure preserves and the timing question each one answers.

A coarser, forward-safe variant, the *state-preserving* (regime-preserving) measure, is used as a robustness check in the supplement rather than as a member of the headline menu. It restricts each randomized entry to bars carrying the same forward-safe volatility-regime label (calm, neutral, or stressed) as the bar the trade actually occupied, reusing the candidate-pool sampler with the regime label substituted for the volatility quantile, and falling back to a regime-agnostic uniform draw on windows containing no same-regime bar. Because the regime label is a coarse tercile discretization of the same volatility state that the context-matched measure conditions on continuously, the two are close cousins; we report the regime-preserving result as a sensitivity check, not as an independent member of Θ_0.

4.2. What makes a measure admissible

Not every structure-preserving measure earns a place in the menu. The gap-permutation, leading-gap, and context-matched measures qualify; the uniform-feasible and

Table 1. Structure-preserving measures: preserved features and the placement question each answers. The gap-permutation measure is the minimal admissible member used for the headline analysis; the uniform-feasible and slack-redistribution rows are listed for contrast and are *not* admissible (they relax structure preservation, AM1).

Measure	Preserved features	Interpretation
Gap-permutation $\mathbb{Q}_{\text{gp}}$ (used here)	Price path, ordered durations, gap multiset, weights, directions, non-overlap	Entry placement conditional on the realized spacing burden
Leading-gap / path	As above, plus exclusion of the universal warm-up region	Placement skill once the common warm-up window is removed
Context-matched	As above, plus coarse entry context (trailing volatility)	Incremental entry skill beyond market-state matching
Uniform feasible schedules	Price path, durations, weights, directions, non-overlap (gap multiset relaxed)	A less restrictive spacing null (not admissible)
Slack-redistribution	Price path, durations, weights, directions, approximate turnover	Sensitivity to exact gap preservation (not admissible)

slack-redistribution measures of Table 1 do not, because they relax the exact gap preservation that anchors the structural null. We make the qualification precise through three requirements: the measure must preserve the realized structure, it must condition only on information available at entry, and the features it holds fixed must not be tainted by the realized outcome.

Definition 5 (Admissible conditioning measure)**.** *A structure-preserving measure $\mathbb{Q}_\theta$, holding fixed $(\mathcal{S}, \{P_t\})$ together with an additional feature set M_θ, is* admissible *(written $\theta \in \Theta_0$) if it satisfies:* (AM1) Structure preservation. *$\mathbb{Q}_\theta$-almost surely the schedule preserves the exact structural profile $\mathcal{S} = (\mathbf{h}, \mathbf{g}^{\text{int}}, g^{\text{ext}}, \boldsymbol{\omega}, \mathbf{d})$ and the non-overlap constraint A_s.* (AM2) Entry-measurability. *The extra feature set M_θ held fixed beyond $\mathcal{S}$ is $\mathcal{F}_t^{\text{in}}$-measurable, so that no information unavailable at entry enters placement.* (AM3) Outcome-independence. *M_θ is a function of $(\{P_t\}, t)$ and the template alone; it does not reference any realized return or the value $T(S_0)$.*

Here $\mathcal{F}_t^{\text{in}} = \sigma(\{P_u : u \le t\})$ is the entry filtration. The three conditions do separable work. AM1 keeps every resampled schedule inside the same structural equivalence class as the realized one, so that the re-randomization changes only timing. AM2 forbids conditioning on anything the trader could not have seen at the moment of entry; this is what rules out a measure that, say, matched on the realized forward volatility over the holding window, which would launder hindsight into the placement law. AM3 is the sharpest guard: it bars the conditioning set from referencing the outcome itself, which is what would otherwise let a measure quietly absorb the very edge it is meant to test. The state-preserving measure earns admissibility only through its *forward-safe* construction of the regime label; a backward-looking label would violate AM2 or AM3. The payoff of the predicate is that each admissible measure yields a well-posed null.

Lemma 2 (Admissible nulls are well posed)**.** *For any $\theta \in \Theta_0$, the realized schedule lies in the support of $\mathbb{Q}_\theta$ given $\mathcal{C}$. Consequently the exchangeability hypothesis of Theorem 1(i) is well posed under each H_0^θ, and the intersection–union test of Theorem 2 is applied to a family of individually valid tests.*

Lemma 2 is what connects this section back to the finite-sample guarantee: because the realized schedule is itself a feasible draw from every admissible measure, the super-

uniformity of Theorem 1(i) applies measure by measure, and the per-measure p-values $\hat{p}_M^\theta$ are each valid in finite samples. The proof is given in Appendix A.

4.3. The conditioning-refinement order

The admissible measures are not interchangeable; they sit in a partial order by how much they condition on. Write $\theta \leq \theta'$ when $\sigma(\mathcal{C}, M_\theta) \subseteq \sigma(\mathcal{C}, M_{\theta'})$, that is, when $\mathbb{Q}_{\theta'}$ holds fixed everything $\mathbb{Q}_\theta$ does and more. Under this order the gap-permutation measure $\mathbb{Q}_{\text{gp}}$, with $M_{\text{gp}} = \emptyset$, is the *minimum* element: it conditions on the bare geometry and grants placement the widest feasible support. The leading-gap restriction and the context-matched measure are both strict refinements of $\mathbb{Q}_{\text{gp}}$ but are mutually incomparable, since one removes a warm-up region and the other matches on volatility context, and neither feature set contains the other. Support nests downward under refinement: as a measure conditions on more, the feasible schedule space shrinks, and the alternatives the realized schedule is compared against grow more like it. A refinement therefore asks a more demanding question, "is there placement skill *beyond* the matched context?", and a finding that survives a refinement is the stronger claim. The minimal measure is the most generous test for placement information; the refinements progressively absorb candidate confounds into the conditioning set.

4.4. Reporting the sensitivity range

Because no single measure is privileged by the data, the honest report is the collection of answers across the declared menu. We aggregate the per-measure p-values through an intersection–union test, then read the spread of estimands as a model-risk band.

Definition 6 (Measure-invariant null and skill claim)**.** *The* measure-invariant null *is the union* $H_0^\cup = \bigcup_{\theta\in\Theta_0} H_0^\theta$*: there is some admissible measure under which the realized placement is exchangeable with random placement.* Measure-invariant entry-placement skill *is its complement: rejection of* H_0^θ *for* every $\theta \in \Theta_0$.

Theorem 2 (Intersection–union test over a declared finite menu)**.** *Let* $\hat{p}_M^\theta$ *be the Monte Carlo p-value of Theorem 1 computed under* $\mathbb{Q}_\theta$*, and define*

$$p^{\text{inv}} = \max_{\theta\in\Theta_0} \hat{p}_M^\theta.$$

Rejecting the measure-invariant null $H_0^\cup$ *when* $p^{\text{inv}} \leq \alpha$ *is a level-*α *test: under* $H_0^\cup$, $\Pr(p^{\text{inv}} \leq \alpha \mid \mathcal{C}) \leq \alpha$.

Proof. $H_0^\cup$ holds iff $H_0^{\theta^*}$ holds for at least one $\theta^* \in \Theta_0$. On the event $H_0^{\theta^*}$, and using Lemma 2 so that the exchangeability hypothesis is well posed, Theorem 1(i) gives $\Pr(\hat{p}_M^{\theta^*} \leq \alpha \mid \mathcal{C}) \leq \alpha$. Since $p^{\text{inv}} = \max_\theta \hat{p}_M^\theta \geq \hat{p}_M^{\theta^*}$, the event $\{p^{\text{inv}} \leq \alpha\}$ implies $\{\hat{p}_M^{\theta^*} \leq \alpha\}$, so $\Pr(p^{\text{inv}} \leq \alpha \mid \mathcal{C}) \leq \Pr(\hat{p}_M^{\theta^*} \leq \alpha \mid \mathcal{C}) \leq \alpha$. □

The price of insisting on rejection under every admissible measure is power, and the loss is governed by the least-favorable member of the menu.

Corollary 1 (Power is governed by the least-favorable measure)**.** *Under Assumption 2, suppose the local placement alternative gives* $\mathbb{Q}_\theta$ *a noncentrality* $h_\theta \geq 0$ *as in Corollary 2. Then the asymptotic power of the intersection–union test is bounded by the least-favorable measure,* $\beta^{\text{inv}} \leq \Phi\big(\min_{\theta\in\Theta_0} h_\theta - z_{1-\alpha}\big)$*, with equality when the per-measure rejection events coincide.*

Beyond the binary verdict, the substantive output is the spread of the estimand q_θ across the menu. We report this spread directly as a band rather than collapsing it to a point.

Definition 7 (Sensitivity region)**.** *The* sensitivity region *induced by* Θ_0 *is the finite point set* $R(\Theta_0) = \{q_\theta : \theta \in \Theta_0\}$*, with outer bracket* $[\min_{\theta\in\Theta_0} q_\theta,\ \max_{\theta\in\Theta_0} q_\theta]$*. Under the agnostic stance that any* $\theta \in \Theta_0$ *could be the conditioning partition one would defend,* $R(\Theta_0)$ *is the set of answers the test returns across the defensible questions.*

In practice the band is read together with the intersection–union verdict: a p^{inv} above α says that at least one defensible conditioning choice fails to reject, and the width of $R(\Theta_0)$ says how much the conclusion moves as one travels across the menu. A wide band is itself a finding, an admission that the placement question is sensitive to which structure one chooses to hold fixed.

4.5. A partial-identification reading, not a Manski set

The sensitivity band invites comparison with a sharp identified set, and the comparison is instructive precisely because it fails. $R(\Theta_0)$ is not a Manski identified set, and reading it as one would overstate what the data deliver.

Remark 1 (The sensitivity region is not a Manski identified set)**.** $R(\Theta_0)$ *differs from a sharp Manski identified set in three respects. First, the menu* Θ_0 *is chosen by the analyst, not derived from the data, so enlarging it only widens* $R(\Theta_0)$*. Second,* $R(\Theta_0)$ *is a finite point set and its outer bracket* $[\min_\theta q_\theta, \max_\theta q_\theta]$ *is merely its outer envelope; interior points of the bracket are not claimed attainable. Third, no* θ *is privileged by the data, so* $R(\Theta_0)$ *is "the set of answers each defensible question returns," not "a set containing one fixed structural truth." Precisely:* $R(\Theta_0)$ *is the* image *of the estimand map* $\theta \mapsto q_\theta$ *over the declared, finite menu* Θ_0*, that is, a sensitivity range. Under a worst-case stance that treats* Θ_0 *as a menu over which an adversary selects, the robustified estimand* $q^{\text{rob}} := q_{\theta^*}$ *(with* θ^* *the worst-case selection) is a single number.*

The three differences in Remark 1 have a common root: a Manski set brackets one fixed estimand under weak assumptions about the data, whereas $R(\Theta_0)$ ranges over different estimands, one per question we are willing to ask. The width of $R(\Theta_0)$ thus measures the consequence of a modeling choice, not the residual ambiguity left by the data about a single target. This distinction is not a technicality; it disciplines what unanimous rejection can and cannot establish.

Proposition 1 (Non-sufficiency and non-necessity)**.** *Let* Θ_0 *be admissible. Then:* (i) Non-sufficiency. *Unanimous rejection is not sufficient to establish informative entry placement: a confound that displaces* $T(S_0)$ *in the same direction under every* $\mathbb{Q}_\theta \in \mathfrak{Q}$ *survives the intersection and is reported as "skill" though it is not placement.* (ii) Non-necessity. *Unanimous rejection is not necessary to capture informative state-conditional placement: a measure* $\mathbb{Q}_{\theta'} \in \Theta_0$ *can absorb a real state-conditional edge into its conditioning set, so that* $H_0^{\theta'}$ *is not rejected even though placement genuinely carries information.*

Proposition 1 sets the limits of the method honestly. Part (i) warns that the intersection–union test launders a common-direction confound into an apparent skill claim: if some non-placement feature pushes the realized return up under every measure in the menu, no amount of measure invariance will unmask it. Part (ii) is the dual caution: a sufficiently refined measure can condition away the very edge one hoped to detect, so a non-rejection under the most demanding member is not evidence of no skill. Together they mean the

measure-invariant verdict should be reported as "placement information that is robust across the declared admissible menu," with the menu stated in full, and never as an unqualified discovery of timing skill. The exposure-invariance result of Proposition 2, which shows that the structure-preserving test does not reject pure-exposure alternatives under its stated exchangeability condition while a return-mean benchmark can, sharpens part (i): the structure-preserving construction defends against this exposure confound by design, but it offers no defense against a confound that rides on placement-shaped structure itself. We therefore present the cross-asset results of Section 7 as a sensitivity band over Θ_0 accompanied by the intersection–union p^{inv}, and we read both through the cautions of Proposition 1.

5. Asymptotic behavior

The validity of the procedure is finite-sample and rests on exchangeability alone: under any admissible structure-preserving measure $\mathbb{Q}_\theta$ the Monte Carlo p-value $\hat{p}_M$ is super-uniform conditional on $\mathcal{C} = \sigma(\mathcal{S}, \{P_t\})$, with no appeal to a limit theorem (Section 3.2, Algorithm 1). The present section asks a separate question: as the realized decision count N grows, does the test detect a genuine placement edge, and at what rate? Here evidence accumulates in N, the number of decisions, rather than in the length of the exogenous path, because $\mathbb{Q}_\theta$ rearranges a fixed set of N contributions on a fixed path. The standardized contrast $\mathrm{RCSI}_{z,s}$ is the natural pivot for these statements: it is a descriptive, approximately pivotal effect size on the $\mathbb{Q}_\theta$ scale, and the asymptotic results below describe its behavior, not the basis of any verdict. Throughout, $T(S) = \log(1 + CR)$ is the log cumulative-return statistic induced by a schedule S, S_0 is the realized schedule, and the longer derivations are deferred to Appendix B. Because Assumption 2 posits the Gaussian limit directly, the results of this section should be read as characterizing the *rate* and *direction* of detection under an imposed normal approximation, not as independently establishing that approximation; none of them is used to license the verdict, whose validity is the finite-sample, assumption-free guarantee of Section 3.2.

5.1. A central-limit condition for the standardized statistic

The consistency, detection-rate, and local-power results require a Gaussian approximation for RCSI_z, and no i.i.d. structure is available to supply one: under $\mathbb{Q}_\theta$ a draw is a permutation of the realized gaps that relocates each decision on the fixed exogenous path, so the contributions are functions of the permutation rather than an independent array, and the placements are further coupled by the non-overlap constraint and the shared path. We therefore state the needed limit theorem as an explicit, named high-level condition and prove the propositions *conditional* on it. The condition is imposed, not derived.

Assumption 2 (Combinatorial / finite-population CLT condition)**.** *Write* $\log(1 + CR) = \sum_{j=1}^{N} \xi_j$*, where under* $\mathbb{Q}_\theta$ *the* ξ_j *are the* N *trade log-contributions evaluated at a placement of the fixed duration/gap template drawn from* $\mathbb{Q}_\theta$ *on the fixed path. Let* $\mu_{\mathbb{Q}}, \sigma_{\mathbb{Q}}$ *denote the conditional* $\mathbb{Q}_\theta$ *mean and standard deviation of* $T(S) = \log(1 + CR)$ *and set* $Z_N = (T(S_0) - \mu_{\mathbb{Q}})/\sigma_{\mathbb{Q}}$*. We assume that under* H_0^θ*,*

$$Z_N \;\to_d\; N(0,1) \qquad (N \to \infty),$$

and that this convergence holds under the following two regularity requirements: (a) *a no-dominant-term, Lindeberg-type condition* $\max_j \mathrm{Var}(\xi_j) / \sum_j \mathrm{Var}(\xi_j) \to 0$*, so that no single trade carries a non-negligible share of the return budget; and* (b) *weak (*α*-mixing) dependence of the realized path, so that the autocorrelation entering the permutation variance decays fast enough for the finite-population variance to be the correct standardization. We further assume that under the per-trade location-shift alternative of Theorem 3 the centered statistic* $Z_N - \delta_N \to_d N(0,1)$*, with*

δ_N the deterministic displacement defined there, so that the location shift moves the limit law without altering its Gaussian shape.

Assumption 2 is a high-level condition, not a corollary of a single theorem: a $\mathbb{Q}_\theta$-draw is a permutation of the realized gaps, so the contributions ξ_j are themselves functions of the permutation rather than a fixed array, and no off-the-shelf result delivers normality automatically. It is motivated by the combinatorial central-limit theory for permutation statistics, with Hoeffding's combinatorial central-limit theorem as the canonical reference point [17]: standardized sums over a randomly permuted finite population are asymptotically normal precisely when a no-dominant-term (Lindeberg) condition of the form in part (a) holds. The exchangeability that licenses the conditional construction is the classical theory of Aldous [10], and the finite-population permutation framework is that of Lehmann and Romano [2]. The mixing requirement (b) is what lets the path's autocorrelation be absorbed into the finite-population variance rather than break the approximation; the final clause, normality of the *centered* statistic $Z_N - \delta_N$ under the alternative, is what licenses the residual-normality step in the proofs deferred to Appendix B.

The conditions are sufficient, not necessary, and we do not claim to have verified them for any particular realized path: requirements (a)–(b) are imposed, and the simulation study of Section 6 checks calibration rather than the conditions themselves. The condition is nonetheless falsifiable and can fail: when one or a few decisions dominate the return budget, (a) is violated, the normal approximation degrades, and Z_N is not pivotal. We therefore treat RCSI_z as a descriptive, approximately pivotal effect size only; the test's *validity* rests on the finite-sample exchangeability of the Monte Carlo p-value, not on Assumption 2, which is invoked solely for the asymptotic statements that follow. The adequacy of the approximation is an empirical matter, probed in distribution by the Kolmogorov–Smirnov calibration tests of Section 6, which assess the null p-value's uniformity rather than the Gaussianity of Z_N per se: across the four no-skill worlds the KS test cannot reject uniformity (p between 0.515 and 0.701).

5.2. Consistency and the minimum detectable edge

Validity is the floor; the test must also detect a genuine edge as decisions accumulate. Because $\mathbb{Q}_\theta$ rearranges a fixed set of N contributions on a fixed path, evidence accrues in N.

Theorem 3 (Consistency and the $N^{-1/2}$ detection rate)**.** *Suppose Assumption 2 holds. Consider the per-trade location-shift alternative, a stylized local alternative that adds a common deterministic per-trade increment $\mu_1 > 0$ to each of the N trade contributions to $\log(1 + CR)$ relative to its $\mathbb{Q}_\theta$ baseline, holding the $\mathbb{Q}_\theta$ variance fixed; write $\zeta = \mu_1/\sigma_1$ for the implied common standardized entry edge, with per-trade standard deviation σ_1 and $\sigma_{\mathbb{Q}}^2 = \sum_j \mathrm{Var}(\xi_j)$. This is a deterministic mean perturbation of the contributions at fixed null variance, used to characterize the detection rate; it is a modeling device rather than a placement reachable within $\mathrm{supp}(\mathbb{Q}_\theta)$, since re-placing the template on the fixed path generically perturbs both the mean and the permutation variance. Under it $\sigma_{\mathbb{Q}}$ remains the correct standardization precisely because a location shift does not perturb the variance. Then, under Assumption 2, $Z_N = \delta_N + O_p(1)$ with deterministic mean displacement $\delta_N = N\mu_1/\sigma_{\mathbb{Q}}$, the $O_p(1)$ residual being Gaussian only by the alternative-law clause of that assumption and not otherwise derived, and under the no-dominant-term regime where $\sigma_{\mathbb{Q}} \asymp \sqrt{N}\,\sigma_1$ this gives $\delta_N \asymp \sqrt{N}\,\zeta$. Hence the realized $\mathbb{Q}_\theta$-percentile tends to one and the level-α test is consistent as $N \to \infty$; equivalently, the minimum edge detectable at fixed power shrinks as $\zeta_{\min} \asymp N^{-1/2}$.*

The proof is given in Appendix B. Two consequences match the simulation record. First, low-frequency strategies are weakly identified: a decision count as small as $N = 4$ admits only a large $\zeta_{\min}$, so a near-threshold p-value there is read as exploratory rather

than as evidence of absence, and at such N the asymptotics of Assumption 2 are not in force at all, making the reading doubly cautious. Second, because power is governed by N, a non-rejection at large N is informative while a non-rejection at small N is not. The detection rate is borne out in the controlled power study: on a known-skill design the rejection rate at $\alpha = 0.05$ rises monotonically with the per-event signal, from 0.0425 at 0 bp to 0.130 at 5 bp, 0.475 at 10 bp, 0.925 at 20 bp, and 1.000 at 35 bp, with the mean RCSI_z tracking the signal from near zero to 7.23 at 35 bp (Section 6).

5.3. Local power and the exposure direction

The $N^{-1/2}$ boundary of Theorem 3 is sharp: at exactly that rate the test has nondegenerate, computable power. The following is a direct specialization to the Pitman regime, recorded as a corollary.

Corollary 2 (Local power against placement alternatives)**.** *Suppose Assumption 2 holds, and consider the Pitman sequence $\zeta_N = h/\sqrt{N}$ for fixed $h \geq 0$. Then the mean displacement of Theorem 3 is $\delta_N = \sqrt{N}\,\zeta_N + o(1) = h + o(1)$, a bounded shift, so by Slutsky's theorem $Z_N \to_d N(h, 1)$. The one-sided level-α test then has asymptotic power*

$$\beta(h) = \Phi\big(h - z_{1-\alpha}\big),$$

with $z_{1-\alpha}$ the standard normal quantile. Hence $\beta(0) = \alpha$, $\beta(z_{1-\alpha}) = \frac{1}{2}$, and at $\alpha = 0.05$ power reaches 0.80 at $h = z_{0.95} + z_{0.80} \approx 2.49$.

We close by making precise the sense in which the test is orthogonal to exposure. The claim is generic before it is financial: the test is orthogonal to any exposure shared by all structure-matched placements, in the sense that an alternative profiting only through that shared exposure leaves the conditional placement law unchanged and so cannot be detected by the placement contrast, whereas a mean-against-benchmark test is not orthogonal to it and loads on the shared exposure. The exposure direction is thus the set of alternatives that leave the conditional law of placement unchanged: a decision sequence can be profitable purely through the exposure that every structurally matched placement shares on the path, while its placement on the index set carries no edge. Reality Check and SPA are the finance specialization of the mean-against-benchmark comparator: against such alternatives a return-against-zero test loads on the exposure, whereas the placement test does not.

Proposition 2 (Exposure invariance and shared leading noncentrality)**.** *Let SP denote the structure-preserving test and RC/SPA the single-strategy specialization of White's Reality Check [5] and Hansen's SPA test [6], studentized bootstrap tests of mean decision return against a fixed zero benchmark. Call an alternative a* pure-exposure *alternative if it is generated by a data-generating process under which the per-decision drift loads only on an exposure that is shared by every structure-matched placement on the path, with no component that depends on where the decisions are placed on the index set. We take as the defining* exchangeability condition *of the pure-exposure class that, under such a process, the realized placement leaves the conditional placement law $\mathbb{Q}_\theta(\cdot \mid \mathcal{C})$ invariant, so that $T(S_0)$ remains exchangeable with its $\mathbb{Q}_\theta$-resamples given $\mathcal{C}$. We impose this as an explicit assumption on the exposure model rather than deriving it: an exposure common to every member of the placement orbit makes it the natural formalization of "profit comes from holding, not from timing," but it is a modeling restriction, and the proposition's force is conditional on it. Then:* (i) *against a pure-exposure alternative $T(S_0)$ is distributed as a $\mathbb{Q}_\theta$-draw, so the SP test holds its level in finite samples and with no appeal to a central-limit approximation; under Assumption 2 the implied asymptotic noncentrality of Z_N is, in addition, zero. A return-mean test, by contrast, references the realized mean against a fixed benchmark, on which the exposure loads with strictly*

positive noncentrality, so under Assumption 2 its asymptotic power against the pure-exposure alternative tends to one. (ii) *Against a placement (location) alternative under a shared location model in which the per-decision return variance used to standardize both statistics agrees to leading order, SP and RC/SPA carry the same leading noncentrality h under Assumption 2, so neither test dominates the other locally.*

The proof is given in Appendix B. Part (i) has two halves that should not be conflated. The level-control half is an exchangeability statement: a pure-exposure alternative keeps $T(S_0)$ a $\mathbb{Q}_\theta$-draw, so the finite-sample validity guarantee applies verbatim and SP holds its level, with no central-limit scaffolding and no decomposition of the statistic into an exposure part and a placement part. The zero-noncentrality half is the asymptotic translation of "central draw": it concerns the limiting Gaussian location of Z_N and is meaningful only under Assumption 2. Part (ii) is a local-power statement following from Corollary 2: under the shared location model, with the two standardizing variances agreeing to leading order, a shared location shift enters both standardized statistics as the same Pitman noncentrality h. This exposure-invariance statement is the precise sense in which SP and the return-based tests answer different questions: a pure-exposure alternative leaves the $\mathbb{Q}_\theta$-law of placement unchanged, so SP does not reject, while the return-mean tests inherit the full exposure signal. The finite-sample shadow of this contrast is the head-to-head experiment of Section 6: in the structural-exposure world, where a profitable strategy times its decisions at random, SP holds its nominal size at 0.0500 while Reality Check and SPA reject at 0.660 and 0.637 respectively.

6. Simulation study: calibration, power, and competitors

We validate the procedure on known-truth designs before applying it to data. The synthetic worlds are constructed so that the presence or absence of entry-placement skill is fixed by design, which lets us read size off the no-skill worlds and power off the skill worlds. Although the worked statistic is a cumulative trading return, the designs stand in for any setting in which a finite sequence of decisions is placed on an exogenous process: the no-skill worlds are exogenous paths on which decision timing is uninformative by construction, and the skill world injects a genuine timing edge. The reusable artifact under test is the sampler of Algorithm 1; the theory it instantiates is referenced, not re-derived.

6.1. Calibration: size under four no-skill worlds

Table 2 reports the rejection rate at the nominal $\alpha = 0.05$ across four worlds in which decision timing carries no information by construction (IID returns; volatility clustering; a pure drift; and a structural-exposure world in which a strategy is reliably profitable but times its entries at random). A valid placement test should hold size in all four. It does: the rejection rates are 0.05133, 0.04933, 0.05600, and 0.05467, each within roughly one Monte Carlo standard error (≈ 0.006 at 1,500 replications) of the nominal 0.050. The structural-exposure world is the diagnostic case: the strategy is profitable on 92.87% of paths, yet the placement test rejects only 0.05467 of the time, because re-placing the realized template on the fixed path reproduces the same exposure and so leaves the realized statistic a central draw. The Kolmogorov–Smirnov test cannot reject uniformity of the null p-values in any world, with KS p ranging from 0.5153 to 0.7009 (Figure 2).

6.2. Power: monotone detection of injected skill

The lower block of Table 2 traces power against an injected per-event signal in the entry-skill world. Power is monotone in the signal: the rejection rate rises from 0.0425 at zero signal (at the nominal size, as it must), to 0.1300 at 5 bp, 0.4750 at 10 bp, 0.9250 at 20 bp, and 1.0000 at and beyond 35 bp, with the mean RCSI_z climbing in step from near zero to

Table 2. Calibration and power on synthetic known-truth worlds (1,500 replications for the no-skill worlds, 400 for the power curve; 2,000 null simulations per replication; base seed 9100). Reject rates are one-sided at $\alpha = 0.05$; KS p tests uniformity of the null p-value distribution on $[0, 1]$. The first block holds size; the second traces power against an injected per-event signal. The RCSI_z column is a descriptive standardized summary, not a pivotal z-statistic; its central-limit calibration (Assumption 2) is not assured at the small N of the application.

World / signal	Mean act. ret.	Reject @ 5%	KS p (unif.)	Mean RCSI_z
Size (no-skill worlds)				
IID, no skill	−0.00371	0.05133	0.5153	−0.00924
Vol. clustering, no skill	0.00337	0.04933	0.7009	0.00884
Drift only, no skill	0.15531	0.05600	0.5978	0.01788
Structural exposure	0.62207	0.05467	0.5662	0.00724
Power (entry-skill world, per-event signal)				
0 bp	0.00164	0.0425	0.9424	0.01027
5 bp	0.15833	0.1300	7.65e−34	0.69904
10 bp	0.38461	0.4750	2.17e−117	1.73153
20 bp	0.92064	0.9250	4.66e−280	3.79852
35 bp	2.08394	1.0000	0.0	7.22876
50 bp	3.95214	1.0000	0.0	11.18959

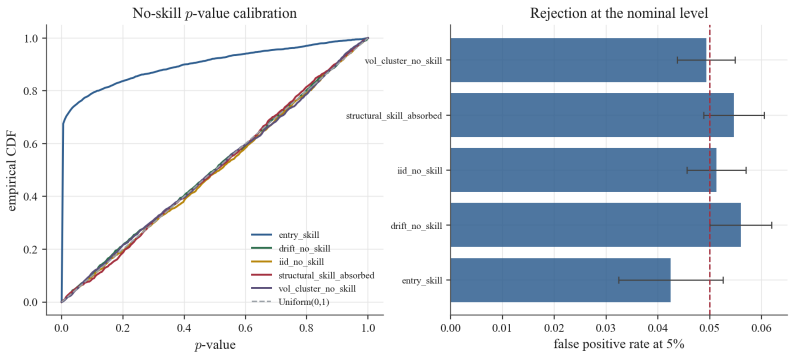


Figure 2. Calibration on synthetic no-skill worlds. The null p-value distribution is close to uniform and the test holds its nominal size, the finite-sample face of the super-uniformity guarantee of Theorem 1(i).

11.19 at 50 bp. This is the finite-sample face of the $N^{-1/2}$ detection rate of Theorem 3: as the standardized edge grows, the realized percentile is driven into the upper tail. The KS uniformity p collapses toward zero as soon as the signal is nonzero, confirming that the null-uniform calibration breaks exactly where it should (Figure 3).

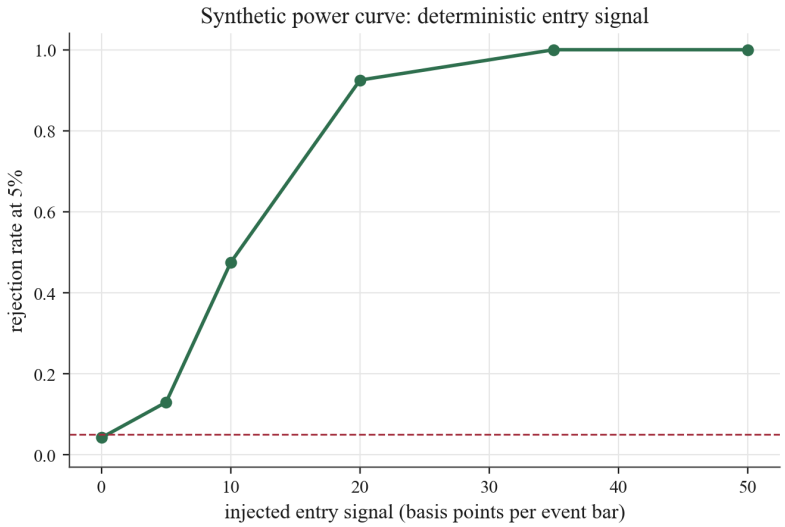


Figure 3. Monotone power against an injected per-event placement edge: the rejection rate rises from nominal size at zero signal to unit power, the finite-sample face of the $N^{-1/2}$ detection rate of Theorem 3.

6.3. Head-to-head with Reality Check and SPA

Table 3 sets the placement test against the single-strategy specializations of White's Reality Check [5] and Hansen's SPA test [6], the canonical data-snooping benchmarks, across the same five worlds. In the three genuinely null worlds the placement test holds close to nominal size (0.0525, 0.0600, 0.0625). The decisive world is structural exposure: there a profitable-but-untimed strategy is rejected by Reality Check 66.00% of the time and by SPA 63.75% of the time, because those benchmarks test the profitability margin rather than the placement margin, while the structure-preserving test holds its nominal size at 0.0500 and the leading-gap path measure rejects 0.0000. This is the finite-sample shadow of Proposition 2(i). On the genuine entry-skill world all four tests reach power 1.0000, consistent with the shared leading noncentrality of Proposition 2(ii): the placement test pays for its exposure-robustness with no loss of power against true timing (Figure 4).

Table 3. Head-to-head rejection rates across five known-truth worlds (400 replications; SP and Path measures, $M = 2{,}000$ draws; RC and SPA, $B = 999$ bootstrap resamples, block length 20). The first four worlds carry no timing skill and a valid timing test should hold size; only the fifth carries genuine entry skill. The Monte Carlo standard error at a true size of 0.05 is ≈ 0.011.

World	SP (placement)	Path (placement)	RC (profit)	SPA (profit)
IID, no skill	0.0525	0.0250	0.1375	0.1275
Drift only, no skill	0.0600	0.0375	0.3875	0.3925
Vol. clustering, no skill	0.0625	0.0350	0.1250	0.1225
Structural exposure	0.0500	0.0000	0.6600	0.6375
Entry placement (skill)	1.0000	1.0000	1.0000	1.0000

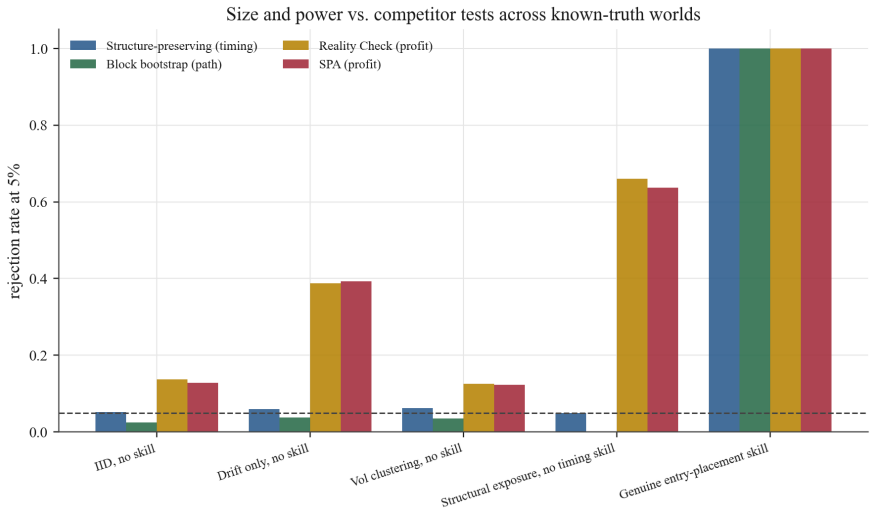


Figure 4. Size and power across five known-truth worlds. In the structural-exposure world (profitable, but with no timing skill) the single-strategy specializations of Reality Check and SPA reject about two-thirds of the time while the structure-preserving placement test holds its nominal size—the finite-sample shadow of Proposition 2(i)—and all tests reach unit power against a genuine placement edge.

6.4. Exact enumeration at small N

Where the feasible orbit is small enough to enumerate, the Monte Carlo p-value can be checked against the exact tail area. For an $N = 4$ Squeeze Breakout instance with durations $[1, 2, 4, 6]$, internal gaps $[834, 4094, 637]$, and external slack 470, the orbit has $3! \times (470 + 1) = 2{,}826$ feasible schedules, spanning the 471 leading offsets and the 3! gap arrangements. Each feasible schedule yields a distinct statistic value on this path because the continuous price grid is fine (resolution 0.000354), so the support is non-atomic here; distinctness follows from the price-grid resolution and the 471 leading offsets, not from the gaps being distinct. The exact one-sided tail for the realized schedule is $q = 0.057325$ (rounding to 0.057; plus-one convention 0.057658). The like-for-like consistency check holds the cost model fixed: a 20,000-draw Monte Carlo under the same deterministic round-trip cost model ($c = 0.000470$) returns 0.059747, within about 1.5 Monte Carlo standard errors (SE ≈ 0.0016) of the exact tail 0.057. This near-agreement of the brute-force exact tail with the same-cost Monte Carlo estimate is consistent with the large-M consistency of $\hat{p}_M$ established in Theorem 1(ii). Separately, and as a sensitivity to the fill model rather than as a third confirming estimate of the same quantity, the canonical stochastic-fill Monte Carlo returns $p = 0.053$; the small gap from 0.057 reflects the change in cost convention, not Monte Carlo error. All fall short of the 5% threshold, so the verdict is a stable non-rejection. Only the $3! = 6$ internal-gap arrangements carry spacing information, so the placement margin is weakly identified at this small N, exactly as the detection-rate analysis predicts.

6.5. A non-finance validation: condition-monitoring windows

To show that the construction references nothing specific to prices, we instantiate Algorithm 1 verbatim in a condition-monitoring setting with a non-return outcome. An exogenous condition process v_t evolves on its own operating timeline; an analyst observes a fixed roster of N monitoring windows of ordered durations $\mathbf{h}$ separated by inter-window gaps $\mathbf{g}^{\text{int}}$, and the outcome is the exogenous condition change captured during the observed windows, $T(\text{schedule}) = \sum_{t\ \text{active}} (v_t - v_{t-1})$. The mapping to the abstraction is exact: the durations are monitoring-window lengths, the gaps are inter-window intervals, the index set is operating time, and "placement skill" is placing observation windows over stretches

where the captured condition change is largest. Because the windows are observational, not corrective interventions, the path-exogeneity assumption remains intact. The structure-preserving null fixes the ordered durations, the internal-gap multiset, and the path, re-randomizing only the leading offset and the gap order (the measure $\mathbb{Q}_{\mathrm{gp}}$); the statistic is not a return and no cost model appears.

Table 4 reports the outcome over 4,000 independent exogenous paths per cell, each with 1,000 admissible re-placements. Under uninformative (randomly placed) scheduling the test holds its nominal 0.050 size across decision counts $N \in \{4, 8, 16, 32\}$, with the mild conservatism at larger N that the discreteness of the placement orbit predicts. When the planner instead schedules against a noisy forecast of the path, power is monotone in scheduling skill: as the forecast noise grows from 0.5 to 4.0 times the path's own scale, rejection falls from 0.996 to 0.206, approaching the nominal size as foresight vanishes. The same algorithm, with a non-return statistic on a non-financial process, is correctly sized and monotonically powered, which is exactly what the generality claim requires.

Table 4. Non-finance validation (condition-monitoring windows): structure-preserving test size and power. Size is the rejection rate at $\alpha = 0.05$ under uninformative placement; power is the rejection rate when the window schedule is chosen against a forecast of the exogenous path corrupted by noise of the stated multiple of the path's scale. 4,000 exogenous paths per cell, 1,000 re-placements each; Monte Carlo standard errors in parentheses.

Experiment	N	Forecast noise	Rejection at 0.05
Size	4	—	0.0505 (0.0035)
Size	8	—	0.0450 (0.0033)
Size	16	—	0.0428 (0.0032)
Size	32	—	0.0490 (0.0034)
Power	16	0.5	0.9955 (0.0011)
Power	16	1.0	0.8433 (0.0057)
Power	16	2.0	0.4492 (0.0079)
Power	16	4.0	0.2057 (0.0064)

7. Application: evaluating trading-strategy timing

We validate the framework on a single, deliberately adversarial finance question: does the calendar *placement* of a strategy's entries carry information, once its realized structure (the ordered holding durations, the inter-trade gaps, the capital weights, and the directions) and the realized price path are held fixed? This is the setting in which the structure-preserving null has real teeth, because a profitable strategy can earn its return entirely through the market exposure that *every* structurally matched placement shares on the path, while its timing adds nothing. Throughout we use the gap-permutation measure $\mathbb{Q}_{\mathrm{gp}}$ (the minimum element of the admissible menu Θ_0 of Definition 4), $M = 5{,}000$ structure-preserving draws per strategy, and the round-trip transaction cost $c = 0.000470$ in return units. Further empirical detail—the agent library, the regime labeling, and the per-asset trade logs—is provided in the supplementary project repository (see Data Availability); here we report only the two results that bear on the general method.

7.1. Exposure invariance: a head-to-head with Reality Check and SPA

Proposition 2 predicts that the structure-preserving test holds level against pure-exposure alternatives satisfying its exchangeability condition, whereas a return-mean test that references a fixed zero benchmark can reject because the benchmark absorbs the exposure premium. We test this directly against the single-strategy specializations of White's Reality Check [5] and Hansen's SPA test [6] across five known-truth worlds (Table 3; 400 replications, $M = 2{,}000$ structure-preserving draws, $B = 999$ bootstrap resamples,

block length 20). In the three genuinely null worlds without structural exposure (IID, drift-only, volatility-clustering), the placement test holds close to its nominal size, rejecting at 0.0525, 0.0600, and 0.0625. The diagnostic world is *structural exposure*: a strategy that is profitable through shared exposure alone, with no timing edge. There White's Reality Check rejects 66.0% of the time and Hansen's SPA 63.75%, while the structure-preserving test holds size at its nominal 0.0500 (and the leading-gap path measure rejects 0.0000). This is precisely the contrast Proposition 2(i) describes: the exposure loads on the fixed benchmark with growing noncentrality, but leaves the conditional placement law invariant. On the genuine entry-skill world (a 35 bp per-event edge) all four tests reach power 1.000, consistent with the shared leading noncentrality of Proposition 2(ii): the placement test pays for its exposure-robustness with no loss of local power against true timing alternatives.

Design of the controlled worlds.

The comparison is reproducible in principle from the construction, not only from the code. There are five known-truth worlds: an IID no-skill world, a drift-only no-skill world, a volatility-clustering no-skill world, the decisive structural-exposure world, and a genuine-entry-skill world. In the decisive structural-exposure world the per-bar returns are drawn IID with a constant positive mean drift and are generated independently of where the trades fall, so the strategy is genuinely profitable through a drift-times-exposure mechanism (its long, unit-weighted positions accumulate the positive drift along the realized path) while its entries are placed uniformly at random under the same structure sampler, giving it zero entry-placement skill by construction. Reality Check and SPA are implemented as a single-strategy test of the mean per-trade net return against a zero benchmark using a stationary block bootstrap [18] ($B = 999$ resamples), evaluated over exactly those trades; the placement test is applied to the identical trades. The whole comparison is run over the same replication count reported for Table 3. The full generator is in the committed code (`synthetic_timing_experiments.py`, driven by `competitor_comparison.py`) and reproduces the reported table. The over-rejection is not engineered: the world is profitable-but-untimed by construction, and Reality Check and SPA reject precisely because the strategy is profitable, which is the point.

7.2. A cross-asset panel under multiplicity control

We then run the test across a convenience sample of 322 strategy-by-asset combinations spanning 47 instruments. At the nominal 5% threshold the panel produces 13 rejections, slightly *fewer* than the 16.1 expected by chance (4.0% versus 5%), and the empirical distribution of p-values departs significantly from uniform (KS statistic 0.114, $p = 0.0004$), with a deficit of small p-values relative to chance (13 observed versus 16.1 expected) and most mass at large p-values (mean and median panel p of 0.567 and 0.580), i.e. in the conservative direction (Figure 5). Under either of two standard multiplicity corrections the panel yields nothing: the Benjamini–Hochberg procedure [19] returns zero discoveries at FDR 0.05, and Bonferroni control returns zero at FWER 0.05 (threshold 1.5528×10^{-4}). The 322 strategy-by-asset cells are drawn from 47 instruments and are therefore not mutually independent, so these Benjamini–Hochberg and Bonferroni counts are applied under an independence approximation, and the small number of path-degenerate, very-low-trade cells are retained but flagged in the count rather than excluded; this dependence only weakens, never creates, the negative finding. The texture of the survivors reinforces the verdict: the smallest p-value in the entire panel, 0.0013997 on DIA, belongs to a *random-entry baseline* rather than a designed strategy, and random baselines occupy two of the top-ten smallest p-values and three of the 13 nominal rejections. The framework's negative finding is therefore stable: after multiplicity control there is no evidence of entry-placement skill under the neutral

gap-permutation measure anywhere in the panel. The panel is evaluated under the neutral gap-permutation measure $\mathbb{Q}_{\text{gp}}$; the measure-invariant extension $p^{\text{inv}} = \max_\theta \hat{p}_\theta$ across the admissible menu is exercised on the synthetic worlds (Table 3) but is not computed for the 322-strategy panel, which we note as a limitation. We flag one caveat the diagnostics make explicit. For small trade counts the orbit of feasible schedules is thin and the null becomes weakly identified; the path-degeneracy scan labels the low-trade rows degenerate (for instance the USDC-USD, INTL, and RSPN strategies), whereas higher-count tests (e.g. VTEB mean-reversion at 29 trades) are not. The exact $N = 4$ enumeration of Section 6.4 (orbit size 2,826, exact one-sided tail $q = 0.057$) shows the same weak identification of the placement margin at the small-N boundary, and is consistent with the canonical stochastic-fill Monte Carlo $p = 0.053$.

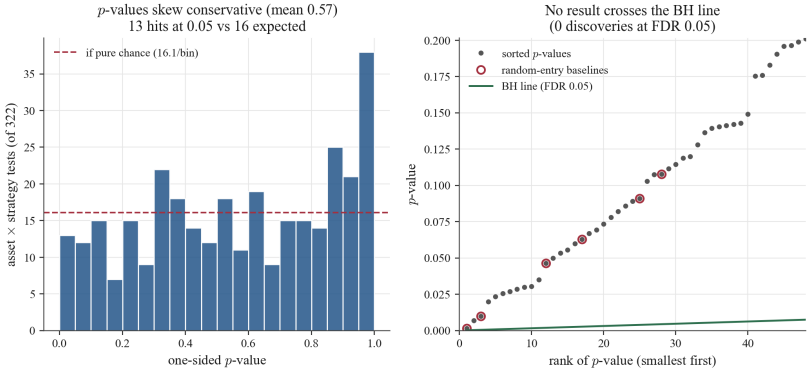


Figure 5. The 322-cell cross-asset panel of p-values under the neutral gap-permutation measure. The distribution is shifted toward large p-values, with a deficit of small p-values relative to uniform (13 observed versus 16.1 expected at the 5% threshold), and nothing survives Benjamini–Hochberg or Bonferroni control.

7.3. Beyond finance

Finance is one validation, not the object of the method. Section 6.5 gives a fully worked non-finance validation, a condition-monitoring window problem in which the same algorithm holds its nominal size across decision counts and is monotonically powered with a non-return statistic. The abstract requirement is a finite sequence of intervals overlaid on an exogenous stochastic process, with a realized structure (durations, inter-decision gaps, magnitudes, signs, and a non-overlap constraint) whose informativeness one wishes to separate from the placement of the intervals on the index set. In event-time analysis the durations, gaps, and non-overlap constraint of a marked point process are preserved while the event placement is re-drawn from an admissible conditional law. In passive monitoring problems, observation windows can be re-placed against a condition path precisely because the windows do not alter that path. In each case the finite-sample validity of Theorem 1, the admissible-menu sensitivity range of Definition 7, and the intersection–union test of Theorem 2 carry over only after the domain-specific exchangeability and exogeneity conditions are defended; the reusable artifact is Algorithm 1, not the trading example.

8. Discussion and limitations

8.1. Scope of validity

The guarantee the procedure earns is narrow and worth stating precisely. The finite-sample validity of Theorem 1(i) is validity against a *sharp* null: the Monte Carlo p-value is super-uniform under the hypothesis that the realized statistic $T(S_0)$ is exchangeable with its $\mathbb{Q}_\theta$-resamples given the conditioning σ-field $\mathcal{C}_s$, which holds in particular when $S_0 \overset{d}{=} \mathbb{Q}_\theta$ given $\mathcal{C}_s$. This is a property of the realized schedule relative to a *declared* reference

law, not a property the procedure confers on arbitrary data. Under any alternative the schedule was produced by a rule, so it is not a $\mathbb{Q}_\theta$-draw, and the test then measures the discrepancy between the realized placement and the reference law rather than certifying that the reference law is the correct counterfactual. We therefore say the test is valid against H_0^θ, and we reserve the word "valid" for that conditional sense; the contrast with the model-X construction of Candès et al. [7], where the conditional law is known by design and exchangeability is thereby guaranteed, is the contrast between a law one knows and a law one names. Concretely, unlike the model-X conditional randomization test and knockoffs of Candès et al. [7], which condition on a model for the covariate distribution, our test conditions on the realized decision structure and the exogenous path and re-randomizes only placement, so it requires no covariate model.

Two further boundaries separate what the verdict rests on from what merely decorates it. First, the verdict rests on exchangeability alone. The asymptotic results, the consistency and $N^{-1/2}$ detection rate of Theorem 3 and the local-power statement of Corollary 2, are conditional on the combinatorial central-limit condition of Assumption 2, which we impose rather than verify: the per-trade contributions are functions of the permutation, not a fixed array, so the classical finite-population central-limit theorem is a motivating analogy and not a sufficient theorem for a statistic built from one fixed, autocorrelated, heavy-tailed path. The standardized effect size $\mathrm{RCSI}_{z,s}$ is accordingly descriptive and only approximately pivotal under Assumption 2; it is never the carrier of the decision. Second, calibration evidence speaks only to exchangeability, not to normality. The Kolmogorov–Smirnov uniformity checks on the no-skill worlds (Table 2, KS p from 0.5153 to 0.7009) probe whether the p-value is uniform, which follows from exchangeability; they cannot confirm the Gaussian shape of Z_N and we do not read them as doing so. Where the orbit is small enough to enumerate, the Monte Carlo estimate is replaced outright by the exact tail area, as in the $N = 4$ Squeeze Breakout case whose 2,826 feasible schedules yield an exact one-sided tail of 0.057 in agreement with the canonical Monte Carlo $p = 0.053$.

8.2. Choosing the conditioning law

The most consequential decision the analyst makes is the choice of conditioning measure $\mathbb{Q}_\theta$ from the admissible menu Θ_0 of Definition 4, and we take that dependence as the central methodological point rather than a nuisance to be hidden. Every admissible measure fixes the realized path $\{P_t\}$ and the structural profile $\mathcal{S} = (\mathbf{h}, \mathbf{g}^{\mathrm{int}}, g^{\mathrm{ext}}, \boldsymbol{\omega}, \mathbf{d})$; the members differ only in which additional features M_θ of the schedule they hold fixed, and that choice is a modeling decision on the same footing as the specification of a regression or the selection of an instrument. It is not pinned down by the data. Each $\mathbb{Q}_\theta$ carries its own null H_0^θ and its own estimand $q_\theta = \mathbb{Q}_\theta\big(T(S) \geq T(S_0) \mid \mathcal{C}_s\big)$, so the object the procedure returns is not a scalar but the sensitivity range $R(\Theta_0) = \{q_\theta : \theta \in \Theta_0\}$ of Definition 7, with its outer bracket $[\min_{\theta\in\Theta_0} q_\theta, \max_{\theta\in\Theta_0} q_\theta]$. The practical guidance follows from the admissibility predicate (AM1)–(AM3) of Definition 5: hold fixed any feature of the schedule that is part of the structure the decision rule was not free to choose, but require that the held-fixed feature be measurable at entry (AM2) and a function of the path and template alone, never of the realized outcome (AM3). The gap-permutation measure, which conditions on the geometry $\mathcal{S}$ alone, is the minimum element of the refinement order and the neutral default; the leading-gap restriction additionally forbids placement in the universal warm-up region; the context-matched measure restricts entries to bars of comparable trailing volatility. Coarser relaxations that break structure preservation, such as the uniform-feasible and slack-redistribution measures, violate (AM1) and lie outside Θ_0.

The width of $R(\Theta_0)$ is the robustness statement. A verdict constant across the admissible class is robust to the partition of the realized record into structure and placement;

a verdict that flips is reported as a range whose elements disagree. We therefore adopt measure invariance, rejection of H_0^θ for every $\theta \in \Theta_0$, as the bar a claim of informative placement must clear, and the intersection–union test of Theorem 2 delivers a valid level-α test of the measure-invariant null through $p^{\text{inv}} = \max_{\theta\in\Theta_0} \hat{p}_M^\theta$. The reason for the conjunctive standard is identification rather than conservatism: because the measures differ precisely in where the structure/placement line is drawn, a rejection under one measure but not another is confounded with the conditioning choice, and only invariance isolates placement from the analyst's partition. Two honest boundaries on this standard are recorded once. By Proposition 1, unanimous rejection is neither sufficient (a confound shared by every admissible measure survives the intersection) nor necessary (a measure can absorb a genuine state-conditional edge into its conditioning set and fail to reject); and by Corollary 1 the power of the intersection–union test is bounded by its least-favorable admissible measure, so the standard is deliberately low-power and its non-rejection is correspondingly weak evidence. The substantive, powered statement is the single-measure one under the neutral gap-permutation measure, against which the real-path positive control is shown to be powered. This worst-case-over-a-menu posture is the inferential analogue of robust risk measurement; we emphasize that Θ_0 is a finite, analyst-enumerated menu and $R(\Theta_0)$ a sensitivity range rather than a sharply identified set in the Manski sense, as Remark 1 sets out.

8.3. Where the method breaks

We set out the failure modes plainly. Some are scope limitations we accept; none moves a reported number.

Near-flat processes drive the null dispersion toward zero.

The statistic is the cumulative return of the template repriced on the realized path, so a randomized decision that lands on a locally flat stretch contributes near zero by construction. For short-duration templates on a path with flat patches many draws return nearly identical statistics, the conditional dispersion σ_s^{sim} approaches zero, and the percentile becomes sensitive to negligible differences. The effect size $\text{RCSI}_{z,s}$ handles exact zero dispersion by definition, being set to zero when $CR_s^{\text{actual}} = \mu_s^{\text{sim}}$ and otherwise undefined, but the p-value under near-zero dispersion warrants a separate guard. We flag any test whose null dispersion is near zero as path-degenerate and decline to interpret its percentile; in the finance application the path-degeneracy scan shows the dispersion bounded away from zero for all but a thin tail tied to dollar-pegged and very-low-trade-count instruments, and none of those degenerate tests is a discovery under multiplicity control.

Tiny decision counts carry little placement information.

When the number of decisions N_s is small the feasible orbit is dominated by the leading-gap draw, which on a trending path is close to a monotone function of calendar position and so behaves as an exposure channel rather than a placement channel. The internal-gap permutation, the component that actually tests spacing, is degenerate at $N_s = 2$ and weak for small N_s. Three penalties then compound: tied or few internal gaps make the test strictly conservative through the right-tail tie convention of Theorem 1(i), the small count inflates the minimum detectable edge through the $N^{-1/2}$ rate of Theorem 3, and the leading-gap draw dominates the estimand. We therefore read small-N_s non-rejections as exploratory on identification grounds, not merely on power grounds, and draw no inference from them. Atomicity of the statistic's support is a separate matter: on a fine continuous price grid (resolution 0.000354) each feasible schedule yields a distinct statistic value, so the support is non-atomic here even at small N_s, as the $N = 4$ enumeration

with 2,826 distinct statistic values shows. This is a property of the price grid and the leading-offset range on this path, not a general implication of the gaps being distinct.

Exchangeability honesty.

The conditioning set can absorb skill, so non-rejection is silence about the conditioned dimensions, not a floor on total skill. The ordered holding-duration vector $\mathbf{h}$ and internal-gap multiset $\mathbf{g}^{\text{int}}$ are themselves outputs of the decision rule, so an edge residing in exit logic, in duration selection, or in the joint choice of when to act and how long to hold is charged to the conditioned structure and is neither manufactured nor recovered by the placement contrast. We accordingly avoid the phrase "lower bound on skill," which overstates what holds, and read non-rejection as the absence of a detectable edge *within the placement margin the chosen measure exposes*. The asymmetry whereby Algorithm 1 keeps durations in realized order while permuting gaps is itself a modeling choice that a reader could replace with a duration-permuting or joint duration-gap-pairing measure; each such measure is a further admissible member of Θ_0 and, by menu monotonicity, would only widen the reported range. Finally, validity is conditional on the path-exogeneity requirement stated up front as Assumption 1: the construction presumes the decisions do not themselves move the exogenous path, a presumption that confines the execution-scheduling application to the small-impact regime, that fails where the path is endogenous (treatment that alters the trajectory, servicing that alters degradation), and that any new domain must verify before the guarantee of Theorem 1 transfers.

9. Conclusions

The relevant object of inference in evaluating a finite sequence of decisions overlaid on an exogenous stochastic process is not the realized payoff but the marginal contribution of the *placement* of those decisions, measured against a counterfactual that carries the same structural burden on the same realized path. We have given that counterfactual an explicit construction. Holding fixed the realized exogenous path and the structural profile, the ordered durations, the inter-decision gaps, the magnitudes, the signs, and the non-overlap constraint, and re-randomizing only the placement from an admissible structure-preserving law, the procedure refers the realized statistic to the distribution it induces and returns a Monte Carlo p-value that is finite-sample valid under the no-placement-skill null (Theorem 1), consistent at the $N^{-1/2}$ detection rate under a combinatorial central-limit condition (Theorem 3, Corollary 2), and exposure-invariant in the sense made precise by Proposition 2. The reusable artifact is the sampler of Algorithm 1: a single, self-contained randomization of one decision schedule that preserves the exact duration profile and the multiset of realized gaps, from which the entire inference is assembled.

The contribution is methodological before it is empirical. Because the conditioning law is a declared modeling choice and not a datum, the procedure reports a sensitivity range of measure-specific verdicts rather than a single number, with measure invariance, survival across the admissible menu via the intersection–union test of Theorem 2, the strongest claim the design supports and the honest bar a placement claim must clear. We validate the procedure on synthetic known-truth worlds, where it holds nominal size across four no-skill designs (reject rates from 0.04933 to 0.05600 against a target of 0.050) and attains monotone power against injected placement skill (from 0.0425 at zero signal to 1.0000), and against competing data-snooping tests, which by design answer the different question of profitability and so reject a profitable-but-untimed strategy at 0.660 (Reality Check) and 0.637 (SPA) precisely where the placement test holds size at 0.050. The finance application is one validation, not the point: across a convenience sample of 322 strategy-by-asset tests on 47 instruments, nothing survives Benjamini–Hochberg or Bonferroni control, profit

is common while entry-placement skill under the neutral gap-permutation measure is absent, and the smallest panel p-value belongs to a random-entry baseline. This negative result is conditional on the realized duration and gap profile that the structure-preserving construction holds fixed; it is not a global no-skill verdict, since an edge residing in duration selection, exit logic, or the joint choice of when to act and how long to hold is charged to the conditioned structure and is neither manufactured nor detectable within the placement margin.

The same three ingredients, a fixed exogenous path, a preserved structural profile, and a randomized decision margin, recur beyond the trading instance, but only where the path is not changed by the decisions being re-placed. The construction applies naturally to event-time and announcement-relative analysis, passive monitoring windows, the timing of factor and allocation rules when market impact is negligible, and the evaluation of learned or agentic policies scored against a fixed environment trajectory. In each case the finite-sample validity, local-power, and exposure-invariance results transfer subject to a domain-specific feasibility and exchangeability check, and the admissible measure family returns as the object the analyst must specify and defend. What the method offers, and what is portable, is a discipline: state the structure to be held fixed, randomize only the decision margin, declare the menu of conditioning laws over which the verdict is to be read, and report the sensitivity range honestly rather than a single number the data do not identify.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, and project administration: A.P. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study does not involve humans or animals.

Informed Consent Statement: Not applicable. This study does not involve humans.

Data Availability Statement: All code, synthetic-world generators, simulation outputs, and the cross-asset panel logs that reproduce every number in this paper are openly available in a FAIR repository: https://github.com/ampatel355/FORTUNAFRAMEWORK (permanently archived at https://doi.org/10.5281/zenodo.20724999). The exact-enumeration, calibration, competitor-comparison, and panel result files, together with the two synthetic-summary files that back the calibration table (`synthetic_null_validation_summary.csv` and `synthetic_power_curve_summary.csv`) and the non-finance validation file `nonfinance_validation_summary.csv` are released under the MIT License (code) and CC BY 4.0 (this article).

Acknowledgments: During the preparation of this manuscript, the author used OpenAI Codex and Claude (Anthropic) for editorial organization, formatting assistance, and journal-specific manuscript preparation. The author reviewed and edited the content as needed and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A Proof of Finite-Sample Validity

This appendix gives the full argument behind the finite-sample guarantee of the structure-preserving placement test. The validity of the verdict rests on exchangeability alone and invokes no central-limit approximation; the asymptotic refinements (consistency, local power, and the exposure-direction comparison) are deferred to Appendix B. Throughout, $\mathcal{C}$ denotes the conditioning σ-field $\mathcal{C}_s$ of Section 2.2, fixing the realized price path $\{P_t\}_{t=0}^{T}$, the structural profile $\mathcal{S} = (\mathbf{h}, \mathbf{g}^{\text{int}}, g^{\text{ext}}, \boldsymbol{\omega}, \mathbf{d})$, and the feasibility constraints

$\mathcal{A}_s$. We write S_0 for the realized schedule and $T(S)$ for the cumulative-return statistic, and we fix a structure-preserving measure $\mathbb{Q}_\theta$ from the admissible menu $\mathfrak{Q}$ of Definition 4.

Appendix A.1 Well-posedness of the admissible nulls

Before stating the validity result we record that each admissible null is well posed, in the sense that the realized schedule is a possible draw from the measure against which it is referred. This is what licenses the exchangeability hypothesis of Theorem 1(i), and, through it, the intersection–union test of Theorem 2. The following is Lemma 2 of the main text.

Proof of Lemma 2. Admissibility condition (AM1) requires that $\mathbb{Q}_\theta$ place all its mass on schedules preserving the exact structural profile $\mathcal{S}$ and the non-overlap constraint $\mathcal{A}_s$. The realized schedule S_0 generated the profile $\mathcal{S}$ and is non-overlapping by construction, so it satisfies every constraint that defines the support of $\mathbb{Q}_\theta$. Conditions (AM2) and (AM3) only restrict which additional features M_θ are held fixed, and by (AM3) those features are functions of $(\{P_t\}, t)$ and the template alone and therefore agree on S_0 and on its $\mathbb{Q}_\theta$-resamples; they shrink the support but never exclude S_0. Hence $S_0 \in \operatorname{supp}(\mathbb{Q}_\theta \mid \mathcal{C})$, so the conditional law of $T(S_0)$ given $\mathcal{C}$ under H_0^θ is the same as that of a generic draw $T(S)$, $S \sim \mathbb{Q}_\theta$, and the exchangeability hypothesis can be asserted. □

Appendix A.2 Proof of Theorem 1

The estimand attached to $\mathbb{Q}_\theta$ is the upper-tail probability $q_\theta = \mathbb{Q}_\theta\big(T(S) \geq T(S_0) \mid \mathcal{C}\big)$; the Monte Carlo p-value of Algorithm 1 estimates it with the plus-one correction of Dwass [16] and Phipson and Smyth [13].

Proof of Theorem 1. *(i) Validity.* Condition on $\mathcal{C}$. By hypothesis $T(S_0)$ is exchangeable with $\{T(S_m)\}_{m=1}^M$; this is guaranteed whenever $S_0 \overset{d}{=} \mathbb{Q}_\theta$ given $\mathcal{C}$, since then $S_0, S_1, \ldots, S_M$ are themselves exchangeable and T is a fixed measurable map, so the values $T(S_0), T(S_1), \ldots, T(S_M)$ inherit exchangeability. Lemma 2 ensures this hypothesis is well posed, S_0 lying in the support of $\mathbb{Q}_\theta$. Absent ties, the rank of $T(S_0)$ among the $M+1$ values $\{T(S_m)\}_{m=0}^M$ is therefore uniform on $\{1, \ldots, M+1\}$. Writing R for the number of resamples with $T(S_m) \geq T(S_0)$, the upper-tail rank is $R+1$ and $\hat{p}_M = (1+R)/(M+1)$ is uniform on the grid $\{1/(M+1), \ldots, 1\}$, so $\Pr(\hat{p}_M \leq \alpha \mid \mathcal{C}) = \alpha$ on the grid. Ties only ever enlarge the numerator $1+R$: a resample with $T(S_m) = T(S_0)$ is counted into the right tail, which can raise $\hat{p}_M$ but never lower it. Hence ties move the bound strictly toward conservatism and can never inflate the rejection probability, giving $\Pr(\hat{p}_M \leq \alpha \mid \mathcal{C}) \leq \alpha$ in general. This is the standard finite-sample validity of Monte Carlo randomization tests [13–16], here conditioned on the structural profile; it uses exchangeability alone and no distributional approximation.

(ii) Consistency in M. Fix S_0, and do not assume H_0^θ. Given $\mathcal{C}$ the draws $S_1, S_2, \ldots$ are i.i.d. from $\mathbb{Q}_\theta$, so the indicators $\mathbf{1}\{T(S_m) \geq T(S_0)\}$, $m \geq 1$, are i.i.d. Bernoulli with mean $q_\theta = \mathbb{Q}_\theta(T(S) \geq T(S_0) \mid \mathcal{C})$. The strong law of large numbers gives $M^{-1} \sum_{m=1}^M \mathbf{1}\{T(S_m) \geq T(S_0)\} \to q_\theta$ almost surely, and the deterministic factors converge to the same limit, so $\hat{p}_M \to q_\theta$ almost surely. This limit does not invoke H_0^θ: it holds for every fixed schedule and is the engine that lets the simulated p-value track the population tail probability in the consistency and power results of Appendix B. □

The separation between the two parts is deliberate. Part (i) is the entire basis of the reported verdict and is finite-sample: it requires only that the realized placement be exchangeable with its structure-preserving resamples, a property the sampler of Algorithm 1 enforces by construction. Part (ii) is a statement about Monte Carlo error in the number of draws M and is used only as a bridge in the asymptotic analysis. In the implementation we

set $M = 5{,}000$ draws per strategy; the exact small-N enumeration of Section 6.4 confirms that this Monte Carlo p-value agrees with the exact tail probability to within Monte Carlo error.

Appendix B Asymptotics

This appendix collects the large-N analysis: the consistency of the test together with its $N^{-1/2}$ detection rate, the local-power function against Pitman placement alternatives, and the exposure-invariance comparison with pure-exposure alternatives. None of these statements is needed for the validity of the verdict, which is finite-sample (Appendix A); they characterize when the test has power and how it relates to the studentized return-mean tests of White [5] and Hansen [6]. Throughout we work under a fixed admissible measure $\mathbb{Q}_\theta$, write $T(S) = \log(1 + CR)$ for the log-return statistic, and standardize by the conditional $\mathbb{Q}_\theta$ mean and standard deviation $(\mu_\mathbb{Q}, \sigma_\mathbb{Q})$, $Z_N = (T(S_0) - \mu_\mathbb{Q})/\sigma_\mathbb{Q}$. The analysis is conditional throughout on the combinatorial central-limit condition stated as Assumption 2 in the main text.

Appendix B.1 Proof of Theorem 3

Proof of Theorem 3. Under Assumption 2, $Z_N = (T(S_0) - \mu_\mathbb{Q})/\sigma_\mathbb{Q} \to_d N(0,1)$ when H_0^θ holds. The location-shift alternative adds a *deterministic* mean to each contribution: $E[\log(1 + CR)]$ moves from $\mu_\mathbb{Q}$ to $\mu_\mathbb{Q} + N\mu_1$, while the centering and scaling $(\mu_\mathbb{Q}, \sigma_\mathbb{Q})$ are fixed functionals of $\mathbb{Q}_\theta$ and $\mathcal{C}$. Subtracting the null mean and dividing by $\sigma_\mathbb{Q}$ therefore displaces the standardized statistic by the deterministic amount $\delta_N = N\mu_1/\sigma_\mathbb{Q}$, leaving an $O_p(1)$ fluctuation governed by the same limit law; this last step is licensed by the alternative-law clause of Assumption 2, under which $Z_N - \delta_N \to_d N(0,1)$, so the residual is genuinely $O_p(1)$ and Gaussian rather than assumed so. Writing $\delta_N = \sqrt{N}\,(\mu_1/\sigma_1) \cdot (\sqrt{N}\,\sigma_1/\sigma_\mathbb{Q})$ and using $\sigma_\mathbb{Q} \asymp \sqrt{N}\,\sigma_1$ (the no-dominant-term regime of Assumption 2(a), under which the variance accumulates linearly in N) gives $\delta_N \asymp \sqrt{N}\,\zeta \to \infty$. The upper-tail probability $q_\theta = \Pr_\mathbb{Q}(Z \geq Z_N)$ then tends to zero, and Theorem 1(ii) makes $\hat{p}_M$ track it. Fixing the noncentrality $\sqrt{N}\,\zeta$ at the value delivering a target power inverts to $\zeta_{\min} \asymp N^{-1/2}$. No circularity arises: validity (Theorem 1) is used only through part (ii), the deterministic $M \to \infty$ limit of $\hat{p}_M$, which holds for any fixed S_0 and does not presuppose the alternative. □

Appendix B.2 Proof of Corollary 2

Proof of Corollary 2. By Theorem 3 the deterministic displacement is $\delta_N = N\mu_1/\sigma_\mathbb{Q} = \sqrt{N}\,\zeta_N \cdot (\sqrt{N}\,\sigma_1/\sigma_\mathbb{Q}) \to h$ along $\zeta_N = h/\sqrt{N}$, because the second factor tends to 1 under the no-dominant-term scaling of Assumption 2(a). Adding a deterministic mean shift $\delta_N \to h$ to a statistic whose centered form converges to $N(0,1)$ under the alternative clause of Assumption 2 shifts the limit to $N(h,1)$ by Slutsky's theorem. Therefore $\Pr(Z_N \geq z_{1-\alpha}) \to 1 - \Phi(z_{1-\alpha} - h) = \Phi(h - z_{1-\alpha})$. The three stated values follow by substitution: $h = 0$ gives $\Phi(-z_{1-\alpha}) = \alpha$; $h = z_{1-\alpha}$ gives $\Phi(0) = \frac{1}{2}$; and $\Phi(z_{0.80}) = 0.80$ with $z_{0.95} \approx 1.645$, $z_{0.80} \approx 0.842$ gives $h \approx 2.49$. □

The detection rate $\zeta_{\min} \asymp N^{-1/2}$ is borne out by the synthetic power curves of Section 6: reject rates climb monotonically with the per-event signal, from nominal size at zero signal to unit power once the standardized edge is large, while size is held at the nominal level across the no-skill worlds.

Appendix B.3 Proof of Proposition 2

Proof of Proposition 2. *(i)* By definition, under a pure-exposure alternative the conditional placement law $\mathbb{Q}_\theta(\cdot \mid \mathcal{C})$ is unchanged, so the realized schedule S_0 is exchangeable with

$S_1, \ldots, S_M \sim \mathbb{Q}_\theta$ given $\mathcal{C}$: profitability comes from the exposure common to all structurally matched placements on the path, not from where the trades sit, so the realized statistic $T(S_0)$ is distributed as a $\mathbb{Q}_\theta$-draw. Theorem 1(i) then gives $\Pr(\hat{p}_M \leq \alpha \mid \mathcal{C}) \leq \alpha$: the SP rejection probability stays at the nominal level. This level control is finite-sample and uses only exchangeability; we do not decompose $T(S_0)$ into an exposure component and a placement component, and the conclusion needs no such additive split. Under Assumption 2 the asymptotic statement is then immediate: a central draw has limiting law $N(0,1)$, so the noncentrality of Z_N is zero. The return-mean statistic $\sqrt{N}\,\bar{d}$, by contrast, is centered at the realized mean trade return, which the exposure raises above the fixed zero benchmark; under Assumption 2 its noncentrality is therefore strictly positive and grows like $\sqrt{N}$, so RC/SPA reject with probability tending to one. *(ii)* Under a shared per-trade location shift the placement carries a common mean displacement; provided the per-trade return variance standardizing the two statistics agrees to leading order under that model, Corollary 2 delivers it into the z-standardized SP statistic and the studentized return-mean statistic as the same leading noncentrality h, so to first order their local powers coincide and neither dominates. That equal-leading-variance condition is exactly what licenses the *same* h rather than merely two positive noncentralities. Relative efficiency beyond this shared leading noncentrality is left for future work: equality of the full local-power functions would require matching the two variance functionals against placement, which holds only under the shared-location model. □

The exposure-invariance result of part (i) is the formal counterpart of the simulation headline: in the structural-exposure world (profitable, but with no timing skill) the return-mean tests of White [5] and Hansen [6] reject roughly two-thirds of the time (66.0% and 63.75% respectively), because their benchmark is a profitability benchmark, while the placement test holds its nominal size (0.0500). Part (ii) shows this separation costs nothing against genuine placement alternatives, where all four tests reach unit power.

References

1. Fisher, R.A. *The Design of Experiments*; Oliver & Boyd: Edinburgh, 1935.
2. Lehmann, E.L.; Romano, J.P. *Testing Statistical Hypotheses*, 3rd ed.; Springer: New York, 2005.
3. Manski, C.F. *Partial Identification of Probability Distributions*; Springer Series in Statistics, Springer: New York, 2003.
4. Tamer, E. Partial identification in econometrics. *Annual Review of Economics* **2010**, *2*, 167–195. https://doi.org/10.1146/annurev.economics.050708.143401.
5. White, H. A reality check for data snooping. *Econometrica* **2000**, *68*, 1097–1126.
6. Hansen, P.R. A test for superior predictive ability. *Journal of Business & Economic Statistics* **2005**, *23*, 365–380.
7. Candès, E.; Fan, Y.; Janson, L.; Lv, J. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B* **2018**, *80*, 551–577.
8. Ernst, M.D. Permutation methods: A basis for exact inference. *Statistical Science* **2004**, *19*, 676–685.
9. Besag, J.; Clifford, P. Generalized Monte Carlo significance tests. *Biometrika* **1989**, *76*, 633–642.
10. Aldous, D.J. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*; Springer: Berlin, 1985; Vol. 1117, *Lecture Notes in Mathematics*, pp. 1–198.
11. Pesarin, F.; Salmaso, L. *Permutation Tests for Complex Data: Theory, Applications and Software*; John Wiley & Sons: Chichester, 2010.
12. Bailey, D.H.; López de Prado, M. The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management* **2014**, *40*, 94–107.
13. Phipson, B.; Smyth, G.K. Permutation *p*-values should never be zero: Calculating exact *p*-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* **2010**, *9*, Article 39.

14. Hope, A.C.A. A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society: Series B* **1968**, *30*, 582–598.
15. Hemerik, J.; Goeman, J. Exact testing with random permutations. *TEST* **2018**, *27*, 811–825.
16. Dwass, M. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics* **1957**, *28*, 181–187.
17. Hoeffding, W. A combinatorial central limit theorem. *The Annals of Mathematical Statistics* **1951**, *22*, 558–566.
18. Politis, D.N.; Romano, J.P. The stationary bootstrap. *Journal of the American Statistical Association* **1994**, *89*, 1303–1313.
19. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **1995**, *57*, 289–300.

https://doi.org/10.3390/stats1010000