MDPI

Article

Separating Backtest Profitability from Entry-Timing Skill: A Structure-Preserving Randomization Test

Aryan Patel

[1] Independent Researcher; ampatel355@gmail.com

Abstract

Profitable trading rules can earn returns through market exposure without adding value through entry timing. This paper evaluates that distinction with a structure-preserving randomization test: conditional on a declared placement law, the realized trade structure and price path are held fixed while only the calendar placement of trades is re-randomized. Under the sharp null that the realized placement is exchangeable with those re-placements, the plus-one Monte Carlo p-value is finite-sample valid and measures whether the observed schedule is unusually favorable relative to structurally matched alternatives. In synthetic worlds, White's Reality Check and Hansen's SPA reject a profitable-but-untimed exposure strategy about two-thirds of the time (0.660 and 0.637), while the placement test holds nominal size (0.050) and retains power against genuine timing. On an eleven-rule gold-futures panel and a 322-test cross-asset scan, profitability is common but no rule shows robust, measure-invariant entry-placement skill after multiplicity control. The result is not a claim that the rules are unprofitable; it is evidence that the timing component should not be priced as active skill without clearing a structure-matched counterfactual.

Keywords: trading-strategy evaluation; market timing; timing skill; data snooping; multiple testing; conditional randomization; backtest overfitting; Reality Check; superior predictive ability; technical trading rules

JEL Classification: C12; C15; C58; G11; G14; G17

1. Introduction

A profitable backtest is routinely read as evidence that a strategy's decisions were well timed, but common data-snooping tests answer a different question. In a controlled world where a strategy holds a real structural exposure that makes it profitable but places its entries with no entry-placement skill, White's Reality Check and Hansen's SPA reject about two-thirds of the time, while a structure-preserving placement test holds its nominal size. The strategy is genuinely profitable; the tests are answering their own profitability question. The allocation problem is that profitability may be the wrong object when the investor needs to know whether the entry and exit timing deserves an active fee.

The profit a strategy earns on a single price path confounds three economically distinct things: the structural exposure its positions imply, the drift of the market it happened to trade, and the quality of its timing, whether its entries and exits landed at informative moments. A rule can be reliably profitable while supplying no entry-placement skill, simply by holding a rising exposure. The standard data-snooping tools, Reality Check (White, 2000), SPA (P. R. Hansen, 2005), the data-snooping framing of Sullivan, Timmermann, and White (Sullivan et al., 1999), and the multiple-testing critique of Harvey, Liu, and

Received:
Accepted:
Published:

Zhu (Harvey et al., 2016), all take a strategy's realized decision sequence as given and ask whether its profitability survives selection. That is a useful question, but it is not the same as asking whether the calendar placement of trades was informative.

This paper makes the conflation visible, quantifies its practical effect, and supplies an instrument for the narrower timing question. We hold the realized trade structure and the exogenous price path fixed and re-randomize only the calendar placement of the trades, comparing the realized performance to its distribution over admissible re-placements. Under the declared null that the realized placement is exchangeable with placements drawn from the chosen law, the resulting plus-one Monte Carlo p-value is finite-sample valid and requires no model for returns.

Contributions.

- **Two questions, one trade log.** On worlds with known ground truth, Reality Check and SPA reject a profitable-but-untimed structural-exposure strategy 66.0% and 63.7% of the time—correctly, as verdicts on profitability—while the placement test, which asks the distinct entry-timing question, holds its nominal 0.050 size. The tests are complementary rather than competing: in exactly the case where an allocator must separate exposure from timing, a profitability verdict and a timing verdict diverge by more than an order of magnitude in rejection rate (Section 3.1).
- **A timing-focused placement test.** A conditional randomization test that holds the realized trade structure and price path fixed and re-randomizes only calendar placement, with a finite-sample-valid p-value under the declared exchangeability law and no return model (Section 2; validity in Appendix B).
- **Model risk made explicit.** A declared family of admissible re-randomization measures, with the verdict reported as a sensitivity range rather than under a single favorable choice (Section 2).
- **Applied evidence under multiplicity control.** A deep eleven-rule gold-futures panel and a cross-asset robustness scan under Benjamini–Hochberg (Benjamini & Hochberg, 1995) and Bonferroni control (Sections 3.3–3.4).

Related literature.

The paper draws on three strands. On the evaluation of trading rules under data snooping, we build on White's Reality Check (White, 2000), Hansen's superior-predictive-ability test (P. R. Hansen, 2005), the technical-trading-rule study of Sullivan et al. (1999), and the stepwise multiple-testing procedures of Romano and Wolf (2005), whose resampling engines are the stationary and block bootstraps (Künsch, 1989; Politis & Romano, 1994). On multiple testing and backtest overfitting in empirical finance, we draw on Harvey et al. (2016), false-discovery-rate control (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Storey, 2002), the deflated Sharpe ratio and the probability of backtest overfitting (Bailey et al., 2017; Bailey & López de Prado, 2014), and the machine-learning treatment of López de Prado (2018), set against market-efficiency and adaptive-markets framing (Fama, 1970; Lo, 2004) and the time-series-momentum exposure that drives much realized profitability (Moskowitz et al., 2012). The instrument itself is conditional randomization inference: classical permutation and randomization testing (Ernst, 2004; Fisher, 1935; Lehmann & Romano, 2005; Pesarin & Salmaso, 2010), Monte Carlo significance tests with the plus-one correction (Besag & Clifford, 1989; Dwass, 1957; Hope, 1968; Phipson & Smyth, 2010), exact testing with random permutations (Hemerik & Goeman, 2018), exchangeability theory (Aldous, 1985), and the model-X conditional randomization test of Candès et al. (2018), whose conditioning law is known by design where ours is analyst-declared; the conditioning is also kin to selective and post-selection inference (Berk et al., 2013; Fithian

et al., 2014; Lee et al., 2016), though we condition on a declared structure rather than on a selection event.

Findings preview.

A profitability test and the placement test diverge sharply on profitable-but-untimed strategies—Reality Check and SPA reject about 66% of the time, answering the profitability question, while the placement test holds its 5% entry-timing size; on gold no rule clears entry-placement significance under the neutral measure, four volatility-filtered rules reject only under volatility-state-matched nulls (a measure-dependent, non-robust effect), and nothing survives multiplicity control across the cross-asset tests. Profit is common; isolable entry-placement skill is not. Section 2 sets out the data, the test, and the measure family; Section 3 reports the head-to-head, the gold panel, and the cross-asset scan; Section 4 discusses interpretation, costs, and limits; Section 5 concludes, with proofs in the appendices.

2. Materials and Methods

Our object of study is not whether a trading rule made money but whether the *timing* of its decisions made money. The two are easy to confuse and economically distinct: a rule can profit by holding a structural exposure, by riding a drifting market, or because its entries and exits landed at informative moments. We therefore build the method around a single discipline, namely holding everything about a strategy fixed except the calendar placement of its trades, and asking whether that placement was special under a declared placement law. This section lays out the data (Section 2.1), the rules and the trade-structure object they produce (Section 2.2), the conditional randomization test (Section 2.3), the conditional placement estimand (Section 2.4), the effect-size statistics we report (Section 2.5), the family of re-randomization measures we treat as model risk (Section 2.6), the multiplicity control for the cross-asset panel (Section 2.7), and the Reality Check and SPA benchmarks we run head-to-head (Section 2.8). All validity theory and proofs are deferred to Appendix B; the body gives intuition and a schematic.

2.1. Data

The primary instrument is the front-month gold futures contract (GC=F). We use a common evaluation window running from 2002-03-04 to 2026-04-02, which after applying a universal signal warm-up leaves 6,049 daily bars for the strategy panel and 6,169 bars for the buy-and-hold benchmark. Over this window gold delivered a cost-free buy-and-hold cumulative return of 15.944 (a gross multiple of roughly 16.9×, inclusive of roll), with an annualized Sharpe ratio of 0.733 and a maximum drawdown of −44.4%. The benchmark is computed over the full 6,169-bar series and the strategy panel over the 6,049-bar series, the 120-bar difference being the realized gap before the first strategy signal can fire. Gold is a deliberate choice for a timing study: it carries a strong, persistent structural drift over the sample, exactly the kind of exposure that inflates profitability statistics without implying any entry-placement skill, so it is a setting where conflating profit with skill is most tempting and most costly.

Beyond the gold panel we assemble a cross-asset universe of 47 instruments spanning equity index proxies, single-name equities, commodity futures, currencies, and a stablecoin, sampled at the same daily frequency. Running the eleven rules across this universe yields 322 strategy-asset tests, which we treat jointly under multiplicity control (Section 2.7). The universe is intentionally heterogeneous, including near-degenerate price paths such as the USDC-USD stablecoin, because such cases stress the test's null-dispersion behavior and are diagnostic rather than incidental. Per-instrument provenance and the full price series are listed in the Data Availability statement.

2.2. Trading Rules and the Trade-Structure Object

We evaluate eleven trading rules drawn from the standard practitioner repertoire: a trend-pullback rule, a breakout rule combining volume and momentum confirmation, a mean-reversion rule with a volatility filter, a trend-momentum verification rule, an ADX trend-following rule, an uptrend oversold-reversion rule, a volatility-squeeze breakout, a Connors RSI(2) pullback, a Donchian trend-reentry, a turn-of-month seasonality rule, and a random-entry baseline. The random baseline is included as a calibration anchor: it has no economic content, so any test that flags it as skilled is mislabeling luck. Each rule is run with a fixed parameterization and a uniform signal warm-up, and trades are subject to an expected round-trip transaction cost of $c = 0.000470$, applied identically to the realized trades and to every simulated schedule. The parameterizations are the standard values associated with each named rule family (for instance the 200-period moving average, the Connors RSI(2), and the Donchian channel) and are held fixed across the entire study; in particular the same configuration is applied to every one of the 47 instruments in the cross-asset panel without per-asset re-optimization, so that panel is out of sample with respect to parameter choice.

Running a rule on a price series produces a realized track record, and from that track record we extract a *trade-structure profile*

$$S = \left(h,\ g^{\text{int}},\ g^{\text{ext}},\ \omega,\ d\right), \tag{1}$$

where h is the *ordered* sequence of trade durations (holding periods), g^{int} the multiset of internal gaps between consecutive trades, g^{ext} the external (leading and trailing) gap budget, ω the position sizes, and d the trade directions. The ordering of the durations in h is load-bearing and is preserved throughout: h is an ordered sequence of holding periods, not a multiset, so that the realized shape of the schedule, namely which trade is long and which is short and which is held briefly and which at length, is carried intact into the null. The internal gaps g^{int} enter as a multiset because re-ordering the spacing between trades is part of the placement degree of freedom; the durations themselves are never re-ordered. What S does *not* encode is when the schedule sits on the calendar. That placement is precisely the degree of freedom the test will interrogate. The number of trades N per rule ranges from $N = 4$ (Volatility Squeeze Breakout) to $N = 215$ (random baseline) on gold, and the same extraction is applied unchanged across the cross-asset universe.

2.3. The Conditional Randomization Test

The intuition is a thought experiment. Take a strategy's realized trades, freeze the exogenous price path exactly as history delivered it, and freeze the trade structure S of Equation (1), then ask: of all the ways this same schedule could have been slid along the calendar, did the realized placement perform unusually well? If timing carries no information, the realized placement is just one admissible placement among many and should look ordinary against them. If the entries were genuinely well timed, the realized placement should sit in the right tail. Figure 1 renders this schematically: the top track shows the realized trade schedule against the fixed price path; the lower tracks show admissible re-placements that share the identical ordered durations, internal-gap multiset, sizes, directions, and non-overlap constraint, differing only in calendar position; a histogram on the right collects the performance statistic over re-placements with the realized value marked.

Concretely, the procedure is:

1. **Fix** the realized price path and the trade-structure profile S (ordered durations, internal-gap multiset, weights, directions, non-overlap).

2. **Draw** M alternative schedules from a re-randomization law Q that re-places the schedule on the calendar while honoring S (Section 2.6).
3. **Recompute** the realized performance statistic T for each drawn schedule against the same fixed path, yielding a null distribution $\{T_1, \ldots, T_M\}$.
4. **Compare** the realized T_{obs} to that distribution.

Held fixed: the price path, the ordered durations, the internal-gap multiset, the sizes, the directions, and non-overlap. Randomized: only the calendar placement, namely the leading gap and the ordering of the internal gaps. Structural exposure and the realized path affect both the observed statistic and the null distribution. The inferential question is whether the observed placement is extreme relative to placements with the same structure on the same path, not whether realized profit can be uniquely decomposed into exposure and timing components.

Letting $k = \#\{m : T_m \geq T_{\text{obs}}\}$ be the number of re-placements matching or beating the realized value, we report the Monte Carlo p-value

$$p = \frac{1+k}{M+1}. \tag{2}$$

Under the null that timing is uninformative, the realized placement is exchangeable with the admissible alternatives drawn under Q, so Equation (2) is a finite-sample-valid p-value. The $+1$ in numerator and denominator is the standard correction that keeps the test valid for finite M. When N is small the placement orbit can be enumerated rather than sampled: for the $N = 4$ Squeeze Breakout the orbit comprises $3! \times (470 + 1) = 2{,}826$ feasible schedules (six internal-gap permutations crossed with leading gaps in $\{0, \ldots, 470\}$), all yielding distinct statistic values, against which the one-sided tail probability is 0.057 under a deterministic cost model and the canonical Monte Carlo value is 0.053 under the stochastic-fill model. We use M in the thousands for the main panels and report enumeration where it is feasible. The exchangeability argument and the formal validity statement are proved in Appendix B.

2.4. The Conditional Placement Estimand

The test does not require a literal algebraic decomposition of realized returns into invariant exposure and timing terms. It defines a conditional comparison. Given the fixed conditioning information $\mathcal{C}_s = \sigma(S, \{P_t\})$, the placement law Q induces a distribution of the statistic $T(S)$ over feasible schedules that preserve the realized structure. The right-tail estimand is

$$q_Q(S_0) = Q\{T(S) \geq T(S_0) \mid \mathcal{C}_s\},$$

and the associated effect-size estimand is

$$\Delta_Q(S_0) = T(S_0) - \mathbb{E}_Q[T(S) \mid \mathcal{C}_s].$$

The p-value in Equation (2) estimates $q_Q(S_0)$, while RCSI in Equation (3) estimates $\Delta_Q(S_0)$ in return units. Both quantities are conditional on the chosen Q. Validity follows from rank exchangeability under the sharp null H_0^Q that the realized placement is itself a draw from Q given $\mathcal{C}_s$; the data do not prove Q correct. Structural exposure and the realized path enter the observed and re-randomized schedules alike, so the comparison asks whether the realized calendar placement was unusually favorable relative to structurally matched alternatives, not whether all exposure effects have disappeared from the statistic.

2.5. The Statistic: RCSI and $RCSI_z$

The performance statistic T is the realized cumulative return of the strategy, and the headline inferential object is the exact null percentile of $T(S_0)$ within its placement orbit, equivalently the p-value of Equation (2): it is fully nonparametric, finite-sample valid, and the effect size we report and interpret in the gold table. Alongside it we report the Realized-versus-Counterfactual Skill Index, the raw gap between the realized return and the average return over re-placements,

$$\text{RCSI} \;=\; r_{\text{realized}} \;-\; \bar{r}_{\text{sim}}, \tag{3}$$

which estimates the conditional placement advantage $\Delta_Q(S_0)$ of Section 2.4 in return units. Its standardized counterpart divides by the dispersion of the null, $\text{RCSI}_z = \text{RCSI} \,/\, \text{sd}(r_{\text{sim}})$, and we demote it to an appendix diagnostic (Appendix B.3) rather than a headline figure. RCSI_z is approximately pivotal only where a no-dominant-term (Lindeberg) condition holds, so that no single trade carries a non-negligible share of the return budget; it fails exactly where one or a few trades dominate, as at the four-trade Squeeze Breakout, and there the standardization is not interpretable. We therefore read the exact null percentile, never RCSI_z, as the primary measure of how far the realized placement sits from its null, and we scan for path-degeneracy in the cross-asset panel rather than reading a large RCSI_z as evidence of skill.

2.6. The Re-Randomization Measure as Model Risk

The re-randomization law Q is not unique, and we treat that non-uniqueness as model risk rather than papering over it, in the spirit of robust and coherent risk measurement (Artzner et al., 1999; Cont, 2006; Gilboa & Schmeidler, 1989; L. P. Hansen & Sargent, 2008), reporting a sensitivity range over an admissible family read in the spirit of partial identification (Manski, 2003; Tamer, 2010) rather than a single favorable number. Any admissible Q must preserve the trade-structure profile S and the non-overlap constraint; within that constraint several choices are defensible. Our neutral baseline measure factors the placement as a uniform draw of the leading gap on $\{0, \ldots, g^{\text{ext}}\}$ together with a uniform draw over permutations of the internal-gap multiset g^{int}. We also consider a leading-gap restriction that forbids any simulated first entry inside the 200-bar window of the longest indicator lookback (the 200-period moving average), a context-matched measure that conditions re-placements on local market context, and a volatility-state / regime-preserving measure that re-places trades only into bars sharing the realized volatility regime. Each measure encodes a different null about what counts as a fair comparison.

We make the decision rule explicit: the neutral gap-permutation measure is the canonical null against which the headline verdict is read, because it is the member of the family that conditions on the trade structure alone and adds no further conditioning on market context, so it is least likely to encode the strategy's own signal into the comparison set. The other measures are reported as a declared sensitivity analysis. Crucially, each measure defines a *separate* sharp hypothesis H_0^θ (Appendix A), and the procedure does *not* combine them into a single level-α test: we report the full per-strategy p-value matrix across measures (Table 4) and describe a result as "robust" only as a heuristic, meaning it survives the family rather than a single favorable choice of Q. "Robust" in this sense is a descriptive shorthand, not a multiplicity-corrected joint level statement, and a rejection that appears under one admissible measure and vanishes under the canonical neutral one is reported as measure-dependent, not as standalone evidence of timing skill.

2.7. Multiplicity Control

The cross-asset panel runs 322 strategy-asset tests, so individual p-values cannot be read at face value. We control multiplicity two ways. The Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) controls the false discovery rate and is our primary screen for whether any entry-placement skill survives selection. The Bonferroni correction controls the family-wise error rate and gives a conservative second view; with 322 tests the Bonferroni threshold is $\alpha/322 \approx 1.55 \times 10^{-4}$. Because the 322 cells reuse a common set of eleven rule definitions across 47 instruments and are therefore dependent rather than independent, we additionally report the Benjamini–Yekutieli procedure (Benjamini & Yekutieli, 2001), the dependence-robust counterpart of Benjamini–Hochberg that controls the false discovery rate under arbitrary dependence. We also compare the observed count of nominally significant tests against its expectation under a uniform null and assess the full p-value distribution against uniformity with a Kolmogorov-Smirnov test, so that a panel-level reading does not hinge on any single cell. This panel-level multiplicity control addresses search across the strategy-asset grid; it is distinct from the within-strategy placement inference of Section 2.3.

2.8. Benchmarks: Reality Check and SPA

To show what the placement test isolates that profitability tests do not, we run White's Reality Check (White, 2000) and Hansen's SPA test (P. R. Hansen, 2005) head-to-head against our structure-preserving (SP) test and a path-resampling block bootstrap. These benchmarks are standard data-snooping tools (Romano & Wolf, 2005; Sullivan et al., 1999), and they are designed to ask whether a strategy is profitable after accounting for selection, not whether its timing added value. We compare all four on simulated worlds with known ground truth: an IID no-skill world, a drift-only no-skill world, a volatility-clustering no-skill world, a structural-exposure world that is profitable but carries no entry-placement skill, and a genuine entry-placement-skill world in which an informative per-event signal of 0.0035 is injected. Each world is replicated 400 times, and we report the rejection rate of each test in each world. The discriminating case is the structural-exposure world: a test that respects the profit-versus-timing distinction should hold its size there, whereas a profitability test should over-reject because the strategy really is profitable, just not because of timing. Results are reported in Section 3.

3. Results

We report results in five parts. Section 3.1 comes first because it carries the paper's central contrast: it compares the placement test with White's Reality Check (White, 2000) and Hansen's SPA (P. R. Hansen, 2005) in a controlled world where a strategy is profitable but its profit comes entirely from structural exposure rather than timing, and shows that the standard tools reject about 13× more often than the timing (placement) null there, while still correctly answering their own profitability question. Section 3.2 then establishes that the placement test does what it claims on ground truth we control: it holds its nominal size when timing is uninformative and gains power as genuine timing is injected. Sections 3.3 and 3.4 apply it to real markets, first to a deep single-asset panel on gold futures and then to a 322-test cross-asset panel under multiplicity control. Section 3.5 treats the choice of re-randomization law as model risk and reports how verdicts move across an admissible family of placement measures. Throughout, the economic question is narrower than profitability: conditional on the realized trade structure (the ordered sequence of trade durations, the position sizes, the directions, and the non-overlap constraint) and the realized price path, did the calendar placement of the trades add value?

3.1. Head-to-Head Against Reality Check and SPA

The placement test answers a different question from the standard data-snooping tests, and the difference is sharpest in a world where a strategy is genuinely profitable but earns nothing from timing. We construct five synthetic worlds, 400 replications each, and compare four tests: the structure-preserving placement test (SP), a path-based block bootstrap (Path), White's Reality Check (RC) (White, 2000), and Hansen's SPA (SPA) (P. R. Hansen, 2005). RC and SPA test profitability and selection; SP and Path test timing. Table 3 and Figure 4 report rejection rates by world.

The first three worlds carry no skill of any kind, IID returns, returns with drift only, and returns with volatility clustering. Here all four tests should hold nominal size, and the placement test does: SP rejects at 0.052, 0.060, and 0.062 respectively, hugging the nominal 0.05. RC and SPA also stay near size in the IID and volatility-clustering worlds, but already over-reject under pure drift (RC $= 0.388$, SPA $= 0.393$), because a profitable-looking strategy riding a drifting market trips a profitability test even though nothing was timed.

The fourth world is the decisive one. Here a strategy holds a real structural exposure that makes it profitable, but its entries are placed with no entry-placement skill whatsoever. A test that isolates timing should hold its size; a test that targets profitability will tend to reject. In this world the placement test rejects at 0.050, exactly nominal in this sample of 400 replications, and the path-based timing test rejects 0% of the time. White's Reality Check and Hansen's SPA, by contrast, reject this profitable-but-untimed strategy about two-thirds of the time, at 0.660 and 0.637 respectively. That is a difference in estimand: RC and SPA answer whether the strategy is profitable beyond what selection explains, and in this constructed world the answer is yes. But that is precisely the question an allocator who already owns the exposure does not need answered. Read as a verdict on timing, then, a profitability test diverges from the placement null by more than an order of magnitude here—not because either test is wrong, but because the two answer different questions.

Crucially, this is not the placement test trading away power for size. In the fifth world, where genuine entry-placement skill is injected (per-event signal 0.0035), all four tests reject on every replication: SP, Path, RC, and SPA all at 1.000. The placement test holds nominal size against profitable exposure *and* retains full power against real timing; in the exposure-only world RC and SPA answer the profitability question rather than the placement question.

Design of the controlled worlds.

The 66%/64% rejection pattern is reproducible in principle from this 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 (Politis & Romano, 1994) ($B = 999$ 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 rejection pattern

follows from the estimands: the world is profitable-but-untimed by construction, and Reality Check and SPA reject precisely because the strategy is profitable.

3.2. Synthetic Calibration and Positive Controls

Having shown what the placement test isolates that the standard tools do not, we now verify that it does so correctly: a test that isolates timing is only useful if it is silent when there is no timing to detect and vocal when there is. We calibrate on synthetic markets where the data-generating process is known, so that "entry-placement skill" has an operational definition we can dial up and down.

Size under the null.

When the entry signal carries no information about forward returns, the test must reject at its nominal level. At a per-event signal-to-forward-return correlation of $\rho = 0.0001$, effectively zero, the rejection rate is 0.00 and the mean p-value is 0.52, indistinguishable from the uniform null a valid test should produce (Figure 2). The synthetic power curve confirms this at the decision threshold: with no injected signal, the empirical rejection rate is 0.043, within Monte Carlo error of the nominal 0.05. The placement test does not manufacture entry-placement skill out of exposure or drift.

Power as timing is injected.

As genuine timing is introduced, power rises monotonically and steeply (Figure 3). Expressed as edge per event bar, the rejection rate climbs from 0.043 at no signal to 0.130 at 5 bps, 0.475 at 10 bps, 0.925 at 20 bps, and 1.000 at 35 bps. The correlation-parameterized experiment tells the same story: at $\rho = 0.10$ the test rejects 48% of the time (mean $p = 0.099$), and by $\rho \approx 0.20$ it rejects on every replication (reject 1.00, mean $p = 0.005$). The practically relevant summary for an analyst is the minimum detectable edge: at the trade counts realized on gold futures, the test attains roughly 80% power once the signal-forward-return correlation reaches $\rho \approx 0.15$. Below that, modest timing edges are simply not separable from luck at this sample size, and the minimum detectable edge shrinks at the expected $N^{-1/2}$ rate as the number of trades grows.

Real-data positive control.

To confirm calibration outside the synthetic world, we run a positive control on the real price path with a deliberately neutral, no-skill schedule. The calibrated baseline lands at the 50.5th percentile of its own null with a mean RCSI_z of 0.03, the centered, null-consistent behavior expected of an untimed schedule placed against the realized path. The instrument reads zero when it should read zero.

3.3. The Gold Eleven-Rule Panel

We apply the test to eleven trading rules on gold futures (GC=F) over the common window 2002-03-04 to 2026-04-02, 6,049 bars per strategy, with an expected round-trip transaction cost of $c = 0.000470$. The rules span trend, breakout, mean-reversion, and seasonal families, plus a random-entry baseline. Table 1 reports, for each rule, the realized cumulative return, the number of trades N, the placement p-value under the neutral gap-permutation measure, the percentile of the realized statistic within its null distribution (the primary effect size), the descriptive effect sizes RCSI and RCSI_z, and the resulting classification.

The economic reading is the point of the table. Ten of the eleven rules earned a positive cumulative return, spanning -0.010 to 0.803, against a cost-free buy-and-hold benchmark of 15.944 over the same window. By the ordinary standard of a profitable backtest, most of these rules "work." Yet under the neutral measure, *not one* rule crosses $p \leq 0.05$: the

smallest p-value in the entire panel belongs to the Volatility Squeeze Breakout rule, which has only $N = 4$ trades: 0.053 in Table 1 under the stochastic-fill model, with 0.057 by exact enumeration of its placement orbit under a deterministic cost model. Read through the primary effect size, the exact null percentile, the most profitable rule, ADX Trend Following at a cumulative return of 0.803, sits at the 85.9th percentile of its null with $p = 0.141$: respectable, but not separable from the placement luck of its realized trade structure. The classifier resolves the panel into four rules above the null median but not significant, five indistinguishable from random luck, and two below the null median, with no rule reaching moderate or strong skill. The standardized RCSI_z (here 1.08 for ADX) is reported alongside but, per Section 2.5, is not the quantity on which any verdict turns.

The random-entry baseline is instructive precisely because it is designed to have no skill. On its single seed-42 realization it earned a cumulative return of 0.601, larger than nine of the ten rule-based returns (all but ADX Trend Following), and landed at the 29.0th percentile of its null. But that single-seed percentile is an artifact: averaged across $R = 100$ outer seeds, the random baseline's percentile is 48.3 ± 25.6, centered near 50 with the wide dispersion of a process that is, by construction, pure noise. A strategy can be both profitable and demonstrably untimed, and on gold every rule in the panel is consistent with that description under the neutral measure.

3.4. The Cross-Asset Panel Under Multiplicity Control

A single-asset panel cannot rule out that the absence of entry-placement skill is specific to gold. We therefore run the same placement test across 47 assets, yielding 322 rule-by-asset tests, and ask whether the field of p-values shows any concentration of entry-placement skill once multiplicity is controlled. Table 2 summarizes the panel.

If entry-placement skill were widespread, the p-value field would be left-shifted relative to uniform and significant tests would survive correction. Neither happens. The observed number of tests with nominal $p \leq 0.05$ is 13, *below* the 16.1 expected under a uniform null, so the raw count of nominal hits is if anything sparser than chance. The p-values are not left-shifted toward skill: the p-value distribution 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), so the empirical CDF lies below the diagonal, the departure being in the conservative direction rather than concentrated in the left tail. Under Benjamini–Hochberg (Benjamini & Hochberg, 1995), Bonferroni control (the latter at a threshold of 1.55×10^{-4}), and the dependence-robust Benjamini–Yekutieli procedure (Benjamini & Yekutieli, 2001)—which is strictly more conservative than Benjamini–Hochberg, so that no test it discovers can exceed the (already empty) Benjamini–Hochberg set—*zero* tests survive; the smallest p-value anywhere in the panel is 0.0014. The 322 strategy-by-asset cells are drawn from 47 instruments and are therefore not mutually independent, so the 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. There is no measure-invariant, multiplicity-robust entry-placement skill anywhere in the panel.

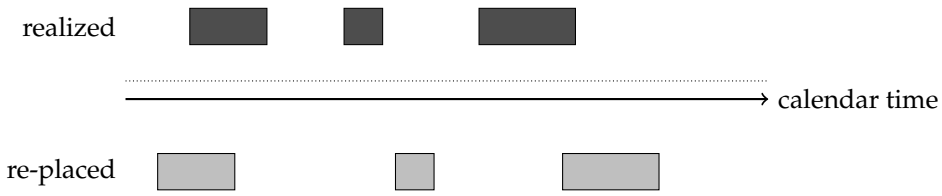
The composition of the smallest p-values makes the same point from the other direction. The single smallest p-value in all 322 tests, 0.0014, belongs not to a real trading rule but to a *random* baseline (on DIA, at the 99.88th percentile of its null), and two of the ten smallest-p tests are random baselines. When noise processes populate the extreme tail of the p-value field as readily as designed strategies, the extreme tail is measuring multiplicity, not skill. Profitability is common across these assets; isolable entry-placement skill is absent.

3.5. Schedule-Measure Sensitivity

The re-randomization law is itself a modeling choice, and we treat it as model risk rather than a fixed assumption. We re-run the gold panel under an admissible family of placement measures: the neutral gap-permutation baseline; a leading-gap restriction that forbids any simulated first entry inside the 200-bar window of the longest indicator lookback (the 200-period moving average); a context-matched measure (one that draws the comparison set to match the local market context at each realized entry); and a volatility-state, regime-preserving measure (one that preserves the volatility regime occupied at each realized entry). Table 4 reports per-rule p-values across the family. As stated in Section 2.6, the neutral measure is the declared canonical null; the others are sensitivity members, and each defines its own separate hypothesis rather than a combined level-α test.

Under the neutral baseline and the leading-gap restriction, the conclusion of Section 3.3 is unchanged: no rule rejects at 5%, no verdict changes, and nothing crosses the threshold. The 200-bar leading-gap restriction, set by the longest indicator lookback (the 200-period moving average) and the most defensible mechanical correction (it forbids placements that exploit a region where the signal could not yet have fired), changes zero verdicts and crosses zero thresholds. Whatever the gold rules earned, it is not robust entry-placement skill that survives a neutral re-randomization.

A specific and economically interpretable pattern emerges only under the two volatility-state-aware measures. Exactly four rules reject at 5%, and they are the *same* four under both the context-matched and the regime-preserving measure: Connors RSI(2) Pullback ($p < 0.001$ context-matched, $p \approx 0.0002$ regime-preserving), ADX Trend Following ($p = 0.010$ and 0.012), Breakout + Volume + Momentum ($p = 0.022$ and 0.037), and Volatility Squeeze Breakout ($p = 0.030$ and 0.032). Under the same volatility-state measures, four other rules fall to the null floor ($p \approx 1$). We note that volatility-state matching raises the null mean sharply, so the RCSI_z magnitudes under these measures (which reach large negative values for the random and turn-of-month rules) are not comparable across measures and only the tail p-value is interpreted. The interpretation is deliberately cautious. These four rules are all volatility-filtered by construction, and they reject only once the placement null is forced to match the volatility state at entry. That is consistent with the rules having learned *when* the market is in an exploitable volatility regime, but it is equally consistent with the volatility-state measure encoding into the null exactly the conditioning the rules use, so that the "skill" is an artifact of how the comparison set is drawn. For these two measures validity is moreover only approximate: the realized schedule lies in the support of the measure, but the relocated configuration is a genuine Q_θ-draw only insofar as the local-context match is exact, and the orbits can be near-degenerate, so we report a per-measure Kolmogorov-Smirnov uniformity check of the null p-values for each measure rather than relying on the neutral-measure calibration alone (Appendix B). Because the rejection appears under one admissible family of measures and vanishes under the canonical neutral one, we do not read it as robust evidence of entry-placement skill. The honest summary across the model-risk family is that gold entry-placement skill is measure-dependent, present only under volatility-state-matched nulls and absent under neutral ones, which is a far weaker claim than a profitable backtest would invite.



Same ordered trade durations, sizes, directions, and non-overlap, on the same price path; only the calendar placement is re-randomized.

Figure 1. The structure-preserving randomization. The realized trade schedule (top) is re-placed on the calendar (bottom) while its ordered durations, sizes, directions, and non-overlap, and the underlying price path, are held fixed. The null distribution is the performance over such re-placements.

Table 1. Gold-futures eleven-rule panel (GC=F), placement test under the neutral measure. Cumulative returns are net of the round-trip transaction cost $c = 0.000470$. p is the one-sided structure-preserving randomization p-value; RCSI_z is the standardized excess return; percentile is the realized statistic's rank in the null. The random-entry row is a single seed-42 realization (seed-averaged percentile 48.3 ± 25.6 over $R = 100$ outer seeds), and RCSI_z is a descriptive standardized summary, not interpretable as a z-statistic at the small N of the low-trade rules. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Rule	Trades	Cum. return	p	RCSI_z	Pctile
Trend Pullback	77	0.083	0.870	-1.03	13.0
Breakout Vol+Mom	46	0.297	0.176	0.91	82.4
Mean Rev Vol Filter	13	-0.010	0.662	-0.44	33.8
Validation AVM	35	0.219	0.345	0.32	65.5
ADX Trend	110	0.803	0.141	1.08	85.9
Oversold Reversion	30	0.082	0.341	0.35	65.9
Squeeze Breakout	4	0.065	0.053*	1.50	94.7
Connors RSI2	41	0.118	0.143	1.07	85.7
Donchian Reentry	60	0.071	0.647	-0.46	35.3
Turn-of-Month	141	0.323	0.234	0.65	76.7
Random control	215	0.601	0.710	-0.64	29.0
Buy and hold		15.944			

Table 2. Cross-asset panel under multiplicity control: 322 asset-by-rule tests across 47 instruments. BH = Benjamini–Hochberg (FDR 0.05); Bonferroni at FWER 0.05. The KS test is against a uniform null of p-values.

Quantity	Value
Instruments	47
Asset $\times$ rule tests	322
Nominal $p \leq 0.05$ (observed / expected by chance)	13 / 16.1
BH discoveries (FDR 0.05)	0
Bonferroni discoveries (FWER 0.05)	0
KS statistic vs. uniform (p-value)	0.114 (0.0004)
Smallest panel p-value	0.0014
Random-entry baselines in the smallest-10	2

Table 3. Head-to-head on known-truth worlds: rejection rates of White's Reality Check (RC), Hansen's SPA, and the placement (structure-preserving) test. In the profitable-but-untimed exposure world, RC and SPA reject while the placement test holds its nominal size of 0.050.

World	Reality Check	SPA	Placement test
IID, no skill	0.138	0.128	0.052
Drift only, no skill	0.388	0.393	0.060
Vol clustering, no skill	0.125	0.122	0.062
Structural exposure, no timing skill	0.660	0.637	0.050
Genuine entry-placement skill	1.000	1.000	1.000

Table 4. Schedule-measure sensitivity: one-sided placement p-values for each rule under each admissible re-randomization measure. The four volatility-filtered rules reject only under the volatility-state-matched (regime-preserving) measure. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Rule	Gap permutation	Leading gap	Context matched	Regime preserving
Trend Pullback	0.870	0.910	1.000	1.000
Breakout Vol+Mom	0.176	0.128	0.022**	0.037**
Mean Rev Vol Filter	0.662	0.668	0.395	0.186
Validation AVM	0.345	0.303	0.993	0.986
ADX Trend	0.141	0.055*	0.010***	0.012**
Oversold Reversion	0.341	0.383	0.146	0.117
Squeeze Breakout	0.053*	0.100*	0.030**	0.032**
Connors RSI2	0.143	0.155	0.000***	0.000***
Donchian Reentry	0.647	0.671	1.000	1.000
Turn-of-Month	0.234	0.320	1.000	1.000

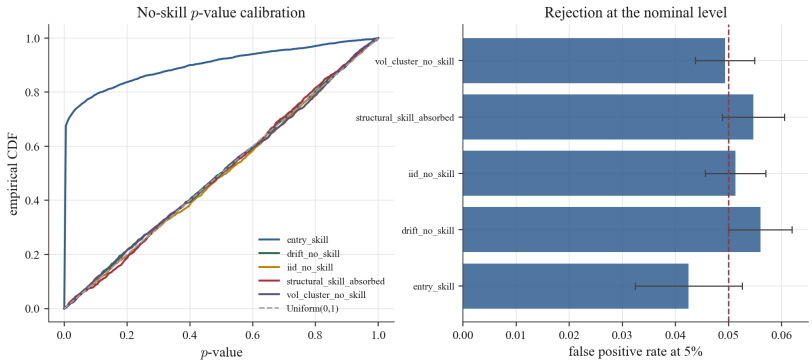


Figure 2. Calibration of the placement test on synthetic worlds: the p-value distribution under the no-timing null is close to uniform and the size is controlled at the nominal level.

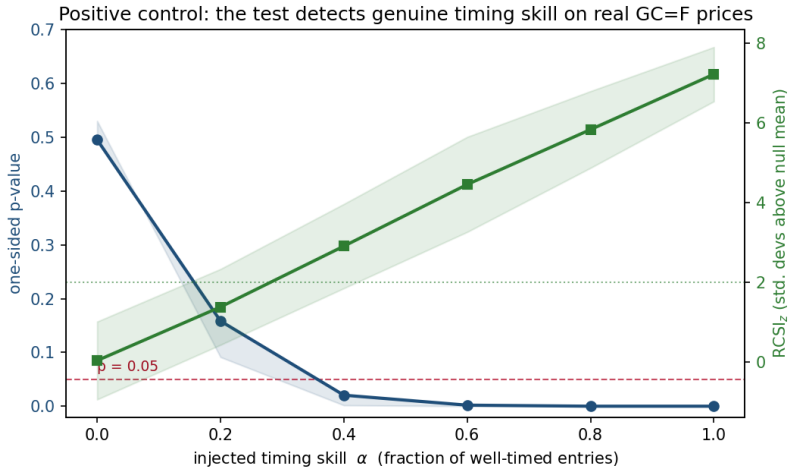


Figure 3. Power on a real-data positive control: as genuine timing skill is injected, the placement test's rejection rate rises, confirming it detects timing when present. The horizontal axis is the injected timing-skill fraction α (the share of entries relocated to well-timed bars); the body reports the same experiment equivalently in per-event edge (basis points) and in signal–forward-return correlation ρ.

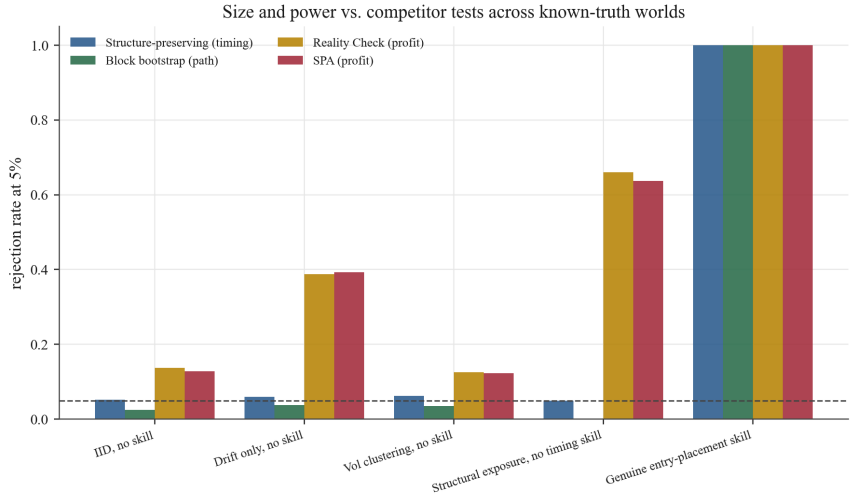


Figure 4. Head-to-head with Reality Check and SPA across known-truth worlds (cf. Table 3).

4. Discussion

4.1. Economic interpretation: profitability and entry-placement skill are distinct

The central economic lesson of this study is that a profitable backtest and a well-timed one are not the same object, and an allocator who treats them as interchangeable will systematically misprice the strategies in front of her. In the gold panel of Table 1, ten of the eleven rules earn a positive cumulative return, spanning roughly -0.010 to 0.803, and the buy-and-hold benchmark itself returns 15.944 in cumulative terms over the common window. Profitability, in other words, is abundant. Timing value is not. Under the neutral gap-permutation null no rule clears the five percent threshold; the smallest p-value in the panel belongs to the four-trade Volatility Squeeze Breakout at 0.053 (0.057 by exact orbit enumeration), and the classifier resolves the panel into four above-median (but not significant) rules, five indistinguishable from luck, and two below the null median, with nothing reaching moderate or strong skill. The practical reading is that the returns these rules earn are compensation for holding an exposure to gold along a rising path, not payment for choosing *when* to hold it.

For capital allocation, "no robust timing value" has a concrete and somewhat deflationary meaning. It does not say the strategies lose money, and it does not say gold was a poor place to be invested over 2002 to 2026; the size of the buy-and-hold cumulative return makes the opposite case. It says that the entry and exit decisions, the part of the strategy an active manager is actually paid to supply, did not improve on a structurally matched schedule placed without skill. An allocator should therefore price these rules as exposure vehicles and compare them to the cheapest available way of holding the same exposure, rather than paying an active fee for a timing edge the evidence does not support. This is the allocation-relevant distinction that profit-based tests cannot draw: in our controlled exposure world (Table 3), White's Reality Check and Hansen's SPA flag a profitable-but-untimed strategy about two-thirds of the time (0.660 and 0.637 respectively), while the placement test holds its nominal size at 0.050. A manager who reads a Reality Check or SPA verdict as a verdict on *timing*—rather than the verdict on profitability that it is—will repeatedly price exposure as timing skill in exactly the setting where the distinction matters most.

A residual signal does appear, but only at a specific and economically narrow margin. When the re-randomization law is allowed to condition on the volatility state, four of the gold rules cross the five percent threshold under both the context-matched and the regime-preserving measures: Connors RSI(2) Pullback ($p < 0.001$), ADX Trend Following (≈ 0.010 to 0.012), Breakout plus Volume plus Momentum (0.022 to 0.037), and Volatility Squeeze Breakout (0.030 to 0.032). These are the volatility-filtered rules, and the pattern is internally coherent: their apparent edge lives in the conditioning information about volatility regimes, not in calendar placement per se. We read this honestly as a candidate margin rather than a finding of skill, and, as set out in Section 2.6, each measure is a separate hypothesis and the canonical verdict is the neutral one fixed in advance. The same four rules show no rejection under the neutral baseline or the leading-gap restriction, where no verdict changes and nothing crosses threshold, so the effect is contingent on the analyst's choice of which features of the environment to hold fixed. For an allocator this is a hypothesis worth conditioning future evidence on, not a green light, and it underscores that the re-randomization law is itself a modeling decision with economic content.

4.2. Transaction costs, turnover, and implementability

The test conditions on the realized trade structure, which means turnover and holding periods are held fixed across the realized schedule and its re-randomized alternatives. Within a given test, every alternative schedule incurs the *same number* of round trips as the realized one at the same expected round-trip cost $c = 0.000470$, because the ordered sequence of trade durations and the non-overlap constraint are preserved. We are deliberately cautious about how far to push this. The cost charge is identical in *count* across the orbit, so a flat per-round-trip cost is a constant that nets out of the comparison between the realized placement and its alternatives; it does not differentially advantage the realized placement, and the timing verdict is not an artifact of a cost wedge between the strategy and its null. The one caveat is that this exact-netting holds for a cost that is a fixed charge per round trip; if costs are instead price-proportional, the per-trip charge depends on the price level at which each trade is placed, and relocating the template along a trending path can move the cost term slightly, so the netting becomes approximate rather than exact. On gold's strongly trending path this is a second-order effect at $c = 0.000470$, but it is a caveat we flag rather than a property we claim in general. In all cases the level of costs still governs net profitability and still determines whether a nominally positive rule is investable after frictions, and a rule that survives the placement test on gross timing grounds can remain unattractive once c is charged against a thin edge.

Implementability cuts the other way for the low-trade rules that supply most of the residual signal. The Volatility Squeeze Breakout fires only four times over more than two decades, and its exact enumeration (Section 2.3, with $6 \times 471 = 2826$ feasible schedules) makes the fragility visible: the one-sided tail probability sits at 0.057 under a deterministic cost model and 0.053 under the stochastic-fill model, straddling the conventional threshold. A four-trade rule is not a deployable allocation on its own terms regardless of its p-value; the turnover is too low to diversify placement risk, and a single mistimed entry would dominate realized performance. The broader point for practitioners is that the placement test isolates timing on a gross, structure-matched basis, and a favorable timing verdict is a necessary input to an allocation decision, not a sufficient one. Net-of-cost profitability, turnover-driven capacity, and the number of independent trades all remain binding constraints that the test does not adjudicate.

4.3. Limitations and when not to use the test

The placement test is informative only when the price path admits meaningful variation in where the trades could have been placed, and there are identifiable regimes where this condition fails. The first is path degeneracy on near-flat instruments. In the cross-asset dispersion scan, eight of the 322 tests are flagged path-degenerate, and the lowest-dispersion cases concentrate on the USDC-USD stablecoin, whose near-constant price path collapses the null dispersion of the performance statistic almost regardless of trade count: its breakout-volume-momentum panel records a null dispersion of 0.00126 over 42 trades, and its trend-pullback panel records 0.00371 over more than a thousand trades. When every admissible re-placement yields essentially the same payoff, the test has no signal to work with, and a small p-value in that regime reflects a degenerate path rather than entry-placement skill. The second failure mode is tiny-N. Two tests in the panel have $N = 1$ trade (both INTL), where no placement variation exists by construction, and the four-trade gold example shows that even slightly larger N leaves the verdict sensitive to the cost model. We therefore recommend the test be reported only where the null distribution carries non-trivial dispersion and the trade count is large enough that no single placement dominates.

The validity claim rests on exchangeability of the realized placement with its admissible alternatives under the chosen null, and we state this honestly rather than as a formality. The p-value is finite-sample valid, taking the form $(1 + k)/(M + 1)$, but it is valid *for the null encoded by the chosen measure*; the data do not prove that null. A scope condition deserves emphasis: the clean size statement holds when the realized placement is exchangeable under the declared law and the realized holding durations are placement-independent. When durations are adaptive (outcome-dependent), so that the strategy's exits and trade-skipping respond to realized returns, conditioning on the realized ordered durations no longer yields a clean exchangeability of the template under relocation, and the guarantee that survives is the conservative one proved in Appendix A.4: non-rejection is silence about the conditioned dimensions rather than a clean level-α size statement, and any predictive content in adaptive exits is assigned to the conditioned structure rather than to the placement test. The volatility-state results in Section 4.1 are the clearest illustration that the verdict can move with the measure, and we do not claim a single canonical answer where the measures disagree.

Finally, the evidence base is asymmetric in a way that should temper any general conclusion. The gold panel is a single asset studied at depth, with synthetic calibration, a real-data positive control, and an outer-loop robustness analysis over $R = 100$ seeds, and it is the setting in which we have powered evidence: the synthetic power curve holds size at 0.043 under no signal and rises to 1.000 at a thirty-five basis point per-event edge. The

322-test cross-asset panel is breadth rather than depth, and it functions as a cross-asset robustness scan of 47 instruments rather than a designed cross-section. Its message is appropriately modest and entirely negative: the observed count of nominal rejections (13) falls below the uniform expectation (16.1), nothing survives Benjamini–Hochberg (Benjamini & Hochberg, 1995) or Bonferroni control, the p-value distribution 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), so the empirical CDF lies below the diagonal, and the single smallest p-value across all 322 tests belongs to a random baseline on DIA. We read the cross-asset panel as evidence *against* pervasive entry-placement skill, not as a representative survey from which broad positive claims about any asset class could be drawn, and the genuinely powered conclusions of this paper are the single-asset and controlled-world results, not the breadth scan.

5. Conclusions

The object of inference in evaluating a trading strategy is not its realized payoff but the contribution of the timing of its decisions relative to a counterfactual that carries the same structural burden on the same price path. A profitable backtest can be earned by a structural exposure, by a drifting market, or by genuinely well-placed entries and exits, and a track record alone cannot separate these sources. We have developed a conditional randomization test that answers the narrower question an allocator actually faces: holding the realized trade structure fixed, namely the ordered durations, the position sizes, the directions, and the non-overlap constraint, together with the exogenous price path, was the calendar placement of the trades unusually favorable relative to structurally matched alternatives? The test re-randomizes only that placement, so under the declared sharp null that the realized schedule is exchangeable with admissible alternatives, it yields a finite-sample valid Monte Carlo p-value. The choice of re-randomization law is treated as model risk, an admissible family of measures over which we report a sensitivity range rather than a single number, with the neutral gap-permutation measure declared as the canonical null and measure invariance the strongest claim the design supports.

The empirical verdict is that profit is common but robust, measure-invariant entry-placement skill is absent. In the eleven-rule gold-futures panel (Table 1), ten of eleven rules earn positive cumulative returns, yet no rule clears $p \leq 0.05$ under the neutral gap-permutation measure (the minimum is the Squeeze Breakout at $p = 0.053$ (0.057 by exact enumeration)), and the classifier assigns no rule moderate or strong skill. Four volatility-filtered rules reject only once the conditioning measure is matched to the volatility state (Table 4), an effect that survives no neutral measure and so does not constitute measure-invariant evidence. Across a cross-asset robustness scan of 322 strategy-by-asset tests on 47 instruments, the observed count of nominal rejections (13) falls below its uniform expectation (16.1), nothing survives Benjamini–Hochberg (Benjamini & Hochberg, 1995) or Bonferroni control, and the single smallest p-value belongs to a random baseline. The synthetic power curves and the real-path positive control confirm the test is correctly sized and adequately powered, and the head-to-head comparison (Table 3) shows that White's Reality Check (White, 2000) and Hansen's SPA (P. R. Hansen, 2005), which ask about profitability rather than placement, flag a profitable-but-untimed structural-exposure strategy about two-thirds of the time, whereas the placement test holds its nominal size of 0.050. The reading is not that these strategies are bad but that, within the scope of the timing question, their realized profitability is not separately identifiable from structural exposure.

For backtest evaluation the implication is direct. A profitable track record is on its own uninformative about entry-placement skill, so a timing claim should clear a structure-

matched counterfactual before it is paid for; the test supplies that instrument once the analyst has declared the trade structure, path, statistic, and placement law. Because the procedure refutes rather than certifies, it is a screening diagnostic rather than an allocation signal, and a non-rejection is evidence of absence only where the design is powered. The construction is not tied to entry timing, long-only trading, or gold: wherever a decision sequence is overlaid on an exogenous process, holding the structural footprint fixed while randomizing only the decision margin separates the marginal contribution of the decisions from the exposure they create, with the validity, local-power, and measure-sensitivity results transferring subject to a domain-specific feasibility check. Profitability and entry-placement skill are distinct quantities, and standard data-snooping tests, by asking only about profitability, conflate them.

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, derived data, simulation outputs, and cross-asset panel logs needed to reproduce the reported results are available at https://github.com/ampatel355/FORTUNAFRAMEWORK and archived at https://doi.org/10.5281/zenodo.20724999. Raw vendor market data are not redistributed and remain subject to the terms of the original provider; the repository includes the derived files used by the reproduction scripts.

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. The conditional randomization test: formal construction and validity

This appendix gives the formal construction of the conditional randomization test and the admissibility predicate for the conditioning measure. All heavy notation is collected here; the body uses only the resulting estimand and p-value. Throughout, $\{P_t\}_{t=0}^{T}$ is the realized open-price path, treated as fixed.

Appendix A.1. Estimand and conditioning set

A strategy realizes a trade log $\mathcal{T} = \{(\tau_j^{\text{in}}, \tau_j^{\text{out}}, \omega_j, d_j)\}_{j=1}^{N}$, with entry and exit bars $\tau_j^{\text{in}} < \tau_j^{\text{out}}$, capital weight ω_j, and direction sign $d_j \in \{-1, +1\}$. Its *structural profile* is the tuple

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

the *ordered* holding-duration vector $\mathbf{h}$, the internal-gap multiset $\mathbf{g}^{\text{int}}$, the external slack g^{ext} (the unused calendar room before the first and after the last trade), the weights $\boldsymbol{\omega}$, and the directions $\mathbf{d}$. A *schedule* S is any placement of this fixed trade template on the calendar; it is *feasible* if it preserves $\mathcal{S}$ and induces no overlapping positions, the constraint set we denote $\mathcal{A}_s$. Let $\mathcal{F}(\mathcal{C}_s)$ be the space of feasible schedules and $S_0 \in \mathcal{F}(\mathcal{C}_s)$ the realized one.

The test conditions on the σ-field

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

equivalently $\mathcal{C}_s = \sigma(\mathcal{S}, \{P_t\})$. For any schedule S, let $T(S)$ be the cumulative return obtained by applying the preserved weights, directions, and ordered holding durations to the fixed path; T is a fixed $\mathcal{C}_s$-measurable map of S. Given a conditioning measure $\mathbb{Q}_\theta$ over $\mathcal{F}(\mathcal{C}_s)$, the *conditional entry-placement advantage* is

$$\Delta_\theta(s) = T(S_0) - \mathbb{E}_{S\sim\mathbb{Q}_\theta}\big[T(S) \mid \mathcal{C}_s\big], \qquad q_\theta = \mathbb{Q}_\theta\big(T(S) \geq T(S_0) \mid \mathcal{C}_s\big),$$

and the test targets the right-tail probability q_θ. The inferential question is narrow by design: conditional on $\mathcal{C}_s$ and the declared law $\mathbb{Q}_\theta$, durations, spacing, sizing, direction, and the price path are held fixed while calendar placement varies. Thus q_θ measures whether the realized template was favorable relative to the comparison set generated by that law. The sharp null is

$$H_0^\theta: \quad S_0 \overset{d}{=} \mathbb{Q}_\theta \text{ given } \mathcal{C}_s \qquad \text{(equivalently, } T(S_0) \text{ exchangeable with } \mathbb{Q}_\theta\text{-draws)}.$$

Entry-placement skill is, by construction, a favorable violation of H_0^θ. Each $\mathbb{Q}_\theta$ in the admissible menu defines its *own* H_0^θ; these are separate hypotheses, not pooled into one level-α test, and the "robust" label of the body is the heuristic that a rejection survives the menu.

Appendix A.2. The reference measure and its sampler

The primary measure is the gap-permutation null $\mathbb{Q}_{\text{gp}}$, which relocates the template by (i) drawing a leading gap g^{lead} uniformly on $\{0, \ldots, g^{\text{ext}}\}$ and (ii) drawing a uniform permutation of the internal-gap multiset $\mathbf{g}^{\text{int}}$, then reassembling the trades with their original ordered durations, weights, and directions. Because the internal span $\sum_j h_j + \sum_j g_j^{\text{int}}$ is invariant to the gap permutation, the leading-gap range $\{0, \ldots, g^{\text{ext}}\}$ is feasible for *every* permutation, so $\mathbb{Q}_{\text{gp}}$ is the uniform law on the product orbit $\{0, \ldots, g^{\text{ext}}\} \times \{\text{permutations of } \mathbf{g}^{\text{int}}\}$, and its induced marginals are uniform on $\{0, \ldots, g^{\text{ext}}\}$ for the leading gap and uniform over the permutations of $\mathbf{g}^{\text{int}}$ for the internal spacing. $\mathbb{Q}_{\text{gp}}$ is not the uniform measure over all non-overlapping schedules consistent with $\mathcal{S}$ (which is combinatorially expensive when gaps differ in size), but it preserves the realized spacing multiset exactly and is invariant under relabeling of trade indices. The realized schedule lies in its support, so H_0^θ is well posed. It is the declared canonical null because, among the admissible menu, it adds no conditioning on market context beyond the trade structure itself.

Appendix A.3. Admissibility of the conditioning measure

The boundary between conditioned "structure" and tested "placement" is a modeling choice, so we treat θ as part of the inferential object and restrict attention to a checkable admissible class. Let $\mathcal{F}_t^{\text{in}} = \sigma(\{P_u : u \leq t\})$ be the entry filtration. A structure-preserving measure $\mathbb{Q}_\theta$, holding fixed $(\mathcal{S}, \{P_t\})$ together with an extra feature set M_θ, is *admissible* ($\theta \in \Theta_0$) when:

- **(AM1) Structure preservation.** $\mathbb{Q}_\theta$-almost surely the schedule preserves the exact profile $\mathcal{S}$, in particular the ordered holding-duration vector $\mathbf{h}$, and the non-overlap constraint $\mathcal{A}_s$; the realized internal-gap *multiset* is preserved, not merely a feasibility envelope.

- **(AM2) Entry-measurability.** The feature set M_θ is $\mathcal{F}_t^{\text{in}}$-measurable, so no information unavailable at the entry bar enters the placement law.
- **(AM3) Outcome-independence.** M_θ is a function of $(\{P_t\}, t)$ and the template alone; it references no realized return, nor the value $T(S_0)$.

The reported admissible menu Θ_0 comprises the gap-permutation, leading-gap, context-matched, and regime-preserving measures, each of which fixes the realized internal-gap multiset and ordered durations (AM1) and conditions only on $\mathcal{F}^{\text{in}}$-features (AM2)–(AM3); the context-matched and regime-preserving (volatility-state) members additionally match each relocated entry to its realized local context. The uniform-feasible and slack-redistribution variants relax (AM1) by coarsening the profile and are therefore members of the wider family Ω but *not* of Θ_0. Admissibility is exactly what guarantees well-posedness.

Lemma A1 (Admissible nulls are well posed)**.** *For any $\theta \in \Theta_0$, the realized schedule lies in the support of $\mathbb{Q}_\theta$ given $\mathcal{C}_s$. Hence the exchangeability hypothesis of Proposition A1 (i) is well posed under each H_0^θ, and the per-measure tests of the body are individually valid.*

Proof. By (AM1) the realized schedule carries the profile $\mathcal{S}$ and satisfies the non-overlap constraint, which is the support condition defining the $\mathbb{Q}_\theta$-orbit. By (AM2) the bars the realized entries occupy are $\mathcal{F}^{\text{in}}$-events on the fixed path, so a context-matched measure contains S_0 because each realized entry matches its own context. Thus $S_0 \in \text{supp}(\mathbb{Q}_\theta \mid \mathcal{C}_s)$ and the conditional right tail at $T(S_0)$ is non-degenerate. Membership of S_0 in the support is necessary but not sufficient for the realized *joint* configuration to be a genuine $\mathbb{Q}_\theta$-draw: for the context-matched and regime-preserving measures the relocated configuration is exchangeable with S_0 only insofar as the local-context match is exact, so for those two members the validity is approximate and we report a per-measure uniformity check (Appendix B) rather than relying on the neutral-measure calibration. For the gap-permutation and leading-gap members the orbit is exact and the exchangeability is exact. □

Appendix A.4. Conditioning is conservative

The ordered holding-duration vector $\mathbf{h}$ and internal-gap multiset $\mathbf{g}^{\text{int}}$ are themselves outputs of the strategy's exit and turnover logic, not exogenous primitives. Conditioning on them assigns any predictive content in exits, holding-period selection, or trade-skipping to $\mathcal{S}$ rather than to the placement test. The design is therefore deliberately conservative: a strategy whose edge lives in adaptive exits or volatility-scaled holding periods will not reject H_0^θ even when those decisions are informative. This is also the scope condition behind the clean size statement: when durations are placement-independent the realized template is exchangeable with its relocations under $\mathbb{Q}_\theta$ and the level guarantee is exact; when durations are adaptive (outcome-dependent), conditioning on the realized $\mathbf{h}$ is the conservative choice, so rejection is strong evidence of entry-placement edge while non-rejection is silence about the conditioned dimensions, not a clean size statement and not a bound on total skill.

Appendix B. Exchangeability and finite-sample validity

This appendix states the exchangeability principle underlying the test, proves the finite-sample validity of the Monte Carlo p-value, and records the one high-level limit condition under which the standardized effect size RCSI_z is approximately pivotal. The validity result rests on exchangeability alone and invokes no model for returns and no central-limit approximation. For the context-matched and regime-preserving measures, where the exchangeability is approximate rather than exact (Lemma A1), we report a

per-measure Kolmogorov–Smirnov check of the null p-value's uniformity, not only for the neutral measure, so that each measure's calibration is verified on its own terms.

Appendix B.1. Exchangeability under the conditioning law

Fix $\theta \in \Theta_0$ and let $S_1, \ldots, S_M$ be i.i.d. draws from $\mathbb{Q}_\theta$ given $\mathcal{C}_s$. The construction is conditional in the sense of Lehmann and Romano (2005): everything is held fixed except the calendar placement, which is re-randomized under $\mathbb{Q}_\theta$. Under H_0^θ the realized schedule is itself a $\mathbb{Q}_\theta$-draw, so $S_0, S_1, \ldots, S_M$ are exchangeable given $\mathcal{C}_s$, and because T is a fixed measurable map the realized statistic $T(S_0)$ is exchangeable with $\{T(S_m)\}_{m=1}^M$. This is the only probabilistic input the level guarantee uses; it is an exchangeability of *schedules* on a fixed path, rather than of returns, which is what frees the test from any data-generating model (Fisher, 1935; Hemerik & Goeman, 2018).

Appendix B.2. Finite-sample validity of the Monte Carlo p-value

Proposition A1 (Validity and large-M consistency)**.** *Fix $\theta \in \Theta_0$ and i.i.d. draws $S_1, \ldots, S_M \sim \mathbb{Q}_\theta$ given $\mathcal{C}_s$. Define the one-sided Monte Carlo p-value*

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

(i) Validity. *If H_0^θ holds, in the statistic-level form that $T(S_0)$ is exchangeable with $\{T(S_m)\}_{m=1}^M$ given $\mathcal{C}_s$, then $\hat{p}_M$ is super-uniform: for every α on the grid $\{1/(M+1), \ldots, 1\}$, $\Pr(\hat{p}_M \leq \alpha \mid \mathcal{C}_s) \leq \alpha$, 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, $\hat{p}_M \to q_\theta = \mathbb{Q}_\theta(T(S) \geq T(S_0) \mid \mathcal{C}_s)$ almost surely as $M \to \infty$.*

Proof. (i) Exchangeability of $T(S_0)$ with $\{T(S_m)\}_{m=1}^M$ given $\mathcal{C}_s$ makes the rank of $T(S_0)$ among $\{T(S_m)\}_{m=0}^M$ uniform on $\{1, \ldots, M+1\}$ absent ties. Dividing the upper-tail rank by $M+1$ and counting ties into the right tail gives $\Pr(\hat{p}_M \leq \alpha \mid \mathcal{C}_s) \leq \alpha$; ties only enlarge the numerator of $\hat{p}_M$, moving the bound strictly toward conservatism and never inflating the rejection probability. (ii) For fixed S_0 the indicators $\mathbf{1}\{T(S_m) \geq T(S_0)\}$, $m \geq 1$, are i.i.d. Bernoulli(q_θ) given $\mathcal{C}_s$; the strong law of large numbers gives the limit, which does not invoke H_0^θ. □

The $+1$ in numerator and denominator is what makes $\hat{p}_M$ valid at finite M: the naive $\sum \mathbf{1}\{\cdot\}/M$ is downward biased and can be invalidly zero, whereas the corrected form is finite-sample valid for a constrained, dependence-preserving resampling of a single realized path. Validity is *against the sharp null* H_0^θ: $\mathbb{Q}_\theta$ is analyst-chosen, and the guarantee transfers only insofar as the realized schedule is exchangeable with its $\mathbb{Q}_\theta$-resamples and only on the attainable p-grid. Two granularities should not be conflated. The first is intrinsic to the orbit: when N is small the placement information is carried by the $(N-1)!$ internal-gap permutations, the leading-gap draw mainly translates the template rigidly along the path and contributes little placement-discriminating information at small N, so the placement margin is weakly identified and the minimum detectable edge is large; the orbit itself need not be coarse, as the $N = 4$ exact enumeration shows. The second is the Monte Carlo budget M, which only refines the attainable grid. In both cases the consequence is a loss of power, never of size under the declared sharp null: small N is an identification problem, not a validity problem, because part (i) bounds the conditional rejection probability by α on the whole attainable grid. In the pure-exposure world used in the calibration study, entry placement is generated from the same structure sampler used by the null, so $T(S_0)$ remains

a $\mathbb{Q}_\theta$-draw and the test holds its nominal level with no appeal to any limit law, while a return-mean benchmark inherits the full profitability signal.

Appendix B.3. A central-limit condition for RCSI_z

The standardized effect size is $\mathrm{RCSI}_z = (T(S_0) - \mu_{\mathbb{Q}})/\sigma_{\mathbb{Q}}$, with $\mu_{\mathbb{Q}}, \sigma_{\mathbb{Q}}$ the conditional $\mathbb{Q}_\theta$-mean and standard deviation of $T(S)$, and we treat it as an appendix diagnostic only; the headline effect size in the body is the exact null percentile, which is always valid. Writing $\log(1 + CR) = \sum_{j=1}^{N} \xi_j$ for the N trade log-contributions of a $\mathbb{Q}_\theta$-placement, and $Z_N = (T(S_0) - \mu_{\mathbb{Q}})/\sigma_{\mathbb{Q}}$, asymptotic pivotality of RCSI_z requires that under H_0^θ, $Z_N \to_d N(0,1)$ as $N \to \infty$. We state this as an *assumption*, not a theorem: under $\mathbb{Q}_\theta$ a draw permutes the realized gaps on one fixed, autocorrelated path, so the ξ_j are functions of the permutation, neither independent nor identically distributed and coupled by the non-overlap constraint, and no off-the-shelf result delivers normality. The condition is motivated by the combinatorial central-limit theory for permutation statistics, with Hoeffding's theorem as the canonical instance (Hoeffding, 1951), under which standardized sums over a randomly permuted finite population are asymptotically normal precisely when (a) a no-dominant-term Lindeberg condition $\max_j \mathrm{Var}(\xi_j)/\sum_j \mathrm{Var}(\xi_j) \to 0$ holds, so no single trade carries a non-negligible share of the return budget, and (b) the realized path is weakly (α-mixing) dependent, so its autocorrelation is absorbed into the finite-population variance rather than breaking the standardization. These conditions are sufficient, not necessary, and we do not verify them for any particular path; they are falsifiable and fail when one or a few trades dominate the budget, which is exactly the case at the four-trade Squeeze Breakout, where a single trade carries a large share of the return budget so the Lindeberg ratio is far from zero and Z_N is not pivotal. By contrast, on the higher-N rules (for instance ADX and the random baseline, with tens to hundreds of trades) the ratio is small and the normal approximation is reasonable. We therefore retain RCSI_z as a descriptive, approximately pivotal effect size only, trustworthy on high-N rules and not on the low-N ones. The *validity* of the test rests on the finite-sample exchangeability of Proposition A1 (i), not on this condition, which is invoked solely for asymptotic statements; its empirical adequacy is checked by the per-measure Kolmogorov–Smirnov calibration of the null p-value's uniformity reported in the body, not by direct verification of Gaussianity.

Appendix B.4. Consistency in trade count

Evidence accumulates in the trade count N, not in the length of the price path, because $\mathbb{Q}_\theta$ rearranges a fixed set of N contributions on a fixed path. Under the central-limit condition above and a per-trade location-shift alternative that adds a common standardized increment $\zeta = \mu_1/\sigma_1$ to each contribution at fixed null variance, the noncentrality of Z_N grows like $\zeta\sqrt{N}$, so the test is consistent and its minimum detectable edge shrinks at the $N^{-1/2}$ rate. The practical reading is that detection is governed by the number of trades, not the sample span: low-N panels carry little placement information regardless of how long the path is, which is why near-threshold small-N verdicts are reported as exploratory and a robust timing claim is required to hold across the admissible menu Θ_0 rather than under any single measure.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *école d'été de probabilités de saint-flour xiii—1983* (Vol. 1117, pp. 1–198). Berlin: Springer.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2017). The probability of backtest overfitting. *Journal of Computational Finance*, *20*(4), 39–69.

Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, *40*(5), 94–107.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, *41*(2), 802–837.

Besag, J., & Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika*, *76*(4), 633–642.

Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, *80*(3), 551–577.

Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments. *Mathematical Finance*, *16*(3), 519–547.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, *28*(1), 181–187.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, *19*(4), 676–685.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

Fithian, W., Sun, D., & Taylor, J. (2014). *Optimal inference after model selection.* arXiv:1410.2597.

Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, *18*(2), 141–153. https://doi.org/10.1016/0304-4068(89)90018-9.

Hansen, L. P., & Sargent, T. J. (2008). *Robustness*. Princeton, NJ: Princeton University Press.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, *23*(4), 365–380.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, *29*(1), 5–68.

Hemerik, J., & Goeman, J. (2018). Exact testing with random permutations. *TEST*, *27*(4), 811–825.

Hoeffding, W. (1951). A combinatorial central limit theorem. *The Annals of Mathematical Statistics*, *22*(4), 558–566.

Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society: Series B*, *30*(3), 582–598.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907–927.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York: Springer.

Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, *30*(5), 15–29.

López de Prado, M. (2018). *Advances in financial machine learning*. Wiley.

Manski, C. F. (2003). *Partial identification of probability distributions*. New York: Springer.

Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. *Journal of Financial Economics*, *104*(2), 228–250.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software*. Chichester: John Wiley & Sons.

Phipson, B., & Smyth, G. K. (2010). Permutation *p*-values should never be zero: Calculating exact *p*-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, *9*(1), Article 39.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, *64*(3), 479–498.

Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance*, *54*(5), 1647–1691.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, *2*(1), 167–195. https://doi.org/10.1146/annurev.economics.050708.143401.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126.