How Many Backtest Winners Survive Deflation?

A Controlled Study of the Deflated Sharpe Ratio and Multiple-Testing Haircuts

Eugen Soloviov

July 2026

Abstract

A parameter search over a backtest grid is a multiple-testing machine: the best of many zero-skill configurations reliably produces an impressive Sharpe ratio. We run four seeded, fully reproducible experiments with known ground truth to measure how well the standard corrections repair this. Under a global null (selecting the best of 1000 independent zero-skill strategies), the naive single-test verdict certifies a discovery in every repetition (false-discovery rate 1.000) at an average best annualized Sharpe of 1.63, while the Deflated Sharpe Ratio (DSR), the Harvey–Liu haircuts, and White's Reality Check bring the false-discovery rate to 0.001–0.057, at or near the nominal $\alpha = 0.05$. The deflated benchmark SR_0 averages 1.63 annualized — exactly the noise ceiling: the bar a strategy must clear is not zero but the best Sharpe that luck alone buys for the given search size and sample length. A planted-signal sweep shows DSR power rising from 0.005 to 1.000, with the 50%-power point just above that ceiling (about 1.73 annualized). Two realistic moving-average searches (640 correlated trials each, bootstrap tests on 5000 resamples) then probe the correlated-search regime. On pure noise every estimator rejects the seductive winner (annualized Sharpe 0.81). On a genuine regime-switching edge, whose in-sample *selected* maximum is an annualized Sharpe of 3.92, the Reality Check ($p = 0.0024$), the studentized SPA-type test ($p = 0.0038$), and the Harvey–Liu haircuts (0.148–0.198) all certify the edge — but DSR fed the *raw* trial count wrongly rejects it ($0.748 < 0.95$). The natural fix, an effective number of trials, turns out not to be a single number: five standard estimators computed on the same trial matrix disagree by two orders of magnitude (1.6 to 370.0), and at the smallest estimates the deflation is nearly inert — there the correct verdicts are inherited from the winners' un-deflated significance, not produced by deflation. The honest deliverable is a robustness band: the real edge is retained for every effective count below 144.8 (four of the five estimators, including genuinely deflating mid-range bars of annualized 1.85–2.00), while the noise winner is rejected at every effective count. Bootstrap-based tests sidestep the choice entirely by resampling the whole search jointly.

1 Introduction

Every systematic trading study contains a hidden hypothesis-testing problem. A researcher who sweeps a parameter grid, keeps the configuration with the highest backtest Sharpe ratio, and then reports that Sharpe ratio as if it were a single experiment has run a multiple-testing procedure while quoting single-test statistics. The reported performance is the maximum of many noisy estimates, and the maximum of many zero-mean draws is emphatically not zero-mean. Bailey et al. (2014) call the resulting literature pseudo-mathematical; Harvey et al. (2016) document the same selection mechanism operating at the scale of an entire academic field, with hundreds of published factors mined from broadly the same data.

Several corrections are in wide circulation. Bailey and López de Prado (2012) introduced the Probabilistic Sharpe Ratio (PSR), which converts a sample Sharpe ratio into the probability that

the true Sharpe ratio exceeds a chosen benchmark, accounting for sample length, skewness, and kurtosis. Bailey and López de Prado (2014) extended it to the Deflated Sharpe Ratio (DSR), which sets that benchmark to the expected maximum Sharpe ratio a search of N independent zero-skill trials would produce — so the winner is measured against the best that luck alone could have delivered. Harvey and Liu (2015) approach the same problem through classical multiple-testing adjustments (Bonferroni, Holm, Benjamini–Hochberg–Yekutieli), translating adjusted p-values back into a "haircut" on the reported Sharpe ratio, with the economic framing developed in Harvey and Liu (2014). A third lineage tests the entire search at once: White's Reality Check (White, 2000) bootstraps the distribution of the *best* performance across all trials under the null, and Hansen (2005) sharpened it against irrelevant alternatives via studentization. López de Prado (2018) surveys these tools alongside the non-parametric Probability of Backtest Overfitting of Bailey et al. (2017).

What is harder to find in this literature is a *controlled* comparison: experiments in which the ground truth is known by construction — because the data-generating process is synthetic and seeded — so that false-discovery rates and detection power can be measured directly rather than argued about. That is the gap this paper fills. We make no claim of new estimators; the contribution is calibration evidence and an honest account of a failure mode and of what can, and cannot, be salvaged from it.

Concretely, we run four experiments (Section 3) against the estimators of Section 2:

1. **Null calibration.** Select the best of $N = 1000$ independent zero-skill strategies, 2000 times. The naive single test certifies a discovery in every repetition; DSR, the Harvey–Liu haircuts, and the Reality Check bring the false-discovery rate to between 0.001 and 0.057, at or near the nominal level. The deflated benchmark SR_0 reproduces the empirical noise ceiling to two decimals (both 1.63 annualized).

2. **Planted power.** With 25 genuine strategies hidden among 1000, DSR's detection power traces an S-curve that crosses just above the noise ceiling: edges below the ceiling are statistically indistinguishable from luck; edges above it are retained with power approaching one, at a measured false-positive rate of 0.000.

3. **A realistic search on noise.** A moving-average crossover grid of 640 correlated configurations, applied to a pure random walk, yields a winner with annualized Sharpe 0.81. Every estimator rejects it, at every choice of effective trial count.

4. **The same search on a real edge.** On a regime-switching series the in-sample selected maximum is an annualized Sharpe of 3.92 and the bootstrap tests certify it decisively — but DSR, fed the raw trial count of 640, *wrongly rejects* the genuine edge. The obvious repair, an effective number of trials, is not a single number: five standard estimators computed on the same trial matrix span 1.6 to 370.0. Four of the five retain the edge — including mid-range estimators whose deflated bar is a genuine annualized 1.85–2.00 — and DSR retains it for every effective count below 144.8 while rejecting the noise winner at every effective count. We argue that this robustness band, not any point estimate, is the honest deliverable of effective-N deflation, and we document where the band's lower anchor is misleading.

Section 4 reports the numbers, Section 5 draws the practical lessons, and Section 6 states the limitations plainly. All results derive from one seeded run of one public harness; a companion script verifies every numeric claim in this manuscript against the run's JSON output.

2 Estimators

Notation. Sharpe ratios are per-observation, $\widehat{SR} = \hat{\mu}/\hat{\sigma}$ computed on T return observations, unless explicitly annualized; annualized values multiply the per-observation value by $\sqrt{252}$ (we treat 252 observations as one year of daily data). Φ and Φ^{-1} denote the standard normal CDF and quantile function. Tests are one-sided (skill means $SR > 0$) at level $\alpha = 0.05$ throughout. Equation-number citations below refer to the numbering printed in the primary sources, which we re-extracted from the authors' posted full texts for Bailey and López de Prado (2012, 2014); Harvey and Liu (2015); the tests of Section 2.4 are described per their standard characterization, as discussed there.

2.1 Probabilistic Sharpe Ratio

The PSR of Bailey and López de Prado (2012), their Eq. (11), with the estimated variance of $\widehat{SR}$ from their Eq. (8) (a result they attribute to Mertens), is the probability that the true Sharpe ratio exceeds a benchmark SR^* given the sample evidence:

$$\widehat{\mathrm{PSR}}(SR^*) = \Phi\left[\frac{(\widehat{SR} - SR^*)\sqrt{T-1}}{\sqrt{1 - \hat{\gamma}_3 \widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}}\right], \tag{1}$$

where $\hat{\gamma}_3$ is the sample skewness and $\hat{\gamma}_4$ the sample kurtosis of the strategy's returns in the *non-excess* convention (normal returns give $\hat{\gamma}_4 = 3$). The convention matters: for normal returns the denominator reduces to the classical asymptotic $\mathbb{V}[\widehat{SR}] \approx (1 + SR^2/2)/T$, which only happens with non-excess kurtosis; substituting excess kurtosis silently corrupts the deflation. All quantities are computed in the returns' native frequency — the source is explicit that PSR is invariant to calendar conventions, so no annualization enters the test itself. The same expression is restated as Eq. (2) of Bailey and López de Prado (2014).

2.2 Deflated Sharpe Ratio

The DSR of Bailey and López de Prado (2014) is PSR evaluated at a benchmark that is not zero but the expected maximum Sharpe ratio that a search of N independent zero-skill trials would produce. Let $\{\widehat{SR}_n\}_{n=1}^{N}$ be the Sharpe estimates of all trials in the search, with empirical mean $\hat{\mu}_{SR}$ and variance $\widehat{V}_{SR}$. Their Eq. (1) (derived in their appendix as Eqs. (5)–(6) via an extreme-value approximation to the expected maximum of N iid standard normal draws) gives the deflated benchmark

$$SR_0 = \hat{\mu}_{SR} + \sqrt{\widehat{V}_{SR}}\left[(1-\gamma)\,\Phi^{-1}\Big(1 - \frac{1}{N}\Big) + \gamma\,\Phi^{-1}\Big(1 - \frac{1}{Ne}\Big)\right], \tag{2}$$

where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant and e is Euler's number. Their Eq. (1) states the benchmark under the null $\mathbb{E}[\widehat{SR}_n] = 0$, in which the mean term vanishes; we carry the empirical mean $\hat{\mu}_{SR}$ to match the authors' own reference implementation (function `getExpMaxSR` in the published paper), which does the same. Then

$$\widehat{\mathrm{DSR}} = \widehat{\mathrm{PSR}}(SR_0), \tag{3}$$

their Eq. (2), computed with the *selected* strategy's own $\widehat{SR}$, T, $\hat{\gamma}_3$, and $\hat{\gamma}_4$ in Eq. (1). We flag a discovery when $\widehat{\mathrm{DSR}} > 1 - \alpha = 0.95$.

Three properties of Eq. (2) drive everything that follows. First, SR_0 grows with the search: roughly like the square root of $\log N$ through the quantile terms, and proportionally to the cross-trial dispersion $\sqrt{\widehat{V}_{SR}}$, which itself shrinks like one over the square root of T under the null. The deflated bar is therefore a *noise ceiling* specific to the search size and sample length. Second, the dispersion term treats the observed spread of trial Sharpes as pure luck: when part of that spread is produced by genuinely skilled trials, Eq. (2) reads real skill as more luck to deflate against. Third — the caveat printed in the source itself — the derivation assumes the N trials are *independent*. The paper's Appendix 3 addresses determining N when they are not; Section 2.5 returns to this point, which turns out to be the single most consequential implementation decision in our experiments.

2.3 Harvey–Liu haircuts

Harvey and Liu (2015) start from the exact algebraic link between the Sharpe ratio and the single-test t-statistic (their Eqs. (1)–(2)): $t = \widehat{SR}\sqrt{T}$, from which a one-sided p-value follows, $p = 1 - \Phi(t)$ under the normal approximation. Testing M strategies yields ordered p-values $p_{(1)} \le \cdots \le p_{(M)}$, which are adjusted by one of three classical procedures — Bonferroni, Holm (Holm, 1979), and Benjamini–Hochberg–Yekutieli (BHY; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001):

$$\text{Bonferroni:}\quad p_{(i)}^{\text{Bonf}} = \min\big[M\,p_{(i)},\,1\big], \tag{4}$$

$$\text{Holm:}\quad p_{(i)}^{\text{Holm}} = \min\Big[\max_{j\le i}\,(M-j+1)\,p_{(j)},\,1\Big], \tag{5}$$

$$\text{BHY:}\quad p_{(M)}^{\text{BHY}} = p_{(M)},\qquad p_{(i)}^{\text{BHY}} = \min\Big[p_{(i+1)}^{\text{BHY}},\,\frac{M\,c(M)}{i}\,p_{(i)}\Big], \tag{6}$$

where the step-up recursion in Eq. (6) runs from $i = M-1$ down to the smallest p-value, with the normalizing constant *multiplying* the adjustment,

$$c(M) = \sum_{j=1}^{M}\frac{1}{j}. \tag{7}$$

Benjamini and Hochberg (1995) originally set $c(M) = 1$, valid under independence or positive dependence; Harvey and Liu adopt the Benjamini and Yekutieli (2001) harmonic-sum choice precisely because it controls the false-discovery rate under *arbitrary* dependence among the test statistics — the situation of correlated backtest trials. The constant grows slowly: $c(1000) \approx 7.49$, so the price of dependence-robustness is less than one order of magnitude. Bonferroni and Holm control the familywise error rate FWER = Pr(at least one false rejection); BHY controls the false-discovery rate FDR = $\mathbb{E}$[false rejections/rejections], a more lenient criterion for the *bulk* of hypotheses. One rank-specific subtlety matters for everything we report: the best-strategy haircut evaluates only the *top-ranked* hypothesis, and at rank one Holm coincides with Bonferroni exactly — $(M-1+1)\,p_{(1)} = M\,p_{(1)}$ — so the two never disagree about a search winner and should not be read as independent corroboration. Moreover, BHY multiplies the top p-value by $M\,c(M) \ge M$, so for the top pick BHY is the *most* conservative of the three, even though it controls the more lenient criterion overall.

The adjusted p-value of the search winner is then converted back through the t-statistic link into a haircut Sharpe ratio SR_{adj} — "the Sharpe ratio that would have resulted from a single test" — and the headline quantity is the haircut fraction

$$\text{hc} = 1 - \frac{SR_{\text{adj}}}{\widehat{SR}}. \tag{8}$$

2.4 Reality Check and the studentized (SPA-type) variant

The tests above operate on summary statistics of the trials. White's Reality Check (White, 2000) instead tests the whole search at once, using the full $T \times K$ matrix of trial returns (in excess of a benchmark; ours is zero, i.e. not trading). With $\bar{f}_k$ the sample mean performance of trial k, the null is $H_0\colon \max_k \mathbb{E}[f_k] \le 0$ and the statistic is the scaled best average, $V = \max_k \sqrt{T}\,\bar{f}_k$. Its distribution under the null is approximated by the stationary bootstrap of Politis and Romano (1994): resampled series are built from blocks of geometrically distributed random length (expected block length $1/p$), wrapping circularly so the resample is itself stationary. Each bootstrap replication recenters every trial by its own full-sample mean, so the resampled world satisfies the null by construction, and the p-value is the add-one fraction of bootstrap statistics $V^{*b} = \max_k \sqrt{T}\,(\bar{f}^{*b}_k - \bar{f}_k)$ that reach the observed V, i.e. $\hat{p} = (1 + \#\{V^{*b} \ge V\})/(B+1)$. Because the bootstrap resamples all K trial series *jointly*, the cross-trial correlation structure is preserved exactly; no independent-trials assumption enters anywhere.

Hansen (2005) showed that the unstandardized maximum can be dominated by poor, high-variance alternatives, making the Reality Check conservative, and proposed the Superior Predictive Ability (SPA) test with two modifications: studentizing each trial by an estimate of its own standard error before taking the maximum, and a sample-dependent (consistent) recentering of the null distribution. Our implementation applies only the first modification — each trial's statistic is divided by its sample standard deviation, with White's full recentering retained — so it is a *studentized Reality Check*, not Hansen's full SPA; we label it "SPA-type" everywhere in this paper, and note that omitting Hansen's recentering can only make it more conservative. One honesty flag: unlike Sections 2.1–2.3, whose equations we re-extracted from the primary texts, this subsection follows the standard characterization of White (2000), Hansen (2005), and Politis and Romano (1994); we cite no equation numbers from those papers.

2.5 Effective number of trials: a spread, not a number

Equation (2) needs the number of *independent* trials, and a parameter grid does not supply that number directly: neighboring configurations trade nearly identical positions. Three strands of the literature converge on the same warning. Bailey and López de Prado (2014) state the independence assumption and point to their Appendix 3 for an effective-N construction under dependence. Harvey and Liu (2015) adopt the BHY constant of Eq. (7) specifically for dependence robustness. Harvey et al. (2016) model the correlation among the factors they audit rather than counting them, arriving at a t-hurdle of roughly three for contemporary discoveries.

There is, however, no canonical estimator of the effective trial count, and the candidates disagree — in our experiments, by two orders of magnitude on the same matrix. We therefore compute five standard ones and report the whole spread. With $\bar{\rho}$ the average pairwise correlation among the K trial return series, the average-correlation estimate is

$$N_{\text{eff}}^{\text{corr}} = \frac{K}{1 + (K-1)\,\bar{\rho}}, \tag{9}$$

the variance-reduction factor of the *mean* of K equicorrelated variables — exact in that setting, and the smallest of the five in practice. A functional mismatch deserves flagging here: DSR's benchmark is an expected *maximum* (an extreme-value quantity), not a mean, and there is no reason the effective count appropriate for averaging transfers to extremes; Eq. (9) is best read as a lower anchor. The remaining four work from the eigenvalues $\lambda_1 \ge \lambda_2 \ge \dots$ of the $K \times K$ trial correlation matrix (which sum to K): the *participation ratio* $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$; *PCA-95*, the

number of leading eigenvalues needed to explain 95% of total variance; the *Kaiser criterion*, the count of eigenvalues exceeding one; and the *Cheverud–Nyholt* estimate $1 + (K-1)\big(1 - \mathrm{Var}(\lambda)/K\big)$, a standard recipe imported from the multiple-testing literature in statistical genetics, known to over-count effective tests under strong equicorrelation and in practice the largest of the five (the upper anchor). We cite no primary sources for these four; they are used here as widely known recipes, not endorsed estimators.

Because $\widehat{\mathrm{DSR}}$ is monotonically decreasing in N, feeding each estimate into Eq. (2) traces a curve, and the entire analysis can be compressed into one number found by bisection:

$$N^* = \max\big\{N \geq 2 : \widehat{\mathrm{DSR}}(N) > 1 - \alpha\big\}, \tag{10}$$

the largest effective trial count at which the winner still survives. A claimed discovery then comes with a robustness band — "retained for every $N_{\mathrm{eff}} < N^*$" — instead of a verdict conditioned on one contestable point estimate. One domain note: the expected-maximum formula is defined for $N \geq 2$ (the $\Phi^{-1}(1-1/N)$ term diverges as $N \to 1$), so fractional estimates below two are evaluated at the $N = 2$ floor.

3 Experimental design

All experiments are generated and analyzed by a single Python harness (`scripts/run_all.py`, estimators in `scripts/deflate.py`), under Python 3.14.6 with NumPy 2.4.3. Every random stream is seeded, so every number in this paper is bit-reproducible by one command. Ground truth is known by construction in each experiment: we know which strategies have skill because we planted them. A verification script (`scripts/check_paper_numbers.py`) checks every numeric claim in this manuscript against the harness output `results/results.json`; values quoted below are rounded from that file, and quantities labeled *derived* are arithmetic combinations of its entries (annualization by $\sqrt{252}$, differences, products, interpolation).

3.1 Experiment 1: null calibration

Each repetition draws a $T \times N$ matrix of iid standard normal returns with $T = 1000$ observations (about four years of daily data) and $N = 1000$ strategies — every true Sharpe ratio is exactly zero. We select the winner $\widehat{SR}_{\max} = \max_n \widehat{SR}_n$ and ask each procedure whether it certifies a discovery at $\alpha = 0.05$:

- *Naive:* one-sided single test on the winner, $p = 1 - \Phi(\widehat{SR}_{\max}\sqrt{T}) < \alpha$, ignoring the search;
- *DSR:* Eq. (3) > 0.95, with all N trial Sharpes informing SR_0;
- *Harvey–Liu:* adjusted p-value of the winner below α under each of Eqs. (4)–(6) with $M = N$;
- *Reality Check:* bootstrap $p < \alpha$ on the full return matrix.

The closed-form procedures run over $M = 2000$ repetitions. The Reality Check, being bootstrap-based, runs over 400 repetitions with 500 stationary bootstrap resamples each (expected block length 20; the null data are iid, so blocking is harmless). Under a global null every certified discovery is false, so the empirical discovery rate *is* the false-discovery rate.

3.2 Experiment 2: planted power

Calibration is cheap if a test never fires; the counterpart question is power. We repeat the setup of Experiment 1 but add a constant per-observation drift s_{true} to $n_{\text{true}} = 25$ randomly chosen strategies per repetition, giving them true per-observation Sharpe $s_{\text{true}} \in \{0.05, 0.08, 0.12, 0.16, 0.20\}$ (annualized 0.79 to 3.17), with 1000 repetitions per level. We record whether DSR certifies the search winner, and whether that winner is genuinely one of the planted strategies (a true positive) or an impostor (a false positive). For reference we also record how often the naive test certifies a winner that happens to be planted. Two scope notes, stated up front: the power figures are specific to this configuration (25 planted among $N = 1000$, $T = 1000$) — we did not sweep these parameters — and Experiment 2 measures the power of DSR only; the retention behavior of the other procedures is probed through the single case study of Experiment 4.

3.3 Experiments 3–4: a realistic search, without and with an edge

The first two experiments use independent trials — exactly the world Eq. (2) assumes. Real searches are not like that, so we build one. A moving-average crossover strategy computes fast and slow simple moving averages of the log price and holds the position $\text{sign}(\text{MA}_f - \text{MA}_s) \in \{-1, 0, +1\}$; the grid sweeps $f \in \{3, 6, \ldots, 48\}$ (16 values) and $s \in \{55, 60, \ldots, 250\}$ (40 values), $K = 640$ configurations in all. Positions are strictly causal — the position formed at observation t earns the return at $t + 1$ — and frictionless (no transaction costs; the point here is selection inference, not net profitability). The price series has $T = 756$ daily returns (three years at 252 observations per year), leaving 755 tradable strategy returns per configuration.

Two data-generating processes, identical in every respect except the presence of skill to find:

- *Noise (Experiment 3):* iid Gaussian returns with zero mean and standard deviation 0.01 per observation — a pure random walk in log price. No configuration has true skill; the correct verdict is rejection.
- *Regime edge (Experiment 4):* the same Gaussian noise plus a regime-switching drift $d_t = s_t \cdot 0.006$, where the state $s_t \in \{-1, +1\}$ flips with probability 0.02 per observation (expected regime length 50 observations). Trends persist for weeks at a time; a crossover rule genuinely can extract this. An omniscient regime-follower would earn per-observation Sharpe 0.6 (about 9.5 annualized — derived), so a good deal of real skill is available; the correct verdict is retention.

Each search is analyzed exactly as a practitioner would analyze a live sweep: naive p-value and un-deflated PSR against zero for the winner; DSR under the raw trial count $N = K$ and under each of the five effective-count estimators of Section 2.5, together with the survival crossing N^* of Eq. (10); Reality Check and the SPA-type test on the full 755×640 return matrix (5000 resamples for these case studies, expected block length 20); and the three Harvey–Liu haircuts with $M = K$.

4 Results

4.1 Null calibration: the naive verdict is always wrong

Table 1 shows the false-discovery rates. Selecting the best of 1000 zero-skill strategies produced an average best per-observation Sharpe corresponding to an annualized 1.63, and the naive analyst's median p-value on that winner was 0.00069 — apparently overwhelming evidence, in every single repetition (false-discovery rate 1.000). This is the multiple-testing machine at work: the naive test

Table 1: Experiment 1 (global null): empirical false-discovery rate at $\alpha = 0.05$, i.e. the share of repetitions in which each procedure certifies the best of $N = 1000$ zero-skill strategies as a discovery. Closed-form procedures: 2000 repetitions; Reality Check: 400 repetitions of 500 stationary-bootstrap resamples. Bonferroni and Holm are reported as one row because they coincide exactly for the top-ranked winner (Section 2.3). Monte Carlo standard error of a proportion near the nominal level is about 0.005.

Procedure	Discovery rule	Error criterion	FDR
Naive single test	$p < 0.05$	none (ignores search)	1.000
DSR	$\widehat{\mathrm{DSR}} > 0.95$	deflated benchmark	0.001
Harvey–Liu, Bonferroni = Holm	adj. $p < 0.05$	FWER	0.057
Harvey–Liu, BHY	adj. $p < 0.05$	FDR	0.007
White Reality Check	bootstrap $p < 0.05$	test of the max	0.0225

answers "could *this one* strategy be zero-skill?" when the question posed by the search is "could the *best of a thousand* be the product of zero skill?"

Every correction restores control, each with its own personality. Bonferroni and Holm — identical for the winner by construction — sit at 0.057, statistically indistinguishable from the nominal level (within about two Monte Carlo standard errors of 0.005; the small excess is consistent with the normal approximation to the t-statistic being slightly optimistic in the extreme tail where $\min_n p_n$ lives). BHY lands at 0.007 — close to $\alpha/c(1000) \approx 0.007$ (derived), which is exactly the conservatism its harmonic-sum constant buys as insurance against dependence that this iid experiment does not actually contain, and consistent with BHY being the most conservative of the three at rank one. The Reality Check comes in at 0.0225, conservative in the direction Hansen (2005) predicted for the unstandardized statistic. DSR is the strictest of all at 0.001, and the reason is instructive: its bar is not a significance level but the noise ceiling itself. The average deflated benchmark SR_0 across repetitions was 1.63 annualized (derived from the per-observation value) — identical to two decimals with the average realized maximum. Eq. (2) is doing precisely its job: predicting what the best of N lucky draws looks like, and refusing to be impressed by it. Consistent with calibration, the DSR value of the null winner averaged 0.495 across repetitions, indistinguishable from the ideal 0.5 — the estimator judges the typical lucky winner exactly as likely as not to beat the luck benchmark, which is the correct answer.

4.2 Planted power: the S-curve crosses just above the noise ceiling

Table 2 reports DSR's detection power against the planted true Sharpe ratio. The pattern is a sharp S-curve positioned just above the noise ceiling of Experiment 1: linear interpolation between the two adjacent grid levels puts the 50%-power point near an annualized 1.73 (derived), slightly above the ceiling of 1.63 rather than at it. A genuine annualized Sharpe of 0.79 — respectable in production — is essentially undetectable (power 0.005): it lives so far below the ceiling that the search's own noise drowns it, and indeed at that level the search winner is one of the planted strategies in only about two-thirds of repetitions (0.670). At annualized 1.27 the power is still only 0.090. Crossing the ceiling changes everything: 0.651 at annualized 1.90, then 0.998 at 2.54 and 1.000 at 3.17. Throughout the sweep the measured false-positive rate of DSR is 0.000 — in no repetition at any level did DSR certify an impostor. The naive column is a foil: it fires on essentially every winner (compare Table 1), so its apparent sensitivity carries no information about skill. We reiterate the scope note from Section 3.2: these are DSR power figures at one configuration, not a

Table 2: Experiment 2 (planted signal): DSR detection power over 1000 repetitions per level, with 25 planted strategies among $N = 1000$ and $T = 1000$. "Naive hit" is the share of repetitions in which the naive test certifies the winner *and* the winner is genuinely planted; since the naive test essentially always fires, this column mostly measures how often the best-by-Sharpe trial is a planted one.

True SR (per obs.)	True SR (annual)	DSR power	DSR false-pos. rate	Naive hit
0.05	0.79	0.005	0.000	0.670
0.08	1.27	0.090	0.000	0.984
0.12	1.90	0.651	0.000	1.000
0.16	2.54	0.998	0.000	1.000
0.20	3.17	1.000	0.000	1.000

general power study.

The practical reading is sobering and useful: given this search size and sample length, the deflated bar sits at an annualized Sharpe of roughly 1.63, and *no statistical procedure operating on the same evidence can do fundamentally better* — a sub-ceiling edge is not distinguishable from the best of a thousand lucky draws, however real it is. Deflation does not destroy value; it prices the evidence.

4.3 A realistic search on noise: everything correctly rejects

Table 3 (left column) reports the moving-average sweep on a pure random walk. The winner — crossover $(45, 120)$ — earned an annualized Sharpe of 0.81 over three years. Plotted, such an equity curve looks eminently fundable. Every layer of scrutiny disagrees. Even before any deflation, the winner is not single-test significant: the naive p-value is 0.081 and the un-deflated PSR against a zero benchmark is 0.918, short of the 0.95 threshold (the sample is short and the Sharpe modest). Deflation-aware diagnostics reject with progressively more context. DSR with the raw $N = 640$ gives 0.431. The Reality Check and the SPA-type test, which resample the entire 640-trial matrix jointly (5000 resamples), return p-values of 0.570 and 0.569: the observed best is entirely typical of a searched random walk. The Harvey–Liu adjusted p-value of the winner is 1.00 under Holm, and all three haircuts are total (1.00): the certified Sharpe after multiple-testing adjustment is zero.

The effective-count band (Table 4, left half) adds the robustness statement: the five estimators span 1.6 to 379.9, the implied deflated bars run from an annualized 0.31 to 0.87, and the implied DSR values run from 0.805 down to 0.455 — *every* choice rejects, and the survival crossing of Eq. (10) confirms there is no effective trial count at which this winner survives. One honesty note that Section 5 develops: at the smallest estimates the bar has collapsed to near zero and DSR nearly coincides with the un-deflated PSR (0.918), so the rejection *there* rests on the winner's own modest raw significance rather than on any real deflation; the genuinely deflating verdicts are the mid-range and upper rows.

4.4 The same search on a real edge: the effective-N band

The right column of Table 3 is the reason this paper exists. The regime-switching series contains genuine, exploitable trend persistence, and the search finds it: the winner $(3, 55)$ has an in-sample selected maximum of 3.92 annualized — selection-inflated by construction, but resting on real skill — with a naive p-value of 6.1×10^{-12} and an un-deflated PSR of 1.000. The bootstrap tests agree

Table 3: Experiments 3–4: the same 640-configuration moving-average crossover search ($T = 755$ tradable returns) on a pure random walk (left) and on a regime-switching series with a real trend edge (right). Winner Sharpe ratios are in-sample *selected* maxima and therefore selection-inflated by construction. Bootstrap tests use 5000 resamples. Bonferroni and Holm haircuts coincide for the top-ranked winner (Section 2.3) and are reported as one row. The DSR discovery threshold is 0.95; for RC, SPA-type, and adjusted p-values the threshold is $\alpha = 0.05$.

	Noise (no skill)	Regime edge (real skill)
Winner (f, s)	$(45, 120)$	$(3, 55)$
Winner SR, in-sample selected (annualized)	0.81	3.92
Naive one-sided p	0.081	6.1×10^{-12}
Un-deflated $\widehat{\mathrm{PSR}}(0)$	0.918	1.000
Mean pairwise correlation $\bar{\rho}$	0.61	0.62
SR_0 with raw $N = 640$ (annualized)	0.91	3.51
DSR with raw N	0.431	0.748
Reality Check p	0.570	0.0024
SPA-type (studentized RC) p	0.569	0.0038
Harvey–Liu haircut, Bonferroni = Holm	1.00	0.148
Harvey–Liu haircut, BHY	1.00	0.198
Holm-adjusted p of winner	1.00	3.9×10^{-9}
Verdict, DSR (raw N)	reject (correct)	reject (**wrong**)
Verdict, RC and SPA-type	reject (correct)	retain (correct)
Verdict, Harvey–Liu (all three)	reject (correct)	retain (correct)

emphatically: with 5000 resamples the Reality Check returns $p = 0.0024$ and the SPA-type test $p = 0.0038$ — interior p-values, not the resolution floor of the bootstrap. The Harvey–Liu haircuts are mild: 0.148 under Bonferroni and Holm (one number, by the rank-one identity), 0.198 under BHY — the most conservative for a top pick, exactly as Section 2.3 predicts — leaving certified annualized Sharpes of 3.33 and 3.14 (derived), with a Holm-adjusted p-value of 3.9×10^{-9}. Real skill, correctly retained, even after paying full freight for 640 trials.

DSR with the raw trial count reaches the opposite verdict. Fed $N = 640$, Eq. (2) sets the deflated benchmark at an annualized 3.51 — demanding that the winner beat the expected best of 640 *independent* zero-skill trials — and Eq. (3) returns 0.748, below the 0.95 threshold. A practitioner following the raw-N recipe would discard a genuine edge worth an annualized Sharpe of nearly four. Two distinct mechanisms inflate that bar. The first is multiplicity: the grid's 640 trials have mean pairwise correlation 0.62 — neighboring crossovers hold nearly the same positions — so 640 vastly overstates the number of independent chances luck was given. The second is subtler: the dispersion term of Eq. (2) treats the cross-trial spread of Sharpe estimates as luck, but in this search much of the spread is real — the per-observation trial-Sharpe dispersion $\sqrt{\widehat{V}_{SR}}$ is 0.079 here versus 0.014 on the noise search (derived) because genuinely skilled regions of the grid pull the distribution apart, and the formula reads that skill dispersion as more luck to deflate against.

The effective-count band (Table 4, right half) is the honest repair, and it must be read whole. The five estimators again span two orders of magnitude (1.6 to 370.0). Four of the five retain the edge. At the bottom anchor (average correlation 1.6, evaluated at the $N = 2$ floor) the bar collapses

Table 4: The effective-trial-count band for both searches: five standard estimators of N_{eff}, the deflated benchmark and DSR each implies, and the survival crossing N^* of Eq. (10). Annualized SR_0 values are reported by the harness; the raw grid count is included for reference. The average-correlation estimates fall below the $N = 2$ domain floor of Eq. (2) and are evaluated at that floor. The five estimators span two orders of magnitude on the same matrix — the reason we report a band rather than a point verdict.

	Noise (no skill)			Regime edge (real skill)		
Estimator of N_{eff}	N_{eff}	SR_0 (ann.)	DSR	N_{eff}	SR_0 (ann.)	DSR
Average correlation, Eq. (9)	1.6	0.31	0.805	1.6	0.25	1.000
Participation ratio	2.4	0.35	0.785	2.4	0.43	1.000
PCA (95% of variance)	17	0.61	0.633	16	1.85	1.000
Kaiser criterion	22	0.64	0.616	21	2.00	0.999
Cheverud–Nyholt	379.9	0.87	0.455	370.0	3.31	0.845
Raw grid count K	640	0.91	0.431	640	3.51	0.748
Survives (DSR > 0.95) for $N_{\text{eff}} < N^*$	no N_{eff} (never survives)			$N^* = 144.8$		

to an annualized 0.25 and DSR returns 1.000 — but at that anchor the deflation is nearly inert, and the retention is simply the winner's un-deflated PSR (1.000) shining through. The informative rows are the mid-range estimators: PCA-95 ($N_{\text{eff}} = 16$) and Kaiser ($N_{\text{eff}} = 21$) set genuinely deflating bars of an annualized 1.85 and 2.00 — comparable to the noise ceiling of Experiment 1 — and the edge still clears them decisively (DSR 1.000 and 0.999). Only the near-independence Cheverud–Nyholt estimate (370.0, close to the raw count of 640) flips the verdict, rejecting at 0.845. The survival crossing summarizes the whole curve: this winner is retained for every effective trial count below $N^* = 144.8$ — more than a fifth of the raw grid — and the noise winner of Section 4.3 is retained at none. That pair of statements, robust across every defensible mid-range estimate, is what this experiment licenses; a single "corrected" DSR at one hand-picked N_{eff} is not.

5 Discussion

5.1 Deflation restores calibration, and the bar is the noise ceiling

The headline result of Experiments 1–2 is that the machinery works exactly as advertised when its assumptions hold. The naive verdict on a searched winner is not slightly optimistic — it is always wrong under the null (false-discovery rate 1.000), with median apparent significance of 0.00069. Every principled correction restores the error rate to the neighborhood of the nominal $\alpha = 0.05$ (0.001–0.057). And the deflated benchmark is not an abstract penalty: SR_0 reproduced the empirically realized noise ceiling to two decimals (1.63 annualized for a search of 1000 trials over 1000 observations). The practical form of this statement deserves emphasis in any research process: *a search has a Sharpe budget that luck will pay regardless of skill*, computable in closed form from the trial count, trial dispersion, and sample length, and no winner below that budget is evidence of anything. Power against real edges is then a question of where the edge sits relative to the ceiling (Table 2): negligible below it, near-perfect above it, with the 50%-power point just above the ceiling (about an annualized 1.73, derived) and false positives at 0.000 throughout.

5.2 Effective N: report a band, and anchor it honestly

Experiment 4 documents the failure mode that we believe practitioners most need to know about, and Experiments 3–4 together show what can honestly be salvaged. It is tempting — it feels *conservative* — to feed DSR the raw size of the parameter grid. On a correlated search this over-deflates twice over: the raw count overstates the independent chances luck was given, and the dispersion term simultaneously reads real-skill spread as luck (0.079 versus 0.014 per observation across our two searches, derived). In our regime-edge search the resulting benchmark (annualized 3.51) exceeded what even a strong genuine edge could clear on three years of data, and DSR rejected a strategy whose reality is not in doubt — we planted it.

The obvious repair — substitute an effective trial count — immediately runs into the fact that there is no such number, only estimators that disagree by two orders of magnitude on the same matrix (1.6 to 379.9 on noise, 1.6 to 370.0 on the edge). Worse, the band's ends are individually treacherous. At the bottom, the average-correlation estimate (Eq. (9)) lands below two on both searches, where the expected-maximum benchmark barely deflates at all: the bars are an annualized 0.25 and 0.31, and DSR effectively degenerates to the un-deflated PSR against zero. Both verdicts at that anchor are *inherited* — the edge is retained because its raw PSR is 1.000, and the noise winner is rejected because its raw PSR happens to be 0.918, short of 0.95. The latter is luck of the seed: this particular noise winner is modest (annualized 0.81). A luckier noise draw, strong enough to clear single-test significance — and Experiment 1 shows searched noise winners routinely reach an annualized 1.63 — would sail over a bar of 0.31 unchallenged. An effective count below two does not deflate; it abdicates. We also note the functional mismatch of Section 2.5: Eq. (9) is the variance-reduction factor for a *mean* of equicorrelated variables, imported into a formula about the *maximum*; nothing guarantees those two notions of "effectively independent" coincide. At the top, the Cheverud–Nyholt estimate — a published, widely used recipe, known to over-count under strong equicorrelation — lands at 370.0 and reproduces the raw-count error almost exactly (DSR 0.845 versus 0.748), rejecting the genuine edge. So the claim "any defensible dependence adjustment beats the raw count" is false, and we do not make it.

What survives scrutiny is the band itself. The mid-range estimators — participation ratio, PCA-95, Kaiser, spanning roughly 2.4 to 22 — impose genuine deflation (bars up to an annualized 2.00 on the edge search) and deliver the correct verdict on both searches. The survival crossing N^* compresses the whole curve into one auditable statement: the edge is retained for every $N_{\text{eff}} < 144.8$, the noise winner for none. A discovery claim of the form "retained across the entire defensible range of effective counts, with a survival margin documented" is weaker than a single certified number, but it is the strongest statement this evidence actually supports — and unlike the point estimate, it does not silently hinge on which effective-count recipe the analyst happened to choose.

5.3 What the bootstrap tests know that DSR does not

The Reality Check and the SPA-type test never faced any of this because they never needed a trial count: resampling the full return matrix jointly carries the entire cross-trial dependence structure into the null distribution automatically. Their separation on the two searches is stark — $p = 0.570$ and 0.569 on noise versus $p = 0.0024$ and 0.0038 on the real edge, with 5000 resamples — and required no judgment calls. This robustness has a price and a scope. The price is data and computation: RC and SPA-type need the full $T \times K$ matrix of trial returns and thousands of bootstrap replications, whereas DSR needs only the list of trial Sharpe ratios — often the only thing archived from a historical search, and cheap at any scale. The scope difference is subtler: the two families answer different questions. DSR asks, "is the winner exceptional *relative to this search's*

own luck distribution?" RC and SPA-type ask, "after accounting for the full search, does the best rule beat the benchmark at all?" The first is a selection-adjusted quality bar on the champion; the second is an existence test for skill anywhere in the family. On our clean constructions they agree — on noise everything rejects at every effective count, and on the real edge the bootstrap tests, the Harvey–Liu haircuts, and DSR across its entire defensible mid-range all retain (Table 3, Table 4) — and their agreement across such different mechanisms is itself evidence; in messier settings they are complements, not substitutes.

5.4 A practical recipe

Our results suggest a concrete checklist for anyone auditing a backtest search. Log every trial, not just the winner — every estimator here consumes the whole search, and the common practice of archiving only the champion destroys exactly the information deflation needs. Report the deflated benchmark SR_0 alongside the winner's Sharpe, so readers see the noise ceiling the search itself created, and label the winner's Sharpe as the in-sample selected maximum that it is. Do not report DSR at a single effective trial count: compute the spread of standard estimators and the survival crossing N^*, and distrust both ends of the band — the smallest estimates barely deflate (the verdict is just un-deflated PSR), and the largest reproduce the raw-count over-deflation. A verdict that flips inside the defensible mid-range is not a discovery. When trial return series are available, run the Reality Check or its studentized SPA-type variant as an assumption-light cross-check. And treat the Harvey–Liu haircut as the communication layer: a statement like "after Holm adjustment across 640 trials the certified Sharpe drops by 0.148" is legible to any investment committee in a way that a bootstrap p-value is not — remembering that for a single winner Holm *is* Bonferroni, and BHY is the most conservative of the three.

6 Limitations

- **Synthetic data, by design.** Both data-generating processes — iid Gaussian returns and a two-state regime-switching drift — were chosen so that ground truth is known exactly, which is the point of a calibration study and also its boundary. Real returns have fat tails, volatility clustering, and structural breaks; our experiments say nothing about how large T must be for Eq. (1)'s moment corrections to absorb such features, only that the selection-bias machinery works when its assumptions hold.

- **No consensus effective-N estimator, and we do not supply one.** The five estimators we report are standard recipes, not derivations matched to DSR's extreme-value benchmark; the average-correlation form has the functional mismatch discussed in Section 2.5 (a mean-variance quantity applied to a maximum), and Cheverud–Nyholt is known to over-count under strong equicorrelation. Bailey and López de Prado's Appendix 3 clustering construction is the treatment we did not implement. Our claim is deliberately limited to the band: on these two searches the verdicts are stable across the defensible mid-range and summarized by N^*, and the band's ends are individually misleading in the ways documented in Section 5.2.

- **Power evidence is DSR-only and configuration-specific.** Experiment 2 measures the power of DSR at one configuration ($n_{\text{true}} = 25$, $N = 1000$, $T = 1000$) with no sensitivity sweep; the retention behavior of the Reality Check, the SPA-type test, and the haircuts is probed only through the single case study of Experiment 4.

- **RC/SPA described, not re-derived; SPA-type is not Hansen's full test.** As flagged in Section 2.4, we did not re-extract the equations of White (2000), Hansen (2005), or Politis and Romano (1994) from the primary texts; our description follows their standard characterization. Our implementation is a studentized Reality Check: it adopts Hansen's studentization but retains White's full recentering, and does not implement Hansen's consistent (sample-dependent) recentering, so it is labeled SPA-type throughout and is, if anything, conservative relative to Hansen's SPA.
- **Experiments 3–4 are case studies.** The two searches are single seeded draws — deliberately, since they play the role of worked examples with known truth. The calibration and power claims rest on the Monte Carlo experiments (2000 and 1000 repetitions); the search experiments demonstrate existence of the raw-count failure and the shape of the effective-count band, not their frequency across seeds.
- **No cost model, no execution claims.** The searches are frictionless and long–short unconstrained; wall-clock performance of the harness is likewise not a claim of this paper.

7 Conclusion

Under controlled conditions with known ground truth, the deflation toolkit does what it promises. The naive single-test verdict on a searched winner is worthless — wrong with certainty under the null — while the Deflated Sharpe Ratio, the Harvey–Liu haircuts, and the Reality Check all restore error control to the neighborhood of the nominal level, and the deflated benchmark equals the noise ceiling of the search to two decimals. Power is a function of where an edge sits relative to that ceiling, so deflation converts an unanswerable question ("is this backtest good?") into a computable one ("does this edge clear what luck alone would have bought?").

The toolkit's one sharp edge is the trial count inside DSR. On a realistic correlated grid, the raw count rejected a genuine edge that every other estimator certified — and the seemingly obvious repair, an effective number of trials, is not a number but a contested range spanning two orders of magnitude, whose smallest values barely deflate at all and whose largest reproduce the raw-count error. What the evidence does support is a robustness band: on these searches the genuine edge survives every effective count below 144.8, including mid-range bars that deflate as hard as the noise ceiling itself, while the noise winner survives none. The bootstrap tests sidestep the issue entirely by construction, at the cost of requiring the full trial return matrix. Used together — DSR reported as a band with its survival crossing, RC and the SPA-type test for the search, a haircut for the committee — the estimators delivered a coherent verdict on every question our experiments could pose, and their agreement across such different mechanisms is the most reassuring calibration result of all.

Reproducibility. All code, tests, and outputs accompany this paper: `scripts/run_all.py` regenerates `results/results.json` from fixed seeds (Python 3.14.6, NumPy 2.4.3); `scripts/check_paper_numbers.py` verifies every numeric claim in this manuscript against that file and fails on any mismatch; `tests/` contains deterministic invariant tests for every estimator used.

References

David H. Bailey and Marcos López de Prado. The sharpe ratio efficient frontier. *Journal of Risk*, 15 (2):3–44, 2012. doi: 10.21314/JOR.2012.255. Also available as SSRN Working Paper No. 1821643. Introduces the Probabilistic Sharpe Ratio (PSR), Eq. (11).

David H. Bailey and Marcos López de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014. doi: 10.3905/jpm.2014.40.5.094. Special 40th Anniversary Issue. Also available as SSRN Working Paper No. 2460551 and at `davidhbailey.com/dhbpapers/deflated-sharpe.pdf`. Defines the Deflated Sharpe Ratio (DSR), Eq. (2), and the expected maximum Sharpe ratio under N independent trials, Eq. (1)/Eq. (6).

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5):458–471, 2014. doi: 10.1090/noti1105.

David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. The probability of backtest overfitting. *Journal of Computational Finance*, 20(4):39–69, 2017. doi: 10.21314/JCF.2016.322. Online first 19 Sep 2016; print April 2017. Also available as SSRN Working Paper No. 2326253. Sibling paper to the DSR line of work: introduces the Probability of Backtest Overfitting (PBO) via Combinatorially Symmetric Cross-Validation (CSCV), a non-parametric companion diagnostic to the (parametric) DSR.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005. doi: 10.1198/073500105000000063.

Campbell R. Harvey and Yan Liu. Evaluating trading strategies. *Journal of Portfolio Management*, 40(5):108–118, 2014. doi: 10.3905/jpm.2014.40.5.108. Special 40th Anniversary Issue. Winner, Bernstein Fabozzi/Jacobs Levy Award for Best Article in JPM, 2014.

Campbell R. Harvey and Yan Liu. Backtesting. *Journal of Portfolio Management*, 42(1):13–28, 2015. doi: 10.3905/jpm.2015.42.1.013. Winner, Bernstein Fabozzi/Jacobs Levy Award for Best Article in JPM, 2016. Derives the Bonferroni, Holm, and BHY multiple-testing adjusted p-values and the haircut Sharpe ratio, Eqs. (1)–(5).

Campbell R. Harvey, Yan Liu, and Heqing Zhu. ...and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016. doi: 10.1093/rfs/hhv059. Editor's Choice. Also circulated as NBER Working Paper No. 20592 and SSRN Working Paper No. 2249314. Documents $\geq$316 factors tested in the cross-sectional-returns literature; motivates the $t > 3.0$ significance hurdle used in Harvey and Liu (2015).

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. No CrossRef-registered DOI found (pre-DOI-era journal article); JSTOR

stable URL: https://www.jstor.org/stable/4615733. Bibliographic details corroborated independently across JSTOR, Scientific Research Publishing reference database, and multiple citing papers, but the full text itself was not directly fetched – flagged as bibliographically VERIFIED / full-text UNVERIFIED.

Marcos López de Prado. *Advances in Financial Machine Learning*. John Wiley & Sons, Hoboken, NJ, 2018. ISBN 978-1-119-48208-6. Part III (backtesting) discusses backtest overfitting, PBO/CSCV, and DSR in the context of ML-driven strategy selection.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152.