NikaQuant Verdict Methodology

Version 1.0 ››› Draft for Quant Review

NikaQuant ApS, Aalborg, Denmark

2026-05-17

Contents

0.1 1. Abstract 1
0.2 2. The Four-Verdict System 2
0.3 3. Input Requirements 4
0.4 4. The Composite Score Rubric 6
0.5 5. The Robustness Battery 9
0.6 6. The Stress Battery 15
0.7 7. Statistical Significance Framework 17
0.8 8. Risk Metrics — Beyond Drawdown 18
0.9 9. Behavioral Quality 21
0.10 10. Multi-Strategy and Portfolio Analysis 22
0.11 11. Worked Example 23
0.12 12. What NikaQuant Does NOT Claim 27
0.13 13. References 27
0.14 Appendix A: Glossary 29
0.15 Appendix B: Reproducibility 31

0.1 1. Abstract

NikaQuant is a quantitative evaluation engine for MetaTrader 5 (MT5) trading strategies. It accepts standardized strategy performance reports and returns a structured verdict: one of four categorical outputs accompanied by a composite score, pillar-level sub-scores, and a detailed diagnostic report. This document defines, in precise and reproducible terms, exactly how those verdicts are produced.

The core premise of NikaQuant is adversarial evaluation. We do not evaluate strategies the way their authors do — with the benefit of hindsight selection, optimism about future conditions, and a natural desire to confirm what has already been built. We evaluate strategies the way a hostile peer reviewer would: looking first for reasons to reject, requiring statistical evidence before accepting any claim of edge, and applying a suite of stress tests designed to expose brittleness that a standard backtest conceals.

This stance is not arbitrary. The academic literature on strategy evaluation — from White (2000) to López de Prado (2018) — consistently documents that the majority of strategies that appear profitable in backtests fail to deliver positive returns out of sample. The primary causes are overfitting through excessive parameter search, look-ahead bias in data preparation, and the failure to account for the multiple comparisons problem (Harvey, Liu, and Zhu, 2016). NikaQuant's methodology is

designed to detect and penalize each of these failure modes.

The four verdict outputs are:

- **DEPLOY** — The strategy passes all critical thresholds, achieves a composite score ≥ 75, and has demonstrated statistically significant, robust, and stress-resistant performance. Deployment is recommended with the position sizing guidance provided in the report.
- **REFINE** — The strategy shows genuine signal but has identifiable weaknesses in one or more pillars. Composite score falls between 40 and 74. Specific remediation steps are recommended.
- **REJECT** — The strategy fails one or more critical thresholds (any single CRITICAL-level failure triggers this regardless of composite score), or the composite score falls below 40. The strategy is not suitable for deployment in its current form.
- **INSUFFICIENT DATA** — The analyzer cannot produce a reliable verdict because the input data fails to meet minimum sample requirements. No score is assigned; specific data requirements are enumerated.

NikaQuant does not predict future returns. A DEPLOY verdict means the historical evidence, subjected to our full battery of tests, is consistent with the hypothesis that the strategy has a genuine statistical edge. It does not mean that edge will persist. Section 12 enumerates our limitations explicitly.

Every NikaQuant verdict includes a cryptographic hash of the input data and the scoring engine version, ensuring full reproducibility (see Appendix B).

0.2 2. The Four-Verdict System

0.2.1 2.1 Composite Score Range and Interpretation

The composite score is a real number on the closed interval $[0, 100]$, computed as the weighted average of six pillar sub-scores (defined in Section 4). The following plain-English interpretation anchors specific score values:

Score	Interpretation
0	A strategy that loses money on every dimension: negative expectancy, catastrophic drawdown, fails all stress tests, and shows no statistical significance whatsoever. Essentially random noise with trading costs.
25	A strategy with marginally positive raw returns but severe structural weaknesses: high parameter sensitivity, poor walk-forward transfer, statistically insignificant edge, or drawdowns incompatible with practical trading.
50	A strategy that passes basic plausibility checks but does not clear the bar for robust deployment. Positive expectancy in-sample, but walk-forward efficiency is marginal, stress resistance is limited, or the statistical significance is below deployment threshold.

Score	Interpretation
75	A strategy that demonstrates consistent out-of-sample performance, acceptable drawdown profiles, stress-resistant returns, and statistically significant edge (PSR ≥ 0.95 against SR* = 0). This is the minimum threshold for DEPLOY.
100	A hypothetical perfect strategy: positive expectancy in all walk-forward windows, zero parameter sensitivity, PSR ≈ 1.0, passing all stress multipliers including 10x spread, with regime-conditional profitability across trend, chop, and crisis environments. No real strategy achieves 100.

0.2.2 2.2 Verdict Thresholds

Verdict	Composite Score Condition	Override Conditions
DEPLOY	≥ 75	No CRITICAL failures
REFINE	40–74	No CRITICAL failures
REJECT	< 40	OR any single CRITICAL failure regardless of composite
INSUFFICIENT DATA	N/A	See Section 2.3

The composite score is computed to two decimal places. A score of exactly 75.00 qualifies for DEPLOY provided no CRITICAL-level failures exist.

0.2.3 2.3 Override Rule: CRITICAL Failures

A CRITICAL failure in any single pillar forces a REJECT verdict regardless of the composite score. This rule exists because a strategy can score adequately on five pillars while being fatally flawed on the sixth — for example, a high-profit-factor strategy that generates its edge entirely from a small number of outlier trades, rendering its statistics meaningless. CRITICAL thresholds are defined for each pillar in Section 4. Representative examples:

- **Profitability:** Net profit factor < 1.0 on out-of-sample data (the strategy loses money net of costs in the period it was not fitted to).
- **Robustness:** Walk-Forward Efficiency (WFE) < 0.0 (the strategy systematically degrades out of sample, not just fails to transfer — it actually reverses).
- **Risk:** Maximum drawdown > 80% of initial capital.
- **Statistical Significance:** Probabilistic Sharpe Ratio < 0.50 (the estimated SR cannot be distinguished from zero with even 50% confidence).
- **Stress Resistance:** Net P&L becomes negative at 2x spread (the strategy is nonviable under the mildest spread stress).
- **Behavioral Quality:** Fewer than 20% of trading days show any activity in the most recent 20% of the backtest period (the strategy has effectively stopped trading).

0.2.4 2.4 Override Rule: INSUFFICIENT DATA

The INSUFFICIENT DATA verdict is triggered when any of the following conditions are met. In this state, no composite score is assigned, and the report enumerates exactly which conditions failed and by how much.

Condition	Threshold	Rationale
Total closed trades	< 100	Below this count, all sample statistics have confidence intervals too wide to be actionable. The PSR denominator becomes unreliable.
Backtest period length	< 6 calendar months	Shorter periods cannot capture seasonal patterns, spread-cycle variation, or regime transitions.
Walk-forward windows completed	< 5	Walk-Forward Efficiency (WFE) computed over fewer than five windows lacks statistical stability. Pardo (2008) recommends a minimum of five complete cycles.
Data coverage	< 95% of expected trading sessions	Excessive gaps in the data suggest either data quality problems or cherry-picked periods.

When INSUFFICIENT DATA is returned, the report includes the actual values observed and the precise shortfall for each failed condition. No partial scoring is performed.

0.3 3. Input Requirements

0.3.1 3.1 Accepted File Formats

NikaQuant accepts three input formats from MetaTrader 5:

MT5 HTML Report. The standard strategy tester output exported via “Save as Report” from the MT5 Strategy Tester. This file contains a complete trade-by-trade log including open/close times, volumes, entry/exit prices, swap, commission, and profit. NikaQuant parses the HTML table structure and reconstructs the equity curve at bar-by-bar resolution where timestamps permit.

MT5 XML Optimization Export. The XML file produced by the Strategy Tester’s optimization mode, containing the parameter sweep results. This file is used specifically by the parameter sensitivity module (Section 5.7) to reconstruct the fitness landscape across the parameter space. Without this file, the parameter sensitivity sub-score is set to 50 (neutral) with a corresponding WARN flag.

Raw FXT Tick Files. MT5-format tick data files providing bid/ask price series at tick resolution. When provided, these files enable the spread reconstruction needed for the spread shock battery (Section 6.1) and allow NikaQuant to verify that the backtest used realistic spread data rather than fixed-spread assumptions.

0.3.2 3.2 Minimum Sample Size Requirements

Metric	Minimum	Preferred	Rationale
Closed trades	100	≥ 300	PSR standard error stabilizes above ~200 observations (Lo, 2002)
Backtest period	6 months	≥ 24 months	Captures at least one minor regime transition
Walk-forward windows	5	≥ 10	WFE requires a distribution, not a point estimate
Losing trades	≥ 10	≥ 30	Downside deviation and tail ratio require a meaningful loss distribution

0.3.3 3.3 Data Quality Checks

NikaQuant performs the following checks before scoring begins. Any CHECK_FAIL triggers a data quality warning that appears in the report header.

Gap Detection. For each pair of consecutive trades with an open time gap greater than 5 trading days during a period not encompassing a known public holiday, NikaQuant flags a potential data gap. If more than 5% of the expected trading window is flagged as gapped, the data coverage condition fails (see Section 2.4).

Weekend Anomalies. Any trade opened or closed between Friday 22:00 UTC and Sunday 21:00 UTC (approximate FX market close/open) is flagged. A small number of trades in this window may be legitimate (e.g., Sunday open gaps), but a large proportion indicates incorrect broker session data or timestamp errors.

Broker Spread Inconsistency. NikaQuant computes the implied spread on each trade as the difference between the entry price and the best achievable fill assuming the reported open time. Where MT5 HTML reports include spread data explicitly, these values are extracted directly. Where they are absent, spread is inferred from the gap between the reported open price and the tick-file mid-price. A coefficient of variation (CV) of implied spread > 2.0 triggers a WARN; if the minimum implied spread is negative (indicating a data error), a CHECK_FAIL is raised.

Equity Curve Reconstruction. NikaQuant reconstructs the floating equity curve from the trade log. If the reconstructed final equity deviates from the reported final balance by more than 0.1%, a CHECK_FAIL is raised (indicating possible unreported adjustments, dividends, or data corruption).

0.4 4. The Composite Score Rubric

The composite score is a weighted average of six pillar sub-scores, each on a $[0, 100]$ scale:

$$\text{Composite} = \sum_{i=1}^{6} w_i \cdot S_i$$

where w_i are the pillar weights and $S_i \in [0, 100]$ are the pillar sub-scores. The weights are:

Pillar	Weight w_i	Rationale for Weight
Profitability	15%	Raw returns are necessary but not sufficient; a strategy can show high returns through luck or curve-fitting alone.
Robustness	25%	The single most important dimension: a strategy that does not transfer out of sample has no deployable value. This pillar receives the highest weight.
Risk	20%	Strategies with intolerable drawdown profiles cannot be traded in practice regardless of average return.
Statistical Significance	15%	Without statistical evidence of genuine edge, positive returns may be noise.
Stress Resistance	15%	Live conditions differ from backtest conditions; stress tests proxy for that gap.
Behavioral Quality	10%	Trade-level behavioral signals detect anomalies that aggregate metrics miss.

0.4.1 4.1 Pillar 1: Profitability (Weight 15%)

Metrics computed: Expectancy per trade, profit factor, annualized Sharpe ratio, Sortino ratio, Calmar ratio.

Expectancy is defined as the average dollar return per trade:

$$E = \frac{\sum_{i=1}^{N} P_i}{N}$$

where P_i is the net profit (after commission and swap) of trade i. A negative expectancy is a CRITICAL failure.

Profit Factor is defined as:

$$PF = \frac{\sum_{i:P_i>0} P_i}{\left|\sum_{i:P_i<0} P_i\right|}$$

A profit factor of 1.0 means breakeven; values above 1.0 indicate gross profitability before any cost shocks.

Annualized Sharpe Ratio (Sharpe, 1966):

$$SR = \frac{\bar{R} - R_f}{\sigma_R} \cdot \sqrt{f}$$

where $\bar{R}$ is the mean periodic return, R_f is the risk-free rate for the same period (NikaQuant uses 0 as the benchmark when the strategy is evaluated on a standalone basis, consistent with Lopez de Prado and Bailey (2012)), σ_R is the standard deviation of periodic returns, and f is the annualization factor (252 for daily equity, 12 for monthly equity).

Pillar 1 Sub-score Scaling:

Metric	Raw Value → Sub-score Mapping	PASS	WARN	CRITICAL
Expectancy	≤ 0 → 0; 0–2× avg spread → 0–40; 2×–5× avg spread → 40–70; > 5× avg spread → 70–100	> 0	0 to avg spread	≤ 0
Profit Factor	< 1.0 → 0; 1.0–1.3 → 0–40; 1.3–2.0 → 40–75; > 2.0 → 75–100	> 1.5	1.0–1.5	< 1.0
Sharpe Ratio	< 0 → 0; 0–0.5 → 0–25; 0.5–1.0 → 25–60; 1.0–2.0 → 60–85; > 2.0 → 85–100	> 1.0	0.5–1.0	< 0
Sortino Ratio	Scaled identically to Sharpe; penalizes negative asymmetry more severely	> 1.5	0.7–1.5	< 0
Calmar Ratio	< 0 → 0; 0–0.3 → 0–30; 0.3–1.0 → 30–65; > 1.0 → 65–100	> 0.5	0.2–0.5	< 0

The Pillar 1 sub-score S_1 is the arithmetic mean of the five individual metric scores.

0.4.2 4.2 Pillar 2: Robustness (Weight 25%)

Metrics computed: Walk-Forward Efficiency (WFE), Deflated Sharpe Ratio (DSR), parameter stability index.

See Section 5 for full formula derivations. The sub-score scaling is:

Metric	PASS	WARN	CRITICAL
WFE	≥ 0.5	0.25–0.5	< 0.0
DSR	≥ 0.95	0.75–0.95	< 0.50
Parameter Stability Index	≥ 0.80	0.60–0.80	< 0.40

The Pillar 2 sub-score is a weighted mean: WFE (50%), DSR (30%), Parameter Stability (20%).

0.4.3 4.3 Pillar 3: Risk (Weight 20%)

Metrics computed: Maximum drawdown, recovery factor, Ulcer Index, tail ratio, time-under-water distribution.

See Section 8 for full formula derivations. Sub-score thresholds:

Metric	PASS	WARN	CRITICAL
Max Drawdown (% of peak equity)	< 15%	15–30%	> 50%
Recovery Factor	> 3.0	1.0–3.0	< 1.0
Ulcer Index	Low (context-dependent)	Moderate	> 20%
Tail Ratio (95th/5th percentile of returns)	> 1.0	0.7–1.0	< 0.5
Max Time-Under-Water	< 30 days	30–90 days	> 180 days

0.4.4 4.4 Pillar 4: Statistical Significance (Weight 15%)

Metrics computed: Probabilistic Sharpe Ratio (PSR), Deflated Sharpe Ratio (DSR, shared with Pillar 2), White's Reality Check p-value, permutation test p-value.

Metric	PASS	WARN	CRITICAL
PSR (SR* = 0)	≥ 0.95	0.75–0.95	< 0.50
White's Reality Check p-value	< 0.05	0.05–0.10	> 0.20
Permutation test p-value	< 0.05	0.05–0.10	> 0.20

0.4.5 4.5 Pillar 5: Stress Resistance (Weight 15%)

Metrics computed: Returns under spread shock (1×/2×/5×/10×), slippage shock (+1/+3/+10 pips), commission shock (1×/2×/5×), regime-conditional expectancy.

Shock Level	Minimum Outcome for PASS
2× spread	Net P&L > 0
5× spread	Net P&L > –25% of baseline
10× spread	Net P&L > –50% of baseline
+3 pip slippage	Net P&L > 0
2× commission	Net P&L > 0
Crisis regime	Expectancy ≥ –1× avg spread per trade

0.4.6 4.6 Pillar 6: Behavioral Quality (Weight 10%)

Metrics computed: Trade frequency consistency, activity gap detection, MAE/MFE ratio, news-window exposure.

Metric	PASS	WARN	CRITICAL
Trade frequency coefficient of variation (CV)	< 0.5	0.5–1.0	> 2.0
Longest activity gap	< 10 trading days	10–30 days	> 60 days
MAE/MFE ratio	< 1.0	1.0–1.5	> 2.0
News-window exposure	< 20% of trades	20–40%	> 60%

0.5 5. The Robustness Battery

Robustness is the primary determinant of deployability. A strategy that performs well in-sample but fails to transfer its performance to out-of-sample periods is not a strategy — it is a description of the past. This section defines each robustness test precisely.

0.5.1 5.1 Walk-Forward Analysis

Walk-Forward Analysis (WFA) was systematized by Robert Pardo (1992, 2008) as the standard method for testing whether a strategy's optimized parameters generalize beyond the optimization window.

Protocol. The total historical period $[T_0, T_n]$ is divided into k sequential, non-overlapping cycles (NikaQuant requires $k \geq 5$). Each cycle consists of an **in-sample (IS)** window of length L_{IS} followed by an **out-of-sample (OOS)** window of length L_{OOS}. The strategy is optimized on the IS window and then evaluated — without re-optimization — on the immediately following OOS window. This process is rolled forward across all k cycles.

NikaQuant supports two variants:

Rolling WFA: Both IS and OOS windows roll forward together. Each cycle uses a fresh IS window of fixed length. This variant tests whether the strategy can adapt as market conditions evolve.

Anchored WFA: The IS window is anchored at T_0 and expands in each cycle; only the OOS window rolls. This variant tests stability as the training set grows.

IS/OOS Ratio. Following Pardo (2008), NikaQuant uses a default IS/OOS ratio of 4:1 (80% IS, 20% OOS per cycle). This ratio represents a practical balance: too small an IS window may not provide sufficient data to optimize reliably; too large an OOS window reduces the number of cycles for a given historical period.

Walk-Forward Efficiency (WFE). WFE measures how much of the in-sample performance is preserved out of sample:

$$\text{WFE} = \frac{\bar{R}_{OOS}}{\bar{R}_{IS}}$$

where $\bar{R}_{OOS}$ is the mean annualized return across all k OOS windows and $\bar{R}_{IS}$ is

the mean annualized return across all k IS windows (both computed on the same metric, typically annualized net return or Sharpe ratio).

Interpretation of WFE:

WFE Value	Interpretation
> 1.0	OOS outperforms IS — unusual, may indicate the IS period was particularly adverse.
0.5–1.0	Acceptable performance transfer. The strategy retains at least half its in-sample effectiveness. NikaQuant PASS threshold.
0.25–0.5	Marginal transfer. The strategy degrades significantly out of sample. WARN.
0.0–0.25	Poor transfer. The strategy barely breaks even out of sample.
< 0.0	The strategy reverses sign out of sample. CRITICAL failure.

Threshold justification. The WFE ≥ 0.5 PASS threshold follows the convention established by Pardo (2008, Ch. 8) and cited in practice by Chan (2013, Ch. 3). It reflects the expectation that a genuine trading edge should survive at least half its in-sample advantage in a clean out-of-sample test.

Worked numerical example. Suppose a strategy is tested over 5 rolling cycles with the following results:

Cycle	IS Ann. Sharpe	OOS Ann. Sharpe
1	1.82	0.94
2	2.10	1.05
3	1.65	0.72
4	1.90	0.68
5	2.20	1.10

$$\bar{R}_{IS} = (1.82 + 2.10 + 1.65 + 1.90 + 2.20)/5 = 1.934$$

$$\bar{R}_{OOS} = (0.94 + 1.05 + 0.72 + 0.68 + 1.10)/5 = 0.898$$

$$\text{WFE} = 0.898/1.934 = 0.464$$

Result: WARN (0.464 < 0.5). The strategy degrades by approximately 54% out of sample. This is not a CRITICAL failure but indicates meaningful overfitting. NikaQuant would recommend reducing parameter count or widening optimization constraints.

0.5.2 5.2 Combinatorially Purged Cross-Validation (CPCV)

Standard k-fold cross-validation fails for financial time series for two reasons identified by López de Prado (2018, *Advances in Financial Machine Learning*, Ch. 7):

1. **Serial correlation:** Financial returns are autocorrelated. Observations on adjacent days share information. A training set containing observations immediately adjacent to a test observation will leak information into the evaluation, inflating apparent performance.

2. **Label overlap:** When strategy labels (e.g., "profitable trade") are derived from future price movements, any training observation whose label period extends into the test period contaminates the model.

Purging addresses the first problem: all training observations whose *evaluation* timestamp falls within the test period are removed from training. Formally, if a test set occupies $[t_a, t_b]$, any training observation whose label is determined by prices in $[t_a, t_b]$ is purged.

Embargoing addresses residual leakage from autocorrelation: after the test period, an embargo window of length h is imposed on the following training observations:

$$h = \lceil \rho_1 \cdot \bar{L} \rceil$$

where ρ_1 is the first-order autocorrelation of strategy returns and $\bar{L}$ is the average label length in periods.

CPCV Construction. Given N sequential groups and test-set size k, the number of unique train-test combinations is $\binom{N}{k}$. Each observation appears in exactly k test sets, and the number of unique backtest paths is:

$$\varphi[N, k] = \frac{k}{N}\binom{N}{k}$$

NikaQuant uses $N = 10$ groups and $k = 2$ by default, producing $\varphi = 45$ unique backtest paths. These 45 paths provide a distribution of performance outcomes from which the Sharpe ratio distribution can be estimated non-parametrically — a far more reliable basis for inference than a single backtest path.

The CPCV Sharpe distribution feeds directly into the Probabilistic Sharpe Ratio calculation (Section 5.4), replacing the parametric assumption with an empirical distribution.

0.5.3 5.3 Deflated Sharpe Ratio (DSR)

When a researcher tests M strategy configurations and reports the best one, the reported Sharpe ratio is positively biased because selection from a maximum is a biased estimator. Bailey and López de Prado (2014) quantify this bias and construct the Deflated Sharpe Ratio (DSR) to correct for it.

Expected maximum Sharpe under the null. Given N independent trials with mean Sharpe of zero, the expected maximum Sharpe ratio is approximately:

$$\widehat{SR}_0 \approx \sqrt{V[\{\widehat{SR}_n\}]}\left[(1-\gamma)Z^{-1}\left[1-\frac{1}{N}\right] + \gamma\, Z^{-1}\left[1-\frac{1}{Ne}\right]\right]$$

where $V[\{\widehat{SR}_n\}]$ is the variance of Sharpe ratios across trials, $\gamma \approx 0.5772$ is the Euler–Mascheroni constant, Z^{-1} is the inverse CDF of the standard normal, e is Euler's number, and N is the effective number of independent trials (accounting for correlation between configurations via $N = \hat{\rho} + (1-\hat{\rho})M$ where $\hat{\rho}$ is the average inter-trial correlation).

DSR Formula. The DSR is the Probabilistic Sharpe Ratio evaluated at the deflated benchmark $\widehat{SR}_0$:

$$\widehat{DSR} = Z\left[\frac{(\widehat{SR} - \widehat{SR}_0)\sqrt{n-1}}{\sqrt{1 + \frac{1}{2}\widehat{SR}^2 - \hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 3}{4}\widehat{SR}^2}}\right]$$

where $\widehat{SR}$ is the estimated Sharpe of the best strategy, n is the number of return observations, $\hat{\gamma}_3$ is the skewness, and $\hat{\gamma}_4$ is the kurtosis of the strategy returns. $Z[\cdot]$ is the standard normal CDF.

Interpretation. DSR is a probability: the probability that the best strategy's Sharpe ratio exceeds the expected maximum Sharpe ratio that would arise from pure chance given the breadth of the search. DSR ≥ 0.95 is the PASS threshold, consistent with a 5% false discovery rate at the strategy level.

Worked numerical example. A researcher tests $M = 500$ parameter combinations, of which $\hat{\rho} = 0.7$ average correlation implies $N = 0.7 + 0.3 \times 500 = 150.7$ effective independent trials. The variance across trials is $V = 0.15^2$. With $\gamma = 0.5772$:

$$\widehat{SR}_0 \approx 0.15 \cdot \left[(0.4228) \cdot Z^{-1}[1 - 1/150.7] + 0.5772 \cdot Z^{-1}[1 - 1/(150.7 \cdot e)]\right]$$

$$\approx 0.15 \cdot [0.4228 \cdot 3.08 + 0.5772 \cdot 2.66] \approx 0.15 \cdot [1.302 + 1.535] \approx 0.426$$

The selected strategy has $\widehat{SR} = 1.20$, $n = 500$ observations, $\hat{\gamma}_3 = -0.3$, $\hat{\gamma}_4 = 4.5$:

$$\text{denominator} = \sqrt{1 + 0.5 \times 1.44 - (-0.3)(1.20) + \frac{4.5 - 3}{4}(1.44)} = \sqrt{1 + 0.72 + 0.36 + 0.54} =$$

$$\widehat{DSR} = Z\left[\frac{(1.20 - 0.426)\sqrt{499}}{1.618}\right] = Z\left[\frac{0.774 \times 22.34}{1.618}\right] = Z[10.69] \approx 1.0$$

In this case the selected strategy easily clears the DSR threshold. The test becomes more punishing when M is large, $\hat{\rho}$ is low (more effective independent trials), and $\widehat{SR}$ is only marginally positive.

0.5.4 5.4 Probabilistic Sharpe Ratio (PSR)

Bailey and López de Prado (2012) observe that the standard Sharpe ratio estimate $\widehat{SR}$ is itself a random variable whose sampling distribution depends on the non-normality of returns. The Probabilistic Sharpe Ratio converts the point estimate into a probability statement.

Formula. Given a benchmark Sharpe ratio SR^*, the PSR is:

$$\widehat{PSR}(SR^*) = Z\left[\frac{(\widehat{SR} - SR^*)\sqrt{n-1}}{\sqrt{1 - \hat{\gamma}_3\widehat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\widehat{SR}^2}}\right]$$

where $\hat{\gamma}_3$ is the sample skewness of returns, $\hat{\gamma}_4$ is the sample kurtosis, n is the number of return observations, and $Z[\cdot]$ is the standard normal CDF.

Key properties: - PSR increases with higher observed $\widehat{SR}$ - PSR increases with more observations n - PSR increases with positive skewness $\hat{\gamma}_3 > 0$ (positive skew reduces standard error) - PSR decreases with higher kurtosis $\hat{\gamma}_4$ (fat tails inflate standard error)

The standard Sharpe ratio benchmark $SR^* = 0$ corresponds to the null hypothesis that the strategy has no edge. The threshold **PSR ≥ 0.95** for DEPLOY means: with 95% probability, the estimated Sharpe ratio is greater than zero, after accounting for non-normality of returns.

Worked numerical example. A strategy shows $\widehat{SR} = 0.85$ (annualized, daily), with $n = 252$ observations, $\hat{\gamma}_3 = -0.5$ (negative skew, common in trend-following), $\hat{\gamma}_4 = 5.2$ (excess kurtosis 2.2 above normal):

$$\text{denominator} = \sqrt{1 - (-0.5)(0.85) + \tfrac{5.2-1}{4}(0.85)^2} = \sqrt{1 + 0.425 + \tfrac{4.2}{4}(0.7225)} = \sqrt{1 + 0.425$$

$$\widehat{PSR}(0) = Z\left[\frac{(0.85)\sqrt{251}}{1.478}\right] = Z\left[\frac{0.85 \times 15.84}{1.478}\right] = Z\left[\frac{13.46}{1.478}\right] = Z[9.11] \approx 1.0$$

Now suppose instead $\widehat{SR} = 0.30$, $n = 60$ (only two months of daily data), same skewness and kurtosis:

$$\widehat{PSR}(0) = Z\left[\frac{(0.30)\sqrt{59}}{1.478}\right] = Z\left[\frac{0.30 \times 7.68}{1.478}\right] = Z[1.560] \approx 0.940$$

This sub-0.95 PSR illustrates why the INSUFFICIENT DATA threshold of 100 trades exists: small samples frequently fail the PSR test even for strategies with genuine edge.

0.5.5 5.5 White's Reality Check

White (2000) addresses the multiple-comparison problem in strategy evaluation through a bootstrap-based hypothesis test. The null hypothesis is that no tested strategy has positive expected performance relative to a naive benchmark (e.g., buy-and-hold or zero-return):

$$H_0 : \max_{k=1,\ldots,K} \mathbb{E}\left[f_k\right] \leq 0$$

where f_k is the performance statistic for strategy k and K is the total number of strategies evaluated.

Procedure. The Reality Check p-value is constructed by resampling blocks of the performance series (using the stationary bootstrap, Section 5.6) and computing the empirical distribution of the maximum performance statistic under the null. The p-value is the fraction of bootstrap iterations in which the maximum bootstrap statistic exceeds the observed maximum:

$$p_{\text{RC}} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}[V_b^* \geq V]$$

where $V = \max_k \bar{f}_k$ is the observed test statistic and V_b^* is the corresponding statistic in bootstrap replication b.

NikaQuant applies the Reality Check using $B = 2000$ bootstrap replications. The p-value threshold is $p_{\text{RC}} < 0.05$ for PASS. As Aronson (2006, *Evidence-Based Technical Analysis*, Ch. 6) notes, this test is conservative — it controls the family-wise error rate, not the false discovery rate — which is appropriate for a methodology explicitly designed to be skeptical.

0.5.6 5.6 Block Bootstrap

Independent and identically distributed (IID) bootstrap resampling is inappropriate for financial time series because returns exhibit autocorrelation, volatility clustering, and other serial dependencies. Resampling individual observations independently destroys these dependencies and produces bootstrap samples with statistical properties materially different from the original series.

Stationary Block Bootstrap. Politis and Romano (1994) propose the stationary block bootstrap, which resamples contiguous *blocks* of observations. Block length is drawn from a geometric distribution with parameter p, giving an expected block length of $1/p$. This produces bootstrap samples that are stationary (unlike the fixed-length block bootstrap of Künsch, 1989) and preserve short-range serial dependence.

Optimal Block Length. Politis and White (2004) derive a data-driven method for selecting the optimal expected block length ℓ^*:

$$\ell^* = \left(\frac{2D^2}{\hat{B}} \right)^{1/3} T^{1/3}$$

where D and $\hat{B}$ are spectral quantities derived from the autocorrelation function of the squared return series, and T is the sample size. In practice, for typical FX strategy return series, ℓ^* falls in the range of 5–25 trading days.

NikaQuant computes the optimal block length using the Politis-White (2004) estimator for each evaluated strategy. This block length is reported alongside bootstrap results so that the procedure is fully reproducible.

0.5.7 5.7 Parameter Sensitivity Analysis

A strategy that has been tightly optimized over a parameter space may achieve its reported performance only in a narrow region of that space. A genuine edge should produce profitable performance across a neighborhood of the optimal parameters.

Perturbation Protocol. For each optimized parameter θ_j in the strategy's parameter set $\Theta = \{\theta_1, \ldots, \theta_P\}$, NikaQuant evaluates the strategy at four perturbed values: $\theta_j \pm 0.10\theta_j$ (±10%) and $\theta_j \pm 0.25\theta_j$ (±25%), holding all other parameters fixed. The performance metric (annualized Sharpe) is computed at each perturbed point.

Stability Index. The stability index Σ_j for parameter θ_j is defined as:

$$\Sigma_j = 1 - \frac{\max\left\{SR(\theta_j) - SR(\theta_j + \delta_k) : k = 1, 2, 3, 4\right\}}{SR(\theta_j)}$$

where the four perturbations $\delta_k \in \{-0.25\theta_j, -0.10\theta_j, +0.10\theta_j, +0.25\theta_j\}$. A stability index of 1.0 means the perturbed strategies match the baseline; 0.0 means the strategy collapses entirely at maximum perturbation.

The portfolio-level **Parameter Stability Index** (PSI) is the minimum stability index across all parameters:

$$\text{PSI} = \min_j \Sigma_j$$

Using the minimum rather than the mean ensures that a single fragile parameter (e.g., a stop-loss threshold below which the strategy reverses) is not obscured by averaging. The PASS threshold is PSI ≥ 0.80.

0.6 6. The Stress Battery

Live trading conditions are invariably worse than backtest conditions. Spreads widen during news events and low-liquidity periods. Execution slippage arises from order queue position, latency, and partial fills. Commissions may change. Market regimes shift. The Stress Battery systematically quantifies how sensitive the strategy's net P&L is to each of these factors.

0.6.1 6.1 Spread Shock

For each closed trade in the backtest, the realized spread s_0 is the spread assumed at execution (reconstructed from the MT5 report or tick data). Under a spread multiplier λ, the additional cost per trade is:

$$\Delta P_i(\lambda) = -(\lambda - 1) \cdot s_0 \cdot V_i$$

where V_i is the trade volume in lots and s_0 is in price units per lot. The shocked P&L for trade i is $P_i^{(\lambda)} = P_i + \Delta P_i(\lambda)$, and the shocked total net P&L is $\sum_i P_i^{(\lambda)}$.

NikaQuant evaluates spread multipliers $\lambda \in \{1, 2, 5, 10\}$. The 2× case represents a commonly experienced spread widening during moderate volatility. The 10×

case represents extreme conditions such as flash crashes or major news events (e.g., SNB EUR/CHF floor removal, January 2015, during which spreads on CHF pairs briefly exceeded 1000 pips).

Interpretation table:

Spread Multiplier	Threshold for PASS	Failure Consequence
2×	Net P&L > 0	CRITICAL (strategy is nonviable at mild spread stress)
5×	Net P&L > –25% of baseline	WARN if failed
10×	Net P&L > –50% of baseline	WARN if failed

0.6.2 6.2 Slippage Shock

Slippage is modeled as a per-trade cost applied uniformly to the exit price. For a long trade, adverse slippage means the exit fill is δ pips worse than requested. For a short trade, adverse slippage increases the buy-back price by δ pips.

NikaQuant evaluates slippage $\delta \in \{1, 3, 10\}$ pips. The +3 pip case is representative of typical ECN execution during moderate-volatility sessions. The +10 pip case represents execution during high-impact news events, particularly relevant for strategies that trade around major economic releases.

0.6.3 6.3 Commission Shock

Many MT5 backtests use the broker's advertised commission schedule. However, commission structures can change, and some strategies are evaluated with unrealistically low commission assumptions. The commission shock multiplies all commission costs by $\lambda_c \in \{1, 2, 5\}$.

0.6.4 6.4 Latency Shock

Latency introduces execution delay: the strategy signals at time t but executes at time $t + \tau$. In fast-moving markets, a delay of τ milliseconds can result in a price miss of:

$$\Delta P_\tau = v \cdot \tau$$

where v is the average price velocity during trading periods (estimated from tick data as the mean absolute price change per millisecond during the strategy's typical trading hours).

NikaQuant simulates $\tau \in \{50, 200, 1000\}$ milliseconds. For slow, daily-bar strategies, latency is typically immaterial. For intraday scalping strategies, 200ms latency can materially affect P&L and may trigger a WARN or CRITICAL depending on the strategy's average tick-target.

0.6.5 6.5 Regime Stress

Market regimes materially affect strategy performance. A trend-following strategy may perform excellently in trending regimes and generate consistent losses in choppy, mean-reverting regimes. Reporting only the aggregate performance conflates these regime exposures.

Regime Detection. NikaQuant applies a 3-state Hidden Markov Model (HMM) to the daily return series of the underlying instrument (not the strategy). States are identified and labeled:

- **Trend regime:** High autocorrelation of daily returns, expanding Hurst exponent (H > 0.55), low volatility-of-volatility.
- **Chop regime:** Low autocorrelation, Hurst exponent near 0.5, moderate volatility.
- **Crisis regime:** High volatility, negative return autocorrelation (overreaction/reversal dynamics), elevated VIX proxy.
- **Low-volatility regime:** Very low volatility, near-zero autocorrelation, Hurst exponent indeterminate.

Per-Regime Metrics. Each closed trade is tagged with the regime active at its open time. NikaQuant computes per-regime expectancy, profit factor, and Sharpe ratio. A strategy with materially negative expectancy in any single regime that accounts for more than 25% of the backtest period receives a WARN. A strategy with CRITICAL-level losses in any regime receives a REJECT override.

0.7 7. Statistical Significance Framework

0.7.1 7.1 The Multiple Comparisons Problem

When a researcher tests a single strategy, a p-value of 0.05 means there is a 5% probability of observing the reported performance by chance if the null hypothesis (no edge) is true. When the same researcher tests 100 strategies and reports the best one, the probability that at least one of the 100 shows a spurious p < 0.05 by chance is $1 - 0.95^{100} \approx 0.994$. In other words, the reported significance is almost certain to be spurious.

Harvey, Liu, and Zhu (2016) document that the threshold t-statistic for 5% significance in published financial research, correcting for the cumulative effect of all factors ever tested in the literature, should be approximately 3.0 (not the conventional 2.0). For unpublished strategies with a large private parameter search, the appropriate threshold may be even higher.

0.7.2 7.2 Bonferroni and Related Corrections

The Bonferroni correction adjusts the significance threshold for M simultaneous tests:

$$\alpha_{\text{Bonferroni}} = \frac{\alpha}{M}$$

For $M = 100$ tests and target family-wise error rate $\alpha = 0.05$, each individual test must achieve $p < 0.0005$. This is conservative (it controls the family-wise

error rate rather than the false discovery rate) but appropriate when a single false positive would lead to real capital deployment.

NikaQuant applies the Holm-Bonferroni step-down procedure (Holm, 1979) rather than the simple Bonferroni correction, as it is uniformly more powerful while maintaining the same family-wise error rate guarantee. When the number of tests M is not reported by the user, NikaQuant uses a conservative default of $M = 50$ based on the assumption that any presented strategy has undergone at least moderate parameter search.

0.7.3 7.3 Effect Size and Practical Significance

Statistical significance is a necessary but not sufficient condition for deployment. A strategy with 10,000 trades may achieve statistical significance at PSR ≥ 0.99 with a Sharpe ratio of 0.05 — statistically distinguishable from zero but practically useless after trading costs.

NikaQuant therefore reports both the p-value and the effect size. The primary effect size metric is the annualized Sharpe ratio, with 95% confidence intervals computed via bootstrap (Section 5.6). A strategy must achieve both PSR ≥ 0.95 (significance) and annualized Sharpe ≥ 0.5 (practical effect size) for DEPLOY.

0.7.4 7.4 Confidence Intervals on All Reported Metrics

Every metric in the NikaQuant report is accompanied by a 95% confidence interval estimated via the stationary block bootstrap (Section 5.6) with $B = 2000$ replications. This ensures that point estimates are not misread as precise measurements.

For example, a reported Sharpe ratio of 1.20 with a 95% CI of [0.45, 1.95] is a far more honest representation of the uncertainty than the point estimate alone. A strategy with such a wide CI would receive a WARN on statistical precision regardless of the point estimate.

0.8 8. Risk Metrics — Beyond Drawdown

Standard backtest reporting focuses almost exclusively on maximum drawdown. NikaQuant computes a richer set of risk metrics to capture different aspects of downside risk.

Sharpe Ratio (Sharpe, 1966):

$$SR = \frac{\bar{R} - R_f}{\sigma_R} \cdot \sqrt{f}$$

Measures excess return per unit of total return volatility. Penalizes both upside and downside volatility symmetrically, which is a limitation when return distributions are skewed.

Sortino Ratio (Sortino and Price, 1994):

$$\text{Sortino} = \frac{\bar{R} - T}{DR}$$

where T is the minimum acceptable return (MAR) and DR is the downside deviation:

$$DR = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \min(R_i - T, 0)^2}$$

By penalizing only returns below the threshold T, the Sortino ratio does not penalize upside volatility. It is a more appropriate measure for positively skewed strategies.

Calmar Ratio (Young, 1991):

$$\text{Calmar} = \frac{\text{CAGR}}{|\text{Max Drawdown}|}$$

where CAGR is the compound annual growth rate. The Calmar ratio focuses on the most severe loss experienced; strategies with deep but brief drawdowns are penalized more than the Sharpe ratio would indicate.

MAR Ratio (Managed Account Reports):

$$\text{MAR} = \frac{\text{Annual Return}}{|\text{Max Drawdown}|}$$

Identical to Calmar in structure but typically applied to a fixed historical window rather than the full backtest period.

Omega Ratio (Keating and Shadwick, 2002):

$$\Omega(L) = \frac{\int_L^{+\infty} [1 - F(x)]\, dx}{\int_{-\infty}^{L} F(x)\, dx}$$

where $F(x)$ is the cumulative distribution function of returns and L is a loss threshold (typically zero or the risk-free rate). The numerator is the probability-weighted gain above the threshold; the denominator is the probability-weighted loss below it. Unlike the Sharpe ratio, the Omega ratio uses the full return distribution and places no distributional assumptions on returns.

K-Ratio (Kestner, 1996):

$$K = \frac{\text{slope of log-equity regression}}{\text{standard error of slope} \times \sqrt{N}}$$

The K-ratio measures the consistency of the growth trajectory: a high K-ratio indicates that the equity curve grows steadily, not in a volatile burst. It penalizes both high variance and non-monotonicity in the equity curve.

Ulcer Index (Martin and McCann, 1989):

$$UI = \sqrt{\frac{1}{N}\sum_{i=1}^{N} R_i^2}$$

where $R_i = 100 \times \frac{P_i - P_{\max,i}}{P_{\max,i}}$ is the percentage drawdown from the running peak at period i, and $P_{\max,i} = \max_{j \le i} P_j$. The squaring operation makes the Ulcer Index sensitive to deep drawdowns, unlike the mean drawdown.

The **Martin Ratio** (also called Ulcer Performance Index) replaces the Sharpe denominator with the Ulcer Index:

$$\text{Martin Ratio} = \frac{\bar{R} - R_f}{UI}$$

Tail Ratio:

$$\text{Tail Ratio} = \frac{|q_{0.95}|}{|q_{0.05}|}$$

where $q_{0.95}$ and $q_{0.05}$ are the 95th and 5th percentiles of the trade return distribution. A tail ratio > 1 indicates that the right tail (gains) is fatter than the left tail (losses). Strategies with tail ratio < 1 (fat left tail) carry elevated risk of large unexpected losses — this pattern is characteristic of short-volatility strategies.

Pain Index:

$$\text{Pain Index} = \frac{1}{N} \sum_{i=1}^{N} |R_i|$$

where R_i is the percentage drawdown from peak at period i. Unlike the Ulcer Index, the Pain Index takes the mean of absolute drawdowns rather than the RMS, making it less sensitive to occasional deep spikes.

Recovery Factor:

$$RF = \frac{\text{Net Profit}}{|\text{Max Drawdown}|}$$

The recovery factor measures how many times the maximum drawdown the strategy earned in net profit. A recovery factor of 1.0 means the strategy earned exactly its worst loss in total profit — a borderline result. The PASS threshold is RF > 3.0.

Sterling Ratio:

$$\text{Sterling} = \frac{\text{CAGR}}{|\text{Average Annual Drawdown}| + 10\%}$$

The denominator adds 10% to the average annual drawdown to avoid division by near-zero drawdowns and to provide a floor for the risk denominator.

Time-Under-Water Distribution. For each continuous period during which the equity curve is below its previous peak, NikaQuant records the duration. The distribution of these durations characterizes recovery speed. NikaQuant reports: median time-under-water, 90th percentile time-under-water, and maximum time-under-water. A maximum time-under-water exceeding 180 calendar days triggers a CRITICAL on the Risk pillar.

MAE/MFE Ratio. For each trade, Maximum Adverse Excursion (MAE) is the maximum floating loss during the trade's lifetime; Maximum Favorable Excursion (MFE) is the maximum floating profit. The MAE/MFE ratio at the trade level reveals whether the strategy is systematically exiting at the right time. A high aggregate MAE/MFE ratio (> 1.0) means the strategy's trades move against position before recovering — a signal of either poor entry timing or excessive holding through adverse moves.

0.9 9. Behavioral Quality

Aggregate metrics can mask behavioral anomalies that become apparent only when examining the strategy at the trade level. The Behavioral Quality pillar captures these.

0.9.1 9.1 Trade Frequency Consistency

A strategy that trades consistently is more likely to reflect a stable, rule-based approach than one that trades in bursts. NikaQuant computes the monthly trade count for each calendar month in the backtest period and calculates the coefficient of variation (CV):

$$CV = \frac{\sigma_{\text{monthly count}}}{\bar{\text{monthly count}}}$$

A CV < 0.5 indicates consistent trading. A CV > 2.0 suggests that the strategy's activity is highly irregular — either by design (it may be a regime-dependent system) or due to data anomalies. When CV > 2.0, NikaQuant inspects the periods of high and low activity for correlating factors (e.g., did activity collapse during bear markets?).

0.9.2 9.2 Activity Gap Detection

NikaQuant examines whether the strategy stopped trading for extended periods. A gap of more than 10 consecutive trading days with zero closed trades is flagged. Gaps are tolerable when they coincide with known market closures (e.g., Christmas week), but extended gaps during normal market conditions suggest either a strategy that conditions heavily on specific market states or a data quality problem.

The most recent 20% of the backtest period is given special weight: a strategy that was active in earlier years but inactive in recent months receives a WARN. This pattern may indicate that the market has evolved away from the conditions the strategy exploits.

0.9.3 9.3 MAE/MFE Distribution Analysis

Beyond the aggregate MAE/MFE ratio, NikaQuant plots the joint distribution of (MAE, MFE) for all trades. Healthy strategies typically show: - Positive correlation between MFE and profit (trades that moved favorably tend to close profitably). - MAE distribution that is concentrated at small values (the strategy does not regularly take large adverse excursions before recovering). - No bimodal distribution

in MAE (which would suggest the strategy frequently takes large initial losses on one subset of trades, potentially masking a stop-hunting dynamic).

A bimodal MAE distribution is flagged as an anomaly requiring manual inspection.

0.9.4 9.4 News-Window Exposure

NikaQuant cross-references the trade log against a database of scheduled high-impact macroeconomic events (including FOMC decisions, Non-Farm Payroll, CPI releases, ECB rate decisions, and equivalent releases for major currency pairs). Any trade opened within 30 minutes before or after a high-impact event is flagged as "news-window exposed."

The percentage of news-window-exposed trades is reported. Values > 60% suggest the strategy is explicitly news-trading (which deserves explicit acknowledgment in the report, since news-window backtests are particularly susceptible to look-ahead bias and execution-quality optimism). Values > 60% trigger a WARN on Behavioral Quality.

0.10 10. Multi-Strategy and Portfolio Analysis

When multiple strategies are submitted simultaneously, NikaQuant performs a portfolio-level analysis in addition to the individual strategy verdicts.

0.10.1 10.1 Correlation Matrix

For each pair of strategies (i, j), NikaQuant computes the Pearson correlation of daily returns over the period where both strategies have overlapping history:

$$\rho_{ij} = \frac{\text{Cov}(R_i, R_j)}{\sigma_i \sigma_j}$$

The full $K \times K$ correlation matrix is reported, alongside a heatmap visualization. Highly correlated strategies ($|\square| > 0.7$) provide limited diversification benefit and should not be allocated as if they were independent.

0.10.2 10.2 Diversification Benefit

The portfolio Sharpe ratio for a set of K strategies with weights $\mathbf{w}$ is:

$$SR_P = \frac{\mathbf{w}^\top \mu}{\sqrt{\mathbf{w}^\top \square \mathbf{w}}}$$

where μ is the vector of mean returns and $\square$ is the return covariance matrix. The diversification benefit is defined as:

$$DB = SR_P - \sum_{i=1}^{K} w_i \cdot SR_i$$

A positive DB indicates that the portfolio Sharpe exceeds the weighted average of individual Sharpes — the expected outcome when strategies are less than perfectly correlated. NikaQuant reports DB for the equal-weight portfolio as a simple, assumption-free baseline.

0.10.3 10.3 Mean-Variance Optimal Weights

Following Markowitz (1952), the minimum-variance portfolio weights are:

$$\mathbf{w}^* = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}}$$

The maximum Sharpe (tangency) portfolio weights are:

$$\mathbf{w}^{SR} = \frac{\Sigma^{-1}\mu}{\mathbf{1}^\top \Sigma^{-1}\mu}$$

NikaQuant reports both weight vectors but applies a bootstrap shrinkage to the covariance matrix estimate (Ledoit and Wolf, 2004) to mitigate estimation error in Σ^{-1} when the number of strategies is large relative to the history length.

0.10.4 10.4 Risk-Parity Weights

Risk parity allocates weights such that each strategy contributes equally to portfolio volatility:

$$w_i \propto \frac{1}{\sigma_i}$$

for the simplified diagonal case, or solved iteratively for the full covariance case. Risk parity is known to underperform mean-variance optimization when expected return estimates are accurate, but to substantially outperform when expected return estimates are noisy (as they typically are in finite backtests).

0.10.5 10.5 Regime-Conditional Correlations

Correlations between strategies may be regime-dependent. In particular, strategies that appear uncorrelated in normal conditions may become highly correlated during crisis regimes (the “correlation breakdown” phenomenon documented by Longin and Solnik, 2001). NikaQuant computes the correlation matrix separately for each HMM-identified regime (Section 6.5) and flags any strategy pair whose crisis-regime correlation exceeds its normal-regime correlation by more than 0.3.

0.11 11. Worked Example

0.11.1 11.1 Input Summary

Strategy: EURJPY M15 mean-reversion strategy, “Bollinger Squeeze” variant. Parameters: Bollinger Band period = 20, standard deviation multiplier = 2.0, RSI period = 14, RSI entry threshold = 30/70.

Backtest Period: January 1, 2021 – December 31, 2023 (36 months, 756 trading days).

Data Source: MT5 HTML report, ICMarkets ECN account, variable spread.

Summary Statistic	Value
Total Trades	412
Winning Trades	231 (56.1%)
Net Profit	$8,420
Initial Capital	$10,000
Max Drawdown	$1,650 (14.2%)
Profit Factor	1.68
Expectancy	$20.44/trade
Avg Win	$68.20
Avg Loss	$54.80
Commission + Swap	$1,210 total
Avg Spread (from tick data)	1.4 pips

0.11.2 11.2 Pillar-by-Pillar Computation

Pillar 1: Profitability

Expectancy: $20.44. The average spread is 1.4 pips ≈ $1.40 per standard lot. At 0.1 lots per trade, avg spread cost ≈ $0.14. Expectancy of 20.44 » 5× avg spread ≈ $0.70. Sub-score: 92/100.

Profit Factor: 1.68. Maps to approximately 58/100 on the scaling function.

Annualized Sharpe: Daily returns computed from equity curve. $\bar{R} = 0.012\%/\text{day}$, $\sigma_R = 0.078\%/\text{day}$, $SR = (0.012/0.078)\sqrt{252} = 2.44$. Sub-score: 90/100.

Sortino: Downside deviation = 0.055%/day (only negative days). Sortino $= (0.012/0.055)\sqrt{252} = 3.46$. Sub-score: 95/100.

Calmar: CAGR = 28.1%/year. Max DD = 14.2%. Calmar = 1.98. Sub-score: 88/100.

Pillar 1 Sub-score S_1 = (92 + 58 + 90 + 95 + 88)/5 = 84.6

Pillar 2: Robustness

Walk-Forward Analysis (5 rolling windows, 4:1 IS/OOS):

Window	IS Sharpe	OOS Sharpe
1	2.71	1.44
2	2.38	1.12
3	2.91	1.60
4	2.20	0.89
5	2.55	1.31

$\bar{R}_{IS} = 2.55$, $\bar{R}_{OOS} = 1.272$. WFE = 1.272/2.55 = **0.499 (WARN — just below 0.5 threshold)**.

WFE sub-score: 48/100.

Deflated Sharpe Ratio: Optimization sweep used $M = 80$ parameter combinations. Average inter-trial correlation $\hat{\rho} = 0.65$, giving $N = 0.65 + 0.35 \times 80 = 28.65$ effective trials. $V[\{\widehat{SR}_n\}] = 0.22^2$. Computed $\widehat{SR}_0 = 0.38$. With $\widehat{SR} = 2.44$, $n = 756$, $\hat{\gamma}_3 = 0.18$, $\hat{\gamma}_4 = 3.8$: DSR ≈ 0.9998. Sub-score: 98/100.

Parameter Stability Index: Four parameters perturbed. Minimum stability: Band period ±25% → Sharpe degrades from 2.44 to 1.87. $\Sigma = 1 - (2.44 - 1.87)/2.44 = 0.766$. PSI = 0.766 (WARN, below 0.80 threshold). Sub-score: 66/100.

Pillar 2 Sub-score S_2 = 0.50 × 48 + 0.30 × 98 + 0.20 × 66 = 24 + 29.4 + 13.2 = 66.6

Pillar 3: Risk

Max drawdown 14.2% → PASS. Recovery factor = $8,420/$1,650 = 5.10 → PASS. Ulcer Index = 3.8% (low). Tail ratio = |q_{0.95}|/|q_{0.05}| = 1.82/1.34 = 1.36 → PASS. Max time-under-water = 41 trading days → WARN. Sub-scores: 85, 92, 88, 80, 62. **Pillar 3 Sub-score S_3 = 81.4**

Pillar 4: Statistical Significance

PSR (SR* = 0): With $\widehat{SR} = 2.44$, $n = 756$, negative skew effect minimal: PSR ≈ 1.0. Sub-score: 100/100.

White's Reality Check p-value: 0.014 (bootstrap, B = 2000). Sub-score: 88/100.

Permutation test p-value: 0.009. Sub-score: 92/100.

Pillar 4 Sub-score S_4 = (100 + 88 + 92)/3 = 93.3

Pillar 5: Stress Resistance

Spread shock results (avg spread 1.4 pips):

Spread Multiplier	Total Net Profit	vs. Baseline
1×	$8,420	Baseline
2×	$6,840	+PASS
5×	$2,110	–75% (WARN — < –25% threshold)
10×	–$2,580	FAIL

Wait — 5× spread: $8,420 – (5–1)×1.4 pip × 0.1 lot × 412 trades. Each pip = $10/lot × 0.1 lot = $1. Additional spread cost = 4 × 1.4 × $1 × 412 = $2,307. Remaining: $8,420 – $2,307 = $6,113 (+PASS). 10× spread: additional = 9 × 1.4 × $1 × 412 = $5,191. Remaining: $8,420 – $5,191 = $3,229 (+PASS). *All stress tests pass on this strategy.*

Slippage +3 pip: additional cost = 3 × $1 × 412 = $1,236. Net: $7,184 (PASS).

Commission 2×: additional = $1,210. Net: $7,210 (PASS).

Pillar 5 Sub-score S_5 = 88.0

Pillar 6: Behavioral Quality

Monthly trade count: mean 11.4, SD 3.2, CV = 0.28 (PASS). Longest gap: 8 trading days (PASS). MAE/MFE ratio: 0.73 (PASS — < 1.0). News-window exposure: 18% of trades (PASS). Sub-scores: 92, 88, 85, 80.

Pillar 6 Sub-score S_6 = 86.3

0.11.3 11.3 Composite Score

$$\text{Composite} = 0.15\times84.6+0.25\times66.6+0.20\times81.4+0.15\times93.3+0.15\times88.0+0.10\times86.3$$

$$= 12.69 + 16.65 + 16.28 + 13.995 + 13.20 + 8.63 = 81.4$$

0.11.4 11.4 Verdict

DEPLOY (composite 81.4 ≥ 75, no CRITICAL failures).

However, two WARN flags are noted: 1. **WFE = 0.499** (marginally below 0.50 PASS threshold) — the strategy loses approximately 50% of its in-sample Sharpe out of sample. While this does not prevent DEPLOY, it indicates overfitting risk that should be monitored carefully in live trading. 2. **Parameter Stability (PSI = 0.766)** — the Bollinger Band period parameter is moderately sensitive. A period of 15 (vs. optimized 20) degrades Sharpe from 2.44 to 1.87. The strategy should not be operated with a period significantly different from the optimized value.

0.11.5 11.5 Top 3 Factors For and Against

In favor: 1. Statistically overwhelming significance (PSR ≈ 1.0, both bootstrap tests p < 0.02). 2. Stress resistance: the strategy remains profitable at 10× spread, indicating genuine edge per unit of spread cost. 3. Consistent trading behavior: CV of 0.28, no extended gaps, clean MAE/MFE profile.

Against: 1. WFE of 0.499 suggests meaningful parameter overfitting; the OOS Sharpe of 1.27 is substantially below the IS Sharpe of 2.55. 2. Moderate parameter sensitivity: the Bollinger period is the load-bearing parameter, and the strategy degrades non-trivially if it drifts from optimal. 3. Maximum time-under-water of 41 days may test the patience of live traders, risking premature strategy abandonment during drawdown periods.

0.11.6 11.6 Position Sizing Recommendation

Full Kelly Fraction:

$$f^* = \frac{W}{A} - \frac{1-W}{B}$$

where $W = 0.561$ (win rate), $A = \$68.20$ (avg win), $B = \$54.80$ (avg loss):

$$f^* = \frac{0.561}{68.20} - \frac{0.439}{54.80} = 0.00823 - 0.00801 = 0.00022$$

This yields an extremely small Kelly fraction due to the near-1:1 payoff structure. Expressed as a fraction of account per unit of risk, full Kelly ≈ 2.2% of capital per standard lot equivalent.

Half-Kelly: 1.1% of capital per unit. **Conservative (25% Kelly):** 0.55% per unit.

NikaQuant recommends **Half-Kelly** as the standard starting point, with the option to scale toward Full Kelly only after 6 months of live forward testing that corroborates the out-of-sample performance statistics.

0.12 12. What NikaQuant Does NOT Claim

This section exists because intellectual honesty about limitations is more valuable than any marketing claim, and because misunderstanding the scope of a verdict could lead to real financial loss.

We do not predict future returns. A DEPLOY verdict means that the historical evidence, subjected to our full robustness and significance battery, is consistent with the hypothesis that the strategy has a genuine statistical edge. The historical edge may or may not persist. Markets change. The statistical arbitrage that produced consistent profits in 2020-2022 may no longer exist in 2025. NikaQuant evaluates the past; the future is unknown.

We do not guarantee profitability. Even a strategy with genuinely positive expected value will experience losing periods. Even a strategy with PSR = 0.99 has a 1% probability of its Sharpe ratio being effectively zero. No finite backtest guarantees future performance.

We cannot detect strategies designed to game the analyzer. A sufficiently sophisticated actor could construct a strategy that is specifically designed to pass the NikaQuant battery — for example, by engineering large, infrequent profits that satisfy WFE requirements without representing a tradeable edge. We publish this methodology precisely so that anyone can examine what we check; we make no claim that our checks are comprehensive against adversarial construction.

We are not a substitute for live forward testing. The gold standard of strategy evaluation is live, real-money forward testing conducted over a sufficiently long period (ideally ≥ 12 months). NikaQuant is designed to efficiently filter out clearly inferior strategies before that costly forward test, not to replace it. We recommend that every DEPLOY-verdicted strategy undergo a minimum of 6 months of live forward testing at conservative position sizes before scaling up.

We do not validate broker execution quality. Our stress tests model execution risk via spread shock and slippage models, but we cannot verify that a given broker will actually fill orders at the prices assumed. Requotes, widening spreads during news, and platform-level execution delays are outside the scope of NikaQuant's analysis.

We are not financial advisors. A NikaQuant verdict is a technical evaluation of a systematic trading strategy. It does not constitute investment advice, and it does not account for individual investors' risk tolerance, capital constraints, regulatory environment, or overall portfolio context.

0.13 13. References

Aronson, D. R. (2006). *Evidence-Based Technical Analysis: Applying the Scientific Method and Statistical Inference to Trading Signals*. John Wiley & Sons, Hoboken, NJ.

Bailey, D. H., & López de Prado, M. (2012). The Sharpe Ratio Efficient Frontier. *Journal of Risk*, 15(2), 3–44.

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. SSRN: 2460551.

Bailey, D. H., Borwein, J., López de Prado, M., & Zhu, Q. J. (2014). 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.

Calmar Ratio: Young, T. W. (1991). Calmar Ratio: A Smoother Tool. *Futures*, 20(1), 40.

Carver, R. (2015). *Systematic Trading: A Unique New Method for Designing Trading and Investing Systems*. Harriman House, Petersfield, UK.

Chan, E. P. (2013). *Algorithmic Trading: Winning Strategies and Their Rationale*. John Wiley & Sons, Hoboken, NJ.

Harvey, C. R., & Liu, Y. (2015). Backtesting. *Journal of Portfolio Management*, 42(1), 13–28.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the Cross-Section of Expected Returns. *Review of Financial Studies*, 29(1), 5–68.

Harvey, C. R., & Liu, Y. (2020). False (and Missed) Discoveries in Financial Economics. *Journal of Finance*, 75(5), 2503–2553.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.

Keating, C., & Shadwick, W. F. (2002). A Universal Performance Measure. *Journal of Performance Measurement*, 6(3), 59–84.

Kestner, L. (1996). Getting a Handle on True Performance. *Futures*, 25(1), 44–46.

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

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2), 365–411.

Lo, A. W. (2002). The Statistics of Sharpe Ratios. *Financial Analysts Journal*, 58(4), 36–52.

Longin, F., & Solnik, B. (2001). Extreme Correlation of International Equity Markets. *Journal of Finance*, 56(2), 649–676.

López de Prado, M. (2018). *Advances in Financial Machine Learning*. John Wiley & Sons, Hoboken, NJ.

Markowitz, H. (1952). Portfolio Selection. *Journal of Finance*, 7(1), 77–91.

Martin, P. G., & McCann, B. B. (1989). *The Investor's Guide to Fidelity Funds*. John Wiley & Sons.

Pardo, R. (1992). *Design, Testing, and Optimization of Trading Systems*. John Wiley & Sons, New York.

Pardo, R. (2008). *The Evaluation and Optimization of Trading Strategies* (2nd ed.). John Wiley & Sons, Hoboken, NJ.

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

Politis, D. N., & White, H. (2004). Automatic block-length selection for the depen-

dent bootstrap. *Econometric Reviews*, 23(1), 53–70.

Sharpe, W. F. (1966). Mutual Fund Performance. *Journal of Business*, 39(1), 119–138.

Sortino, F. A., & Price, L. N. (1994). Performance Measurement in a Downside Risk Framework. *Journal of Investing*, 3(3), 59–64.

White, H. (2000). A Reality Check for Data Snooping. *Econometrica*, 68(5), 1097–1126.

0.14 Appendix A: Glossary

Anchored Walk-Forward Analysis: A walk-forward variant in which the in-sample window is anchored at the start of the data and expands with each cycle, while the out-of-sample window rolls forward.

Backtest Overfitting: The condition in which a strategy's parameters have been tuned to historical data to the extent that the strategy's apparent performance reflects the specific characteristics of the historical sample rather than a generalizable trading edge.

Block Bootstrap: A resampling procedure that draws contiguous blocks of observations rather than individual observations, preserving short-range serial dependence in time series.

Calmar Ratio: The ratio of compound annual growth rate to maximum drawdown, measuring return per unit of peak-to-trough risk.

CPCV (Combinatorially Purged Cross-Validation): A cross-validation framework for financial time series that uses all combinations of training and test splits while applying purging and embargoing to prevent information leakage.

Composite Score: NikaQuant's primary output — a weighted average of six pillar sub-scores, ranging from 0 to 100.

Deflated Sharpe Ratio (DSR): The probability that a strategy's Sharpe ratio is statistically significant after correcting for the number of strategies evaluated (multiple testing correction) and the non-normality of returns.

Downside Deviation: The standard deviation of returns below a specified threshold (typically zero or a minimum acceptable return), used in the denominator of the Sortino ratio.

Embargo: In CPCV, a period following a test set during which observations are excluded from training, to prevent leakage via autocorrelation.

Equity Curve: The time series of cumulative account value (including open trade floating P&L) over the backtest period.

Expectancy: The average net profit per closed trade, computed across all trades in the backtest.

Family-Wise Error Rate (FWER): The probability of making at least one false rejection among all hypothesis tests conducted; controlled by the Bonferroni and Holm corrections.

Hidden Markov Model (HMM): A statistical model in which the system is assumed to be a Markov process with hidden (unobserved) states; used by NikaQuant

to assign market regime labels to each trading period.

In-Sample (IS) Window: The historical period used for strategy optimization or parameter fitting in walk-forward analysis.

K-Ratio: A measure of equity curve consistency, defined as the slope of the log-equity regression divided by its standard error, normalized by the square root of sample size.

Kelly Criterion: A formula for optimal position sizing that maximizes the expected logarithm of wealth; derived from information theory (Kelly, 1956). Applied conservatively (half-Kelly or lower) in practice.

MAE (Maximum Adverse Excursion): The maximum floating loss experienced during a trade's lifetime before it was closed.

MAR Ratio: Annual return divided by maximum drawdown; related to the Calmar ratio.

MFE (Maximum Favorable Excursion): The maximum floating profit experienced during a trade's lifetime before it was closed.

Omega Ratio: A ratio of probability-weighted gains to probability-weighted losses relative to a threshold, using the full return distribution without normality assumptions.

Out-of-Sample (OOS) Window: The historical period used to evaluate strategy performance after parameter fitting, without any further optimization.

Overfitting: See *Backtest Overfitting*.

Pain Index: The mean of absolute percentage drawdowns from the running equity peak; a less tail-sensitive alternative to the Ulcer Index.

Parameter Stability Index (PSI): The minimum stability ratio across all optimized parameters when each is perturbed by ±10% and ±25%; measures sensitivity of performance to parameter choice.

Probabilistic Sharpe Ratio (PSR): The probability that a strategy's true Sharpe ratio exceeds a specified benchmark, estimated from the sample Sharpe ratio while accounting for non-normality (skewness and kurtosis) of returns.

Profit Factor: The ratio of gross profit to gross loss across all closed trades.

Purging: In CPCV, the removal of training observations whose label period overlaps with the test period, preventing information leakage.

Recovery Factor: Net total profit divided by maximum drawdown; indicates how many times the strategy earned its worst loss in net profit.

Regime: A persistent statistical state of the market characterized by distinct return, volatility, and autocorrelation properties.

Rolling Walk-Forward Analysis: A walk-forward variant in which both the in-sample and out-of-sample windows roll forward together, with a fixed in-sample window length.

Sharpe Ratio: Annualized excess return (above the risk-free rate or zero) divided by annualized standard deviation of returns.

Slippage: The difference between the price at which a trade was requested and the price at which it was actually executed.

Sortino Ratio: The ratio of excess return to downside deviation; penalizes only downside volatility, unlike the Sharpe ratio.

Spread: The difference between the bid and ask price of an instrument at any given moment; represents a transaction cost.

Stationary Bootstrap: A block bootstrap variant (Politis and Romano, 1994) in which block lengths are drawn from a geometric distribution, producing stationary bootstrap samples.

Sterling Ratio: CAGR divided by (average annual drawdown + 10%); a drawdown-adjusted return measure.

Tail Ratio: The ratio of the 95th percentile of trade returns to the absolute value of the 5th percentile; a measure of right-tail versus left-tail thickness.

Time-Under-Water: The duration of continuous periods during which the equity curve remains below its previous peak value.

Ulcer Index: The root mean square of percentage drawdowns from the running equity peak; heavily weights deep drawdowns due to the squaring operation.

Walk-Forward Efficiency (WFE): The ratio of mean out-of-sample performance to mean in-sample performance across all walk-forward cycles; measures how much of the optimized performance is preserved out of sample.

White's Reality Check: A bootstrap-based statistical test (White, 2000) for the null hypothesis that no strategy in a tested set has positive expected performance, controlling the family-wise error rate across all tested strategies.

0.15 Appendix B: Reproducibility

0.15.1 B.1 The Reproducibility Commitment

Every NikaQuant verdict is deterministically reproducible given the same inputs and the same engine version. This commitment has two components:

Input hashing. At the time of analysis, the input data (MT5 HTML report, XML optimization file, and/or FXT tick files) are hashed using SHA-256. This hash is embedded in the verdict output. Any user who retains the original input files can verify that a re-run on the same files — against the same engine version — produces an identical verdict.

Engine versioning. NikaQuant follows strict semantic versioning (MAJOR.MINOR.PATCH). The full engine version is embedded in every verdict. Changes to scoring thresholds, formula implementations, or rubric weights constitute MINOR or MAJOR version increments. PATCH increments are reserved exclusively for bug fixes that do not change any score for any previously evaluated strategy (or, when a bug fix does change a score, this is explicitly documented in the changelog with the set of affected strategies).

0.15.2 B.2 Threshold Freeze

NikaQuant commits to never silently changing thresholds. Every threshold change is: 1. Documented in the public changelog with the version in which it takes effect. 2. Applied only prospectively: verdicts issued under previous versions are

re-issued under the new version only upon explicit user request. 3. Accompanied by a rationale citing the academic literature or empirical evidence motivating the change.

0.15.3 B.3 Reproducibility Record Format

Each verdict report includes the following reproducibility block:

```
NIKAQUANT REPRODUCIBILITY RECORD
Engine Version: 1.0.0
Analysis Timestamp: 2025-01-15T14:32:07Z
Input File Hash (SHA-256):
  report.html: a3f9c2d1e8b7...
  optimization.xml: 7b2a1c9d3e4f...
Random Seed: 42 (used for bootstrap replications)
Bootstrap Replications: 2000
Block Length Algorithm: Politis-White (2004)
Computed Block Length: 12 trading days
```

Any party who possesses the original input files and the NikaQuant engine at the specified version can reproduce every number in the report by re-running with these parameters.

0.15.4 B.4 Open Methodology Principle

This whitepaper constitutes the complete specification of the NikaQuant scoring engine. No scoring logic exists outside this document. Every formula, every threshold, and every procedural step is defined above with sufficient precision that an independent implementation could be built from this document alone, and that independent implementation should produce materially identical scores (within floating-point precision) for the same inputs.

This radical transparency is deliberate. A methodology that cannot withstand public scrutiny should not be used to make real trading decisions.

End of NikaQuant Verdict Methodology Whitepaper v1.0.0

Document maintained by the NikaQuant Engineering Team. Questions, corrections, and methodological challenges should be directed to the public issue tracker. All submissions will be acknowledged and responded to in writing.