EASYALGOS TECHNICAL WHITE PAPER

AlgoScore

A Quantitative Framework for Evaluating
Expert Advisor Track Records

Valeriia Mischenko
EasyAlgos · easyalgos.ai
Version 1.4 · July 2026

6 Risk-Adjusted Metrics | Manual Overlay | Score Calibration | Full Formulas

Abstract

This paper presents AlgoScore, a deterministic scoring engine that produces a 0–100 quality rating for Foreign Exchange (FX) Expert Advisor (EA) track records. Developed by EasyAlgos—drawing on over a decade of Expert Advisor development, thousands of live client accounts, and tens of billions of dollars in monthly trading volume—AlgoScore evaluates six risk-adjusted metrics across dual time windows, applies Ljung–Box significance testing, and incorporates Bayesian-inspired confidence shrinkage to produce a single, transparent, fully auditable *engine score*. In version 1.4, two transparent layers sit on top of that engine score: a **manual overlay** of up to three multipliers that corrects for risks Myfxbook cannot capture, and a fixed **calibration** curve that re-expresses the result on an intuitive, rank-preserving display scale. The methodology penalizes hidden risks including drawdown persistence, negative tail events, serial return smoothing, and floating loss exposure.

Keywords: risk-adjusted performance, Expert Advisor evaluation, Sortino ratio, Ulcer Index, autocorrelation testing, manual overlay, score calibration, algorithmic trading

1. Executive Summary

AlgoScore is a deterministic scoring system that evaluates Foreign Exchange Expert Advisor track records on a scale of 0 to 100—analogous to a credit score for trading strategy quality. It was developed

by EasyAlgos, drawing on over a decade of Expert Advisor development, thousands of live client accounts, and the processing of tens of billions of dollars in monthly trading volume.

The system analyzes six risk-adjusted metrics across two time windows (all-time and recent 365 days), applies statistical significance testing via the Ljung–Box Q statistic, and adjusts for track record length using Bayesian-inspired confidence shrinkage to produce the *engine score*. A manual overlay then corrects for risks the automation and Myfxbook cannot see, and a fixed calibration curve maps the result onto an intuitive published scale. Every intermediate value is published. Unlike return-based rankings, AlgoScore penalizes hidden risks: drawdown persistence, crash-like tail events, artificial equity curve smoothing, and floating loss exposure.

Feature	Description
6 Risk-Adjusted Metrics	Sortino, Calmar, Ulcer Index, Neg. Skewness, Autocorrelation, Float Stress
Dual Time Windows	All-time + Recent (365d), adaptively blended by track length
Statistical Rigor	Ljung–Box significance testing; autocorrelation penalized only when $p < 0.05$
Confidence Adjustment	Bayesian shrinkage toward neutral (50) for short track records
Manual Overlay (v1.3)	Up to three transparent multipliers for risks Myfxbook cannot capture
Score Calibration (v1.4)	Fixed, rank-preserving display curve anchored to the retail EA population
Full Transparency	Every intermediate value published; all parameters externalized
Deterministic	Identical inputs always produce identical scores

2. Background and Motivation

The Expert Advisor market in Foreign Exchange trading presents a fundamental transparency problem. Thousands of Expert Advisors are marketed with impressive return figures, yet these numbers are incomplete. A strategy showing 200% annual returns may carry catastrophic tail risk, artificially smooth its equity curve through correlated position layering, or benefit from a short lucky streak with no statistical significance. EasyAlgos, with over 10 years in the space, thousands of active clients, and dozens of billions in monthly trading volume, developed AlgoScore to address the specific challenges of objective Expert Advisor comparison. The system was designed by Valeriia Mischenko, leveraging deep domain expertise in quantitative trading metrics and the specific data characteristics of Myfxbook reporting.

2.1 Limitations of Conventional Metrics

- **Total return ignores risk.** A 500% return with 80% drawdown differs fundamentally from 100% return with 15% drawdown.
- **Profit factor conflates frequency with quality.** One large win yields infinite PF with no statistical evidence of edge.

- **Maximum drawdown captures only the single worst episode**, ignoring frequency and duration of underwater periods.
- **Sharpe ratio penalizes upside volatility equally**, inappropriate for trend-following or momentum strategies.

3. Design Principles

3.1 Use Myfxbook time-weighted returns (TWR) directly—deposits/withdrawals already excluded.

3.2 Reconstruct calendar-daily series: fill Myfxbook's missing zero-change days with 0% to avoid upward bias in risk metrics.

3.3 Source max drawdown from Myfxbook Stats panel (includes intraday), never from chart data which excludes intraday drawdown.

3.4 Penalize autocorrelation only when Ljung–Box confirms significance at $\alpha = 0.05$; short tracks may show spurious autocorrelation.

3.5 Fully deterministic: same inputs → same outputs. All parameters externalized in versioned configuration files.

4. Data Acquisition and Preprocessing

Performance data is sourced via the Myfxbook API. For each Expert Advisor, the engine retrieves daily return series, account-level statistics (including the authoritative maximum drawdown value), daily floating profit/loss data, and where available, additional trade-level details from Advanced Statistics.

Data is normalized into a canonical model: track identifier, date range, daily percentage changes, Stats-reported max drawdown, and daily floating loss values. Calendar reconstruction then creates a complete date array [start, ..., end], filling any missing days with 0.0% and converting to decimal returns ($r = pct/100$). This step is critical: without it, risk metrics are systematically biased upward because zero-return periods are excluded from variance and drawdown calculations.

5. The Six Metrics

Metric	Category	Weight	Direction
Sortino Ratio	Performance	30%	Higher = better
Calmar Ratio	Performance	20%	Higher = better
Ulcer Index	Risk	20%	Lower = better
Neg. Skewness	Risk	10%	Lower = better
Autocorrelation	Consistency	10%	Lower = better*
Float Stress	Risk	10%	Lower = better

*Penalty only when Ljung–Box p < 0.05

5.1 Sortino Ratio (30%)

Risk-adjusted return using only downside deviation (MAR = 0). Unlike Sharpe, does not penalize upside volatility.

```
AR = mean(r_d) × 365.25
DD = sqrt(mean(min(0, r_d)²)) × sqrt(365.25)
Sortino = AR / DD    (capped at 10.0 if DD = 0)
```

5.2 Calmar Ratio (20%)

Compound annual growth vs. maximum drawdown. Drawdown always from Myfxbook Stats.

```
CAGR = E_end^(1/Y) − 1      Y = days/365.25
Calmar = CAGR / MaxDD_stats
```

5.3 Ulcer Index (20%)

Measures depth and duration of drawdowns (Peter Martin). Captures sustained underwater pain, not just the worst single episode.

```
Peak_t = max(E_s) for s ≤ t
UI = sqrt(mean((100 × (Peak_t − E_t)/Peak_t)²))
```

5.4 Negative Skewness (10%)

Penalizes left-tail fat: occasional large losses disproportionate to typical returns. Only negative skewness is penalized.

```
NegSkew = max(0, −Skewness_FisherPearson)
```

5.5 Autocorrelation (10%)

Detects serial dependence (position averaging, grid strategies). Uses weekly compounded returns to avoid weekend-zero artifacts. Ljung–Box Q test for joint significance across lags 1–5:

```
r_w = ∏(1 + r_d) − 1  per ISO week
Q = n(n+2) × ∑(r_k²/(n−k))   k = 1..5
p = 1 − χ²_cdf(Q, df = 5)
```

If p ≥ 0.05: subscore = 100. If p < 0.05: AutoPos = sqrt(mean(max(0, r_k)²)) mapped to subscore. Minimum 26 weeks required.

5.6 Float Stress (10%)

Measures the root mean square (RMS) of daily floating loss as a percentage of equity, computed from Myfxbook's daily floating P/L data. This metric captures how much unrealized drawdown a strategy holds overnight, directly penalizing strategies that carry large floating losses—a hallmark of grid and martingale systems.

```
FloatStress = sqrt(mean(dailyFloatingLoss²))
```

where each daily floating loss is expressed as a percentage of account equity. Lower values indicate strategies that close positions quickly with minimal unrealized exposure; higher values indicate strategies that routinely hold large underwater positions.

Float Stress replaced the previous Side Balance metric (which measured long/short trade count balance) because it is far more informative: it directly measures hidden risk from unrealized losses rather than merely assessing directional concentration by trade count.

6. Metric-to-Subscore Mapping

Each raw value is mapped to 0–100 via piecewise linear interpolation between configurable anchors (Appendix A). Values outside anchor range are clamped. Anchors calibrated against a historical Expert Advisor corpus.

7. Time-Window Architecture

Metrics computed on two windows: All-time (full record) and Recent (last 365 days; equals all-time if shorter). Blended adaptively:

```
w_recent = 0.30 × min(1, days/365)  if days ≥ 90; else 0
w_alltime = 1 − w_recent
```

Tracks < 90 days: all-time only. Full year: 70/30 blend. Annualization: 365.25 for all tracks.

8. Confidence Layer and Track Length

8.1 Reliability Shrinkage

Bayesian-inspired shrinkage toward neutral (50). Shorter records pulled more strongly toward 50:

```
Reliability = sqrt(days / (days + 365))
AdjustedBase = 50 + (CombinedBase − 50) × Reliability
```

365d → 71%, 730d → 82%, 2000+d → 92% reliability.

8.2 Length Score (Deprecated)

In v1.2, LengthScore is computed for analytics only and is no longer included in the final score formula. Independent reward for proven track length with logarithmic diminishing returns:

```
LengthScore = 20 × clamp(1 + log2(days/182.625), 0, 5)
```

Track Length	~6 mo	~1 yr	~2 yr	~4 yr	~8+ yr
Length Score	20	40	60	80	100

9. Engine Score Computation

The engine score is the deterministic output of Sections 5–8:

```
EngineScore = AdjustedBase ∈ [0, 100]
```

Integer for internal use; float retained for downstream layers. The engine score is determined entirely by risk-adjusted metric quality, with duration influence limited to the reliability shrinkage in Section 8.1. In versions 1.3 and 1.4 the engine score is *not* the number published directly; it is the input to the manual overlay (Section 10) and calibration (Section 11).

10. Manual Overlay

10.1 Motivation

The engine (Sections 3–9) is fully deterministic, but it can only see what the data source reports. In practice Myfxbook is imperfect in three recurring ways: (1) the true maximum drawdown is sometimes not captured, as the equity curve and the drawdown panel can disagree and either may understate it; (2) deposits and withdrawals made *during* a drawdown can flatter a time-weighted return series, and some developers exploit this to produce smoother equity curves; and (3) real-world scalability — a strategy trading gold with a 30-minute median hold, or a night scalper — is invisible to statistics. The manual overlay is a thin, transparent correction layer that addresses exactly these gaps. It never improves a score; it can only leave it unchanged or reduce it.

10.2 Definition

Let `S_engine ∈ [0, 100]` be the engine score. The raw public score is:

```
S_raw = clamp( S_engine × m_cashflow × m_grid × m_scalability , 0, 100 )
```

Multiplier	Range	Source	Corrects
m_cashflow (cash-flow / true-TWR)	0.00 – 1.00	Transcribed (external true-TWR service)	Return after removing deposits/withdrawals during drawdowns
m_grid (grid / martingale)	1.00, else 0.50 – 0.80	Manual judgement	Residual grid/averaging risk not captured by Float Stress + autocorrelation
m_scalability (scalability)	1.00, else 0.50 – 0.80	Manual judgement	Real-world scalability concerns

Allowed-value rule. Manual multipliers (grid, scalability) are only ever exactly 1.00, or a value in the closed interval [0.50, 0.80]; values in the open gap (0.80, 1.00) are not permitted. The cash-flow multiplier is transcribed, not judged, and may take any value in [0.00, 1.00] (values above 1.00 are clamped to 1.00).

10.3 Cash-flow / True-TWR multiplier

Transcribed, not judged. An external true time-weighted-return service recomputes the account's return while neutralising deposits and withdrawals made during drawdowns; the ratio of that corrected return to the reported return is entered directly. The analyst copies the value and its run metadata exactly and exercises no discretion.

10.4 Grid / Martingale multiplier

A *residual* adjustment. The engine already penalises grid and averaging behaviour through Float Stress and the autocorrelation metric; the overlay only adds the portion of that risk Myfxbook fails to capture:

0.75–0.80 when the drawdown is well captured, 0.60–0.70 when understated, 0.50–0.60 when materially wrong. It is 1.00 for non-grid systems.

10.5 Scalability multiplier

A manual judgement of how well live results are likely to scale: 0.70–0.80 for moderate concern (e.g. a ~30-minute median trade duration on a slippage-prone instrument such as gold), 0.50–0.60 for severe concern (e.g. night scalping with low per-trade expectancy). Each penalty records one or more reason flags (night scalper, ultra-short holding time, spread/latency sensitivity, small-account-only evidence, and others).

10.6 Governance and transparency

- Every applied penalty is published with its multiplier value and a written reason on the strategy's detail page.
- The overlay is applied to EasyAlgos' own EAs on the same terms as third-party systems; where the data warranted it, in-house strategies received some of the heaviest penalties.
- A clean account is never penalised — all three multipliers are 1.00 and `S_raw = S_engine`.
- The backend is authoritative: overlay values are stored server-side and the published score is recomputed there.

11. Score Calibration

11.1 Motivation

The raw public score (Section 10) is internally consistent and rank-correct, but its absolute values are conservative: after the overlay, even the strongest verified retail systems land near 60, and the median lands near 40. Readers intuitively interpret a 0–100 scale the way they read familiar ratings: 90+ exceptional, ~70 typical, below 40 poor. Calibration resolves this as an explicit *presentation* step, rather than by inflating any component of the underlying measurement.

11.2 Definition

```
S_public = C( S_raw )
```

where `C` is a fixed, strictly increasing, piecewise-linear mapping through the anchor points below (full table in Appendix B). `S_public` is rounded half-up to the nearest integer and is the number published on the leaderboard and detail pages; the raw score remains visible in each strategy's transparency breakdown.

S_raw	0	10	20	30	40	50	60	70	85	100
S_public	0	30	45	58	70	78	87	92	97	100

11.3 Anchor derivation

The anchors were derived from percentiles of the raw-score distribution across the tracked retail EA population, then frozen: the median raw retail score (≈40) displays as 70 ("the typical retail EA"); the ~90th-percentile raw score (≈60) displays as 87; the strongest observed systems display in the high 80s; and heavily penalised systems (raw below ~17) still display below 40. Because the curve is frozen rather than recomputed, a strategy's published score can never change due to another strategy entering or leaving the population. Recalibration, if ever warranted, is a versioned methodology change.

11.4 Properties

- **Rank preservation (exact).** C is strictly increasing, so S_raw(A) > S_raw(B) ⟺ S_public(A) > S_public(B) (up to integer-rounding ties). Calibration cannot reorder strategies.
- **Determinism.** Same raw score → same published score, always.
- **Endpoints fixed.** C(0) = 0 and C(100) = 100; the scale's bounds are unchanged.
- **Transparency.** The full chain — engine score, each overlay multiplier with its reason, the raw score, and the calibrated public score — is published on every strategy detail page.

Worked example. A clean breakout system with S_engine = 61 and no overlay has S_raw = 61, so S_public = 87 + (61−60)/(70−60) × (92−87) = 87.5 → **88**. A grid system with S_engine = 62 and a ×0.336

> overlay has S_raw = 20.83, so S_public = 45 + (20.83−20)/(30−20) × (58−45) = 46.08 → **46**. Same relative story as the raw scores, told on a scale people read correctly.

12. Score Interpretation

The bands below apply to the published (calibrated) score:

Range	Rating	Interpretation
80–100	Excellent	Strong risk-adjusted returns, low drawdown pain, proven track record.
60–79	Good	Solid performance, acceptable risk. May have 1–2 areas for improvement.
40–59	Fair	Mixed results or insufficient history. Elevated risk or short duration.
0–39	Poor	Significant concerns: excessive drawdown, tail risk, smoothing, or short record.

AlgoScore is a quantitative tool, not investment advice. A high score does not guarantee future performance.

13. Transparency and Auditability

Every output includes: raw metrics, subscores per window, effective weights, blend weights, reliability coefficient, adjusted base, length score (analytics only), engine score, each overlay multiplier with its written reason and evidence, the raw public score, the calibrated public score, and all processing flags. All parameters — including the overlay values and the calibration anchors — are externalized in versioned configuration.

14. Limitations and Future Work

Limitations: Relies on Myfxbook-reported data; broker-level manipulation propagates where the overlay does not correct it. Float Stress depends on availability of daily floating P/L data. The manual overlay reflects analyst judgement and transcription from third-party services; it is documented and evidenced but not itself automated. The calibration curve is anchored to the current retail EA universe. Weights and anchors are calibrated against the current Expert Advisor universe. Scores track records, not strategies — different brokers/leverage may differ.

Future: Trade-level analysis (holding time, single-trade contribution). Regime-aware scoring. Peer normalization. Iterative weight optimization based on predictive accuracy. Periodic, versioned review of the overlay and calibration anchors.

Appendix A: Subscore Anchor Tables

Piecewise linear interpolation; clamped outside range.

Sortino (↑ better)

Value	0.0	0.5	1.0	1.5	2.0	3.0	4.0
Score	0	35	55	70	80	92	100

Calmar (↑ better)

Value	0.0	0.5	1.0	1.5	2.0	3.0	5.0
Score	0	35	55	68	78	90	100

Ulcer Index (↓ better)

Value	2	4	6	10	15	25	35
Score	100	90	80	60	40	20	0

Neg. Skew (↓ better)

Value	0.0	0.3	0.6	1.0	1.5	2.0	3.0	4.0
Score	100	90	80	65	45	30	10	0

AutoPos (↓ better)

Value	0.0	0.05	0.1	0.15	0.25	0.35	0.5
Score	100	90	75	60	35	20	0

Float Stress (↓ better)

Value	0.2%	1%	3%	6%	12%	20%	30%
Score	100	85	70	50	30	15	0

Appendix B: Calibration Anchor Table

Fixed, strictly increasing piecewise-linear mapping from raw public score to published display score (Section 11).

Raw score	0	10	20	30	40	50	60	70	85	100
Display score	0	30	45	58	70	78	87	92	97	100

