Entropy to the Mean

Winning Streaks Carry a Rising, Measurable Cost: Evidence from Seven Seasons of College Basketball

An Entropic Duel Theory (EDT) White Paper

James Castranova

Independent researcher

Correspondence: jcastra@gmail.com

Preprint. July 2026

Abstract

The hot hand literature asks whether success predicts success. This paper asks the inverse question: conditional on a team continuing to win, does the quality of its performance decay as the streak lengthens? Using 69,022 Division I versus Division I team-game rows across seven college basketball seasons, with point-in-time KenPom opponent adjustment and no future leakage, we estimate the effect of entering win-streak depth on opponent-adjusted margin residual within team-season fixed effects, with standard errors clustered on 2,306 team-seasons. The raw within-team estimate is -0.3207 points per game of depth (SE 0.0280, t = -11.5); an own-construction memoryless null, built at each team's own rate and passed through the identical pipeline, reproduces a mechanical floor of -0.2238; the claim of this paper is the excess beyond that floor, -0.0969 per game of depth (0 of 200 null replicates as negative; lower-tail p at or below .005). The raw estimate carries a pre-registered threshold shape: flat through depths 1-2, onset near depths 3-5, deepening to roughly -3 points at depths 13 and beyond. The naive cross-team specification is opposite-signed, a rating-lag artifact the within-team design removes. The effect survives fatigue, recent-form, and opponent-preparation controls. A loss-streak placebo is opposite-signed but largely mechanical under its own null. In the decisive horse race, accumulated overperformance collapses to insignificance while duration survives. A pre-registered, hash-locked one-shot replication on the held-out 2021-22 season passes both frozen criteria (coefficient -0.4169, t = -4.6). A validated tennis instrument returns a concordant phase-tier residual in two universes, and an EPL construction replicates enter-depth compression locally. The duration-burden interpretation is [supported]; the causal mechanism is [open]; two pre-registered composition tests remain open and gate the strongest public version of the claim.

Keywords: regression to the mean; hot hand; streaks; ordered states; selection bias; within-team identification; entropy; entropression; pre-registration; college basketball.

1. Introduction

Forty years of streak research has been organized around one question: does success predict success? Gilovich, Vallone, and Tversky (1985), building on the small-sample intuitions documented by Tversky and Kahneman (1971), answered no and attributed the perception of momentum to misreading random sequences. Miller and Sanjurjo (2018) showed that the original analysis itself contained a subtle selection bias, and that correcting it revives evidence for a modest hot hand. Reviews of the intervening literature (Bar-Eli, Avugos, and Raab, 2006) document how sensitive this entire debate is to conditioning: the act of selecting sequences bends the statistics of what is selected, and both sides of the hot hand argument have at times been fooled by it.

This paper asks a different question, one the hot hand framing does not reach. It is not whether a winning team is more likely to keep winning. It is whether the same team, while it keeps winning, performs progressively worse relative to expectation as the ordered state ages. The unit of interest is not the next outcome but the conditional quality of performance, opponent-adjusted, indexed by the duration of the streak the team carries into each game.

The distinction matters because the two default deflationary accounts attack different objects. Regression to the mean (Galton, 1886) says that displacement from equilibrium predicts movement back toward it: a team that has banked unusual overperformance should give some back, in proportion to the size of the bank. It says nothing about duration as such. Selection geometry says that conditioning on streak membership mechanically bends conditional averages even for a memoryless process, so the mere presence of within-streak decline proves nothing. A finding that survives as a duration effect must therefore do two separate things: beat displacement in a direct horse race, and exceed the decline that a memoryless process of the same skill produces under the identical construction. This paper does both, reports the size of each subtraction, and states exactly what remains open afterward.

The result, previewed: within team-season, opponent-adjusted margin residual declines by roughly one third of a point per game of entering win-streak depth in the raw estimate (-0.3207), of which -0.2238 is reproduced mechanically by a memoryless null run through the identical construction; the honest per-game burden is the excess, -0.0969, with a threshold-then-deepening shape, replicated out of sample under a hash-locked pre-registration. Accumulated overperformance cannot absorb it. Fatigue, recent form, and operational opponent-preparation proxies cannot absorb it. A loss-streak placebo runs the other way. A memoryless null built at each team's own rate reproduces part of the raw slope mechanically, and that part is subtracted before any claim is made. What survives the subtraction is the object of this paper.

Two disclosures frame everything that follows. First, this is a theory-first research program with an unusual governance structure: frozen specifications, pre-registered verdict logic, synthetic validation gates before any real-data run, and independent adversarial review of every instrument, with the author as sole adjudicator. Several of the program's own earlier claims were killed by its own nulls, and those kills are reported here as part of the evidence that the surviving result is different in kind. Second, two pre-registered identification tests aimed at the composition family of objections are frozen but not yet run. The strongest public version of the claim waits on them, and this paper says so rather than pretending otherwise.

2. Theory: Entropic Duel Theory

Entropic Duel Theory (EDT), formerly Regression Duel Theory (RDT) in the program's internal record, is named for the direction of its central claim: that much of what the literature files under regression to the mean is not a passive statistical settling but the dissolution of ordered competitive states under accumulated load, entropic in the plain sense that order decays, and that the decay pressure scales with how long the order has persisted. The provocation can be stated in four words: entropy to the mean. The name marks a hypothesis about direction and form, not a physical identity: no thermodynamic or information-theoretic entropy is computed or claimed to be measured anywhere in this paper, and the firewall of Section 2.3 governs every interpretive sentence. The duel is empirical and runs throughout: duration of order against distance from the mean. EDT gives the phenomenon it observes a name. Entropression is the excess, duration-indexed compression of sustained ordered competitive performance: the progressive adverse deformation of performance channels with accumulated ordered duration, beyond the deformation generated by an appropriate own-construction memoryless process. The term names the observed form, not its causal mechanism and not a thermodynamic identity. An entropression signal is a substrate-specific statistical registration of that form: a depth coefficient, a channel deformation, a phase-tier rise. The statistics are not entropressing; the ordered performance state undergoes entropression, and the statistics register its signature. Throughout the empirical sections the operational vocabulary stays more specific than the theory-level name: opponent-adjusted margin compression, depth coefficients, channel deformation, excess beyond the memoryless floor. The causal generator of entropression is open.

2.1 The law of form

The theory's most compressed observable statement is a single sentence: *a streak becomes more difficult as it ages inside its own lifespan.* This is not the claim that long streaks eventually end (trivially true), and it is not the claim that unusually good measured performance comes back down (regression to the mean, also true and not in dispute). It is the claim that duration itself, an ordered state persisting inside its own life, accumulates difficulty against continuation, and that this accumulation is measurable as deformation in performance channels while the streak is still alive. Status: [supported], within the evidence assembled below; not [established].

2.2 Vocabulary and the descriptive ledger

EDT holds that observed performance mixes skill, context, variance, and pressure generated by accumulated imbalance. Equilibrium is the resting tendency; drift and flattening are the background condition; streaks are temporary ordered states under load; regression is the observable dissolution of temporary order, not a force. The program's ledger vocabulary (regression debt, regression credit, displacement from equilibrium, compression, enforcement channel, partial discharge) is descriptive bookkeeping for where imbalance has accumulated and through which measurable channels it discharges. Nothing in the ledger is a claim of obligation: no team is ever described as owed a loss, and no process is described as scheduled. Words that project directional compulsion onto free processes (a team is due, the streak must break) are banned from the program's empirical writing, because a memoryless process carries no obligation and no schedule; if a fall is directional, duration-scaling, and exceeds the memoryless benchmark, that is precisely the thing under study, and it earns a measurement, not a slogan.

2.3 The firewall

The program maintains a strict firewall between its observable layer and its causal layer. The observable layer, everything estimated in this paper, is the footprint: compression, channel deformation, duration-indexed decline. The causal layer, the author's position that the pressure is a general dissolutive tendency acting on sustained order, is an inference from recurrence of the same form across substrates, is tagged [speculative] as mechanism, and appears exactly once in this paper, in Section 12. No regression coefficient in this paper is asserted to be a physical quantity, and no thermodynamic identity is claimed. Empirical sections measure the footprint; they do not name the force.

2.4 What the theory stakes in advance

EDT stakes six falsifiable expectations for a substrate like college basketball. (P1) Within the same team-season, opponent-adjusted performance declines with entering streak depth. (P2) The decline is carried by duration, not by accumulated displacement: in a joint model, streak depth survives and banked overperformance does not. (P3) The decline is order-specific: a loss-streak placebo does not reproduce it in the same direction. (P4) Expression is distributed across channels rather than localized in one mundane mechanism, because the pressure discharges through whatever channel is available. (P5) The decline exceeds what an own-construction memoryless process produces, because selection geometry alone bends conditional averages and must be subtracted before anything is claimed. (P6) The same form recurs in other competitive substrates, expressed through substrate-appropriate clocks and channels rather than through an identical statistic. Sections 5 through 9 test each of these. What EDT does not claim: that statistics themselves compress, that skill resists dissolution (skill may extend how long order persists; EDT tests

whether persistence itself creates rising pressure), that markets misprice streaks (market results are excluded from doctrine), or that any of this is fate.

3. Data

The primary frame is 69,022 Division I versus Division I team-game rows across seven seasons: 2016-17, 2017-18, 2018-19, 2019-20, 2022-23, 2023-24, and 2024-25. The 2020-21 season is excluded (COVID irregularity). The 2021-22 season was reserved untouched as the out-of-sample set and analyzed exactly once, under the frozen protocol of Section 6.7.

Opponent adjustment uses point-in-time KenPom adjusted efficiency margin (AdjEM): for each game, the latest weekly archive snapshot on or before the game date, with a preseason fallback for games preceding the first snapshot, and no future leakage anywhere in the pipeline. Expected margin is fit globally as intercept + b1 x (team AdjEM - opponent AdjEM) + b2 x home, with estimated b1 of approximately 0.678, home-court term of approximately 4.18, and residual standard deviation of approximately 13.1 points. The dependent variable throughout is the residual: actual margin minus expected margin.

Streak depth (enter_depth) is the count of consecutive Division I wins the team carries into the game, resetting to zero on any loss. The matched-only definition (Division I opponents only) is primary; Section 6.6 reports the full-schedule alternative and discloses the difference plainly.

4. Identification Design

The load-bearing design choice is team-season fixed effects, implemented by within-team-season demeaning, with standard errors clustered on team-season (2,306 clusters). Each team is its own control. This choice is not cosmetic. A team in mid-streak carries a rating that has not yet caught up to its current run, so a naive cross-team comparison makes streaking teams appear to overperform expectation: the rating-lag artifact. The within-team contrast removes the between-team rating-level artifact that dominates the naive specification and sharply reduces vulnerability to persistent quality differences. Time-varying within-season form remains separately addressed by the recent-form controls.

The diagnostic worth emphasizing: the naive specification and the within-team specification disagree in sign. Streaking teams look like overperformers across teams and are underperformers within team. Any pooled cross-team analysis is vulnerable to rating lag and persistent between-team quality differences and cannot, by itself, identify the within-team depth effect. The sign reversal is itself evidence that the within-team design is necessary, and it supplies a practical lesson for forecasting systems built on lagged ratings, whose evaluation of team strength

under luck and schedule noise is itself a measurement problem (Lopez, Matthews, and Baumer, 2018).

5. Core Result

Within team-season, with clustered standard errors:

```
enter_depth coefficient = -0.3207     clustered SE = 0.0280     t = -11.5
```

Each additional game of entering streak depth costs approximately one third of a point of opponent-adjusted margin in the raw estimate. That raw slope is never this paper's claim by itself: the own-construction memoryless null of Section 7 reproduces -0.2238 of it mechanically, and the headline magnitude is the excess beyond the floor, -0.0969 points per game of depth (0 of 200 null replicates as negative; lower-tail p at or below .005). Raw, floor, and excess are reported together wherever the effect is cited. The raw effect is not linear from zero: the pre-registered bin shape shows a threshold. Depths 0 and 1-2 sit slightly positive; onset arrives at depths 3-5; the burden deepens thereafter. Table 1 reports the pooled out-of-fold shape (each season scored with expected-margin coefficients fit only on the other six; all seven seasons individually negative and significant, mean approximately -0.32, range -0.20 to -0.49) alongside the one-shot out-of-sample shape from Section 6.7.

Depth bin	Pooled seven-season shape (out-of-fold)	Out-of-sample 2021-22 shape
0	+0.23	+0.36
1-2	+0.09	+0.11
3-5	-0.55	-1.00
6-8	-1.75	-1.95
9-12	-1.56	-3.20
13+	-3.10	thin; not separately reported

Table 1. Opponent-adjusted margin residual by entering streak-depth bin. Pre-registered hypothesis: flat through 1-2, negative onset at 3-5, monotone non-positive after. The shape was locked in writing, with a hash, before the out-of-sample season was loaded.

Status of the core result: [established] as a checkpoint finding within this frame, meaning the within-team, opponent-adjusted decline with depth, its threshold shape, and its out-of-sample replication (with the one caveat disclosed in Section 6.7). This tag deliberately does not extend to interpretation.

6. The Falsification Battery

The result above is only as good as the attempts made to kill it. This section reports each attempt, its outcome, and its residual scope. Coefficients are the depth coefficient after adding the named controls.

6.1 Fatigue

Adding rest days and games played in the last nine days: depth remains approximately -0.3062, and the fatigue proxy itself is null. The burden is not tired legs.

6.2 Recent form

Adding the prior-three-game residual: depth remains approximately -0.2717 (t of roughly -9). Recent form carries a small, separately significant negative term but does not absorb depth.

6.3 The loss-streak placebo

Entering-loss-streak depth is opposite-signed: +0.2916 (t = +9.3). Run through the same own-construction memoryless null, the coin at each team's own rate reproduces a null center of +0.2540, leaving an excess of +0.0376 with an upper-tail p of .12: approximately 87 percent of the loss-side mirror is mechanical, and the remainder does not clear conventional significance. Sustained losing reverts upward largely because the construction says it must. The placebo is therefore weak asymmetry evidence: the sign structure is consistent with an order-specific burden, but on its own it argues only weakly against a generic any-streak artifact, and the near-mirror raw magnitudes are, on their own, also consistent with symmetric mean reversion. The horse race below is what discriminates, and it is why the placebo is supporting evidence rather than the decisive test.

6.4 The mean-reversion horse race (decisive)

This is the strongest single test in the paper, because it puts the two rival objects in the same regression and lets them compete. Define streak_cum_resid as the total opponent-adjusted overperformance accumulated during the current ongoing streak: the exact quantity that displacement-based reversion would unwind. It correlates 0.74 with depth, so the competition is real.

Specification	enter_depth	cumulative streak residual
Depth alone	-0.3206 (t = -11.5)	
Cumulative residual alone		-0.0269 (t = -9.4)
Both together	-0.2837 (t = -6.8)	-0.0053 (t = -1.2)

Table 2. The horse race. Each rival is significant alone. Forced to compete, duration survives and accumulated displacement collapses.

The interpretation is sharp. Mean reversion knows only the magnitude of prior overperformance; it cannot explain why two teams with the same banked overperformance differ according to the duration of the streak that banked it. When both are measured, duration carries the burden and distance from the mean does

not. This is the result simple reversion cannot fake, and it is the empirical content of the paper's title: the duel is between duration and displacement, and duration wins. The conditional coefficient also has its own null reference: run through the memoryless construction, the horse-race depth coefficient centers at +0.0117 under the null, so the null-referenced excess is -0.2954, and 0 of 200 memoryless replicates produced a conditional depth coefficient as negative as the observed -0.2837. Unlike the raw slope, essentially none of the conditional coefficient is mechanical. This is the cleanest null-referenced survival in the paper. Status: [supported] at the level of mechanism-discrimination; the coefficient itself remains part of the [established] checkpoint frame.

6.5 Opponent preparation (the target-game story)

The strongest surviving mundane competitor after the first audit was the target-game account: opponents prepare harder for, and play up against, a deep-streak team. The full control (66,178 rows, team-season fixed effects, clustered) adds opponent streak depth, opponent coming off a loss, opponent rest advantage, postseason, road and neutral indicators, and own rest and fatigue.

Term	Coefficient	SE	t
enter_depth	-0.2893	0.0279	-10.4
opponent off a loss	+1.1590	0.1294	+9.0
opponent rest advantage	-0.0876	0.0347	-2.5
postseason	-0.5310	0.2409	-2.2
road	+0.3963	0.1461	+2.7
depth x opponent off a loss	+0.0077		+0.1

Table 3. Full opponent-preparation control. Depth survives essentially unchanged; the interaction that the target-game story requires is null.

Two findings weaken the obvious target-game account. First, the depth coefficient is essentially unchanged with all operational opponent-preparation proxies in the model. Second, the interaction is null: a desperate opponent, coming off a loss, does not amplify the depth burden, contrary to the target-game prediction. A preliminary opponent-strength split points the same way: the burden is present against weak, mid, and strong opponents (approximately -0.30, -0.32, and -0.38 respectively), including weak opponents who cannot capitalize on extra motivation. Honest scope: the operational proxies do not exhaust the construct. Soft channels (media attention, coaching emphasis, rivalry psychology, seeding pressure) are unmeasured. The obvious version of the story took a serious hit; the construct is weakened, not closed. Status: [supported] that the burden is not the operational target-game effect; the soft-channel remainder is [open] and listed in Section 12 as a limitation.

6.6 The depth-definition audit, with a recorded wrong prediction

Streak depth can be defined two ways, and the choice is material: matched-only (consecutive Division I wins) yields -0.3207 (t = -11.5); the true full-schedule definition (all wins including non-Division I) yields -0.1780 (t = -5.8). The original audit predicted that counting non-Division I wins toward depth would strengthen the effect. It did the opposite: the effect roughly halved, though it stays negative and significant. That wrong directional prediction is logged here rather than buried, because a program that hides its misses cannot be trusted on its hits. The interpretation: counting low-resistance blowouts as streak depth dilutes the signal, which argues the burden tracks sustained order against real competitive resistance, not the raw consecutive-win count. The theory-justified matched-only definition is therefore primary, and this choice was frozen before the out-of-sample test. Disclosure, stated plainly: the primary definition yields the larger coefficient; the secondary definition remains negative and significant but smaller.

6.7 Pre-registered out-of-sample replication

Before any out-of-sample data was loaded, the specification was locked in writing and hashed (specification hash cb04061f7e1a00d5): the expected-margin coefficients applied as-is rather than refit, the primary depth definition, the depth bins, the threshold hypothesis, and the pass criteria. The one-shot test on the never-before-analyzed 2021-22 season (9,808 Division I versus Division I rows) returned enter_depth = -0.4169 (clustered SE 0.0907, t = -4.6), passing both frozen criteria: negative with t at or below -2, and magnitude within the pre-registered band of -0.50 to -0.15. The frozen bin shape reproduced (Table 1, right column).

Caveat, stated exactly as it stands in the record: the 2021-22 box scores lacked a venue column, so the frozen residual's home-court term was set to zero for this run. Home-court zeroing may reduce precision and may leave residual venue confounding in this run: within-team demeaning absorbs a constant home-court term only imperfectly when home and away games are unevenly distributed across streak depths within a season. The correct statement is that the frozen replication passed subject to that disclosed limitation, not that it replicated perfectly. A clean re-run with recovered venue data remains open; for this version the gate was dispositioned as waived with disclosure, and no venue-recovery re-run was performed before release.

7. The Null Discipline and the Selection-Geometry Floor

7.1 Why comparing to zero is not enough

Conditioning on streak membership bends conditional averages even for a memoryless process. To be deep in a win streak, a team had to win the preceding games, and winning is correlated with having played well, so the games at the deep

end of a selected streak sit, on average, somewhat closer to the team's own mean than the fortunate games that built the streak. A memoryless coin, flipped at each team's own rate, therefore produces some downward drift across its own selected win streaks. That drift is pure selection geometry: it is not a force, it is not memory, it is a property of conditioned averages, and a spreadsheet produces it. It follows that demonstrating within-streak decline against a flat benchmark demonstrates nothing. The only comparison that isolates memory from selection is real decline against the decline of a memoryless same-skill process run through the identical construction: the same streak-finding code, the same depth variable, the same regression. Selection geometry lives in both and cancels only in the difference.

The mechanism deserves a precise statement, because the intuitive streak-selection account understates it. enter_depth is a sequential regressor constructed from the team's own lagged outcomes, and the within transformation demeans both regressor and outcome over the team-season. A run of wins is, on average, a run of positive residuals, and those games raise the season mean that is subtracted from every observation in that season. Games entered at high depth therefore sit, by construction, inside seasons whose demeaned baseline has been pulled upward by the very wins that created the depth, so their demeaned residuals are pushed downward mechanically even when the underlying process is memoryless. This is the fixed-effects analogue of dynamic-panel (Nickell-type) bias: a regressor built from lagged realizations of the outcome acquires a negative within-panel correlation with the demeaned error. The floor is therefore not merely conditioned-average curvature from selecting streaks; it is generated by the interaction of the sequential regressor with within-demeaning itself. That is why it can be measured only by running a memoryless process through the identical construction, and why it is as large as it is.

7.2 The program rule, applied in both directions

The program's standing rule is symmetric. No signal is called mechanical, artifactual, or mere reversion unless a null model has actually reproduced its shape under the same construction; and no signal is called a finding until it exceeds its own-construction memoryless null. Nulls are clues, not verdicts: each null result informs instrument hygiene rather than adjudicating the theory by itself, and a null produced by a mis-built instrument is a build error, not a verdict.

7.3 The floor and the excess

Applied to the flagship, the own-construction memoryless null (Track 1 of the program's audit sequence) quantifies how much of the raw -0.3207 slope the construction itself generates with no memory present. The verified filed values are a mechanical floor of -0.2238, leaving a floored excess of -0.0969 beyond selection geometry, with 0 of 200 memoryless replicates producing a slope as negative as observed (lower-tail p at or below .005). The same discipline applied to the conditional coefficient of Section 6.4 yields the paper's cleanest null-referenced survival: observed -0.2837 against a null center of +0.0117, an excess of -0.2954,

with 0 of 200 replicates as negative. The reading discipline this imposes on every downstream sentence: the raw coefficient is never cited unqualified. The claim of this paper is the excess, -0.0969 points of opponent-adjusted margin per game of depth beyond what a memoryless same-skill process produces under the identical construction, compounding across the depth ladder. The threshold-then-deepening bin shape and the horse-race result are the reasons to believe the excess is duration-indexed rather than residual selection curvature; the pre-registered tests of Section 10 are the reasons the strongest version of that sentence still waits.

7.4 The program's record with its own null

The credibility of a null discipline is established by what it has killed, and this program's instrument has killed the program's own claims. A three-point-shooting streak-debt claim, once presented internally with a dose-response narrative, was retested against the coin at each team's own rate: the real cooling sat inside the coin band, and the claim was withdrawn as selection geometry. An MLB starter-ramp signal that had survived a first audit pass was retested against a within-pitcher reorder null that holds each pitcher's actual start-outcome distribution fixed and shuffles only the order: the ramp reproduced under reordering and was retired per pre-commitment. An MLB team-game margin channel reported the compression footprint null three carriers deep, and that null stands in the record. These kills are reported because they are the evidence that the surviving basketball result is different in kind: it is the one object in the program that has beaten the same instrument that correctly destroyed the weaker claims. Status of the null discipline itself: [established] as program law.

8. Channel Expression: Distributed and Directional

If the burden were carried by one box-score channel, a single mundane mechanism would be the natural suspect. The channel decomposition (depth effect on each channel, within team, clustered; diagnostic, not verdict-bearing) shows the opposite: distributed expression with a coherent behavioral signature.

Channel	Depth coefficient	t	Effect on margin
Offensive eFG%	-0.00092	-4.7	hurts
Offensive 3P%	-0.00073	-3.2	hurts
Offensive rebound rate	-0.00111	-5.5	hurts
Free-throw generation	-0.00056	-1.7	hurts (weak)
Offensive turnover rate	-0.00072	-6.7	helps (turnovers fall)
Pace	-0.07979	-5.8	slows
Defensive eFG% allowed	+0.00059	+3.2	hurts (mild)
Points scored	-0.1513	-5.5	hurts
Points allowed	+0.0591	+2.3	hurts (mild)

Table 4. Channel decomposition. A correction is on the record: an earlier internal writeup mis-signed the turnover channel; the corrected reading is reported here.

The corrected picture: deep-streak teams tighten on the channel under the most direct volitional control (ball security improves, pace slows) while the less controllable channels erode (shooting efficiency, three-point conversion, offensive rebounding, free-throw generation), and the points decomposition confirms the burden is mostly offensive erosion with mild defensive slippage. The behavioral signature is coherent: under sustained-order pressure, grip the controllable and bleed in the uncontrollable. What this supports is distributed expression consistent with a general strain on the ordered state discharging through available channels, rather than one localized mechanism. What it cannot do, and is not asked to do, is distinguish a law-like account from a sophisticated emergent one; that distinction may not be resolvable from box scores at all. Status: [supported] as a diagnostic pattern.

9. Cross-Substrate Evidence

A duration burden that existed only in one sport, under one construction, would invite the suspicion that it is a construction. The program therefore built substrate-appropriate instruments elsewhere, under one principle: each substrate expresses load through its own clock and channels, and the law-of-form standard does not require the same statistic across substrates, only the same form. Every cross-substrate instrument passes synthetic validation against planted ground truth before touching real data, and negative results are filed with the same weight as positive ones.

The soccer program states this standard at its strongest. Its synthetic-validation instrument computes every coefficient through two independently implemented fixed-effects paths (an explicit block reference and a production within-transform, solved by different factorizations) and hard-stops on any disagreement beyond frozen tolerances. During 2026 pre-registration work that parity gate caught a genuine numerical defect in the instrument itself: an ill-conditioned auxiliary solve on the adversarial synthetic family built to stress it. The repair was verified against 50-digit-precision ground truth on the exact failing designs, production statistics were bit-identical across the repaired versions, and the finished instrument reproduces every coefficient across Windows/MKL and Linux/OpenBLAS platforms to a maximum difference of 2.25e-15. The defect, the failed repair candidate that would have weakened the gate, and the accepted repair are all filed. An instrument that has caught, documented, and survived its own defect is the standard this program means by validated.

9.1 Tennis: a validated phase-tier residual

Tennis is a one-person ordered state with no team-season baseline, so its clock is relative position inside the streak (early, middle, late thirds of win streaks of length six or more) rather than an objective depth ladder. The frozen instrument tests

whether within-streak pressure readouts rise across thirds, against two nulls: a path-erased shuffle (does the shape survive destroying chronology while keeping the bag of wins?) and a structure-matched null with an escalating control ladder (is the shape explained by era, surface, format, and level?). The instrument passed five of five synthetic fixtures with planted ground truth before the real run, including the critical over-control safety case in which a live signal must not be buried by an over-fitted null.

Universe	Eligible streaks	Early / middle / late	Phase rise	Null 1 residual	Null 2 (light) residual	Verdict
Score-only	2,825	2.4102 / 2.5723 / 2.7060	+0.2958	+0.302 (frac 0.000)	+0.0628 (frac 0.000)	C1 + C3
Break-panel	1,359	0.6491 / 0.6816 / 0.7039	+0.0548	+0.0566 (frac 0.000)	+0.029 (frac 0.000)	C1 + C3

Table 5. Tennis phase-tier residual, run v0.1.4, ATP main-draw data (Sackmann dataset, validated atlas construction). C1: survives the path-erased shuffle. C3: a residual remains after non-live structure matching. Concordant across all score readouts (average sets, straight-set rate, deciding-set rate, tiebreak rate) and all break-panel readouts.

Boundaries, kept exactly as filed. The clock is relative phase, a coarser evidentiary shape than basketball's objective depth ladder. Adding opponent tier shrinks the residual (to +0.040 and +0.014 in the two universes), but opponent-tier escalation is flagged as a possibly live expression channel, so per frozen logic that shrinkage is relocated, not closed. The cleanest theoretical cell (late streak phase, early tournament round, non-elite opponent, where the objection that late just means harder rounds has nothing to explain) was deferred on a genuine null-construction question rather than shipped as a guess, and no soft-bracket numbers are part of the filed result. Mechanism unresolved. Status: [supported].

9.2 Soccer: replication of the construction in a different geometry

An English Premier League instrument was frozen to the basketball estimand deliberately: enter-depth on consecutive league wins, team-season fixed effects, opponent strength, home, and rest as covariates, significance assessed only against a within-team-season permutation null. The construction locally replicated enter-depth compression across shot, shots-on-target, and goal channels; a lower-division extension returned a within-football channel-shift result in which the goal-difference channel is robust and process channels are not. Both are filed at [supported], the second explicitly as channel morphology. Soccer's parity structure differs from college basketball's extreme inequality, and the substrate-clock principle predicts exactly this kind of channel-routed, medium-strength expression rather than a copy of the basketball ladder.

A second-stage soccer program is in progress and is disclosed as ongoing rather than folded into the claims above: a frozen multi-league protocol (mean-reversion horse race, two-way studentized inference, outcome-coupled synthetic worlds) with

a fully certified execution instrument, whose large synthetic calibration battery is planned but not yet run. Nothing in this paper depends on that battery; its role is to raise the soccer evidence toward the basketball flagship standard, and its results will be filed whichever way they fall. Status of the filed soccer findings above remains [supported].

9.3 Baseball: honest nulls, filed

Major League Baseball is the program's hardest substrate and its clearest demonstration that the pipeline reports what it finds. The team-game margin channel returned the compression footprint null three carriers deep. The one signal that survived the first full pass, a starter ramp, was subsequently reproduced by the within-pitcher reorder null and retired per pre-commitment, as reported in Section 7.4. The working hypothesis, held as [open], is substrate-structural: baseball's parity and outcome luck may place a real burden below the detection floor of available data, and the honest conclusion may eventually be that the burden is predicted but empirically unreachable there. No basketball result is transferred to baseball by assertion.

10. Alternative Explanations and the Identification Frontier

This section states each surviving rival account, what has already constrained it, and what remains open. The paper's credibility rests on the accuracy of this section more than any other.

10.1 Mean reversion

Constrained decisively by the horse race (Section 6.4): the quantity displacement-based reversion would unwind collapses to insignificance in the joint model while duration survives. A vocabulary point belongs here as well: reversion describes where conditioned averages settle; it does not pull. Any account in which the fall is directional, obligatory, and duration-scaling has stopped being reversion and started being a mechanism, which is the object under study.

10.2 Selection geometry

Constrained by construction: the own-construction memoryless null quantifies it, and the claim is the excess beyond it (Section 7.3). Selection geometry is not a rival explanation of the excess; it is the floor the excess is measured from.

10.3 The inflated-entry (entry-selection) objection

The objection, stated without smuggling: teams that entered a streak at an inflated measured level will post lower opponent-adjusted numbers later, which can appear as a depth slope. The program's synthetic work on this objection surfaced a structural point first: a pre-streak local peak is a level, a starting altitude, and the

flagship footprint is a duration-deepening slope (bins moving from roughly -0.55 to -1.75 to -3.10). Within a fixed episode, a static entry-level offset changes altitude but does not itself generate a within-episode duration slope. A pooled depth ladder may still reflect composition across episodes of different lengths, which is the separate identification question reserved to Section 10.4. In synthetic worlds, the only way an inflated entry level produced a depth slope was to insert, by hand, a term making residuals decline with depth, and that insertion is the tell. The canon sentence of the program's defense: a local-quality peak can explain altitude, not duration-indexed compression; to explain a progressively deepening depth slope, a rival account must introduce a duration clock, and once it does, it has stopped explaining the effect away and started renaming the mechanism this paper measures.

The argument does not substitute for the test, and the test has now been run. The stratification instrument was frozen with pre-registered verdict logic (entry states classified as cold, neutral, and hot on strictly pre-streak local quality; the decisive model a depth ladder within each stratum) and passed its synthetic gating fixtures (a planted all-strata slope, a planted hot-only artifact, a mixed world, and a flat world, all classified correctly) before touching real data. The real-data run returned the strongest pre-registered outcome. Cold entries are real and abundant: 4,044 episodes entered below pre-streak baseline, with mean entry quality -5.791, median -4.696, and 100 percent of the stratum at or below zero, so the objection's raw material, that streaks begin only at inflated peaks, fails on arrival. The depth slope is present in every stratum and is not concentrated in hot entries: cold -1.2752 (t = -10.54), neutral -1.2152 (t = -10.16), hot -1.0035 (t = -12.13), with no significant interaction concentrating the slope in inflated entries. A duration-deepening slope inside genuinely non-inflated entries removes the inflated-entry objection's raw material entirely. Two boundaries are preserved exactly as filed: the strata are defined on the program's chosen pre-streak local-quality measure, and a rival account remains free to propose a materially better stratifier; and the test addresses entry-state inflation only, not during-streak true-quality drift, which remains the business of the Section 10.4 composition tests. Status: [supported]; the entry-selection objection is answered within those stated boundaries.

10.4 Composition across streak lengths

The subtlest rival family: perhaps games inside long streaks differ from games inside short streaks because the teams that produce long streaks differ, so within-streak position effects are really between-population effects. Two pre-registered instruments target this directly. The within-episode separation test asks whether the duration effect survives population matching within team-season, and its kill condition is stated in advance: if every game inside a long streak carries the same pressure profile regardless of position, so that the long streak differs from a short streak only as a whole while pressure does not rise inside it, then an instrument that calls that duration-memory is broken. The within-episode order test asks, by

synthetic construction, whether whole-streak population differences alone can generate within-streak ordering. Both are frozen under the program's governance sequence (specification freeze, independent cross-review, synthetic certification, separately authorized execution) and neither has run its full battery. The evidentiary structure is locked at three levels: form (sustained ordered performance deforms with accumulated ordered duration: [supported]); identification (whether composition alone generates the within-streak ordering: the target of these tests, [open]); and generation (what produces the pressure: [open]). A clean identification result strengthens the first level; it does not settle the third. The strongest public version of this paper's claim waits on these two tests, and the paper says so.

10.5 Fatigue, schedule, and opponent preparation

Constrained as reported in Sections 6.1 and 6.5. The soft, unmeasured preparation channels remain the principal conventional residual. A critic who wishes to retreat to a bare maybe-it-is-fatigue must identify a specific defect in the reported controls or a materially different uncaptured mechanism; the reported controls have materially constrained the principal conventional confounds.

10.6 True quality drifting upward during the streak

A rival in which the team is genuinely improving as the streak ages is addressed by argument, disclosed as argument rather than as an empirical result of any test here. First, a true-quality process that itself scales with streak depth would concede the duration-indexed form while proposing a different generative mechanism; it would not erase the footprint, but it could materially change its interpretation. Second, a team genuinely ascending in true quality with each win should post widening opponent-adjusted margins, the opposite of the observed compression and of the channel signature (tightening, slowing, eroding efficiency).

11. What Would Falsify This

A theory that cannot state its own kill conditions is not offering evidence. The following outcomes, each tied to a frozen or standing instrument, would force retraction or material rewriting of the claim. First: the stratification test of Section 10.3 has now been run and returned its strongest pre-registered outcome (a depth slope in every entry stratum, including 4,044 genuinely cold entries, with no hot-entry concentration); that kill condition is retired on its original terms, and would reopen only if a materially better pre-streak stratifier were demonstrated to reverse the within-stratum ladder. Second: the within-episode separation test finds no rise in pressure across position inside matched episodes; the duration-memory reading is withdrawn. Third: the within-episode order test shows that whole-streak population differences alone generate the within-streak ordering in synthetic worlds matched to the real construction; the identification level fails. Fourth: a refined own-construction null reproduces the full raw slope, collapsing the floored excess toward

zero; the finding reverts to selection geometry, exactly as the program's own three-point and starter-ramp claims did. Fifth: failure to replicate on future sealed seasons under the frozen specification. The program has already retired claims under conditions like the fourth, so these are not hypothetical standards.

12. Discussion and Interpretation Ceiling

What is on the table after the falsification battery of Section 6, the floor, and the open tests is a specific, bounded object: within the same team-season, sustained winning order carries an opponent-adjusted performance cost that scales with the duration of the order, exceeds the own-construction memoryless benchmark, cannot be absorbed by accumulated displacement, is not explained by fatigue or operational opponent preparation, runs opposite in sign for losing order, expresses through distributed and directional channels, and recurs in form, on substrate-appropriate clocks, in tennis and soccer while returning honest nulls in baseball. The order-specific duration-burden interpretation is [supported]. EDT calls this excess duration-indexed compression of sustained ordered performance entropression (Section 2); the measured coefficients and channel deformations are its substrate-specific signals, while its causal generator remains open. The causal mechanism that produces the burden remains [open], with the author's general dissolutive-pressure account tagged [speculative]. The tag on the core empirical effect, within its frame, is [established] as a checkpoint finding.

The interpretation ceiling is stated once and plainly. The author's causal position is that the recurring form reflects a general dissolutive pressure that sustained order accumulates against its own continuation, and the program's name for that inferred pressure is entropic in spirit. That position is tagged [speculative] as mechanism. This paper measures the footprint; it does not measure the force, it asserts no thermodynamic identity, and nothing in Sections 3 through 11 depends on the interpretation. Whether the burden is law-like or is a deep emergent property of competitive systems (adaptation, information leakage, psychological load, or channels not yet named) is [open], and may not be resolvable from box scores. The honest formulation of the research program is that the same form keeps appearing wherever an instrument capable of seeing it has been pointed, and the next tests are aimed at the one family of alternatives that could still unmake it.

Two practical implications stand regardless of interpretation. For forecasting, the rating-lag sign reversal in Section 4 is a warning to any system that evaluates streaking teams with lagged ratings across teams. For the streak literature, the paper's separation of duration from displacement, and its subtraction of an own-construction selection floor before claiming anything, are portable disciplines: the hot hand debate spent decades on exactly the selection subtleties that the floor makes explicit.

13. Claim-Status Ledger

Claim	Status	Basis
Within team-season, opponent-adjusted margin declines with entering win-streak depth; threshold shape; out-of-sample replicated (one documented caveat)	[established] (checkpoint frame)	Sections 5, 6.7
The burden is order-specific: survives quality, fatigue, recent form, displacement, and operational opponent-preparation controls; loss placebo opposite-signed; distributed channel expression	[supported]	Sections 6, 8
Duration, not displacement, carries the burden	[supported]	Section 6.4
The raw slope contains a mechanical selection-geometry component; the claim is the excess beyond the own-construction null	[established] (Track 1 verified: floor -0.2238, excess -0.0969, 0/200 replicates more negative)	Section 7.3
The same form recurs in tennis (phase-tier residual, C1 + C3 both universes) and soccer (enter-depth compression; channel morphology)	[supported]	Section 9
Baseball expresses no detected footprint in tested channels	reported nulls; substrate question [open]	Sections 7.4, 9.3
Within-streak ordering is not a composition artifact	[open]: two pre-registered tests frozen, not yet run	Section 10.4
The depth slope does not require inflated entry to exist	[supported]: stratification run; slope present in all entry strata including 4,044 cold entries; no hot-entry concentration	Section 10.3
The pressure is a general dissolutive force on sustained order	[speculative] as mechanism	Section 12

Table 6. Every substantive claim, its tag, and where it is earned. Nothing above its tag is asserted anywhere in this paper.

14. Reproducibility, Governance, and Disclosures

Pre-registration. The out-of-sample specification was locked in writing and hashed before the held-out season was loaded (hash cb04061f7e1a00d5), covering the applied expected-margin coefficients, the primary depth definition, the depth bins, the threshold hypothesis, and the pass criteria. Downstream instruments follow a standing governance sequence: specification freeze, independent adversarial cross-review, synthetic validation against planted ground truth with gating fixtures,

separately authorized execution, and separately authorized null runs. No construction changes are permitted after synthetic output unless a gate fails and the instrument returns to specification review; no metric selection is permitted after a null.

Adversarial review. Every instrument in this program was reviewed adversarially by two independent large-language-model systems operating under different vendors, one as builder and auditor and one as peer reviewer, with disagreements treated as calibration signals and the author as sole adjudicator of every canon decision. The author further discloses that AI systems assisted in instrument construction, drafting, and audit under that structure; all analytical decisions, adjudications, and this paper's claims are the author's.

Data. College basketball box scores derive from public sports-reference sources; opponent adjustment uses KenPom adjusted efficiency archives under subscription, taken point-in-time with no future leakage. Tennis uses the public Sackmann ATP match dataset (1968-2023 fork). Soccer uses public football-data.co.uk league files. Market prices appear nowhere in the doctrine or the analysis. Artifact inventories, frozen specifications, synthetic-validation reports, and regression outputs are maintained in the program record and are available to reviewers on request.

References

Bar-Eli, M., Avugos, S., and Raab, M. (2006). Twenty years of hot hand research: Review and critique. Psychology of Sport and Exercise, 7(6), 525-553.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology, 17(3), 295-314.

Lopez, M. J., Matthews, G. J., and Baumer, B. S. (2018). How often does the best team win? A unified approach to understanding randomness in North American sport. Annals of Applied Statistics, 12(4), 2483-2516.

Miller, J. B., and Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. Econometrica, 86(6), 2019-2047.

Pomeroy, K. Adjusted efficiency ratings and archives. kenpom.com (subscription data).

Sackmann, J. ATP match results dataset (tennis_atp), community fork covering 1968-2023.

Tversky, A., and Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76(2), 105-110.

football-data.co.uk. English league match data files.