PRE-REGISTERED FALSIFICATION · FBS 2015–2025 · COLLEGEFOOTBALLDATA

Systems Travel, Edges Don't

A coach's system does move with him between schools — that part of the thesis survived every test I threw at it. What did not survive is the claim that a mismatch between system and quarterback shows up in scoring margin the betting market missed. **The effect is bounded below two fifths of what it would take to beat the vig.**

5,022 team-games **770** team-season clusters **405** API calls, all cached

3 falsification tests Analysis run **2026-09-07**

THE CLAIM UNDER TEST

An interaction the market has no machinery to price

Betting lines are built from additive power ratings. A coach's record is in the number and a quarterback's pedigree is in the number, but fit between them is an interaction, and additive models have no term for it. If mismatch costs real points, that cost sits outside the price by construction.

The hypothesis is unusually falsifiable because it forbids main effects. Distance between a system and a player is symmetric: neither the coach being good nor the quarterback being good predicts it. So a signal with no main effect and a live interaction would be the signature of something genuinely unpriced — and a main effect turning up instead would be evidence the mechanism is something else entirely.

Coaching changes make the cleanest test bed. A system arrives intact because it lives in the coach's head; the roster he inherits was assembled by other people for a different system. In the collective era, quarterback acquisition runs through boosters and revenue-share allocation rather than the head coach, so the pairing is not selected on fit.

PREMISE · SYSTEMS TRAVEL	CONSEQUENCE · TEST B	BOUND ON ANY EDGE
55.8%	**+0.01**	**0.95pp**
of play-calling identity is the coach, against 9.1% school and 35.1% year	points per SD, 95% CI [−0.33, +0.35]. p = 0.958	excluded above this after bootstrap correction; break-even at −110 needs 2.38pp

SEQUENCE · EACH TEST GATED THE NEXT

Three falsification tests, run before any model was built

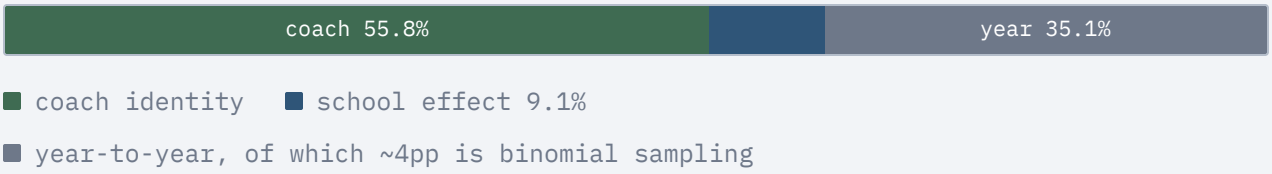
Every one of these was specified with its decision rule fixed in advance. Two killed a component of the design. Running them first is what made the whole project cost a session instead of a season.

01 Does a coach's system travel with him? PREMISE HELD

The benchmark is not zero — some drift across a job change is ordinary year-to-year wobble. So I compared two one-year differences on the same footing: a coach's change across a move against his change across an ordinary season at one school. Writing the axis as $\mu + \alpha_{coach} + \beta_{school} + \varepsilon_{year}$, the ratio of their variances is $1 + \sigma^2_\beta/\sigma^2_\varepsilon$. Rule fixed in advance: **ratio below 2.0 means systems travel**.

Early-down pass rate on neutral downs, z-scored within season, over 62 same-coach school changes and 1,001 same-coach same-school season pairs: **ratio**

1.26, 95% CI [0.83, 1.72] — passing even at the upper bound. Correlation across a move, 0.558.



02 Is designed QB run rate a coach attribute? AXIS KILLED

It led the candidate-axis list, and it is poison. Correlation across a job change: **0.005**. A coach's designed-run rate at his old school tells you nothing about his rate at the new one, because the rate is set by whether the quarterback can run — a roster property. Variance decomposition: coach **0.5%**, school 38.1%, year 61.4%.

Had it entered the system vector, the quarterback would have sat on both sides of the distance measure, shrinking the gap hardest for exactly the worst-fit pairs. That is not attenuation toward zero; it reorders which games look like mismatches.

> **The variance ratio alone would have waved it through.** QB run share scored 1.62 against a 2.0 threshold. The ratio compares the school effect to year noise and is silent on whether a coach effect exists at all — only the near-zero correlation exposed that $\sigma^2_{coach} \approx 0$. Both statistics are needed; the pre-registered rule was incomplete.
>
> **This was a protocol deviation.** The cross-move correlation criterion was added after the variance ratio had already passed — a decision rule changed once the data had been seen, which is the exact move pre-registration exists to prevent. It is defensible here because the addition made the rule *stricter* and because it is disclosed rather than absorbed, but disclosure is what makes it acceptable, not harmlessness. It is named as a deviation for the same reason it is reported at all.

03 Is the quarterback's situational profile measurable? COMPONENT KILLED

The QB side of the fit measure had to be a *contrast* — passing-down efficiency minus early-down efficiency, within player — because any non-contrast measure is

quality, and quality is a main effect the market already prices. The contrast differences quality out, which is exactly what the hypothesis needs.

It differences out everything else too. **Split-half reliability: −0.008.** Split one quarterback's own season randomly in half, compute the contrast on each half, and the halves do not agree. Year-over-year correlation +0.068 [−0.01, +0.15], against +0.283 for overall EPA per dropback on the same players.

Components are individually stable; their difference is not.

MEASURE	YEAR-OVER-YEAR R	SPLIT-HALF
Situational contrast (the intended q_p)	**+0.068**	**−0.008**
Overall EPA/dropback (benchmark)	+0.283	—
Early-down EPA alone	+0.241	—
Passing-down EPA alone	+0.173	—

With a median of eight neutral-script obvious-passing-down plays per quarterback-season against a per-play EPA standard deviation near 1.5, there is nothing left underneath.

An earlier draft claimed the bind was general — that what makes a quarterback measure fit-relevant is what makes it unstable. That is too strong, and the arithmetic says so. For a difference $D = X_1 - X_2$, the classical reliability (Cronbach & Furby 1970) is $\rho_D = (\rho_{11}s_1^2 + \rho_{22}s_2^2 - 2\rho_{12}s_1s_2) / (s_1^2 + s_2^2 - 2\rho_{12}s_1s_2)$. The intercorrelation of the two bins, measured on the same 1,396 quarterback-seasons, is **+0.293** — above the +0.207 at which a difference of components this reliable has exactly zero reliability. Plugging the measured values in predicts ρ_D = −0.08 against the observed +0.068. **The null was forced by the difference operator before any football entered it.** A usable contrast ($\rho_D = 0.30$) would have required the two bins to be *negatively* correlated at −0.13, which two measures of the same passer will never be.

That is a narrower claim than the one it replaces, and deliberately so. It governs *unshrunk single-season differences* and says nothing about estimators that never form one. Pooling seasons raises component reliability by Spearman–Brown and carries the contrast with it — three seasons would put ρ_D near 0.20 on a fixed intercorrelation. Hierarchical shrinkage of the contrast, multi-season pooling and latent-trait treatment with known measurement error were never tried. None of this rescues the thesis: Test B is a hard null on its own terms, and a measurable q_p

would have had to survive it too. It corrects what this document previously called impossible. Reproduced by `scripts/phase0_contrast_arith.py`.

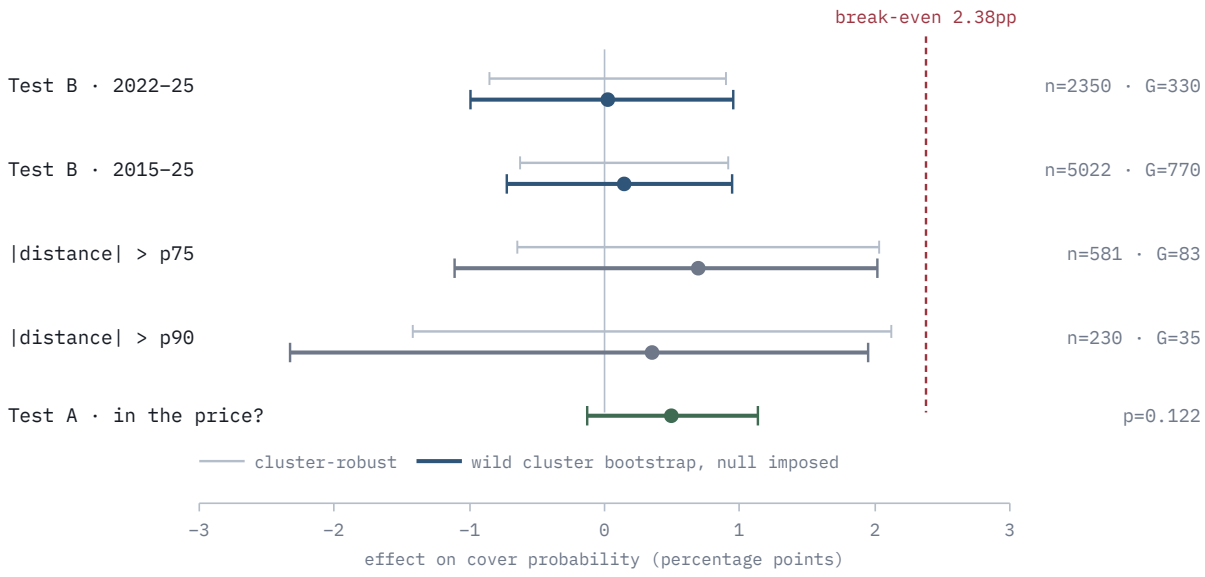
THE GATE

What survived, and what it returned

With no measurable quarterback attribute, the testable form is adaptation cost: a quarterback is now in a system some distance from the one he last played in, both sides measured as the same shrunk multi-season estimate of a coach's play-calling. Split-sample reliability of that estimate is 0.886, and returning quarterbacks under continuing coaches score a distance of exactly zero, as they must.

Every interval falls short of break-even

EFFECT ON COVER PROBABILITY, PERCENTAGE POINTS · 95% CI · CLUSTERED ON TEAM-SEASON, AND BY WILD CLUSTER BOOTSTRAP



Converted from points of margin at $0.4/\sigma_{game}$ with $\sigma_{game} = 15.51$, measured from the residuals rather than assumed. The tail specifications are thin — the top decile has 35 clusters — and the cluster-robust variance estimator is asymptotic in clusters, not observations. On this cluster structure it rejects a true null 9.5% of the time at a nominal 5%. Both estimators are therefore shown: the thin bar is cluster-robust, the heavy bar is a wild cluster bootstrap-t with the null imposed (Rademacher weights, 9,999 replications, interval by test inversion), which rejects at 5.9%. Correction widens every interval — 1.21× at the top decile — but asymmetrically, and to the left: the top-decile upper bound *falls* from 2.12pp to 1.95pp. The widest upper bound anywhere is now 2.02pp. Every one still falls short of the 2.38pp needed to clear vig at −110. Reproduced by `scripts/phase0_wildboot.py`.

It is not that the market got there first

Test A asks whether system distance is already in the price. With a thin first stage — prior-season SP+, recruiting, home field, $R^2 = 0.656$ — it looked significant at +0.365, $p = 0.010$. Adding returning production and a coaching-change indicator moved the first stage to $R^2 = 0.673$ and dropped the coefficient to **+0.192, p = 0.122**. The earlier result was omitted-variable bias.

So both sides return null. The market does not price system distance, *and* system distance does not predict what the market missed. That is a more interesting failure than being front-run: the interaction does not exist at measurable magnitude.

Secondary specifications, thesis window unless noted.

SPECIFICATION	COEFFICIENT	95% CI	P
Test B — (margin − market) ~ Φ	**+0.009**	**[−0.33, +0.35]**	**0.958**
Week-decay interaction Φ × early	−0.250	[−0.80, +0.30]	0.370
Era interaction Φ × collective (long)	−0.238	[−0.99, +0.51]	0.535
Coach quality (must be null)	−0.002	[−0.06, +0.06]	0.945
Recruiting composite (must be null)	−0.001	[−0.01, +0.01]	0.844

The pure-interaction test passes in the uninteresting direction: neither coach quality nor recruiting predicts the residual, which is what a market leaving no simple money on the table should look like. The thesis predicted a decaying edge concentrated in early weeks; the week interaction is null and, if anything, signed the wrong way.

INSTRUMENTATION

Three source defects that would have corrupted the study silently

Each of these is plausible-looking, populated, and wrong. None throws an error. Together they were worth more than any modelling decision in the project.

FIELD	WHAT IT LOOKS LIKE	WHAT IT ACTUALLY IS
`/coaches.hireDate`	98.5% populated, plausible dates	The coach's **first head-coaching hire ever**. Moorhead at Akron 2022 reads 2017-11-28 — his Mississippi State hire. Wrong for every coach on a second job, which is most of the treatment set.
`/coaches/tenures.startYear`	A tenure's first season	**FBS-scoped.** Cignetti's James Madison tenure reads 2022 against a 2018 hire, because JMU played FCS through 2021. Labels every reclassification coach as year-one.
`Coach-season rows`	One row per coach who worked	**Phantom 0-game rows** — an outgoing coach's end year runs a season long. Holgorsen appears in Houston 2024 with zero games, which a "sole coach of record" filter turns into the exclusion of Fritz, who coached all twelve.

Two structural facts also shaped what was testable. The enriched play endpoints carrying air yards and structured passer IDs exist for **2025 only** — advertised in the schema with no coverage caveat — so anything built on them has a one-season backtest. And no scramble play type exists anywhere in the 49-type vocabulary: a scramble and a designed quarterback run are the same record.

Quarterback attribution was therefore parsed from play text, at **99.963%** coverage over 259,446 scrimmage plays. The residual misses are the source's own `N/A` placeholders and kneel-downs.

BOUNDARIES OF THE CLAIM

What this does not establish

A weaker hypothesis than intended

This tested adaptation cost — distance between systems — not coach–quarterback fit. The stronger version was unbuildable because the quarterback attribute had no measurable signal. A charting source with time-to-throw and pressure could revive it, on a one-season backtest.

One axis, k = 1

Systems are represented by a single number: neutral early-down pass rate. It is the axis that survived a stability test, but a richer representation might find structure this cannot. The sample will not support estimating one.

The lines are not verified closes

CFBD carries no timestamp on a line. These are late prices of unknown vintage — adequate to ask whether a signal is in the price at all, not adequate for closing-line value, which needs a timestamped second source.

Test A is a weak test

Its first stage tops out near $R^2 = 0.67$ because prior-season ratings are a poor proxy for what the market knows. Test B carries the verdict precisely because it needs no first stage.

The portal boundary is sharp

Transfer rate among tape-qualified starters jumps from 5.3% to 29.3% at the one-time-transfer rule. Pre-2022 seasons are not a usable control period — the treatment barely existed.

Forty duplicated team-games

Found while making the tail specifications reproducible. Nine team-seasons carry two rows in the Φ panel, from two mechanisms: `is_primary_coach` ties when a firing splits a season evenly, and the prior-coach join fans out when the quarterback's previous team also had two coaches of record. Purdue 2017 and Georgia Southern 2018 survive the measurability filter twice, at two different distances, and drag their opponents' rows in with them — 40 of the 5,022 team-games, all in 2016–18, none in the thesis window. The reported estimates are unmoved, but the panel is not unique on (school, season) and downstream joins on that key must dedupe.

Absence at this magnitude only

An effect below roughly 0.95pp of cover probability is entirely consistent with these data. The claim is that no *tradeable* effect exists, not that the coefficient is exactly zero.

REPRODUCTION

Every API response is cached content-addressed on endpoint and parameters, so the analysis re-runs offline and the backfill is idempotent. Completed seasons cache permanently; the live season carries a twelve-hour TTL. A budget guard refuses calls past 25,000 against the tier's 30,000-per-month ceiling.

```
python scripts/backfill.py --dry-run   # prices the plan before spending
python scripts/build_panel.py          # coach panel, tenure-dated
python scripts/test_systems_travel.py  # test 01
python scripts/test_qp_stability.py    # test 03
python scripts/stage3_gate_game.py     # the gate
python scripts/phase0_wildboot.py      # bootstrap correction, tails
python scripts/phase0_contrast_arith.py
```

Walk-forward discipline throughout: coaching hires, portal moves and recruiting classes filtered by announcement date rather than season label. No pooled cross-validation across seasons. Effect sizes and the minimum detectable effect were fixed before the gate was run.