Interesting Results from NFL Managers

Bayesian Decision Analysis of NFL Fourth Down and Two-Point Decisions

Andrew Bai University of Chicago Booth

---

1 The Question

Two decades after Romer (2006) showed NFL teams are systematically too conservative on fourth down, **are coaches learning?**

- Are coaches making better decisions over time?
- Were optimal decisions *knowable in real-time*, or only with hindsight?
- How many games are teams losing from suboptimal decisions?
- We review other intersting behavioral quirks of NFL management.

2 Data and Win Probability Model

Data: nflfastR play-by-play data, 2006–2024 (71,786 fourth down plays).

Win probability model: I use nflfastR's `vegas_wp`—a pre-computed win probability calibrated to Vegas betting lines. This accounts for:

- Score differential, time remaining, field position, down and distance
- Pre-game point spread (market expectation of team quality)
- Home field advantage, timeouts remaining

The data otherwise gives me:

- Player Stats
- Everything you would want per play like team fixed effects, kicker fixed effects, punter fixed effects.

3 The Framework

State Space

The game state is: $s = (\Delta, \tau, x, d)$ where Δ = score differential, τ = time remaining, x = yards from opponent's end zone, d = yards to go.

Action Space

On fourth down: $\mathcal{A} = \{\texttt{go}, \texttt{punt}, \texttt{fg}\}$

The Decision Rule

Choose the action that maximizes expected win probability:

$$a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[W \mid a, s]$$

For each action, I compute expected WP by:

1. Estimating transition probabilities (conversion rate, punt distance, FG make rate)
2. Looking up win probability at successor states
3. Computing expected WP = $\sum_{s'} W(s') \cdot P(s' \mid s, a)$

Bayes

The transition probabilities are uncertain. The Bayesian expected WP integrates over this uncertainty:

$$\mathbb{E}[W \mid a, s] = \int W(s' \mid a, s, \theta) \cdot p(\theta \mid \mathcal{D})\, d\theta$$

Other Parts of the Model

Conversion model: Hierarchical logistic regression. Yards to go is the main predictor ($\beta_d \approx -0.13$ per yard). Includes team offensive and defensive effects—but team variance is modest ($\tau \approx 0.15$), so shrinkage is substantial.

Field goal model: Logistic regression with kicker random effects. Make probability drops ∼10% per 10 yards. Kicker variance ($\tau \approx 0.03$) is small—kickers are more homogeneous than you'd think.

Punt model: Linear regression with punter random effects. Expected net yards $\approx 33 + 0.15\times$ field_position. Punter variance is modest ($\tau \approx 1.4$ yards).

Team/player effects exist but are small relative to situation effects. The decision depends mostly on field position, yards to go, score, and time, not "we have a great kicker."2 Point conversion outcomes are essentially also coin flips. There is only a few percentage point difference in conversion between the best offense facing the worst defense on a 2 point conversion and vice versa.

4 Results

Key Result

Coaches remain systematically too conservative on fourth down:

- Optimal GO rate: **35.3%**—but coaches only go 14.8%
- 4th & 1 optimal GO rate: **60.0%**—coaches go only ∼25%
- Match rate: **62.8%**—coaches make the wrong call over a third of the time
- Teams lose **1.1 expected wins/season** from fourth down errors

But: 86% of "mistakes" are close calls (<2% WP margin). Only 1.2% are clear errors.

5 Behavioral Findings

Where are coaches most conservative?

Score Situation	Actual GO	Optimal GO	Gap
Up 14+	13.3%	51.0%	**+37.7pp**
Up 1–3	8.8%	35.0%	**+26.7pp**
Losing big (< −14)	39.3%	63.8%	+24.5pp
Down 1–3	12.4%	23.4%	+11.0pp
Tied	8.0%	18.1%	+10.1pp

6 Learning Over Time: The Puzzle

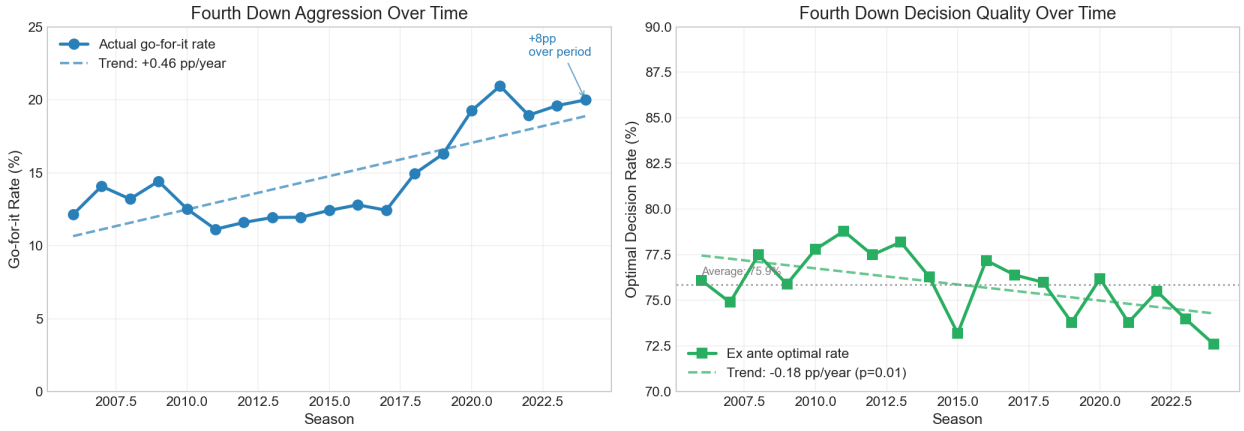

Figure 1: The central puzzle: Fourth-down aggression increased dramatically (+8pp over the period), but decision *quality* stayed flat. Coaches changed behavior but not accuracy.

Fourth down:

- GO rate increased: 12.1% (2006) → 20.0% (2024)
- Optimal is 35.3%—coaches are still way too conservative

7 Timeline: When Did Learning Happen?

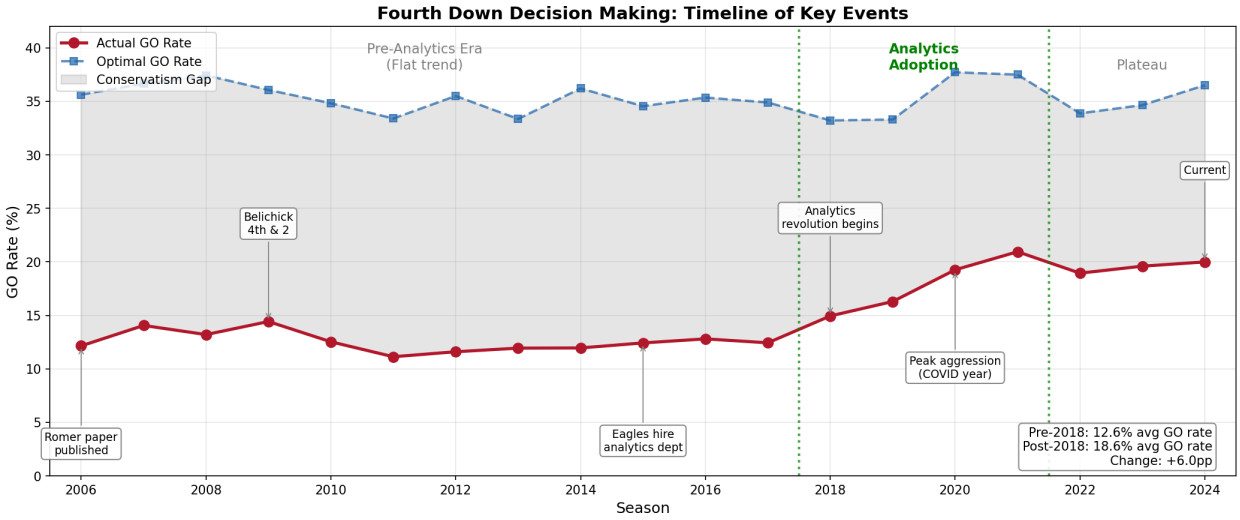

Figure 2: Fourth down decision timeline. The 2018–2021 period shows a clear regime shift (+5.3pp), followed by a plateau. Key events marked: Romer (2006), Belichick 4th & 2 (2009), Eagles analytics hire (2015), analytics revolution (2018).

The "analytics revolution" coincides with: (1) Eagles win Super Bowl LII (Feb 2018) with analytics-heavy approach under Doug Pederson, including the famous "Philly Special" 4th down trick play; (2) Teams begin hiring dedicated analytics staff en masse; (3) Public tools like nflfastR make win probability models accessible.

The Belichick 4th & 2 (2009): Patriots up 34–28, 2:08 left, 4th & 2 at *own 28*. Conventional wisdom: always punt—don't give Peyton Manning short field. Analytics: going for it maximizes win probability (even with Manning, the conversion probability $\times$ win prob if convert $>$ punt win prob. My model confirms Belichick's decision). Belichick went for it, failed, lost, highly controversial and put analytics in the news.

8 Two-Point Conversions: Coaches *Are* Learning

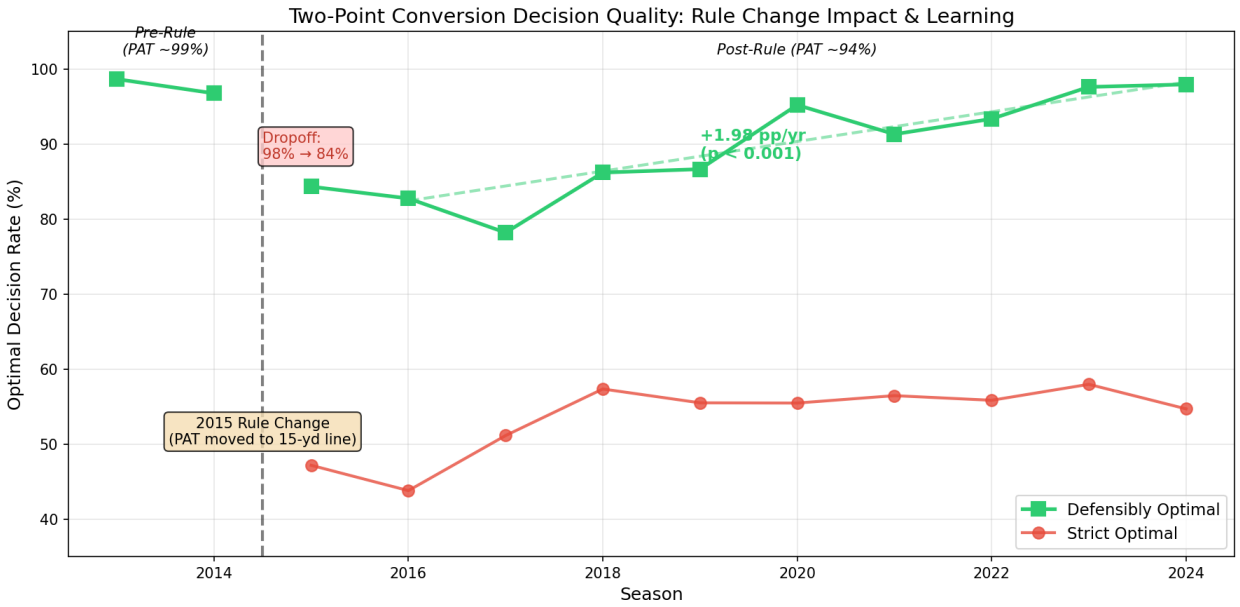

Figure 3: Two-point conversion learning curve around the 2015 rule change. Pre-rule, coaches were ~98% defensibly optimal. The rule change caused a drop to 85%. Coaches then recovered to 99% by 2024.

Two-point conversions:

- Match rate improving: **+1.03 pp/year** (p = 0.06)
- 2016: 61.6% match → 2024: 87.0% match
- Clear learning curve after 2015 rule change

Why do coaches learn here but not on fourth down? Two-point decisions are *simpler*:

- 2 actions (not 3)
- No field position dependence
- Lower stakes (max 5% WP swing vs >30% for fourth down)
- Clearer feedback signal

9 The Down 2 vs Down 3 Paradox

A behavioral puzzle in two-point decisions. After scoring a TD, the team is now down by 2 or 3 points:

Situation	Optimal 2pt Rate	Actual 2pt Rate
Down 2	85%	**79%**
Down 3	91%	**1%**

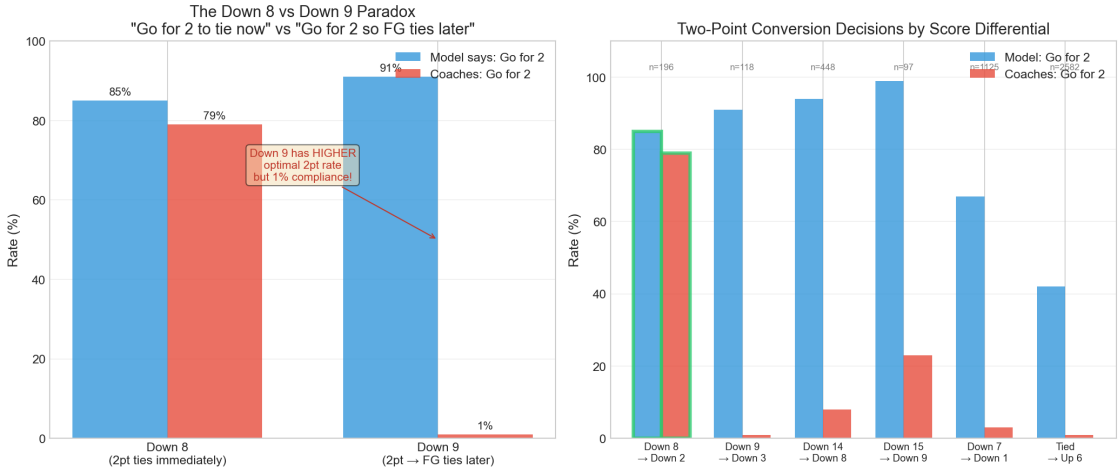

Figure 4: The Down 2 vs Down 3 paradox. When down 2, going for 2 ties immediately—coaches do this 79% of the time. When down 3, going for 2 means a field goal can tie later—coaches do this 1% of the time. But Down 3 has a *higher* optimal 2pt rate (91% vs 85%).

Down 3 has a *higher* optimal 2pt rate, but coaches almost never do it.

Robustness: Is this selection bias? Maybe certain teams systematically end up down 2 vs down 3? No—chi-square test shows team is *not* correlated with deficit type ($p = 0.94$, Cramer's V = 0.13). Being down 2 vs down 3 after a TD is essentially random with respect to team identity.

10 Games Lost

Team	**Avg Wins Lost/Season (2020–24)**
DEN (worst)	1.39
CAR, WAS	1.30
BUF, DET	0.84–0.85
GB (best)	0.76
League average	1.07

Romer (2006) estimated ∼0.4 wins/season.

11 Behavioral Anomaly: The 4th & 10 Effect

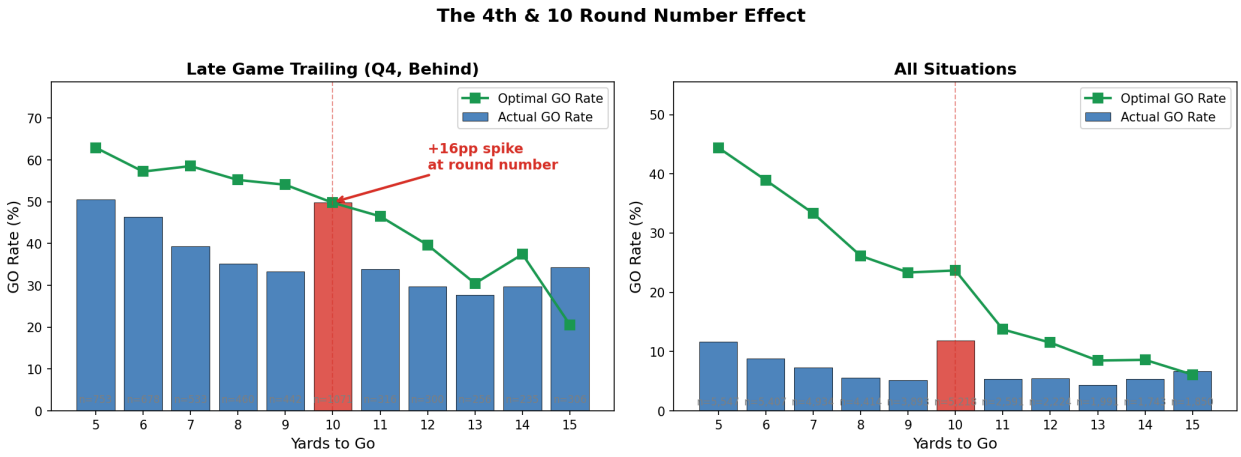

Figure 5: The 4th & 10 effect

Distance	GO Rate	Optimal GO	Gap	Spike
Late Game Trailing (Q4, Behind)				
4th & 8	35.2%	55.2%	+20.0pp	–
4th & 9	33.3%	54.1%	+20.8pp	–
4th & 10	**49.9%**	**49.8%**	**−0.1pp**	**+16.3pp**
4th & 11	33.9%	46.5%	+12.6pp	–
4th & 12	29.7%	39.7%	+10.0pp	–

Is this selection bias? 4th & 10 situations arise from a specific game mechanic: three failed plays from 1st & 10. These drives may systematically occur in more desperate situations. Indeed, 4th & 10 has 51.1% trailing situations vs ∼43–44% for neighbors, and 20.5% Q4+trailing vs ∼11–13%.

But the effect persists after controlling:

Distance	GO Rate	Optimal GO	N
Controlled: Q4, trailing, field position 40–70			
4th & 8	43.3%	60.4%	164
4th & 9	42.2%	52.8%	161
4th & 10	**56.7%**	**56.0%**	432
4th & 11	32.3%	51.1%	133
4th & 12	38.0%	45.5%	121

At 4th & 10, coaches are *near-optimal* (56.7% vs 56.0%). At adjacent distances, they remain too conservative.

12 Behavioral Anomaly: The 50-Yard Line Mental Anchor

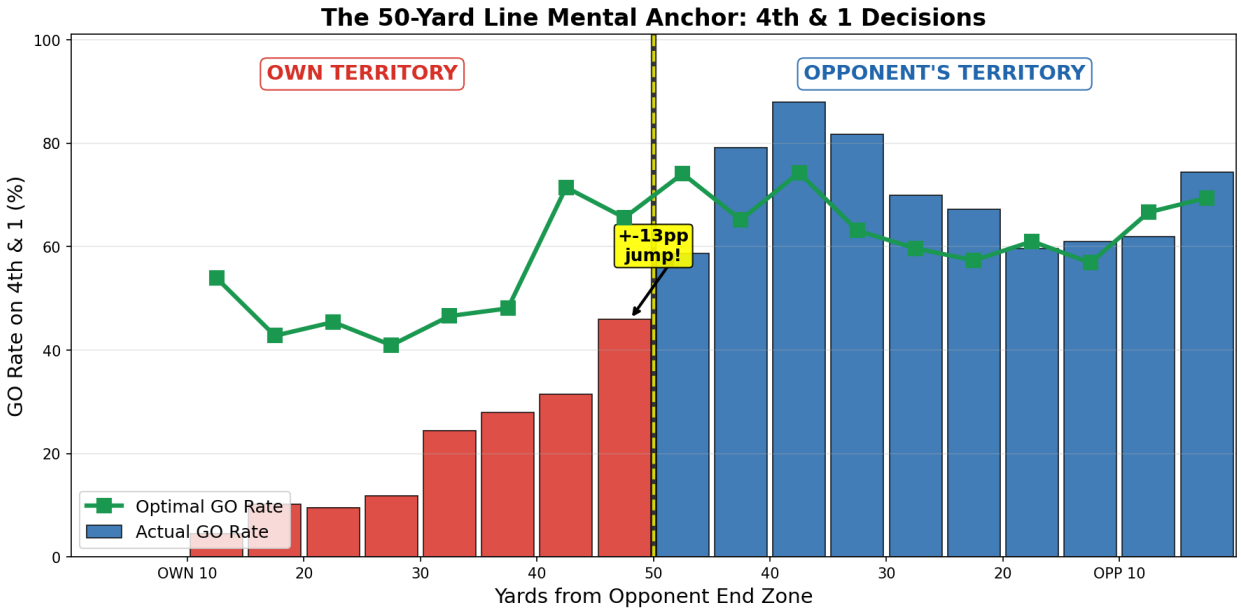

Figure 6: The 50-yard line mental anchor on 4th & 1 decisions.

4th & 1 by territory:

Territory	Actual GO	Optimal GO	Gap
Own territory (51–99)	24.6%	52.3%	+27.7pp
Opponent territory (1–50)	71.5%	65.8%	−5.7pp
Difference	**46.9pp**	**13.5pp**	**33.4pp**