Referee Report — Round 7

Tennis Match Simulator

Referee 2

2026-02-14

Verdict: Accept with Minor Revisions

The project is methodologically sound. Data integrity is confirmed, diagnostics are honest, and 17 analysis scripts cover extensive ground.

Key findings this round:

- **K-factor tested**: K=64 gives +0.4pp accuracy but worse calibration. Trajectory lag not solved by K alone.
- **Calibration curve**: Systematic overconfidence on favorites, underconfidence on underdogs. This is the most exploitable signal.
- **"Agree but less confident"**: +6.4% ROI (N=256, p=0.041) — promising but needs multi-year validation.
- **Multiple testing**: 50+ subsets tested without correction. The p=0.041 signal is not distinguishable from noise in isolation.

Core recommendation: Shift from broad scanning to **depth on one signal**. Implement Platt scaling, validate across years, paper-trade prospectively.

Elo's Calibration Bias Is the Most Exploitable Signal

Range	Predicted	Actual	Error
0–30%	20.3%	26.8%	+6.5pp
40–60%	50.1%	50.1%	±0.5pp
70–100%	76.9%	70.7%	−6.2pp

Pattern: Elo overestimates the gap between favorites and underdogs by 6+ percentage points at the extremes.

Why this matters for betting:

- If Elo says 80%, reality is ∼72%
- If Elo says 25%, reality is ∼32%
- Underdogs are *worth more* than Elo thinks
- Favorites are *worth less* than Elo thinks

Fix: Platt scaling

```
cal = plogis(a + b*qlogis(elo))
```

5 lines of R code. Not yet implemented.

K-Factor Sensitivity: Accuracy vs Calibration Tradeoff

K	Accuracy	Brier
24	63.7%	**0.2162**
32	63.8%	0.2164
48	63.8%	0.2179
64	**64.2%**	0.2203

Higher K $\rightarrow$ more decisive (correct direction more often) but less calibrated (worse probability estimates).

Insight: K-factor alone does not solve trajectory lag.

K=64 gains only +0.4pp accuracy while losing 0.004 Brier. The inertia from accumulated rating capital requires **structural** changes, not parameter tuning:

- Scoreline-weighted updates (Angelini WElo)
- Time-decay on older matches
- Dual short/long-term Elo

The author correctly concluded this. The referee concurs.

Best Signal: "Elo Agrees but Is Less Confident"

Metric	Value
Filter	Agree, Elo 0–5pp less confident
N	256
Win rate	72.3%
Breakeven	67.0%
ROI	**+6.4%**
p-value	0.041
Best sub-segment	Odds 1.7–2.0
Sub-segment ROI	+16.6%
Sub-segment N	49

Logic: When Elo agrees with the market's pick but assigns a lower probability, the market may be overpricing that favorite.

This connects to calibration:
Elo overestimates favorites → when Elo says "favorite, but less confident," it may be reflecting reality more accurately than the market.

Caution:

- p=0.041 on one period
- 50+ subsets tested (no correction)
- N=49 sub-segment is too small
- **Needs multi-year validation**

Edge Signals: Tiered Assessment

Tier 1: Structurally Sound, Needs Validation

Signal	In-Sample	Action Needed
Agree + less confident	+6.4% ROI (N=256)	Multi-year validation
Calibration correction	Implied by cal. table	Implement Platt scaling

Tier 2: Interesting but Fragile

Signal	In-Sample	Issue
Height in R vs R	+1.75% (N=2,390)	Uses ex-post winner
Clay Masters	+13.5% (N=53)	N too small
Heavy favorites	+2.0%	Vig kills margin

Tier 3: Dead Ends (Correctly Identified)

Signal	Conclusion
Fatigue/scheduling	Market prices efficiently
L vs R matchups	Consistent negative Elo ROI
Elo disagreements	Anti-informative (34.7%)

50+ Subset Tests Require Multiple Testing Correction

Subsets tested (partial list):

- 4 K-factor values
- 4+ edge thresholds
- Agreement/disagreement
- 5+ confidence buckets
- 4 tournament levels, 3 surfaces
- 4+ odds ranges
- Handedness, height, fatigue
- Match load, Clay Masters
- Various combinations

At $\alpha = 0.05$ with 50 tests:

- Expected false positives: $\sim$2.5
- Bonferroni threshold: 0.001
- Best p-value found: 0.041
- Does not survive correction

This does not mean the signal is false.

It means single-period evidence is insufficient. The correct response is **multi-year replication**, not tighter p-values from the same data.

Literature Identifies High-Value Upgrades Not Yet Attempted

Paper / Method	What It Does	Why It Matters
Platt scaling	Corrects calibration curve via logistic transform	Directly exploits the 6pp calibration bias. 5 lines of R code.
Angelini WElo	Weights Elo updates by scoreline margin	6-0 6-1 updates more than 7-6 7-6. Addresses trajectory lag.
Ingram Bayesian	Latent serve/return skills with time evolution, surface effects	Direct upgrade for MC engine. Separates serve from return skill.
Gorgi et al.	State-space model for dynamic surface-specific abilities	Proper statistical framework for the dual Elo concept.
Prieto-Lage	Point-win probability varies by rally length and surface	Moves MC engine beyond iid points at low implementation cost.

Priority: Platt scaling (immediate), then WElo scoreline weights (medium-term).

Recommended Path: Depth Over Breadth

The edge-hunting phase has been **broad but shallow**: many dimensions tested, none exhausted.

Phase 1: Calibration Correction (Immediate)

1. Implement Platt scaling: `cal = plogis(a + b * qlogis(elo))`
2. Cross-validate on 2021–2023, test on H1 2024
3. Recalculate all betting signals with calibrated probabilities

Phase 2: Signal Validation (1–2 sessions)

1. Multi-year test of "agree but less confident" across 2021–2024
2. Apply Holm-Bonferroni correction, report adjusted p-values
3. Document results (positive or negative) in `model_analysis.md`

Phase 3: Prospective Test (Definitive)

1. Freeze all parameters at end of H1 2024
2. Paper-trade H2 2024 with zero retrospective adjustment
3. This is the single most important test for claiming an edge

17 Analysis Scripts Demonstrate Thoroughness

Model Variants Tested:

- Standard Elo (K=32)
- K-factor sensitivity (24, 32, 48, 64)
- Dual Elo (long/short-term)
- WElo (tournament-weighted K)
- Elo + H2H, Elo + MC ensemble

Feature Dimensions Explored:

- Surface, tournament level, round
- Handedness, height, fatigue
- Season, confidence, odds ranges

Validation: In-sample, out-of-sample (H2 2024), multi-year (2020–2024)

What's documented vs not:

Analysis	Done	In docs
K-factor	✓	✓
Calibration	✓	✓
Disagreements	✓	✓
Height/fatigue	✓	✓
H2 validation	✓	×
Dual Elo	✓	×
H2H hybrid	✓	×

3 analyses completed but not documented. Results should be reported regardless of sign.

Prioritized Recommendations

1. **Implement Platt scaling** on calibration curve. Fit on 2021–2023, test on 2024. This is the lowest-hanging fruit in the entire project.
2. **Document all out-of-sample results.** The H2 2024 validation scripts exist but results are missing from documentation. Report whether signals held or not.
3. **Multi-year validation** of "agree but less confident" signal across H1 2021–2024. Report consistency across years.
4. **Apply multiple testing correction.** Report total number of subset tests alongside any p-values.
5. **Implement scoreline-weighted Elo updates** per Angelini et al. This addresses trajectory lag structurally.
6. **Define one prospective betting rule**, freeze parameters, and paper-trade H2 2024 with zero adjustments.
7. **Consider Ingram-style upgrade for MC model.** The engine is sound; the parameter estimation is weak.