Do Players Choke in Big Matches?

Evidence from 26 Years of Professional Tennis

Aryamaan Roy Choudhury

Independent Research · Matchup DNA

16 November 2026 · Version 4.0 · Working paper, not peer reviewed

ABSTRACT

Matchup DNA forecasts professional tennis matches with a model built on Elo ratings. I ask whether a player's record in big matches (semi-finals, finals, the late rounds of Grand Slams and of the top non-Slam tier, and the season-ending Finals) contains information the ratings miss. Using 130,460 ATP and WTA main-tour matches from 2000 to May 2026, 14,086 of them big matches, I compare each result with a calibrated Elo forecast. Across players, big-match results vary about as much as binomial chance predicts (observed-to-chance variance ratios of 0.93 to 1.16), the two halves of a career's big matches agree only weakly (split-half r = 0.12 ATP, 0.04 WTA), and none of 18 trait statistics is significant after Holm correction. In a chronological out-of-sample evaluation on 7,476 big matches from 2012 to 2026, adding a player's past big-match or clutch score to the calibrated baseline changed log loss by +0.00003 and +0.00007, with 95% intervals that include zero. The data do not exclude a modest trait: under the simulation model, a trait whose standard deviation is worth up to about 7.1 (ATP) or 4.8 (WTA) percentage points at even odds remains consistent with the results. One tour-level pattern does hold up: Elo is more overconfident in WTA big matches than in regular WTA matches (calibration slope 0.76 against 0.82), the gap is largest in 2018–2026, and calibrating WTA big matches separately improved out-of-sample forecasts slightly. The evidence does not support adding a player-specific big-match adjustment to Matchup DNA's forecasting model.

Keywords: tennis, forecasting, Elo, calibration, shrinkage, out-of-sample evaluation, statistical power

1 Introduction

Matchup DNA produces probabilistic forecasts for ATP and WTA matches. Its ratings treat a final like any other match between the same two players, which raises a practical question: does a player's history in big matches contain information that the ratings do not? If some players consistently fall short in finals and others consistently rise to them, a forecast should adjust for it. If big-match records are mostly chance, an adjustment would only add noise.

The idea that some players "choke" is common in tennis commentary, but it is hard to test from match results. Big matches are rare. A leading player reaches a handful of finals a season, and a record of twenty finals can look unusually good or bad by chance alone. A useful test therefore has to separate differences between players from binomial noise, and it has to say how large an effect it could have missed.

I use 130,460 tour-level matches from 2000 to May 2026 to ask four questions. Do players' big-match results depart from expectation by more than chance? Is a player's record consistent across the two halves of their big matches, and from 2012–2018 to 2019–2026? Would a past big-match record have improved forecasts of later big matches, using only information available at the time? And how large a trait could these tests have missed? A fifth question concerns favourites in general rather than particular players: is Elo calibrated differently in big matches?

In short, big-match records look mostly like chance, past records do not improve out-of-sample forecasts, and a modest trait cannot be ruled out. The one effect that holds up is at the tour level: in WTA big matches, Elo overstates the favourite's chances by more than it does in regular WTA matches.

This is a post-hoc study. Version 1 fixed its definitions in code before it was run, but the question came up after recent results had been seen, and later versions added tests after earlier results were known. Version 3.0 corrected several errors found in an audit of versions 1 and 2; these and later changes are listed in the version history at the end.

2 Related Research

Most quantitative work on pressure in tennis uses point-level data. For modelling purposes, treating points as i.i.d. is usually adequate: Klaassen and Magnus (2001) detected departures in Wimbledon data, with servers slightly less successful on important points and some carry-over from one point to the next, but the departures were small. González-Díaz et al. (2012) found persistent differences between players in how they perform on important points at the US Open, and that these differences are associated with career success. Cohen-Zada et al. (2017) compare men and women directly: the decline under rising stakes within a match is consistent only for men, women's results are inconsistent, and where women's performance does fall, the fall is about half as large. Paserman (2023) finds that both men and women play less aggressively on important points, and Harris et al. (2021) find that double faults and unforced errors become more likely as situational pressure rises.

Two studies measure pressure performance for individual players. Kovalchik and Ingram (2016) modelled player-specific responses to the score and to pressure situations in about three million ATP and WTA points from 2011–2015; allowing for these responses improved predictions of the 2015 men's Grand Slam matches, while for women the player-specific model did no better than one with average effects. Kovalchik and Reid (2020) weighted players' serve and return statistics by the importance of each point; in 2011 Grand Slam matches these "clutch averages" generally separated winners from losers, and predicted the winner from first-set statistics, better than unweighted averages.

These studies ask whether pressure changes how points are played. This paper asks a coarser question that matters for match forecasts: whether such differences add up to persistent differences in big-match results. Results also depend on form. Jetter and Walker (2015) find that recent wins predict the next match, and that top-ranked players do better at Grand Slams than elsewhere, a clutch effect for the group of top players rather than for individuals. Players reach big matches partly by being in form, so a test of pressure has to separate the two. Gorgi et al. (2019) allowed the strength gap between ATP players to widen at Grand Slams and found no evidence that it does once the best-of-five format is modelled, a tour-level question that Section 4.1 asks of Elo's calibration on both tours.

The question has also been examined with match results and Elo expectations outside the academic literature. Sackmann (2019a) compared ATP players' results in each round, from first matches to finals, with Elo's predictions and found significant departures in 29 of about 1,043 player-and-round categories, no more than chance would produce. Sackmann (2025) compared men's best-of-five Grand Slam results with Elo expectations and found virtually no correlation between the first and second halves of players' records. These are the closest precedents for this paper's question and approach.

The statistical framework builds on established approaches from research on clutch hitting in baseball. Cramer (1977) found that players' clutch hitting in 1969 did not predict their clutch hitting in 1970. James (2004) argued that a persistence test that finds nothing says little unless it could have detected a real effect, and Birnbaum (2005) estimated the power of Cramer's test by simulating hitters with clutch talent of known size. Comparing the observed spread of players' records with the spread chance would produce, and treating any excess as true differences, is a standard approach in that literature (Tango et al., 2007). Efron and Morris (1975) showed that shrinking early-season batting averages toward the group mean predicted the rest of the season with less than half the mean squared error of the raw averages, and Brown (2008) compared empirical Bayes methods for predicting batting averages within a season. The variance ratios, consistency tests, power analysis and shrinkage used here are related to these approaches.

The expectation for each match comes from Elo ratings (Elo, 1978), which Kovalchik (2016) found to be among the more accurate published methods for predicting ATP results. Laboratory studies of choking, in which pressure such as competition, money or an audience worsened skilled performance (Baumeister, 1984), motivate the question but measure something different: short tasks, not results over a professional career.

Relative to this work, the paper's contribution is applied. It brings calibrated Elo residuals, variance and persistence tests, simulation-based null distributions with multiple-testing correction, and simulated power and bounds together in one analysis of 130,460 ATP and WTA matches; it evaluates past big-match and clutch scores chronologically out of sample on 7,476 big matches; it compares Elo's calibration in big and regular matches on both tours; and it uses the results to decide whether Matchup DNA's forecasting model should include a big-match adjustment.

3 Data and Methods

3.1 Data

Match results come from Jeff Sackmann's tennis_atp and tennis_wta files (Sackmann, 2026), read from a June 2026 snapshot. Elo ratings are computed from 1968, and the analysis covers tournaments that started between 1 January 2000 and 24 May 2026. The last edition in the snapshot, Roland Garros 2026, started on 25 May and finished in June. Matches from June 2026 onwards are reserved for the project's prospective evaluations, so the whole edition is left out.

Only main-tour singles events are used: the Grand Slams, the ATP Masters and the WTA's top non-Slam tier, other tour-level events, the season-ending Finals and the Olympics. Team competitions (Davis Cup, Billie Jean King Cup, ATP Cup, United Cup, World Team Cup), the Laver Cup, the Next Gen ATP Finals (an under-21 event with shortened sets), WTA 125 events and bronze-medal matches are excluded. Only completed best-of-three and best-of-five matches are rated, so retirements and walkovers are excluded too. Table 1 summarises the sample. Each match is one observation; in the player-level statistics a match contributes one residual to each player, with opposite signs.

Table 1. The sample, 2000 – 24 May 2026. Player counts are for players meeting each test's minimum of 10 big matches.

	ATP	WTA	Both
Matches	68,393	62,067	130,460
Big matches	7,367	6,719	14,086
Big matches, best of five	1,592	0	1,592
Players with 10+ big matches	242	249	491
… with 10+ in each half of their record	148	133	281
… with 10+ in both 2012–18 and 2019–26	26	32	58
Big matches in the out-of-sample test, 2012–2026	3,918	3,558	7,476

3.2 Big matches

A big match is any semi-final or final; a Grand Slam fourth-round or quarter-final match; a quarter-final at the highest tier below the Grand Slams; or any match at the season-ending Finals. Every other main-tour match is a regular match. For the ATP the top non-Slam tier is the Masters. For the WTA, Sackmann's level codes are not consistent across years (Premier Mandatory events are coded as Premier in 2016 and 2020–2022, and from 2023 the Premier Mandatory code covers the whole WTA 1000 tier), so the tier is assigned by event and year: Tier I up to 2008, Premier Mandatory and Premier 5 from 2009 to 2020, and WTA 1000 from 2021. The season-ending Finals are identified by name, because the same code also marks the second-tier WTA Tournament of Champions and Elite Trophy, which are treated as ordinary events. Appendix B.1 gives the rules. I also report results for finals alone and for semi-finals and finals together, which do not depend on the tier assignment.

3.3 Expected results

Player strength is estimated using the project's surface-blended Elo rating, which combines a player's overall rating and their rating on the match surface with equal weights. In the project's model development, the equal weight gave the lowest log loss on 1990–2009 matches among weights of 0.25, 0.5, 0.75 and 1. A similar blending approach has previously been described by Sackmann (2019b). Ratings are taken at the start of each tournament week, so no match is predicted with ratings that include its own result or any other result from that week. Raw Elo probabilities are too confident in best-of-three matches and roughly right in best-of-five, and most best-of-five matches are Grand Slam matches, many of them big. I therefore rescale Elo's log-odds with one slope b_g per group g = (tour, big or regular, best of three or five), fitted by maximum likelihood (equation 1). The residual of player i in match m is the result minus this expectation (equation 2). In the descriptive tests the slopes are fitted on the analysis window itself, so the expectation is in-sample. The out-of-sample evaluation (Section 3.6) fits them only on earlier years.

$$P_{ij} = \sigma\left(b_g \cdot \operatorname{logit} p_{ij}^{Elo}\right) \tag{1}$$

$$e_{im} = y_{im} - P_{im}, \quad y_{im} \in \{0, 1\} \tag{2}$$

3.4 Player scores

A player's big-match score divides the sum of their big-match residuals by the number of big matches plus 20 (equation 3). This is fixed shrinkage toward zero: a player with 20 big matches is weighted halfway between their own record and zero. The constant was set before the version 1 run as a round number; it was not estimated from the data, so the score resembles an empirical Bayes estimate (Efron and Morris, 1975) without being one. Section 4.6 reports constants of 10, 50 and 100, and the out-of-sample evaluation adds an empirical Bayes score whose shrinkage is estimated from earlier data. The clutch score subtracts the same score computed on regular matches (equation 4), so a player Elo under-rates in every match, such as a fast-improving young player, is not counted as clutch.

$$s_i = \Sigma_{m \in big}\, e_{im} / \left(n_i + 20\right) \tag{3}$$

$$c_i = s_i^{big} - s_i^{regular} \tag{4}$$

3.5 Tests for a player trait

Variation beyond chance. For players with at least 10 matches of a kind, I compare the variance of their mean residuals with the average variance that binomial chance would produce given their match probabilities. A variance ratio of 1 means players differ by exactly as much as chance predicts; any excess estimates the variance of a persistent trait, as in the decomposition of observed into true and random

variance used in baseball research (Tango et al., 2007). I compute it for finals, for semi-finals and finals, for all big matches, for regular matches as a reference for Elo's general misfit, and for the difference between each player's big- and regular-match mean residuals (the clutch spread).

Consistency. The split-half test assigns every edition of a tournament to one of two halves (by a hash of its identifier) and correlates the two halves' mean big-match residuals across players with at least 10 big matches in each. The carry-over test correlates scores from 2012–2018 with scores from 2019 – May 2026, refitting the calibration in each period.

Retrospective association. Versions 1 and 2 called this a forecast test. It adds $\beta(s_A - s_B)$ to the calibrated 2019–2026 forecasts (equation 5), fits β on those same matches, and asks whether log loss improves more than when the earlier scores are shuffled among players (1,000 permutations). Because β and the calibration are fitted on the matches being scored, this tests association between the periods; it does not evaluate forecasts.

$$\operatorname{logit} P_m = \operatorname{logit} \hat{P}_m + \beta\,(s_A - s_B) \tag{5}$$

3.6 Out-of-sample evaluation

To test forecasting directly, every big match from 2012 to May 2026 is predicted using only matches from earlier calendar years. For each year Y, the calibration slopes are fitted on all matches from 2000 to the end of $Y-1$, player scores are computed from residuals of those matches, and β is fitted on the big matches from 2004 to $Y-1$, each with the scores and calibration that were available at the start of its own year. A player with no earlier big matches has a score of zero, so no match is dropped. Evaluation starts in 2012 so that every β is fitted on at least eight years of big matches.

All models are scored on the same 7,476 matches. M0 is Elo with one slope per tour and format; M1 has separate big- and regular-match slopes and is the baseline used in the rest of the paper; the player models add the big-match or clutch score to M1. Two variants test the shrinkage and the assumption that a trait lasts a career: empirical Bayes scores, in which each player's mean residual is multiplied by $\tau^2/(\tau^2 +$ its noise variance) with τ^2 estimated from earlier players, and scores built from the previous 5 years only. Differences in log loss and Brier score are paired by match, and their 95% intervals come from 2,000 bootstrap resamples (Efron, 1979) of whole tournament editions.

3.7 Inference

Null distributions. The variance ratios and correlations do not have convenient sampling distributions, so their p-values come from a parametric bootstrap: 2,000 data sets in which every match result is redrawn from its calibrated probability, the calibration is refitted, and every statistic is recomputed. The simulations keep the observed schedule (the same players meet in the same matches), so the null is conditional on who reached each round. They also assume no player differences of any kind beyond Elo, including ordinary form, so a statistic that picks up form can look significant without indicating pressure.

Multiple testing. Nine statistics per tour test for a trait: three big-match variance ratios, the clutch spread, the split-half correlation, two carry-over correlations and two association statistics. I treat these 18 as one family and apply Holm's correction (Holm, 1979). This family was defined in version 3, after the earlier results were known; within-tour adjustments are in Appendix Table A1. Version 2 corrected only the variance ratios. The robustness checks in Section 4.7 re-test the same hypotheses and are not corrected separately.

Power and bounds. Similar simulations were used by Birnbaum (2005) to measure the power of persistence tests for clutch hitting. I add a known trait to the simulations: each player receives $\theta \sim N(0, \tau^2)$ on the log-odds scale in big matches only, constant over the career (equation 6). For each τ from 0.05 to 0.4, 400 simulated data sets give each test's power and the probability of a statistic at or below the observed one. The 95% upper bound on τ is the smallest τ at which that probability falls below 5%, found by linear interpolation between grid points, in the spirit of an equivalence test (Lakens, 2017). I express τ as an edge

at even odds: a player one standard deviation above average in the trait, facing an otherwise equal and average opponent, wins 50 + 100(σ(τ) − 0.5) per cent of big matches. Resampling the simulations gives a Monte Carlo range for each bound.

Calibration. Model-based standard errors of the calibration slopes treat matches as independent. I also report a cluster bootstrap over tournament editions (2,000 resamples).

$$\text{logit } P_m^{\text{true}} = \text{logit } \hat{P}_m + 1_{big}(m) \cdot (\theta_A - \theta_B) \tag{6}$$

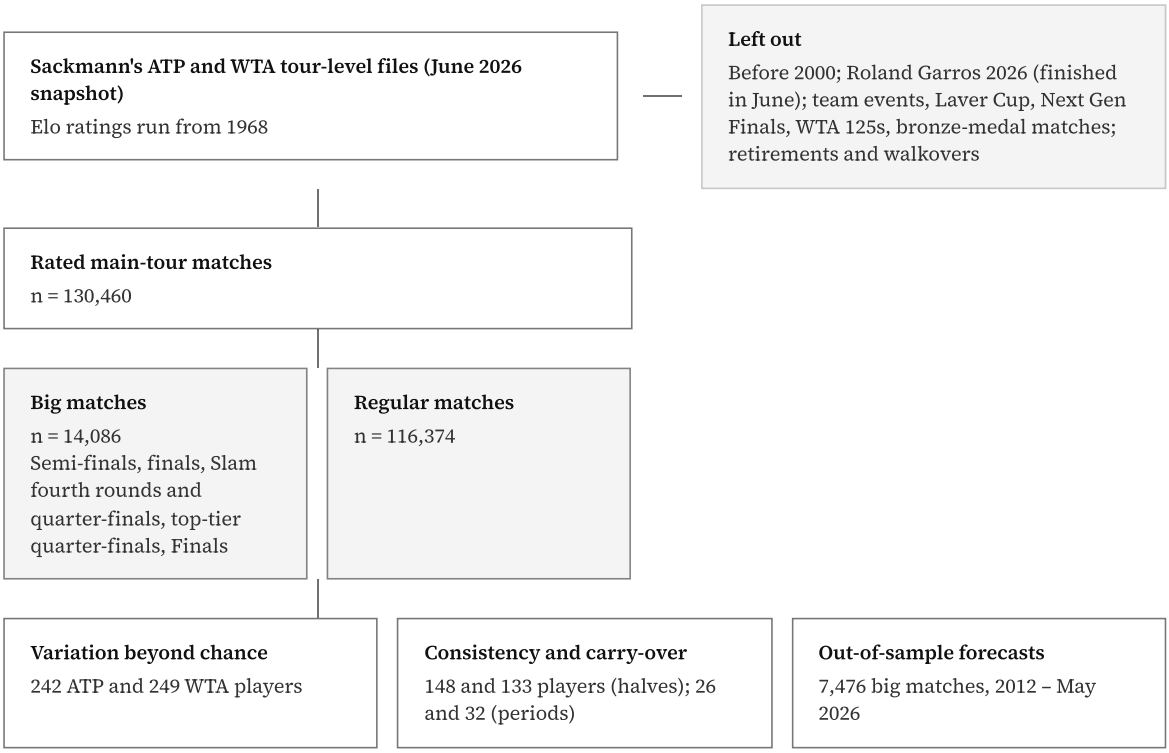


Figure 1. Matches and players in each part of the analysis, 2000 – 24 May 2026.

4 Results

4.1 Calibration in big matches

On the ATP tour, Elo needs the same correction in big and regular matches in both formats (Figure 2, Table 2). On the WTA tour the big-match slope is lower: 0.755 against 0.821, so Elo overstates the favourite's chances more in WTA big matches. The difference survives the cluster bootstrap (p = 0.016, against 0.014 for the model-based test). It is not stable over time (Appendix Table A4): it is absent in 2000–2008 (−0.004), moderate in 2009–2017 (−0.069) and largest in 2018–2026 (−0.177, p < 0.001).

The pattern also helps forecasts out of sample. With calibrations fitted only on earlier years, separate big-match slopes reduced WTA big-match log loss by 0.9 thousandths (95% interval 0.3 to 1.5), from 0.6337 to 0.6328. On the ATP tour they made no clear difference (pooled minus separate calibration: −0.03 thousandths, interval −0.36 to +0.33). The effect is small, and it concerns every WTA favourite in a big match, not particular players.

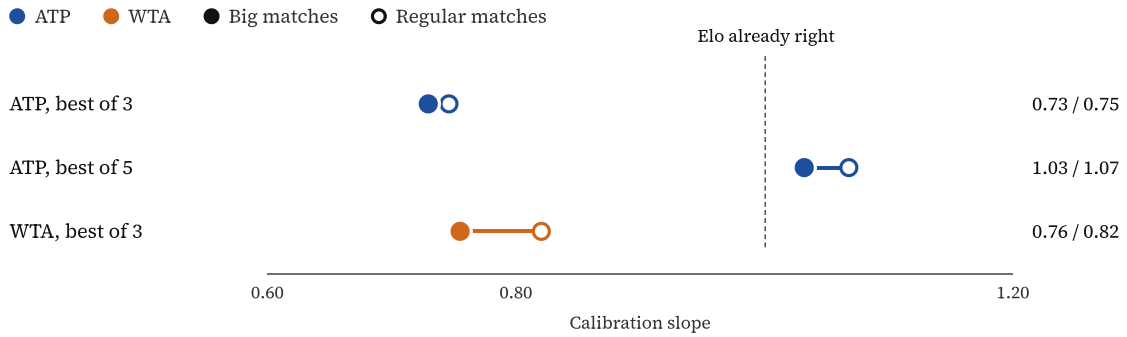


Figure 2. Elo calibration slopes in big and regular matches, 2000 – May 2026. A slope of 1 means Elo's probabilities were already right; below 1, Elo was too confident in the favourite.

Table 2. Calibration slopes with model-based standard errors, and the big-minus-regular difference with model-based (Wald) and edition-cluster bootstrap uncertainty.

Tour and format	Big	Regular	Difference	Wald p	Cluster 95% interval	Cluster p
ATP, best of 3	0.730 (0.027)	0.746 (0.010)	−0.017	0.555	−0.075 to +0.040	0.566
ATP, best of 5	1.032 (0.054)	1.068 (0.022)	−0.036	0.540	−0.138 to +0.080	0.521
WTA, best of 3	0.755 (0.025)	0.821 (0.009)	−0.066	0.014	−0.116 to −0.010	0.016

4.2 Variation beyond chance

Players' big-match results vary about as much as chance predicts (Figure 3, Table 3). Across the six big-match sets the variance ratio runs from 0.93 to 1.16; the largest is for WTA finals, where only 82 players qualify ($p = 0.178$). In standard-deviation terms the excess is smaller still: for all ATP big matches the observed spread is 9.6 points against 9.1 from chance. Where an excess appears, the implied number of matches before a record is half signal is above 100, more than most players play.

The clutch spread is 1.13 on the ATP tour ($p = 0.108$; Holm-adjusted 1.000) and 0.97 on the WTA tour ($p = 0.601$). The only clear excess in Table 3 is in regular WTA matches (1.30, $p < 0.001$), which is consistent with ordinary changes in form that Elo is slow to track rather than with pressure.

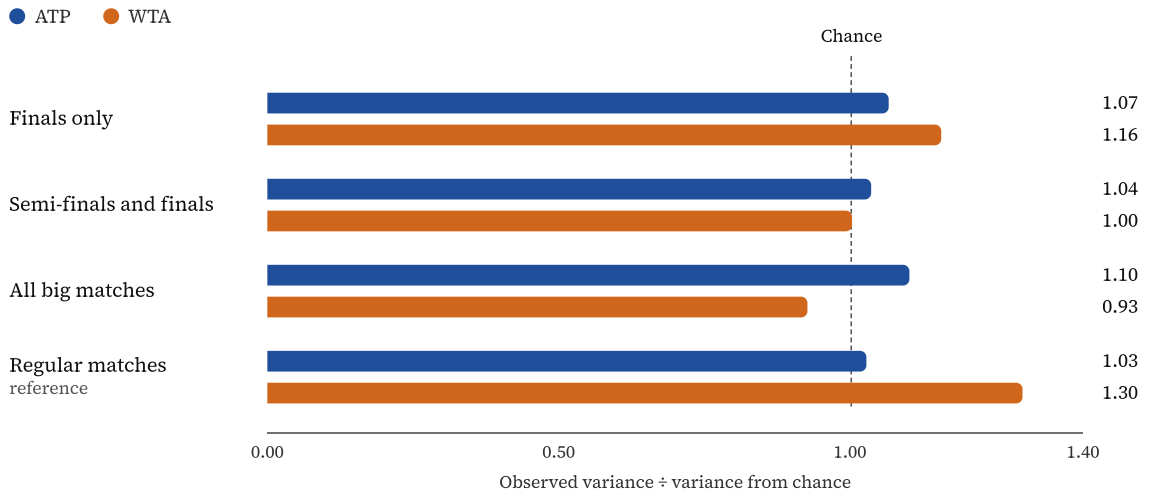


Figure 3. Variance of players' mean residuals divided by the variance chance alone would produce, 2000 – May 2026; players with at least 10 matches of each kind. The dashed line at 1 is chance.

Table 3. Spread of players' mean residuals. Standard deviations are in percentage points of win probability per match; the ratio is of variances. The trait sd is the square root of the excess variance; implied *k* is the number of matches before half of a record is signal. p from 2,000 simulations; Holm across the 18 trait statistics.

Tour and matches	Players	Observed sd	Chance sd	Variance ratio	Trait sd	Implied k	p	Holm p
ATP, finals only	101	11.7	11.3	1.07	2.9	247	0.294	1.000
ATP, semi-finals and finals	220	9.9	9.7	1.04	1.8	636	0.363	1.000
ATP, all big matches	242	9.6	9.1	1.10	2.9	245	0.176	1.000
ATP, regular matches	808	7.3	7.2	1.03	1.2	1,479	0.338	reference
WTA, finals only	82	12.1	11.3	1.16	4.4	109	0.178	1.000
WTA, semi-finals and finals	222	10.4	10.3	1.00	0.6	5,705	0.473	1.000
WTA, all big matches	249	9.2	9.6	0.93	0.0	none	0.749	1.000
WTA, regular matches	790	8.3	7.2	1.30	4.0	135	< 0.001	reference

4.3 Consistency and carry-over

The two halves of players' big-match records barely agree: r = +0.12 (−0.03 to +0.26) on the ATP tour and +0.04 (−0.14 to +0.23) on the WTA tour (Figure 4). Applying the Spearman–Brown formula (Spearman, 1910; Brown, 1910) to these point estimates gives a full-record reliability of 0.22 and 0.08. These are estimates with wide uncertainty, not upper limits, and the formula assumes halves of equal length, which unequal records only approximate.

Between periods, big-match scores carry over weakly (+0.25 ATP, +0.22 WTA), and no more strongly than regular-match scores do (+0.20 and +0.40). The retrospective association test (Table 4) is nominally significant for the WTA big-match score (permutation p = 0.026), but the WTA clutch score, which removes general form, shows no association (p = 0.711). What persists between periods is a player's general tendency to beat or trail their rating, which persists at least as strongly in regular matches. It is the only one of the 18 trait statistics with a raw p below 0.05 (simulation p = 0.006), and after Holm's correction its p-value is 0.108.

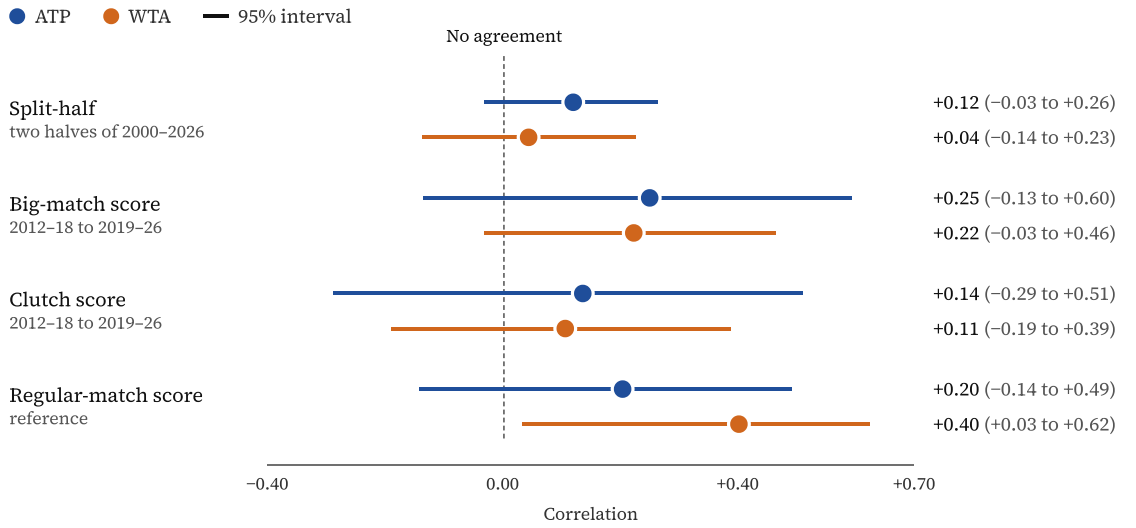


Figure 4. Agreement of player scores with 95% intervals: between the two halves of each player's big matches, and between 2012–2018 and 2019 – May 2026 for the 26 ATP and 32 WTA players with at least 10 big matches in both periods.

Table 4. Retrospective association (equation 5): 2012–2018 scores against 2019–2026 big-match results, with β fitted on the same later matches. Permutation p from 1,000 shuffles of the earlier scores. This is not an out-of-sample test (see Section 4.4).

Tour and score	Later big matches	With an earlier score	β (se)	Log-loss change	Permutation p
ATP, big-match score	1,916	1,516	+1.41 (1.03)	−0.00049	0.183
ATP, clutch score	1,916	1,516	+1.30 (0.86)	−0.00060	0.137
WTA, big-match score	1,713	1,394	+3.11 (1.25)	−0.00184	0.026
WTA, clutch score	1,713	1,394	+0.35 (0.80)	−0.00006	0.711

4.4 Out-of-sample forecasts

Predicting each year only from earlier years, no player score produced a statistically reliable improvement in forecasts of big matches (Table 5, Figure 5). Across both tours, the big-match score changed log loss by +0.03 thousandths and the clutch score by +0.07, against a baseline of 0.6075. Every interval for a player-score model includes zero, including the empirical Bayes and five-year variants and the subset of matches in which both players had at least 10 earlier big matches (Appendix Table A3). The largest point improvement of any variant is 0.4 thousandths.

The fitted β values help explain this. On the ATP tour the coefficient on the clutch score is negative in every evaluation year (between −0.75 and −0.25): a high earlier clutch score, which partly reflects a low regular-match score, is followed by slightly worse big-match results relative to the baseline, not better. The empirical Bayes WTA big-match score is zero in most years because the estimated trait variance is zero, so that model reproduces the baseline.

Table 5. Out-of-sample log loss on every big match, 2012 – May 2026, each year predicted from earlier years only. Differences are from the baseline M1 in thousandths of log loss; negative means better. 95% intervals from 2,000 resamples of tournament editions.

Model	ATP log loss	ATP difference (95%)	WTA log loss	WTA difference (95%)
No big-match calibration	0.5845	−0.03 (−0.36 to +0.33)	0.6337	+0.89 (+0.30 to +1.48)
M1: separate big and regular slopes	0.5845	baseline	0.6328	baseline
Big-match score	0.5845	−0.07 (−0.62 to +0.45)	0.6329	+0.15 (−0.15 to +0.45)
Clutch score	0.5846	+0.03 (−0.41 to +0.48)	0.6329	+0.11 (−0.09 to +0.32)
Big-match score, EB	0.5843	−0.20 (−0.81 to +0.43)	0.6328	0.00 (identical)
Clutch score, EB	0.5846	+0.12 (−0.23 to +0.45)	0.6328	+0.02 (−0.24 to +0.28)
Big-match score, 5 years	0.5841	−0.39 (−1.25 to +0.52)	0.6331	+0.29 (−0.09 to +0.72)
Clutch score, 5 years	0.5846	+0.11 (−0.13 to +0.36)	0.6329	+0.13 (−0.06 to +0.32)

Matches: 3,918 ATP, 3,558 WTA. EB = empirical Bayes shrinkage; 5 years = scores from the previous 5 years only. Brier scores and the subset with long histories are in Appendix Table A3.

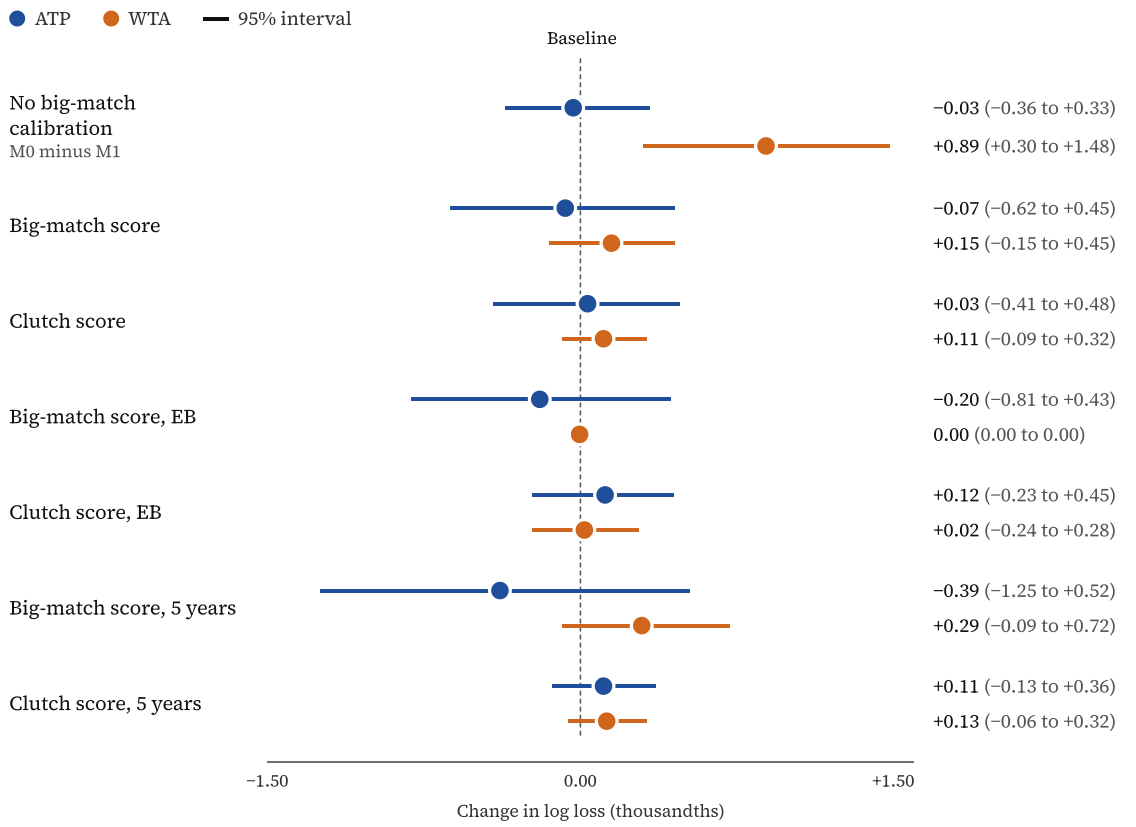


Figure 5. Out-of-sample change in log loss relative to the baseline M1 (separate big- and regular-match calibration), in thousandths, with 95% intervals; negative is better. The first row is the model without the big-match calibration.

4.5 What the data can rule out

The tests can detect a large trait. At $\tau = 0.4$, an edge of 9.9 points at even odds, the clutch-spread and split-half tests detect the trait in essentially every simulation on both tours (Figure 6). At $\tau = 0.15$ (3.7 points) no test has better than a 31% chance.

The clutch spread gives the headline bounds because it is the only test that removes general form. At 95% confidence the trait's standard deviation is at most 7.1 points (Monte Carlo range 6.9 to 7.3) at even odds on the ATP tour, and 4.8 points (Monte Carlo range 4.6 to 5.0) on the WTA tour; the split-half test gives

7.1 points (Monte Carlo range 6.9 to 7.3) and 5.6 points (Monte Carlo range 5.0 to 6.2). These bounds hold only under the simulation model: a normally distributed trait, constant over a career and confined to the defined big matches. They describe the spread of the trait across players, not a maximum for any single player, and an edge of 7.1 points (a 57.1% chance against an equal opponent) would matter for a forecast if it could be measured.

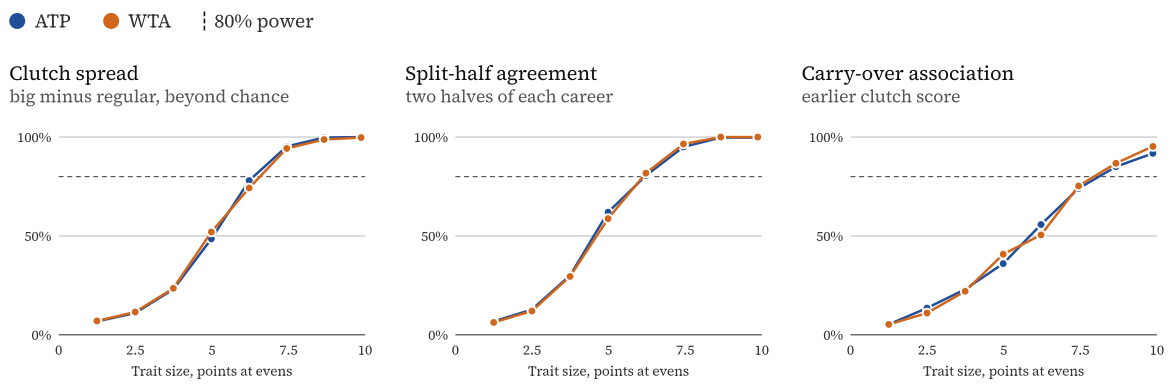


Figure 6. Power of three tests against the size of a simulated clutch trait, 400 simulations per point. The horizontal axis is the edge at even odds, in percentage points, of a player one standard deviation above average. The dashed line is 80% power.

4.6 Players with extreme early records, and sensitivity

Tables 6 and 7 follow the five players on each tour with the highest and lowest big-match scores in 2012–2018, among the 26 ATP and 32 WTA players with enough big matches in both periods. The five highest ATP scores averaged +8.8 points per match in the first period and −0.1 in the second; on the WTA tour the highest five went from +8.1 to +0.6 and the lowest five from −4.8 to −2.4. The ATP lowest five moved from −5.2 to −2.5. The tables illustrate regression toward zero; they are not a test, and with so few big matches per player every individual record in them is dominated by chance. Changes in a player's record can have causes, such as injury or age, that this analysis does not observe.

Across the 8 variants of the carry-over test per tour (shrinkage constants 10, 50 and 100, minimums of 5 and 20 big matches, semi-finals and finals only, and periods split at 2016; Appendix Table A5), no clutch-score association has p below 0.104, and the smallest p for either score is 0.018, which is not significant once the number of variants is considered.

Table 6. ATP players with the highest and lowest big-match scores in 2012–2018, and their records in 2019 – May 2026. Won (exp.) is wins against calibrated expected wins; the score is equation 3 in percentage points.

	Player	2012–18 big	Won (exp.)	Score	2019–26 big	Won (exp.)	Score
Highest five	Karen Khachanov	15	10 (4.8)	+14.9	51	21 (21.2)	−0.2
	Stan Wawrinka	114	66 (54.4)	+8.7	15	5 (6.4)	−4.0
	Nikoloz Basilashvili	12	6 (3.7)	+7.2	14	9 (5.6)	+9.9
	John Isner	78	42 (35.3)	+6.8	24	9 (12.9)	−8.8
	Daniil Medvedev	13	8 (5.9)	+6.5	163	101 (96.5)	+2.5
	Mean of the five			+8.8			−0.1
Lowest five	Pablo Carreno Busta	33	14 (15.4)	−2.5	31	13 (15.3)	−4.4
	Kei Nishikori	119	60 (66.1)	−4.4	12	3 (6.8)	−11.9
	Benoit Paire	27	7 (9.4)	−5.1	12	6 (5.4)	+2.0
	Adrian Mannarino	20	6 (8.5)	−6.2	28	15 (13.0)	+4.1
	Gael Monfils	59	23 (29.1)	−7.8	31	16 (17.2)	−2.3
	Mean of the five			−5.2			−2.5

Table 7. WTA players with the highest and lowest big-match scores in 2012–2018, and their records in 2019 – May 2026. Columns as in Table 6.

	Player	2012–18 big	Won (exp.)	Score	2019–26 big	Won (exp.)	Score
Highest five	Sloane Stephens	50	29 (21.0)	+11.4	16	7 (7.6)	−1.8
	Kiki Bertens	34	21 (15.6)	+10.1	22	12 (11.7)	+0.6
	Anastasia Pavlyuchenkova	42	24 (19.9)	+6.6	27	12 (12.0)	−0.1
	Simona Halep	140	94 (83.8)	+6.4	47	27 (29.9)	−4.3
	Jelena Ostapenko	30	17 (14.0)	+6.0	44	26 (20.6)	+8.4
	Mean of the five			+8.1			+0.6
Lowest five	Daria Kasatkina	23	10 (11.2)	−2.7	44	19 (20.5)	−2.3
	Petra Martic	10	2 (2.9)	−3.1	23	7 (10.3)	−7.6
	Anett Kontaveit	17	6 (7.6)	−4.2	35	21 (19.8)	+2.2
	Shuai Zhang	17	5 (6.9)	−5.1	17	6 (6.7)	−1.8
	Sorana Cirstea	15	2 (5.2)	−9.0	28	12 (13.3)	−2.7
	Mean of the five			−4.8			−2.4

4.7 Robustness checks

Game margins. Replacing each result with the share of games won (Appendix Table A6), players differ in margin far more than Elo predicts in every kind of match (regular-match variance ratios 1.62 ATP and 1.98 WTA), which points to playing style. The clutch spread on game margins is 1.23 on the ATP tour ($p = 0.027$) and 1.05 on the WTA tour ($p = 0.305$).

Format. Best-of-five matches make up a larger share of ATP big matches than of regular ones (22% against 18%), so a difference between big and regular matches could partly reflect format. Comparing big and regular ATP matches within one format (Appendix Table A8), the clutch spread is 1.18 in best-of-three matches ($p = 0.048$; games 1.30, $p = 0.008$) and 0.84 in best-of-five ($p = 0.816$).

These checks leave a small, unresolved ATP pattern. The win-based clutch spread on all matches is not significant ($p = 0.108$), but the game-margin version and the best-of-three comparison give nominal p-values below 0.05. Those checks are not corrected for multiple testing, they re-test the same hypothesis with overlapping data, and the ATP clutch score did not help out-of-sample forecasts. They are consistent with a small ATP excess within the bounds of Section 4.5, and not with a large one.

Form within the tournament. Updating each player's rating through their earlier rounds of the same event (Appendix Table A7) gives an ATP clutch spread of 1.10 and split-half correlations of +0.13 and +0.04. Round-robin matches and others without a complete earlier-round record keep the start-of-week rating (5% of big matches).

5 Discussion and Limitations

5.1 What the evidence supports

Three kinds of significance need to be kept apart. Statistically, no trait statistic is significant once the number of tests is taken into account. The nominal results either reflect general form (the WTA big-match association, which disappears once regular-match form is removed) or come from robustness checks of an ATP clutch spread whose main statistic is not significant. Practically, a trait of a few percentage points cannot be excluded, and the bounds allow an effect large enough to matter in a forecast. Predictively, past big-match records did not improve out-of-sample forecasts, which is the question the forecasting model needs answered.

The evidence therefore does not show that choking or clutch traits do not exist. It shows that tour-level results contain no reliable, measurable player-specific big-match signal of a size that would justify an adjustment. Records of a few dozen big matches are too noisy: at a typical win probability of 70%, chance

alone gives a record of 100 big matches a standard deviation of about 4.6 points per match, comparable to the largest trait the bounds allow.

5.2 The WTA calibration pattern

The lower WTA big-match slope means favourites in WTA big matches win less often than Elo, calibrated on regular matches, would predict. This study cannot tell apart the possible explanations. Players who reach late rounds are selected for being in form that week, which narrows the true gap between them relative to their start-of-week ratings. The players in big matches are stronger on average, and Elo's rating gaps near the top of the WTA may be less informative than elsewhere. The larger gap in 2018–2026 coincides with changes in the WTA calendar and tier structure, including the larger WTA 1000 tier, but nothing here shows that either caused it. Pressure on favourites is one possibility among these, and the data do not single it out.

5.3 What the tests cannot see

The bounds assume a trait that is normally distributed and constant over a career. A few players with large effects among many with none would add little to the variance ratios and could escape them, and a trait that changes over a career would weaken the consistency and carry-over tests; the five-year scores in the out-of-sample test are only a partial check on that. The simulations also keep the matches that were actually played, although under the null different results would have produced different later-round matches.

Other limitations are more ordinary. The study is post-hoc, and the multiple-testing family and several checks were chosen after earlier results were seen. Only 26 ATP and 32 WTA players have enough big matches in both periods, so the carry-over tests have little power. The bootstraps resample players or tournament editions but not both, so they do not capture every form of dependence between matches that share players. The expectation is Elo alone, which misses information such as injuries, and the game-share model is a single-slope fractional logit (Papke and Wooldridge, 1996). Pressure itself is never observed: the analysis measures results against expectation, not players' experience.

5.4 Implications for Matchup DNA

Matchup DNA's live model is a frozen logistic regression on Elo and other pre-match features, under evaluation in a pre-registered prospective protocol. It has no big-match adjustment of any kind, and this exploratory study does not change it. The results argue against adding a player-specific big-match feature. A tour-level adjustment for WTA big matches is a better candidate, but its out-of-sample gain is small, it has varied over time, and it would have to be specified in advance and evaluated on new matches as a new model version.

6 Conclusion

In 26 years of tour-level results, players' big-match records differ from one another by about as much as chance would make them differ, they are only weakly consistent within careers, and they did not improve forecasts of later big matches. The data rule out a large and stable choke or clutch trait, but not a modest one. The available evidence does not establish a sufficiently reliable player-specific big-match adjustment to justify adding one to Matchup DNA's forecasting model. The clearest result is a tour-level one: Elo is more overconfident in WTA big matches than in regular WTA matches, and that pattern is worth testing prospectively.

7 Data and Code Availability

Data and code

Match data are Jeff Sackmann's tennis_atp and tennis_wta files (CC BY-NC-SA 4.0), read from a June 2026 snapshot; the data are not redistributed with the code. The analysis is reproduced by `python -m matchup_dna.posthoc_big_match` followed by the companion modules `posthoc_big_match_robustness`, `posthoc_big_match_players` and `posthoc_big_match_forecast`; `python -m web.publisher.research` exports the figures shown here and refuses if a quoted headline figure differs from the analysis output. The SHA-256 fingerprints of the 118 input files are recorded in the forecast output (combined fingerprint 3910dc824f72926f…). The analysis code is in the project repository in the same commit as this version of the paper. Every random seed is fixed and no result depends on the order in which Python iterates over sets, so rerunning the commands on that commit with the same input files should reproduce the reported results. Step-by-step instructions are in docs/research/big-matches/README.md in the project repository.

Version history

This is version 4.0 (16 November 2026). Version 3.0 corrected errors in versions 1 and 2: the sample had included Roland Garros 2026 and several team and other non-standard events and had used inconsistent WTA tier codes, and Holm's correction had covered only the variance ratios. The corrections changed the sample from 131,649 to 130,460 matches and moved most figures. Versions 3.1 and 4.0 expanded the related research and attribution and made the split-half bootstrap reproducible (ATP interval now −0.03 to +0.26). The full history is in docs/research/big-matches/CHANGELOG.md in the project repository.

Acknowledgements

Claude Code (Anthropic) was used to assist with the implementation of statistical analyses, generation of figures and tables, and drafting and editing of the manuscript. The author takes responsibility for the content, reported results, and any remaining errors.

References

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46(3), 610–620. https://doi.org/10.1037/0022-3514.46.3.610

Birnbaum, P. (2005). Clutch hitting and the Cramer test. *By the Numbers*, 15(1). Reprinted in *Baseball Research Journal*, 37 (2008). https://sabr.org/journal/article/clutch-hitting-and-the-cramer-test/

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Annals of Applied Statistics*, 2(1), 113–152. https://doi.org/10.1214/07-AOAS138

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Cohen-Zada, D., Krumer, A., Rosenboim, M., and Shapir, O. M. (2017). Choking under pressure and gender: Evidence from professional tennis. *Journal of Economic Psychology*, 61, 176–190. https://doi.org/10.1016/j.joep.2017.04.005

Cramer, R. D. (1977). Do clutch hitters exist? *Baseball Research Journal*, 6. Society for American Baseball Research. https://sabr.org/journal/article/do-clutch-hitters-exist/

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350), 311–319. https://doi.org/10.1080/01621459.1975.10479864

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. New York: Arco.

González-Díaz, J., Gossner, O., and Rogers, B. W. (2012). Performing best when it matters most: Evidence from professional tennis. *Journal of Economic Behavior & Organization*, 84(3), 767–781. https://doi.org/10.1016/j.jebo.2012.09.021

Gorgi, P., Koopman, S. J., and Lit, R. (2019). The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *Journal of the Royal Statistical Society Series A*, 182(4), 1393–1409. https://doi.org/10.1111/rssa.12464

Harris, D. J., Vine, S. J., Eysenck, M. W., and Wilson, M. R. (2021). Psychological pressure and compounded errors during elite-level tennis. *Psychology of Sport and Exercise*, 56, 101987. https://doi.org/10.1016/j.psychsport.2021.101987

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. https://www.jstor.org/stable/4615733

James, B. (2004). Underestimating the fog. *Baseball Research Journal*, 33. Society for American Baseball Research. https://sabr.org/research/article/underestimating-the-fog/

Jetter, M., and Walker, J. K. (2015). Game, set, and match: Do women and men perform differently in competitive situations? *Journal of Economic Behavior & Organization*, 119, 96–108. https://doi.org/10.1016/j.jebo.2015.07.017

Klaassen, F. J. G. M., and Magnus, J. R. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, 96(454), 500–509. https://doi.org/10.1198/016214501753168217

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12(3), 127–138. https://doi.org/10.1515/jqas-2015-0059

Kovalchik, S., and Ingram, M. (2016). Hot heads, cool heads, and tacticians: Measuring the mental game in tennis. MIT Sloan Sports Analytics Conference, Research Papers Competition (paper 1464). https://martiningram.github.io/papers/hot-heads-cool-heads.pdf

Kovalchik, S. A., and Reid, M. (2020). Measuring clutch performance in professional tennis. *Statistica Applicata – Italian Journal of Applied Statistics*, 30(2), 255–268. https://doi.org/10.26398/IJAS.0030-011

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362. https://doi.org/10.1177/1948550617697177

Papke, L. E., and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11(6), 619–632. https://doi.org/10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1

Paserman, M. D. (2023). Gender differences in performance in competitive environments? Evidence from professional tennis players. *Journal of Economic Behavior & Organization*, 212, 590–609. https://doi.org/10.1016/j.jebo.2023.05.042

Sackmann, J. (2019a, February 20). Dominic Thiem, Tennys Sandgren, and playing your way in. *Tennis Abstract*. https://www.tennisabstract.com/blog/2019/02/20/dominic-thiem-tennys-sandgren-and-playing-your-way-in/

Sackmann, J. (2019b, December 3). An introduction to tennis Elo. *Tennis Abstract*. https://www.tennisabstract.com/blog/2019/12/03/an-introduction-to-tennis-elo/

Sackmann, J. (2025, January 17). Matteo Berrettini and best-of-five puzzles. *Tennis Abstract*. https://www.tennisabstract.com/blog/2025/01/17/matteo-berrettini-and-best-of-five-puzzles/

Sackmann, J. (2026). Tennis databases, files, and algorithms: tennis_atp and tennis_wta [data sets]. Jeff Sackmann / Tennis Abstract. Licensed CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/). Read from a June 2026 snapshot (archive mirror github.com/Aneeshers/tennis-sackmann-archive); the original repositories at github.com/JeffSackmann were unavailable on 16 November 2026.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Tango, T. M., Lichtman, M. G., and Dolphin, A. E. (2007). *The Book: Playing the Percentages in Baseball*. Washington, DC: Potomac Books.

Appendix A Supplementary Tables

Table A1. Simulation tests: each statistic against 2,000 simulated data sets with no trait. Holm-adjusted p across the 18 trait statistics (both tours) and within each tour.

Statistic	ATP obs.	ATP p	Holm	Holm (tour)	WTA obs.	WTA p	Holm	Holm (tour)
Variance ratio, finals	+1.07	0.294	1.000	0.720	+1.16	0.178	1.000	1.000
Variance ratio, semi-finals and finals	+1.04	0.363	1.000	0.720	+1.00	0.473	1.000	1.000
Variance ratio, all big matches	+1.10	0.176	1.000	0.704	+0.93	0.749	1.000	1.000
Variance ratio, regular (reference)	+1.03	0.338	–	–	+1.30	< 0.001	–	–
Clutch spread	+1.13	0.108	1.000	0.661	+0.97	0.601	1.000	1.000
Split-half r	+0.12	0.094	1.000	0.661	+0.04	0.331	1.000	1.000
Carry-over r, big-match score	+0.25	0.115	1.000	0.661	+0.22	0.108	1.000	0.864
Carry-over r, clutch score	+0.14	0.240	1.000	0.720	+0.11	0.279	1.000	1.000
Association z, big-match score	+1.37	0.082	1.000	0.656	+2.50	0.006	0.108	0.054
Association z, clutch score	+1.51	0.059	1.000	0.535	+0.44	0.318	1.000	1.000

Table A2. Power at each trait size and the 95% upper bound on the trait, with its Monte Carlo range. Trait sizes τ = 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, an edge at even odds of 1.2, 2.5, 3.7, 5.0, 6.2, 7.4, 8.7, 9.9 points.

Tour and test	1.2	2.5	3.7	5.0	6.2	7.4	8.7	9.9	Bound (range)
ATP, variance ratio, finals	5%	6%	10%	23%	34%	53%	71%	87%	8.3 (7.7–8.9)
ATP, variance ratio, semi-finals and finals	8%	10%	25%	47%	79%	96%	99%	100%	5.9 (5.7–6.0)
ATP, variance ratio, all big matches	7%	13%	26%	53%	82%	97%	100%	100%	6.3 (6.1–6.7)
ATP, clutch spread	7%	11%	23%	49%	78%	95%	100%	100%	7.1 (6.9–7.3)
ATP, split-half r	7%	13%	30%	62%	81%	95%	100%	100%	7.1 (6.9–7.3)
ATP, carry-over r, big-match score	5%	10%	17%	29%	49%	63%	74%	89%	–
ATP, carry-over r, clutch score	5%	11%	14%	23%	40%	49%	62%	75%	–
ATP, association z, big-match score	6%	16%	31%	47%	67%	86%	93%	97%	9.1 (8.5–9.5)
ATP, association z, clutch score	5%	14%	23%	36%	56%	74%	85%	92%	–
WTA, variance ratio, finals	3%	7%	9%	20%	30%	41%	65%	78%	–
WTA, variance ratio, semi-finals and finals	6%	10%	20%	42%	68%	90%	99%	100%	5.8 (5.5–6.0)
WTA, variance ratio, all big matches	7%	12%	22%	53%	77%	96%	99%	100%	3.7 (3.5–4.2)
WTA, clutch spread	7%	12%	24%	52%	74%	94%	99%	100%	4.8 (4.6–5.0)
WTA, split-half r	6%	12%	30%	59%	82%	97%	100%	100%	5.6 (5.0–6.2)
WTA, carry-over r, big-match score	5%	13%	21%	39%	51%	72%	85%	95%	9.4 (8.6–9.7)
WTA, carry-over r, clutch score	5%	10%	14%	23%	33%	56%	63%	83%	8.8 (8.5–9.3)
WTA, association z, big-match score	6%	14%	28%	49%	66%	86%	96%	99%	9.8 (9.6–9.8)
WTA, association z, clutch score	5%	11%	22%	41%	51%	75%	87%	95%	8.2 (7.4–8.7)

Table A3. Out-of-sample evaluation in full: differences from M1 in thousandths of log loss (and, for both tours, of Brier score) with 95% edition-bootstrap intervals. The lower block is limited to big matches in which both players had at least 10 earlier big matches.

Model	ATP Δ log loss	WTA Δ log loss	Both Δ log loss	Both Δ Brier
No big-match calibration	−0.03 (−0.36 to +0.33)	+0.89 (+0.30 to +1.48)	+0.41 (+0.08 to +0.74)	+0.15 (+0.04 to +0.27)
Big-match score	−0.07 (−0.62 to +0.45)	+0.15 (−0.15 to +0.45)	+0.03 (−0.30 to +0.34)	+0.04 (−0.10 to +0.18)
Clutch score	+0.03 (−0.41 to +0.48)	+0.11 (−0.09 to +0.32)	+0.07 (−0.18 to +0.32)	+0.02 (−0.10 to +0.13)
Big-match score, EB	−0.20 (−0.81 to +0.43)	0.00 (identical)	−0.10 (−0.43 to +0.22)	−0.01 (−0.14 to +0.13)
Clutch score, EB	+0.12 (−0.23 to +0.45)	+0.02 (−0.24 to +0.28)	+0.07 (−0.14 to +0.30)	+0.05 (−0.05 to +0.15)
Big-match score, 5 years	−0.39 (−1.25 to +0.52)	+0.29 (−0.09 to +0.72)	−0.06 (−0.53 to +0.42)	+0.01 (−0.20 to +0.22)
Clutch score, 5 years	+0.11 (−0.13 to +0.36)	+0.13 (−0.06 to +0.32)	+0.12 (−0.04 to +0.27)	+0.05 (−0.02 to +0.12)
No big-match calibration (long histories)	−0.42 (−0.84 to +0.04)	+0.37 (−0.33 to +1.08)	−0.07 (−0.47 to +0.35)	+0.01 (−0.13 to +0.16)
Big-match score (long histories)	+0.04 (−0.65 to +0.73)	+0.01 (−0.49 to +0.52)	+0.03 (−0.43 to +0.48)	+0.08 (−0.13 to +0.28)
Clutch score (long histories)	−0.17 (−0.69 to +0.38)	+0.02 (−0.27 to +0.32)	−0.08 (−0.40 to +0.23)	−0.06 (−0.20 to +0.08)
Big-match score, EB (long histories)	−0.30 (−1.17 to +0.55)	0.00 (identical)	−0.17 (−0.65 to +0.35)	−0.01 (−0.21 to +0.21)
Clutch score, EB (long histories)	+0.42 (−0.12 to +0.95)	−0.10 (−0.58 to +0.36)	+0.19 (−0.15 to +0.53)	+0.11 (−0.05 to +0.27)
Big-match score, 5 years (long histories)	−0.37 (−1.54 to +0.80)	+0.18 (−0.48 to +0.84)	−0.12 (−0.83 to +0.59)	+0.03 (−0.29 to +0.35)
Clutch score, 5 years (long histories)	+0.13 (−0.19 to +0.46)	+0.04 (−0.27 to +0.34)	+0.09 (−0.12 to +0.31)	+0.04 (−0.06 to +0.14)

Matches: all big matches 3,918 ATP, 3,558 WTA; long histories 2,326 ATP, 1,902 WTA. Baseline log loss (M1), all big matches: ATP 0.5845, WTA 0.6328.

Table A4. Calibration slopes by period, big against regular, with model-based p-values.

Period	Tour and format	Big	Regular	Difference	Wald p
2000–2008	ATP, best of 3	0.746	0.776	−0.030	0.554
2000–2008	ATP, best of 5	1.025	1.092	−0.066	0.513
2000–2008	WTA, best of 3	0.863	0.867	−0.004	0.932
2009–2017	ATP, best of 3	0.803	0.776	+0.027	0.564
2009–2017	ATP, best of 5	1.098	1.094	+0.004	0.968
2009–2017	WTA, best of 3	0.727	0.796	−0.069	0.129
2018–2026	ATP, best of 3	0.625	0.669	−0.044	0.380
2018–2026	ATP, best of 5	0.966	1.012	−0.046	0.656
2018–2026	WTA, best of 3	0.618	0.795	−0.177	< 0.001

Table A5. Sensitivity of the carry-over association, one setting changed at a time: clutch-score correlation with 95% interval, and β with permutation p for each score (1,000 permutations, drawn afresh, so the first rows differ slightly from Table 4).

Tour and variant	Players	Clutch r (95%)	β big (p)	β clutch (p)
ATP, main settings	26	+0.14 (−0.29 to +0.52)	+1.41 (0.180)	+1.30 (0.146)
ATP, shrinkage constant 10	26	+0.16 (−0.32 to +0.54)	+0.94 (0.205)	+0.95 (0.146)
ATP, shrinkage constant 50	26	+0.12 (−0.25 to +0.46)	+2.66 (0.156)	+2.05 (0.151)
ATP, shrinkage constant 100	26	+0.13 (−0.22 to +0.42)	+4.41 (0.122)	+2.98 (0.104)
ATP, minimum 5 big matches	48	+0.07 (−0.21 to +0.34)	+1.41 (0.198)	+1.30 (0.144)
ATP, minimum 20 big matches	14	−0.07 (−0.67 to +0.52)	+1.41 (0.180)	+1.30 (0.130)
ATP, semi-finals and finals only	20	−0.02 (−0.50 to +0.44)	+0.25 (0.856)	+0.51 (0.668)
ATP, periods 2008–15 / 2016–26	34	+0.33 (+0.07 to +0.54)	+1.24 (0.381)	+1.29 (0.250)
WTA, main settings	32	+0.11 (−0.19 to +0.37)	+3.11 (0.018)	+0.35 (0.694)
WTA, shrinkage constant 10	32	+0.11 (−0.19 to +0.38)	+2.17 (0.027)	+0.43 (0.561)
WTA, shrinkage constant 50	32	+0.10 (−0.20 to +0.36)	+4.97 (0.038)	+0.06 (0.964)
WTA, shrinkage constant 100	32	+0.10 (−0.22 to +0.38)	+7.15 (0.046)	−0.22 (0.908)
WTA, minimum 5 big matches	53	+0.16 (−0.08 to +0.38)	+3.11 (0.020)	+0.35 (0.691)
WTA, minimum 20 big matches	15	+0.29 (−0.11 to +0.62)	+3.11 (0.025)	+0.35 (0.680)
WTA, semi-finals and finals only	27	−0.30 (−0.65 to +0.11)	+1.56 (0.332)	+0.02 (0.986)
WTA, periods 2008–15 / 2016–26	33	−0.07 (−0.44 to +0.40)	+0.61 (0.576)	+0.45 (0.566)

Table A6. Game margins: share of games won against a fitted expectation. Variance ratios with permutation p (1,000 permutations within strata of format, match type and closeness).

Statistic	ATP	WTA
Matches	68,393	62,067
Variance ratio, finals only	1.27 (p 0.057)	1.48 (p 0.013)
Variance ratio, semi-finals and finals	1.08 (p 0.270)	0.98 (p 0.557)
Variance ratio, all big matches	1.20 (p 0.070)	1.11 (p 0.158)
Variance ratio, regular matches	1.62 (p < 0.001)	1.98 (p < 0.001)
Clutch spread	1.23 (p 0.027)	1.05 (p 0.305)
Split-half r	+0.18 (0.00 to +0.36)	+0.04 (−0.15 to +0.22)
Carry-over r, clutch score	+0.15 (−0.15 to +0.45)	+0.20 (−0.13 to +0.49)
Association β, clutch score (p)	+4.29 (0.126)	−0.33 (0.881)

Table A7. Within-tournament ratings: the main statistics with each player's rating updated through their earlier rounds of the same event.

Statistic	ATP	WTA
Big matches with a complete earlier-round record	95%	96%
Variance ratio, finals only	1.06	1.15
Variance ratio, semi-finals and finals	1.02	0.99
Variance ratio, all big matches	1.09	0.92
Variance ratio, regular matches	0.99	1.26
Clutch spread	1.10	0.96
Split-half r	+0.13 (−0.02 to +0.27)	+0.04 (−0.13 to +0.21)
Association β, clutch score (p)	+1.18 (0.199)	+0.36 (0.689)

Table A8. ATP clutch spread within one format: big against regular matches of the same length. Wins against 2,000 simulations; games against 1,000 permutations. Added in version 2, after an ATP result had been seen.

Format	Big matches	Regular matches	Players	Wins (p)	Games (p)
Best of 3	5,775	49,874	230	1.18 (0.048)	1.30 (0.008)
Best of 5	1,592	11,152	84	0.84 (0.816)	1.26 (0.070)

Appendix B Technical Details

B.1 Match classification

Events are main-tour levels in Sackmann's files (ATP: Grand Slam, Masters, tour, Finals, Olympics; WTA: Grand Slam, Tier I–V, Premier Mandatory, Premier, International, Finals, Olympics and the 2021 events coded W), minus events whose names identify them as the ATP Cup, United Cup, Laver Cup, Next Gen Finals or a WTA 125 event, and the World Team Cup (Düsseldorf round robin, 2000–2012). The season-ending Finals are the events named Masters Cup and Tour Finals (ATP) and WTA Tour Championships, WTA Championships, WTA Finals and their host-city names from 2016 on (WTA). WTA top-tier events by year are listed in the analysis code (WTA_TOP_TIER in matchup_dna/posthoc_big_match.py) and were checked against season records: Premier Mandatory (Indian Wells, Miami, Madrid, Beijing) and Premier 5 (Dubai or Doha, Rome, Canada, Cincinnati, and Tokyo to 2013 or Wuhan from 2014) for 2009–2020, and the WTA 1000 events actually played from 2021.

B.2 Simulations, power and bounds

Each simulated data set draws one Bernoulli outcome per match from its calibrated probability, so the two players' results are always complementary. The calibration is refitted on the simulated outcomes before any statistic is computed, so the simulated statistics carry the same estimation noise as the observed ones. One-sided p-values are (1 + number at least as large) / (1 + number of simulations). In the power simulations the trait θ_i is drawn once per player and added to the log-odds of every big match (equation 6). The Monte Carlo range of a bound is the 2.5th to 97.5th percentile of the bound over 1,000 resamples of the simulated statistics at each grid point; it reflects simulation noise only, not sampling uncertainty in the data.

B.3 Game margins

Each result is replaced by the share of games the player won, with a tie-break that decides a match counted as one game. The expected share is a fractional logit on the Elo log-odds, fitted per tour, match type and format. Residuals are permuted among matches within strata defined by those groups and by decile of the absolute Elo log-odds, and each stratum's mean squared residual is its noise variance, because close matches have more variable game shares than one-sided ones. Version 2 permuted within the groups only.

B.4 Empirical Bayes scores

In the out-of-sample variant, a player's mean big-match residual m_i, with noise variance $v_i = \Sigma p(1 - p)/n_i^2$, is shrunk to $m_i\,\tau^2/(\tau^2 + v_i)$, where τ^2 is the mean of m_i^2 minus the mean of v_i over players with at least 10 big matches before the year being predicted, floored at zero. The clutch version applies the same formula to the difference between the big- and regular-match means, with the two noise variances added.