Stabilization of NBA Plus-Minus Metrics and the Introduction of Three Context-Adjusted Alternatives: aWOWY, Par ± and HAPM

Ayomide Awoyemi · LUMA Basketball Research · Working paper v1 · November 2026

1. Introduction

At the time of this paper's inception, plus-minus metrics are one of the primary forms of appraising player and lineup value, preferred over box score metrics for their ability to quantify the more latent, intangible aspects of contributing to net margin (winning basketball games).

With each new season, early plus-minus data are presented as findings, often with the caveat that it is too early. The creation of this paper stems from that caveat: at what point does a metric become fit for analysis or further conclusions?

There is a great amount of prior stabilization work, but it is mainly on box score and team statistics (Carleton 2012; Blackport 2014; Narsu 2017; Medvedovsky 2020). As of this paper's creation, there has not been as much work specifically on the stabilization rates of the plus-minus metric family (Section 2).

This paper runs two studies on nine metrics, stabilization and descriptive/predictive power, using every possession-counted stint of 13,404 regular-season games over 11 NBA seasons (2013-14 to 2025-26, without the 2019-20 and 2020-21 seasons) and about 270 rotation players a season. The main finding is that With Or Without You (WOWY), the team's net rating with a player on the court minus its net rating without him, has two structural weaknesses that lead to considerable bias and variance, stemming from its baseline of a player's own replacements and its lack of control and adjustment for opponents and shooting variance. WOWY is the slowest of the nine measured metrics to stabilize, needing 113 games to become provisionally stable, and even over a full season it is the worst of the nine for describing and predicting, at less than half the accuracy of any other established metric.

Due to the structural weaknesses of WOWY, this paper introduces a new method and three new metrics. Adjusted WOWY (aWOWY) addresses WOWY's extreme variance and instability while keeping its outputs in on/off units to preserve its utility for lineup combinations. aWOWY becomes provisionally stable in 78% fewer games than WOWY and predicts next season's games 93% better. Held-out estimation, the method, is an alternative to joint estimation for quantifying player value from stint-level data. It compares each of a player's stints with what the other nine players, rated without him, would be expected to produce. Par ± applies held-out estimation to one season and removes 1-year RAPM's teammate bias (+.00 against +.23). Held-out adjusted plus-minus (HAPM), Par ± shrunk toward a prior, addresses the team-depth bias in WOWY and the teammate bias in RAPM to better isolate a player's value. Compared to 1-year RAPM, it has a near-zero teammate bias, and its time-decay version predicts next season's games about 300% better than WOWY and 12% better than 4-year RAPM.

2. Related Work

Stabilization and sample size.

- *Baseball Therapy: It's a Small Sample Size After All* (Carleton 2012, Baseball Prospectus) re-ran his 2007 study and found for each batting statistic the sample size at which its reliability (KR-21) reaches .70, the threshold most later work adopts.

- *How Long Does It Take For Three Point Shooting To Stabilize?* (Blackport 2014, Nylon Calculus) applied the same method to NBA three-point percentage, taking .70 as the standard threshold, the point after which "the skill aspect will outweigh the noise".
- *When Can We Trust a Team's Stats?* (Narsu 2017, Nylon Calculus) counted, for team statistics, the games until a team's value after k games explains half of the variance in its full-season value. Team net rating takes 8 games, and this paper's data reproduces that count. Narsu notes that this design answers when a statistic can be trusted to reflect the final season, and that a split-half test better answers when it stabilizes.
- *NBA Stabilization Rates and the Padding Approach* (Medvedovsky 2020) projects a player's value for the rest of the season by adding a fixed number X of league-average possessions (the padding) to his own sample:

$$\text{projection} = \frac{\text{his total} + X \cdot \text{league average}}{\text{his possessions} + X}.$$

  X is fitted to predict the rest of the season. It is also the sample size at which his own record and the league average carry equal weight, so a padding constant is a stabilization point. Medvedovsky fitted X for box score statistics and one plus-minus statistic, on-court plus-minus per 100 possessions, which needed 1,007 possessions of padding (about 23 games); this paper's data gives 1,017.
- *When Does an NBA Record Start to Mean Something?* (Shax 2026) measured team records and point differential with split-half reliability, out-of-sample win forecasts and head-to-head rankings. Following Tango, Lichtman and Dolphin (2007), it uses a split-half correlation of .50 as the point where signal equals noise. A team's record reaches it at about game 12 and its point differential at about game 9. It also applies the padding approach to team records.

Noise and evaluation in plus-minus.

- *Measuring How NBA Players Help Their Teams Win* (Rosenbaum 2004, 82games) estimated every player's effect at once by regressing stint margins on the ten players on the court (adjusted plus-minus), with a standard error for each rating.
- *Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing* (Sill 2010, MIT Sloan Sports Analytics Conference) traced the non-intuitive one-season results of adjusted plus-minus to overfitting, added ridge regularization (RAPM), and evaluated each model by predicting the margins of games it was not fitted on. This paper's game-margin R^2 uses his test.
- *NBA Adjusted Plus-Minus: How to Build It, plus Pro Tips and Tricks* (Engelmann 2026), by the author of xRAPM and ESPN's Real Plus-Minus, describes how modern all-in-one metrics combine a lineup component (RAPM) with a box score component; his earlier chapter (Engelmann 2017, Handbook of Statistical Methods and Analyses in Sports) is the formal treatment. A RAPM whose coefficients are pulled toward a prior estimate, such as a box score estimate or earlier seasons, is a prior-informed RAPM. The best known is ESPN's Real Plus-Minus, introduced in *The Next Big Thing: Real Plus-Minus* (Ilardi 2014). Engelmann (2011) built a prior-informed RAPM with priors from earlier seasons: he rated each season with the previous season's ratings as priors and slightly negative priors for rookies, and found that it predicted out-of-sample games better than 1-year and multi-year RAPM. This paper uses a prior-informed RAPM with priors from earlier seasons, instead of a box score estimate, as HAPM's baseline (Section 8).
- *Adjusted Plus-Minus for NHL Players using Ridge Regression with Goals, Shots, Fenwick, and Corsi* (Macdonald 2012, Journal of Quantitative Analysis in Sports) measured the year-to-year consistency of adjusted plus-minus in the NHL and found the ridge estimates more consistent than least squares.

- *Deep Reinforcement Learning for NBA Player Valuation: A Temporal Difference Approach with Shapley Attribution* (Jenkins 2026, MIT Sloan Sports Analytics Conference) values each action by its effect on win probability and divides the credit among the players with Shapley values. It measures stability by correlating each player's ranking after k games with his ranking over the full season, a sample that contains those k games, and finds that RAPM needs about 45 games to reach .80. This paper instead compares the first k games with the rest of the season, and odd with even games, so the two samples never share a game. Measured his way, this paper's RAPM also reaches .80 at 45 games, where a metric with no skill in it already reaches .70; against the rest of the season it does not pass .40.

The plus-minus family.

- *A Comprehensive Review of Plus-Minus Ratings for Evaluating Individual Players in Team Sports* (Hvattum 2019, International Journal of Computer Science in Sport) surveys plus-minus ratings across sports.
- *Thinking Basketball* (Taylor 2016) uses With Or Without You (WOWY) comparisons over the games a player misses, and *Historical Impact: Introducing WOWYR* (Taylor 2016) adjusts for a team's average strength of schedule and home court and regresses on teammates. *The Multi-Year WOWY Database* (RealGM) extends WOWY across seasons, through trades, injuries and retirements, and *WOWY RAPM* (Medvedovsky 2026, DARKO) builds a RAPM-style rating from with-or-without data at the game level.
- *About EPM and Methodology* (Snarr, Dunks and Threes) combines a box score estimate with RAPM, and *LEBRON Introduction* (Narsu and Cranjis McBasketball, BBall Index) combines a box score prior with luck-adjusted RAPM.

Table 1. Prior work on stability and noise, by what each measures.

Work	Metrics	Level	Measure
Carleton (2012)	Batting statistics	Player	Reliability
Blackport (2014)	Three-point percentage	Player	Reliability (KR-21)
Narsu (2017)	Team statistics, including net rating	Team	First k games against the full season
Medvedovsky (2020)	Box score statistics; on-court +/-	Player	Padding, rest of season
Macdonald (2012)	Adjusted plus-minus (NHL)	Player	Year to year
Jenkins (2026)	RAPM; his model	Player	First k games against the full season; year to year
Shax (2026)	Team records, point differential	Team	Split-half; out-of-sample win forecasts
This paper	**Nine metrics: six plus-minus, three box score or box-prior**	**Player and group**	**Split-half, rest of season and next-season game margins, on the same players and games**

3. Data and Definitions

This paper used possession-counted stints (stretches of play with the same ten players on the court) for 13,404 of the 13,530 regular-season games from 2013-14 to 2018-19 and 2021-22 to 2025-26, and per-game box

score counts from the play-by-play. 2019-20 (suspended, then finished in a bubble) and 2020-21 (72 games) are left out of the tests; they still feed the priors and the time-decay pool.

A window of k games is the first $15k$ league games of a season in date order, about k games per team; the rest of the season is every game after it. Using $k = 1, 2, 3, 5, 10, 15, 20, 30, 41$ and 55, a player enters if he played at least 10 possessions per team game in the window and 20 in the rest of the season. These thresholds kept about 270 rotation players a season at $k = 5$.

The nine metrics are listed in Table 2, each computed from the window's games only. Unregularized APM is not included: one season of stints does not estimate it reliably (Sill 2010). ARC and LOKI are LUMA Basketball Research models. Sections 5 and 8 also use:

- *RAPM (k-year)*: RAPM fit on the season and the $k - 1$ before it, each earlier season's possessions weighted .7 per season back (Appendix D).
- *PI RAPM* (prior-informed RAPM): RAPM with each player's coefficient pulled toward a prior from his three previous seasons; Par ± and HAPM use it to rate the other nine players.
- *HAPM (two seasons)*: this season's and last season's Par ± pooled, last season weighted .5.
- *HAPM (time decay)*: every season of a player's Par ±, each weighted .63 per season back, pooled and shrunk toward a center set by playing time (Section 8.3).

Table 2. The nine metrics. All values are points per 100 possessions.

Metric	Definition
On-court +/-	Team net rating with the player on the court: the with-him half of WOWY
WOWY (on/off)	With or without you: team net rating with the player on the court minus team net rating with him off, counting off-court possessions only in games he played
aWOWY	On/off computed at the stint level after adjusting points for the opponents on the floor and setting opponents' mid-range, three-point and free-throw makes to league rates (Section 7)
HAPM (one-year)	Held-out adjusted plus-minus: the margin of the player's stints above what the other nine players, rated without his possessions, would be expected to produce with an average player in his place, shrunk toward a center set by playing time. His stints come from the window only; the other nine's ratings also use the three previous seasons (Section 8.3)
Par ±	Held-out estimation over one season: the same value as one-year HAPM before the shrink (Section 8.2)
RAPM (1-year)	Each player's effect on team point margin per 100 possessions, estimated jointly for all players by ridge regression on the window's stints, with no prior (Sill 2010; settings in Appendix D)
BPM 2.0	The Basketball-Reference specification rebuilt from the window's box score counts; its weights come from a regression of box score statistics on long-run RAPM (Table 23)
EBV	Expected box value: each box score event priced in points by its effect on the possession's expected value, counted from play-by-play; unlike BPM, the prices are not fitted to plus-minus (Table 22)
ARC (1-year)	An all-in-one metric from LUMA Basketball Research, like LEBRON or EPM: a Bayesian adjusted plus-minus fit on the window's stints with each player's EBV as the prior; his estimate moves from the prior toward the stint evidence as his possessions grow

The NBA.com definition of on/off also counts the team's possessions in games the player missed.

Rest-of-season correlation (A). The correlation, across players, between each player's value after k games and his value over the rest of the season (the target of Medvedovsky 2020). The two share no games; a correlation with the full season would overstate stability, because the full season contains the window (for WOWY, .86 against an A that never passes .26).

Split-half correlation (S). The correlation between each player's value on the odd- and on the even-numbered games among the first $2k$ games per team, so each half has k games. Carleton (2012) and Blackport (2014) measure the same reliability with KR-21.

Game-margin R^2 (B). The R^2 of the remaining games' margins on a prediction from the values after k games: in every stint, the five offensive players' values minus the five defenders', times the stint's possessions, summed over the game. Players without a value count as 0, so predicting 0 for every game scores $B = 0$.

Stability criteria. A metric is provisionally stable once its split-half correlation S reaches .50, the point where the differences between players are half skill and half noise, and stable at .75, where the skill part is three times the noise. The .50 point is also where the padding approach (Medvedovsky 2020; Section 2)

gives the player's own value and the league average equal weight: a value measured at that sample size should be shrunk halfway to the league average. A and B carry no threshold: A is limited by the noise in both samples, and single games are noisy for every metric, so both are read across metrics. All measures are averaged over the 11 seasons; intervals are 90% intervals from 2,000 resamples of seasons.

4. How Fast Plus-Minus Metrics Stabilize

Table 3. Games until each metric is provisionally stable (split-half r = .50) and stable (r = .75).

Metric	Provisionally stable (.50)	Stable (.75)	Seasons to stable
Plus-minus			
On-court +/-	30 games (1,320 poss.)	130 games (5,720 poss.)	1.6
WOWY (on/off)	113 games (4,960 poss.)	489 games (21,490 poss.)	6.0
aWOWY	25 games (1,120 poss.)	84 games (3,710 poss.)	1.0
HAPM (one-year)	17 games (760 poss.)	132 games (5,790 poss.)	1.6
Par ±	72 games (3,160 poss.)	316 games (13,920 poss.)	3.9
RAPM (1-year)	57 games (2,500 poss.)	388 games (17,080 poss.)	4.7
Box score and ARC			
BPM	10 games (450 poss.)	31 games (1,380 poss.)	0.4
EBV	9 games (410 poss.)	33 games (1,470 poss.)	0.4
ARC (1-year)	8 games (360 poss.)	276 games (12,140 poss.)	3.4

Split-half r: odd against even games. Within a season, each half holds k of the first $2k$ games; past one season, the halves come from pooled windows of 2 to 12 seasons (so RAPM there is multi-year RAPM). Games: possessions divided by 44, the mean on-court possessions per game. Seasons: games divided by 82. One-year HAPM's early stability comes partly from the playing-time center; Par ±, the same value before the shrink, needs 72 games. Time-decay HAPM carries earlier seasons, so its stability is not a count of games and it is not listed.

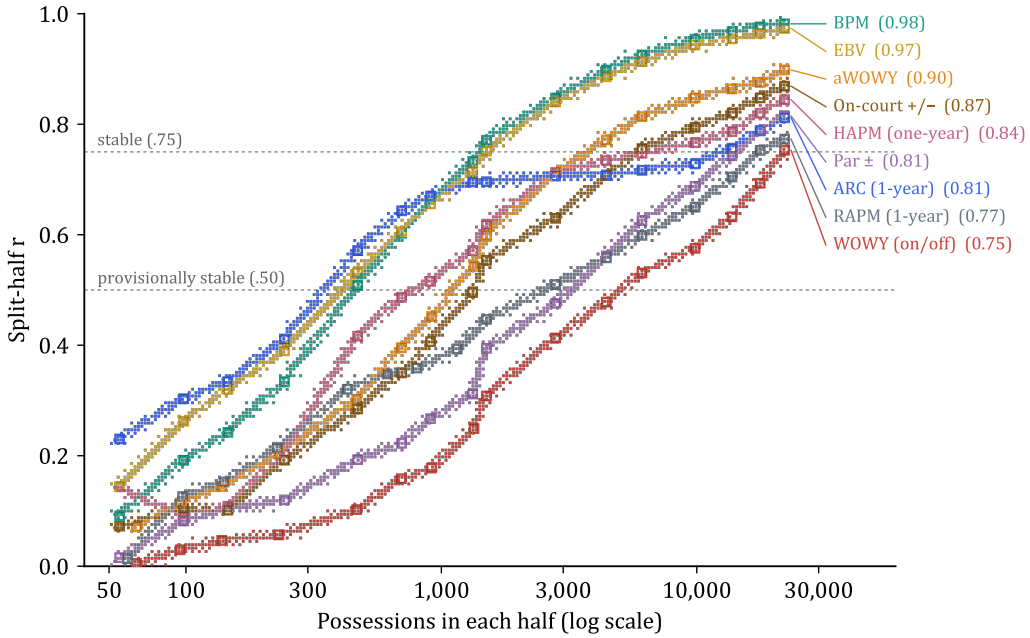

Figure 1. Measured split-half r by possessions in each half (log scale).

Squares are measured means: each season's first k games, then pooled windows of 2 to 12 seasons. The number after each name is its largest measured value.

Limits.

- A player enters the rest-of-season comparison only if he plays at least 20 possessions per remaining team-game, so players who are later benched or injured drop out. The split-half measure needs no future games and is unaffected, and the stability thresholds use it.
- A player's true level can change within a season through trades, injuries and changes of role, so the rest-of-season correlation is a lower bound on how well an early value reflects the player at that time.
- The rest of the season shrinks as k grows, which caps the rest-of-season correlation. It is therefore read across metrics, not against a threshold.
- ARC and EBV are LUMA Basketball Research models, as are the opponent ratings aWOWY uses (LOKI).

5. Descriptive and Predictive Power

Table 4. How well each metric's season-n values describe season n and predict season $n + 1$.

Metric	Describes season n (R^2)	Predicts season $n + 1$ (R^2)	Picks the winner, season $n + 1$
Plus-minus			
On-court +/-	.256	.106	63.0%
WOWY (on/off)	.032	.042	59.2%
aWOWY	.069	.081	61.9%
RAPM (1-year)	.281	.129	64.5%
Par ±	.252	.120	63.6%
HAPM (one-year)	.238	.137	64.5%
Plus-minus with earlier seasons			
RAPM (2-year)	.271	.144	65.5%
RAPM (3-year)	.267	.147	65.7%
RAPM (4-year)	.265	.149	65.9%
HAPM (time decay)	.270	.167	66.7%
Box score and ARC			
BPM	.231	.125	64.7%
EBV	.173	.149	65.1%
ARC (1-year)	.228	.153	65.6%

Nine season pairs, 2013-14 to 2025-26 (2019-20 and 2020-21 left out); every value uses games through season n only. Predicted margin: in each stint, the five players' values minus the opponents'. Describes: R^2 on season n's game margins. Predicts: R^2 on season $n + 1$'s. Picks the winner: share of season $n + 1$ games the predicted margin gets right. Players without a value count as 0 (ARC gives them its no-data value, −2 to −3). RAPM (k-year): the season and the $k - 1$ before it, weighted .7 per season back (Table 25). Par ± and one-year HAPM rate the other nine with a prior from the three previous seasons. Time-decay HAPM: Section 8.3.

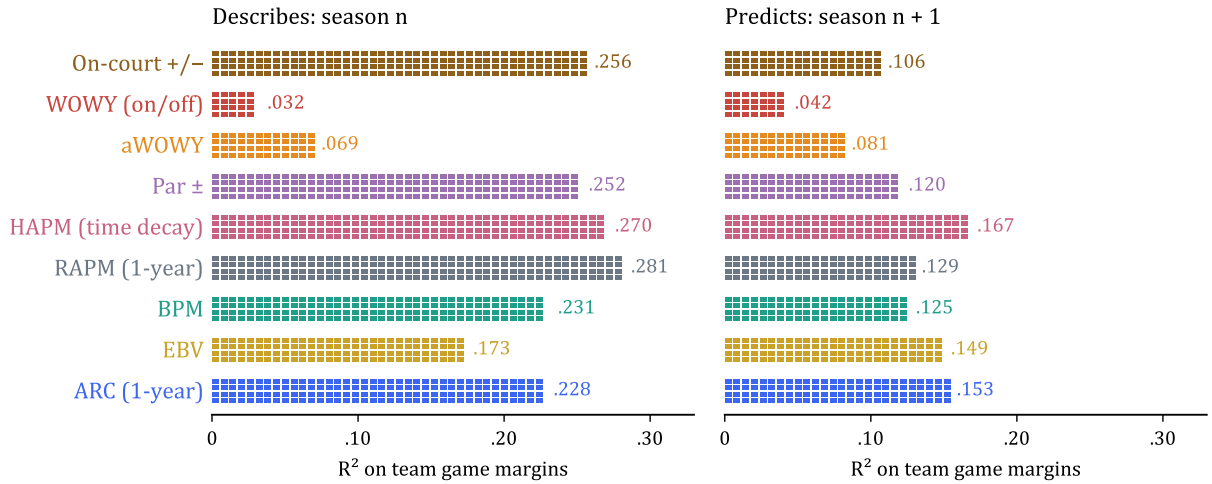

Figure 2. Selected rows of Table 4 as a chart: R^2 on team game margins.

Table 5. Game-margin R^2 (B) on the rest of the season, from each metric's values after the first k games.

k	On-court	WOWY	aWOWY	HAPM	Par ±	RAPM	BPM	EBV	ARC
1	.017	.004	.005	.011	.007	.019	.019	.024	.022
5	.038	.004	.014	.060	.024	.056	.052	.072	.058
10	.077	.004	.023	.074	.049	.089	.090	.099	.091
20	.106	.013	.040	.110	.070	.118	.110	.120	.125
30	.135	.013	.051	.140	.106	.139	.134	.134	.147
41	.147	.015	.061	.155	.114	.154	.153	.148	.165
55	.165	.025	.079	.202	.147	.188	.173	.160	.189

RAPM and ARC: 1-year versions. HAPM: one-year HAPM; Par ± is the same value before the shrink. The HAPM and Par ± columns come from a second run of the same test, in which 1-year RAPM scores .005 to .013 higher than in this table. Players without a value count as 0.

6. WOWY and Its Structural Failings

WOWY, defined as a team's net rating with a player on the court minus its net rating with him off the court in the games he played, is a tool that has increasingly been used as a rawer, quicker estimation of a player's relative value. Gaining popularity for its ease of computability and accessibility.

WOWY is also used in place of the more stable and predictive RAPM because of its utility in quantifying the value of specific lineup combinations: how pairs, trios, or a five-man unit do together. RAPM is structurally a model that allocates credit to individual players and estimates each player's net margin. Because of that, its value for any given combination is only the sum of those prior individual estimates, making it closer to a prediction than a description of how the combination actually played. WOWY, at its core, is an extremely raw description of how any combination played.

Unfortunately, in exchange for this utility, it structurally absorbs an extreme level of bias and variance. Failing even as a rough estimation of a player's value as it is heavily contaminated by variance from opponents, shooting luck, and team-depth bias.

Additionally, WOWY structurally carries a team-depth bias. Functionally, it has increasingly been used as a quick read of a player's ability in a vacuum, with the baseline being the player's own replacements. It is more of a noisy, contaminated read of a player's performance relative to his own team. Thus, players on deep teams are penalized more than players with weaker reserves, who are flattered. RAPM partly struggles with the same bias, though to a lesser degree, as RAPM solves for every player at once, effectively comparing a player alongside his replacements instead of directly substituting him. Staggered and varied rotations, lineup combinations and large sample sizes in general supply information that rounds out the effect and separates them to a degree.

As a result, WOWY fails to become even provisionally stable within a season. Of the nine metrics in this paper, it is the poorest at both description and prediction (Table 4).

This paper introduces three alternatives. aWOWY (Section 7) limits the variance by controlling for opponents and shooting luck while staying in on/off units, so its utility for lineup combinations is preserved. HAPM (Section 8.3) helps isolate a player's actual value by changing the baseline from the lineups that replace him to a counterfactual: the margin his exact stints would be expected to produce with an average player, rated 0, in his place. Par ± (Section 8.2) is HAPM unshrunk and kept in on/off units.

7. aWOWY: Context-Adjusted On/Off

Definition. aWOWY (adjusted WOWY) is WOWY computed at the stint level, with two corrections applied to every stint: opponent adjustment for the strength of the five opponents on the floor, using ratings made before the game, and rim-only luck, which credits all opponents' non-rim shooting (mid-range, three-point and free-throw makes) at league rates (Table 9). Like WOWY, it applies to one player or to any group of players, from a pair to a five-man unit. It becomes provisionally stable in 78% fewer games than WOWY (25 against 113), raises the year-to-year correlation by 69% (.325 to .550) and nearly doubles its accuracy in predicting next season's games (R^2 .042 to .081). This paper uses LOKI, LUMA Basketball Research's player ratings, from the latest weekly release before each game. aWOWY does not require LOKI. Any player rating can be used as a substitute for it (Table 19).

Prior work.

- **Luck-adjusted on/off.** Goldstein (2018a) adjusts both teams' three-point and free-throw makes in a player's season on/off to his team's season rates, and PIPM (Goldstein 2018b) and LEBRON (2020) use a luck-adjusted plus-minus as an input to a player rating. aWOWY adjusts each stint individually rather than the season total, counts every opponent shot away from the rim at the league rate, and stays in units of on/off.
- **Mid-range as luck.** entropyisrude.com (2026) adjusts on/off and two- to five-man combinations for three-point, free-throw and half of mid-range luck on both teams' shots, measured against each shooter's own expected rate. It is the closest prior work to aWOWY. aWOWY differs in that it sets every opponent make away from the rim to the league rate, on defense only, and leaves shooter quality to be absorbed by the opponent ratings.
- **Opponent strength.** RAPTOR's on-off component (Silver 2019) and HARO in hockey (Johnson 2013) adjust a player's on-court results for the opponents he faced, over a season or inside a regression. aWOWY subtracts the five opponents on the floor in every stint, on both the on and the off side, using ratings made before each game.
- **WOWY combos.** Raw on/off for any group of players is a common public tool (for example databallr's WOWY Combos and entropyisrude.com's Combo Search), and L-RAPM (Petridis and Pelechrinis 2026) rates five-man units by ridge regression. aWOWY gives the same group on/off with both corrections applied to every stint and no regression.

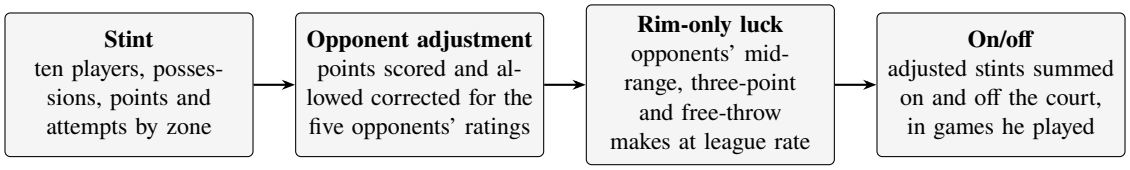

Figure 3. How aWOWY is built from each stint. For a pair or a five-man unit, "on" means all of them on the court.

Table 6. Oklahoma City's possessions, opponents and net rating by which of Gilgeous-Alexander and Hartenstein is on the floor, 2025-26.

	Both	Gilgeous-Alexander without Hartenstein	Hartenstein without Gilgeous-Alexander	Neither
Possessions	1,367	3,289	953	2,513
Net rating, raw	+13.8	**+16.9**	+17.3	−1.3
Opponents faced				
Opposing lineup strength (five opponents' ratings, summed)	+1.1	−1.6	−2.5	−6.2
Opposing players who started the game	83%	63%	59%	49%
Opponents' three-point FG% (league 36.0)	38.0	36.2	37.6	36.9
Opponents' mid-range FG% (league 44.5)	39.5	38.4	40.9	43.4
Opponents' free-throw % (league 78.3)	75.6	74.8	73.0	71.9
Opponents' non-rim points against league rates	−0.9	−4.0	−1.0	−0.7
After aWOWY's two corrections				
Net rating, adjusted	**+14.0**	+11.3	+13.8	−8.2

Every Oklahoma City stint of the 2025-26 regular season. Net ratings are per 100 possessions. Adjusted net rating = raw net rating + opposing lineup strength + opponents' non-rim points against league rates. Opponent ratings: LOKI, latest weekly release before each game.

Table 7. WOWY and aWOWY on the same tests.

Test	WOWY	aWOWY	Change
Stability			
Games to provisionally stable (.50)	113	25	78% fewer games
Games to stable (.75)	489	84	83% fewer games
Split-half r, full season	.31	.60	+94%
Rest-of-season r at 41 games	.26	.53	+104%
Year-to-year r	.325	.550	+69%
Descriptive and predictive power			
Describes season n (R^2)	.032	.069	+116%
Predicts season $n + 1$ (R^2)	.042	.081	+93%
Picks the winner, season $n + 1$	59.2%	61.9%	+2.7 percentage points

Sources: Table 3, Table 4 and Appendix B. Year-to-year: nine season pairs, players with at least 1,000 on-court possessions in both seasons (Table 9).

Table 8. Split-half r of WOWY and aWOWY for pairs and five-man units, odd against even games of the full season.

Group	Poss. per half	Groups	WOWY	aWOWY
Two-man pairs (teammates)				
150 or more	375	1,774	.13	.32
300 or more	564	1,093	.18	.39
Five-man units				
50 or more	85	203	.03	.06
100 or more	165	80	.06	.13

Rows: the minimum possessions a group needs in each half. Poss. per half: the median group's possessions in its smaller half. Groups: number per season. WOWY and aWOWY: the group's on-court net rating minus the team's net rating when the group is not all on the court, in games the group played.

Table 9. On/off with different sets of opponent makes counted at league rates.

	Same season			Next season	
Opponent makes at league rate	**Split-half** r	**Rest-of-season** r	**Game** R^2	**Game** R^2	**Year-to-year** r
No opponent adjustment					
None (WOWY)	.31	.26	.015	.042	.325
Three-point and free throw	.37	.31	.024	.056	.404
Mid-range, three-point and free throw (rim-only)	.40	.35	.028	.064	.416
With the opponent adjustment					
None	.47	.42	.044	.067	.471
Three-point and free throw	.57	.49	.056	.076	.542
Mid-range, three-point and free throw (aWOWY)	.60	.53	.061	.081	.550

Split-half: odd against even games of the full season. Rest of season and game R^2: values after 41 games against the rest of the season. Next season: full-season values on the next season's game margins (nine season pairs). Year-to-year r: players with 1,000+ on-court possessions in both seasons.

8. Introducing Held-Out Estimation, Par ±, and HAPM

8.1 Held-Out Estimation

Prior work: joint estimation. To make held-out estimation easier to understand, this section first defines its counterpart, which this paper calls joint estimation: adjusted plus-minus (APM; Rosenbaum 2004), regularized adjusted plus-minus (RAPM; Sill 2010) and the lineup regressions inside all-in-one metrics. Joint estimation quantifies every player's value at once by fitting one regression of stint margins on the ten players on the floor. Least squares allocates credit among all the players together, choosing the coefficients whose sums best reproduce the stint margins. A player's value then becomes his coefficient in the fit.

Held-out estimation. Held-out estimation quantifies a player's impact as the difference between the margins his stints actually produced and the margins expected from the other nine players in them (rated without his data) with an average player in his place.

How it is built.

1. *Hold out the measured player.* Remove every stint the measured player was on the floor for.
2. *Rate the other players.* Rate every other player from the remaining stints, with any rating model.
3. *Set the expected margin.* For each of his stints, his four teammates' ratings minus the five opponents', with 0 in his place, give the expected margin.
4. *Compare with the actual margin.* His value is the actual margin minus the expected margin, averaged over his stints and weighted by possessions (Figure 4).
5. *Repeat for every player.* Steps 1 to 4 run once for every player.

Changing the question answered. One of the foundational reasons WOWY fails as a measure of a player's value is that it compares the player with his own replacements (Section 6). Joint estimation (APM, RAPM) adjusts for teammates and opponents, but in practice it still asks, to a lesser degree, how much better a player is than the lineups that replace him, and credits him on that basis. Held-out estimation attempts to isolate a player's value from his replacements by structurally asking a different question: how did the actual results of a player's stints compare with what those stints would have been expected to produce with an average player in his place?

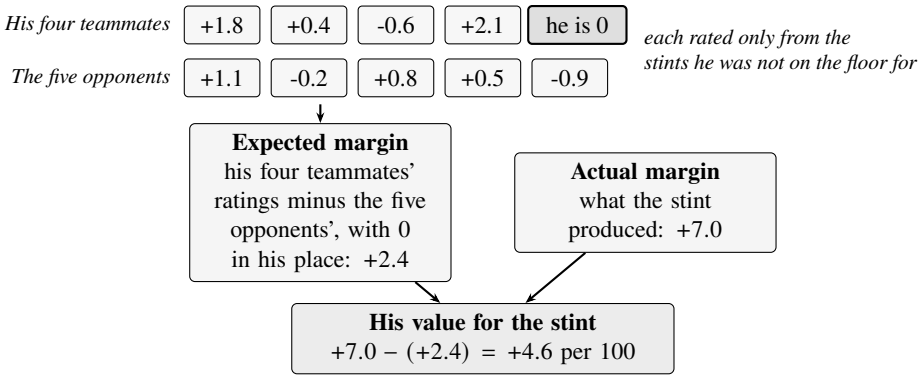

Figure 4. How held-out estimation scores one stint (illustrative numbers).

The nine ratings come from data without any of the measured player's stints. Each of his stints is scored this way; his value is the possession-weighted average of the stint values.

Benefits of held-out estimation.

- Additivity over stints
- Error isolation
- Input versatility

Additivity over stints. When held-out estimation is used properly, each of the measured player's stints is given its own value: the actual margin minus the expected margin. His value over any set of stints is the possession-weighted average of those stint values. A season is one such set, as are his stints with a given teammate. Every such value comes from the same stint values, without needing a new fit.

Conversely, a joint solve has no individual stint values to average, so a value for each subset would need a separate regression. That separate regression would have less data to rate each player.

Table 10. Jaylen Brown's Par ± over different sets of his 2024-25 stints.

His stints	Possessions	Actual	Expected	Par ±	SE
Season	4,318	+8.9	+8.6	+0.3	3.2
With Tatum	2,860	+7.8	+9.0	−1.2	3.7
Without Tatum	1,458	+11.0	+7.8	+3.2	4.8
Close games	125	+7.5	+11.5	−4.0	14.9
Against top-5 defenses	509	+4.6	+4.6	+0.0	7.6
Against top-10 teams	1,149	+6.4	+1.2	+5.3	5.3
Leading by 15 or more	703	+3.0	+2.3	+0.7	6.5
Trailing by 15 or more	110	+14.4	+17.9	−3.5	15.8
First half of season	2,464	+6.7	+9.4	−2.7	3.9
Second half of season	1,854	+11.8	+7.6	+4.2	4.3

Per 100 possessions. Actual: the team's margin in those stints. Expected: the margin expected from the other nine players with an average player in his place. SE: standard error, approximate for subsets. Close games: the last five minutes of the fourth quarter or overtime, within 5 points. Leading or trailing: his team's margin when the stint starts. Top-5 defenses and top-10 teams: by defensive rating and net rating; Boston, his own team, is never his opponent.

Error isolation. With the other nine rated without him, the residual of each of his stints is credited to him alone, not shared among the ten players on the floor. His value, therefore, structurally does not depend on how a regression fit allocates credit within a stint between him and a teammate who shares most of his minutes, the collinearity problem that joint estimation deals with. This does not entirely remove the effect of collinearity; for such a pair, it shows up instead as increased variance in his measured value rather than as an arbitrary allocation of credit (Limits, below).

Input versatility. Held-out estimation is any mechanism that makes an initial estimation of the other nine without the measured player's stints and then attributes the rest of the margin to him. Any rating model can be used for that initial estimation, including RAPM, a box score metric or a statistical plus-minus metric. Par ± and HAPM, the two held-out models introduced in this paper, use a prior-informed RAPM for their initial estimation (Sections 8.2 and 8.3).

All things held equal, held-out estimation is only as good as the rating used to estimate the other nine.

Table 11 compares two versions with the same prior-informed RAPM rating the other nine, one with the hold-out step and one fitted jointly: the hold-out step adds .009 to next-season R^2 and .015 to same-season R^2, in all nine season pairs, and leaves both biases about where they were.

Table 11. The same prior-informed RAPM with the measured player held out or fitted jointly.

Version	Describes season n	Predicts season $n + 1$	Teammate bias	Team-depth bias
Held out	.238	.137	+.00	−.09
One joint solve	.223	.129	−.03	−.08
Difference	+.015 [+.012, +.018]	+.009 [+.006, +.011]	−.02 [−.04, +.03]	+.01 [+.00, +.02]

Describes and predicts as in Table 4, with 90% paired intervals; differences use unrounded values; held out is higher in all nine season pairs. Biases as in Table 28. Bias difference: held out's distance from 0 minus the joint solve's. Both versions rate the other players with the same three-season prior and shrink the measured player toward a playing-time center.

Prior work.

- **Fearnhead and Taylor (2011).** Fearnhead and Taylor use a similar mechanism in the NBA: a player is measured on one season, while prior history estimates the other players. Their prior-history estimation, however, includes the measured player's stints. Held-out estimation structurally derives its initial estimation only from stints without the measured player.
- **Teacher value-added.** Chetty, Friedman and Rockoff (2014) predict a teacher's effect on one year's students from her classes in other years, leaving out the year being judged. Held-out estimation is essentially this framework adapted to the NBA, with the player held out instead of the year.

Limits of held-out estimation.

- *Collinearity in tight pairs.* Holding a player out does not remove collinearity; it moves the error out of the split between the pair and into each player's own measured value. When two players share most of their minutes, joint estimation estimates their combined value relatively accurately. The error stems from how it allocates credit between them, as the ridge penalty in RAPM structurally pulls the split toward an even share (De Bacco, Wang and Blei 2024).

  Held-out estimation instead rates the other nine, including the teammate, only from the stints without him. For a tight pair, said stints are few, making both the estimation of the other nine and the player's eventual measured value less certain.

  At the same playing time, the standard error of Par ± is up to about 10% wider for players who share 75% or more of their minutes with one teammate (Table 12).

- *Rating error.* Given enough stints, the variance in the measured player's stint values starts to stabilize, though any error in the estimation of the other nine does not. It persists, and becomes a larger share of the error that remains.

- *Values do not add up.* Each player is credited with the full residual of his own stints, so the values of players who share the floor do not add up to their stints' margin. In other words, you cannot add up HAPM to get a team's net rating.

Table 12. Par ± standard error by possessions played and by the share of a player's possessions played with one teammate.

Possessions played	Under 60% shared	60–75% shared	75% or more shared
1,500–2,500	4.04	4.06	4.15
2,500–3,500	3.41	3.51	3.59
3,500–4,500	3.10	3.22	3.34
4,500 or more	2.93	3.09	3.24

Mean standard error of Par ±, per 100 possessions; 1,291 player-seasons with 1,500+ possessions, 2022-23 to 2025-26. Shared: the largest share of his possessions played with one teammate. Each cell has 29 to 246 players.

The following two subsections introduce two prior-informed implementations of held-out estimation:

- **Par** ± (Section 8.2): a season-by-season, unshrunk measure of how far a player's stints finished above or below expectation, per 100 possessions, with its standard error.
- **HAPM** (Section 8.3): Par ± shrunk toward a prior value according to its possession sample size, more strongly for players with fewer possessions.

8.2 Par ± Preview

Definition. Par ± is held-out estimation with a prior-informed RAPM (PI RAPM; Engelmann 2011; Section 2) for the initial estimation of the other nine, over one season, unshrunk. It is the difference between the actual results of a player's stints and what they were predicted to produce, per 100 possessions. With the other nine being rated with a ridge regression solved in two steps, both without his stints:

1. A fit on the three previous seasons, with recent seasons weighted more, sets a prior for each of the nine.
2. A fit on the current season's stints pulls each of the nine toward the prior of step one (Appendix D).

Table 13. Denver's 2023-24 starters, rated with Jokić held out and with Caldwell-Pope held out (per 100 possessions)

	Jokić held out		Caldwell-Pope held out	
Player	**Prior**	**Rating**	**Prior**	**Rating**
Nikola Jokić	held out		+6.99	+7.74
Kentavious Caldwell-Pope	+0.34	+0.95	held out	
Aaron Gordon	+2.08	+2.12	+5.52	+5.50
Michael Porter Jr.	+0.91	+0.69	+3.10	+2.19
Jamal Murray	−0.09	−0.59	+0.95	+0.90
The other four starters, summed		**+3.17**		**+16.33**

Prior: the starting value from the three previous seasons. Rating: after the 2023-24 stints. Both leave out every stint of the held-out player. Gordon's prior falls from +5.52 to +2.08 with Jokić held out because most of his previous three seasons were played next to Jokić.

How Par ± is built: a worked example. Figure 5 scores one of Jokić's 1,388 stints with those ratings.

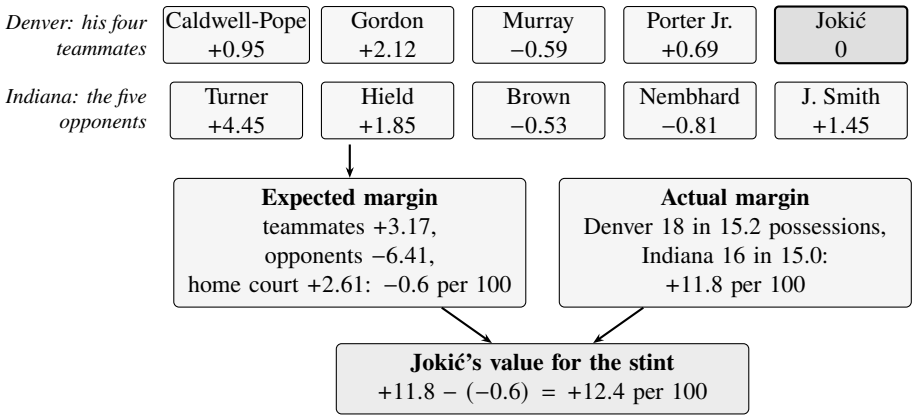

Figure 5. Scoring one Jokić stint: Denver's opening stint at home against Indiana, 2023-24 (7 minutes).

Ratings per 100 possessions, from Table 13 (Jokić held out). Expected margin: the teammates' ratings minus the opponents', plus home court; the score at the start of a stint also enters when it is not tied.

His Par ± is the possession-weighted average of all 1,388 stint values (Figure 6).

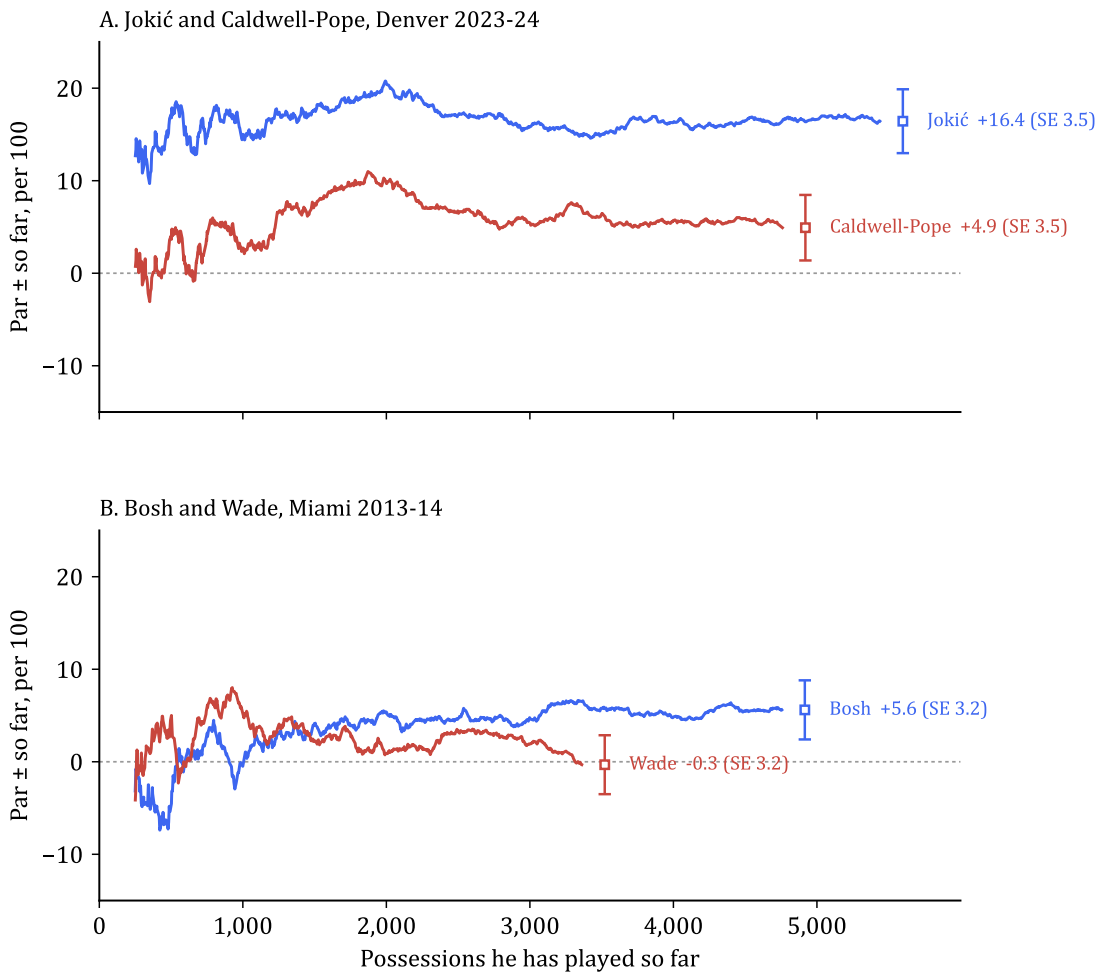

Figure 6. Par ± as it builds over one season, for a tight pair (A) and a pair whose order flipped (B).

Each point is the possession-weighted average of his stint values so far; the other players' ratings use the full season. Square: his final value, with its standard error. Share of his possessions played with his partner: Jokić 80%, Caldwell-Pope 91%, Bosh 49%, Wade 70%.

Jokić and Caldwell-Pope are an edge case; an extremely tight teammate pair whose values rose and fell together with Denver's results. While joint estimation methods place them close together (1-year RAPM +5.2 and +4.4, Table 14), held-out estimation keeps them separate for the entire season. Their values demonstrate just how rare they are. Consider that their values moved together more closely from game to game than those of 99% of 6,582 teammate pairs from 2013-14 to 2025-26. Panel B shows a pair that shared fewer minutes: Wade led Bosh early in 2013-14, and Bosh finished 5.9 ahead.

Table 14. Par ± for Nikola Jokić and Kentavious Caldwell-Pope, Denver 2023-24 (per 100 possessions)

	Both	**Jokić without Caldwell-Pope**	**Caldwell-Pope without Jokić**	**Neither**
Possessions	4,352	1,087	408	1,944
Net rating, raw	+11.8	+8.4	+6.3	−11.4
	Jokić		**Caldwell-Pope**	
Actual margin of his stints	+11.11		+11.33	
Expected with an average player in his place	−5.32		+6.41	
from his four teammates	+1.34		+12.97	
from the opponents and game context	−6.66		−6.56	
Par ± (actual minus expected)	**+16.43** (SE 3.5)		**+4.92** (SE 3.5)	
RAPM (1-year), percentile	+5.2 (99th)		+4.4 (97th)	

Denver stints of the 2023-24 regular season with at least 15 seconds and 0.3 possessions (the stints the Par ± fit uses). Net rating: each team's points per 100 of its own possessions. Expected: each of his stints predicted from the other nine players' ratings, fit without his possessions, plus home court and score. RAPM settings: Appendix D. Percentiles: 2023-24 players with 1,500+ possessions.

Uses.

- *Any set of his stints, without a new fit:* close games, against top-10 teams, leading or trailing by 15 or more, with or without a given teammate (Table 10).
- *Any time period:* a month, a stretch of games or each half of the season, from the same stint values.
- *Input for other models:* an unshrunk value with its standard error.
- *Teammate fit* (exploratory).

Par ±, WOWY and aWOWY. This paper now covers three direct measures of how a player's minutes went: WOWY, aWOWY and Par ± (Table 15). WOWY continues to be the worst of the three at describing games, stability and prediction. aWOWY fixes WOWY's poor stability and variance by adjusting for opponents and shooting, but it does not adjust for teammates. The lack of teammate isolation provides more information to help keep it more stable than Par ±.

Par ± also addresses WOWY's poor stability and variance, though to a lesser degree and with considerably less bias, as it isolates what a player adds to a stint via held-out estimation. Par ± describes the season and predicts the next one much better than both WOWY and aWOWY.

Table 15. How Par ±, WOWY and aWOWY are built and how they test.

	Par ±	**WOWY**	**aWOWY**
Compared with	An average player	His replacements	His replacements
Adjusts for opponents	Yes	No	Yes
Adjusts for shooting luck	No	No	Yes
Adjusts for teammates	Yes	No	No
Measures	A player	A player or a group	A player or a group
Standard error	Yes	No	No
Games to provisionally stable (.50)	72	113	25
Games to stable (.75)	316	489	84
Describes season n (R^2)	.252	.032	.069
Predicts season $n + 1$ (R^2)	.120	.042	.081
Teammate bias	+.00	+.20	+.23
Team-depth bias	−.06	−.21	−.30

Stability as in Table 3, describing and predicting as in Table 4, biases as in Table 28 (0 means none). Shooting luck: opponents' non-rim shooting (Section 7). A group: any set of teammates, from a pair to a five-man unit.

aWOWY still has three uses:

- Any group of players, from a pair to a five-man unit.
- A stable early-season read.
- Simplicity of computation.

Limits.

- *Noisy estimate of player value.* Over a full season, Par ± has a split-half r of .39, measured as in Table 3. A value should be read with its standard error, and rankings built on it should be met with caution.
- *Specific to his stints.* Par ± measures what a player added over an average player in his exact stints. Small sets of stints carry wide standard errors: Brown's −4.0 in close games (Table 10) has a standard error of 14.9.

From Par ± to HAPM. Par ± is a good raw estimate, but it is noisy and unstable. Shrinking it toward a prior gives HAPM (Section 8.3): a more stable, descriptive and predictive estimate of a player's value over a given period.

8.3 HAPM Preview

Definition. HAPM (held-out adjusted plus-minus) is Par ± shrunk toward a prior value. HAPM pulls each measured player's value toward a chosen prior in proportion to its uncertainty, trading bias for decreased variance and increased stability.

How HAPM is built.

1. *Take his Par* ± (Section 8.2) and its standard error for each season.
2. *Choose the seasons.* Use the current season's Par ± alone, or pool several of his seasons, weighting each by its recency and by the inverse of its squared standard error.
3. *Choose the prior.* This paper uses a center set by playing time: the average Par ± of players with the same number of possessions in the other seasons.
4. *Shrink.* HAPM is the prior plus a weight times (Par ± − prior). The weight is larger when Par ± has a smaller standard error.

Versions. This paper uses two main versions of HAPM. One-year HAPM shrinks the current season's Par ±. Time-decay HAPM pools every season of his Par ±, weighting each .63 per season back (a half-life of 1.5 seasons), and shrinks the pooled value. One-year HAPM isolates how a player performed in a given season. Time-decay HAPM gives a live, continuous estimate of his current value, with recent seasons weighted the heaviest.

Table 16. HAPM for Nikola Jokić and Kentavious Caldwell-Pope, Denver 2023-24 (per 100 possessions)

	Jokić	**Caldwell-Pope**
Par ±, 2023-24 (standard error)	+16.4 (3.5)	+4.9 (3.5)
Prior: center for his playing time	+5.1	+3.9
One-year		
Weight on Par ±	.515	.503
HAPM, one-year	+10.9	+4.4
Time decay		
Par ± by season, 2023-24 back to 2020-21	+16.4, +16.8, +12.4, +8.9	+4.9, +3.8, −2.4, +0.7
Share of the pool from 2023-24	37%	33%
Pooled Par ± (standard error)	+13.8 (1.6)	+2.6 (1.6)
Weight on the pooled value	.823	.839
HAPM, time decay	**+12.3**	**+2.8**

HAPM = prior + weight × (Par ± − prior). The pool includes every earlier season (Jokić from 2015-16, Caldwell-Pope from 2013-14); seasons before 2020-21 hold 15% (Jokić) and 17% (Caldwell-Pope) of it.

Table 17. Three HAPM versions and three RAPM versions on the same tests.

	HAPM			RAPM		
	One-year	**Two seasons**	**Time decay**	**1-year**	**2-year**	**4-year**
Split-half r after 41 team games	.62	.75	.85	.45	–	–
Describes season n (R^2)	.238	.260	.270	.281	.271	.265
Predicts season $n + 1$ (R^2)	.137	.158	.167	.129	.144	.149
Teammate bias	+.00	−.02	−.03	+.23	+.08	+.02
Team-depth bias	−.09	−.05	−.03	−.10	−.03	−.01
Value lost or gained with his availability (1 = right)	1.12	1.02	1.00	4.23	2.42	2.01

Split-half r: odd against even games of the first 82; the pooled HAPM versions share earlier seasons in both halves; –: not run. Describes and predicts as in Table 4; biases as in Table 28. Two seasons: last season weighted .5. 2-year and 4-year RAPM: earlier seasons weighted .7 per season back. Value lost or gained: the actual change in team margin when he sits divided by the forecast change (Appendix E).

Two forms of HAPM: one-year HAPM and time decay. The held-out mechanism allows one-year HAPM to use the teammate isolation that prior history provides to quantify a player's performance within a single season. To get a similar teammate isolation effect, joint estimation needs a pooled estimate of every player in the regression, including the measured player, so his value becomes more of a pooled estimate over time than a single-season quantification of his performance. Compared with Par ±, one-year HAPM accepts a small bias toward the chosen prior in return for enough stability for good-faith player comparisons and rankings.

Time-decay HAPM pools his history with a half-life of 1.5 seasons to give a strong, live, isolated estimate of how good a player is at any given time. It predicts the next season's game margins 16% better than 2-year RAPM and 12% better than 4-year RAPM (R^2 .167, against .144 and .149), describes the current season about as well (.270, against .271 and .265), and keeps similar isolation (Table 17).

Uses.

- Approximating the value lost or gained with a player's availability (Appendix E).
- A value to rank players on, with a standard error.
- Input for other models.

Prior work.

- *Improving Plus-Minus* (Fyffe and Vollman 2002, hockey) compared a player's plus-minus with an expected plus-minus, the margin an average player on his team would have in his minutes. The baseline is his own team, which includes him; HAPM's expectation comes from the other nine players, rated without him.
- *Teammates rated apart from the player.* Tango's *Evaluating Catchers* (the with-or-without method), Johnson's relative-to-teammates measures in hockey (HARO, 2013) and RAPTOR's on-off component (Silver 2019) judge a player against teammates measured in their time apart from him. HAPM rates the other nine in a regression fitted without his possessions, with the opponents and the context, stint by stint.
- *On Estimating the Ability of NBA Players* (Fearnhead and Taylor 2011, Journal of Quantitative Analysis in Sports) rated a player from the latest season only while the other players used every season, the closest precedent for one-year HAPM's treatment of earlier seasons. Their teammates' ratings still included the measured player's possessions.

- *Time decay. Time-Decay RAPM* (nbarapm) uses a player's full history but exponentially down-weights older possessions, so the estimate reflects the current player while borrowing stability from the past. DARKO (Medvedovsky) projects players daily from box score and plus-minus data with exponential decay and Kalman filters, weighting recent games most. Time-decay HAPM applies the same idea after the hold-out, decaying his seasons of Par ± rather than the possessions inside a joint fit.
- *Leave-out estimators. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates* (Chetty, Friedman and Rockoff 2014) scores each teacher's year with that year's data left out, so the same errors do not sit on both sides of the comparison. *Leave-Out Estimation of Variance Components* (Kline, Saggio and Sølvsten 2020) uses leave-out estimates to remove the bias that noise adds to estimates of how much units differ. HAPM applies the same motive to a player and his teammates.
- *Players who change teams. Revisiting Relative Shot Metrics* (EvolvingWild 2018, Hockey Graphs) tested whether a relative measure follows the strength of a player's new teammates, and *Total Hockey Rating* (Schuckers and Curro 2013) chose its ridge penalty so that traded players' ratings stay consistent. Table 28 applies the first test to the NBA.
- *Modeling Player and Team Performance in Basketball* (Terner and Franks 2021, Annual Review of Statistics and Its Application) notes that RAPM reduces variance "by biasing the coefficients toward zero". In RAPM, one penalty shrinks every player the same way; HAPM shrinks the held-out player separately, toward a center set by his possessions.

Limits. HAPM is a version of held-out estimation that uses PI RAPM as the initial estimation method, shrunk toward a prior. This paper showcases LUMA Basketball Research's implementation, but the settings are up to discretion. The limits below are those of this implementation.

- *Isolation comes from prior history.* The isolation effects of HAPM are derived from the multi-year prior history used in the initial estimation, rather than the held-out mechanism itself. The 4-year RAPM reaches similar teammate bias without holding anyone out (+.02; Table 17).
- *Sizing and team strength.* The sizing shown in this paper comes from the playing-time center, and the center chosen ignores team strength. Regulars on good teams are undervalued and regulars on bad teams are overvalued (slopes 1.22 and .67; Appendix E). Shrunk toward zero, as RAPM is, the forecasts for top players' absences fall short of the actual change (slopes 2.20 for one-year HAPM and 1.29 for time decay; .94 and .91 with the center).
- *Part of the prediction edge is memory.* Against controls that also remember past seasons, the edge is a tie: a RAPM shrunk toward its own previous season predicts at .165, and a PI RAPM with a stronger history prior at .160 to .161, with absence slopes of .90 to 1.00. A PI RAPM with HAPM's time decay has not been tested.
- *Shooting luck.* This implementation uses actual shooting, so opponents' shooting luck stays in its defensive values. A shooting-adjusted version of HAPM can be built, as aWOWY adjusts for opponents' shooting (Section 7).
- *Values do not add up* to a team's margin, as with any held-out estimate (Section 8.1).

Closing. Questions and replications are welcome at courtshare78@gmail.com or on X at @Starro____. Thank you for reading.

A. aWOWY Checks

Table 18. WOWY, each of aWOWY's two corrections alone, and aWOWY after k games.

k	Poss.	WOWY	Rim-only luck	Opponents	aWOWY (both)
Rest-of-season r					
1	63	.06	.10	.14	.20
5	236	.14	.18	.24	.33
10	464	.17	.24	.30	.41
20	900	.26	.33	.40	.50
30	1,350	.25	.35	.42	.54
41	1,829	.26	.35	.42	.53
55	2,408	.25	.33	.40	.50
Split-half r (k games in each half)					
1	63	.01	.01	.05	.07
5	236	.06	.10	.12	.20
10	464	.10	.16	.19	.30
20	900	.18	.27	.31	.45
30	1,350	.25	.34	.41	.54
41	1,829	.31	.40	.47	.60
Game-margin R^2 on the rest of the season					
1	63	.004	.005	.004	.005
5	236	.004	.007	.007	.014
10	464	.004	.011	.013	.023
20	900	.013	.021	.030	.040
30	1,350	.013	.025	.037	.051
41	1,829	.015	.028	.044	.061
55	2,408	.025	.039	.060	.079

Poss.: the median player's on-court possessions in the first k games. Rim-only luck: opponents' mid-range, three-point and free-throw makes at league rates. Opponents: each stint's margin minus what the opponents' ratings as of that game predict. Split-half at $k = 41$: the full season, odd against even games.

Table 19. aWOWY with other opponent ratings and rim weights.

	After 10 games			After 41 games		
Version	**Rest-of-season** r	**Split-half** r	**Game** R^2	**Rest-of-season** r	**Split-half** r	**Game** R^2
WOWY, no adjustment	.17	.10	.004	.26	.31	.015
Opponent ratings, with rim-only luck						
None (rim-only luck alone)	.24	.16	.011	.35	.40	.028
Opponent team's net rating to date	.25	.16	.010	.36	.41	.028
Opponents' prior-season RAPM (2-year)	.32	.23	.016	.44	.50	.044
Opponents' prior-season RAPM (3-year)	.33	.23	.017	.45	.51	.045
Opponents' LOKI, preseason release	.37	.26	.021	.49	.55	.055
Opponents' LOKI, latest weekly release (aWOWY)	.41	.30	.023	.53	.60	.061
Rim weight, with opponents' weekly LOKI						
.25	.34	.23	.016	.46	.52	.050
.50	.37	.26	.019	.50	.56	.054
.75	.39	.29	.021	.52	.58	.058
1.00 (aWOWY)	.41	.30	.023	.53	.60	.061

Game R^2: on the rest of the season. Split-half: the full season, odd against even games. Every version except WOWY and "None" adjusts each stint for the five opponents. Rim weight w: opponents' non-rim makes counted as $(1 - w)$ times actual plus w times the league rate. Team net rating: the opponent's season to date, split evenly over its five players. Prior-season RAPM: the previous two or three seasons, recent seasons weighted more.

Table 20. aWOWY with opponents' non-rim makes at league rates or at each shooter's own rate.

	Same season			Next season	
Opponents' mid-range, three-point and free-throw makes	**Split-half** r	**Rest-of-season** r	**Game** R^2	**Game** R^2	**Year-to-year** r
League rate (aWOWY)	.603	.533	.061	.081	.550
Shooter's own rate, half shrinkage	.590	.521	.058	.080	.542
Shooter's own rate, standard shrinkage	.594	.524	.059	.081	.544
Shooter's own rate, double shrinkage	.597	.527	.060	.081	.546
Free throws at the shooter's rate, rest at league rate	.601	.532	.060	.081	.549

Same stints, players and tests as Table 9, except split-half r, which uses the 41-game split (aWOWY .603 here, .599 on the full season). Shooter's own rate: his rate before the game (season to date plus .7 times the previous season), shrunk toward the league rate by adding 400 mid-range, 750 three-point and 200 free-throw attempts at league average (standard), or half or double that.

B. Correlations at Every Window

Table 21. Split-half correlation (S) with k games in each half.

k	On-court	WOWY	aWOWY	HAPM	Par ±	RAPM	BPM	EBV	ARC
1	.07	.01	.07	.14	.02	.01	.09	.14	.23
2	.10	.03	.11	.09	.08	.13	.19	.26	.30
5	.19	.06	.20	.20	.12	.21	.33	.39	.41
10	.28	.10	.30	.42	.19	.32	.51	.53	.57
20	.41	.18	.45	.52	.27	.36	.66	.66	.67
30	.50	.25	.54	.57	.31	.39	.73	.71	.70
41	.55	.31	.60	.62	.39	.45	.77	.75	.70

$k = 41$ is the full season split into odd- and even-numbered games. RAPM and ARC are the 1-year versions; HAPM is one-year HAPM; Par ± is the same value before the shrink. At $k = 1$, one-year HAPM's split-half r comes from the playing-time center: Par ± itself correlates .02 across the halves, while possessions in the two halves track each player's role ($r = .49$), because most teams with a game in both halves have exactly one in each. At $k = 2$ and 3, odd and even league games split teams' games unevenly, possessions per half stop tracking role ($r = .06$), and the center's part drops out. Par ± and RAPM rise at every k.

C. Box Score Metrics

Table 22. How EBV prices each box score event.

Event	Side	How it is priced
Counted (play-by-play box score counts)		
Points	O	Face value
Shooting efficiency	O	A charge on true shooting relative to league average, scaled by usage and shrunk for low shot volume; the usage scaling was chosen by a test against later RAPM
Assists	O	Points the pass created beyond the expected value of the shot, from NBA passing tracking (2016-17 on), one price applied to every season
Turnovers	O	The possession's expected value plus the value of the opponent's next attempt
Offensive rebounds	O	Expected value of the team's next attempt
Defensive rebounds	D	A fixed allocation, not measured
Steals	D	The possession gained plus the value of the attempt after a stolen turnover
Blocks	D	The share of blocked shots the defense recovers times the value of an attempt
Personal fouls	D	Free-throw points conceded per foul
Not counted (tracking data)		
Screen assists, potential assists, secondary assists, deflections, loose balls recovered, box outs, charges drawn, contested shots, matchup (who-guarded-whom) data		

Values are per 100 of the player's on-court possessions in the window, with no shrinkage: the same calculation as the published EBV. Points stay with the scorer and the assist credit is added on top, so EBV does not sum to team points.

Table 23. BPM 2.0 as rebuilt from the window's box score counts (Basketball-Reference specification).

Term (per 100 team possessions)	Position 1 / role 1	Position 5 / role 5
Points, adjusted for team context	0.860	
Three-pointers made	0.389	
Assists (by position)	0.580	1.034
Turnovers	−0.964	
Offensive rebounds (by position)	0.613	0.181
Defensive rebounds (by position)	0.116	0.181
Steals (by position)	1.369	1.008
Blocks (by position)	1.327	0.703
Personal fouls	−0.367	
Field goal attempts (by role)	−0.560	−0.780
Free throw attempts × 0.44 (by role)	−0.246	−0.343
Position constant	$-0.818 \times (3 - \text{position})/2$ below position 3	
Role constant	$2.774 \times (\text{role} - 3)/2$	
Team adjustment	Each team's five-man average set to its net rating in the window	

Position (1 = point guard to 5 = center) and offensive role (1 = creator to 5 = receiver) are estimated from the player's share of his team's rebounds, steals, fouls, assists, blocks and points, normalized to a minutes-weighted team average of 3 and clamped to 1–5. Coefficients between 1 and 5 are linear interpolations.

D. RAPM Settings

Table 24. RAPM (1-year) settings.

Setting	Value
Rows	One per stint side: the offense's possessions in a stint
Outcome	Points per 100 possessions on those possessions
Weights	Possessions
Columns	One offensive and one defensive coefficient per player (+1 for the offense five, −1 for the defense five), an intercept and home court
Penalty	Ridge, λ = 3,000 possessions on the player coefficients; intercept and home court unpenalized; no team terms
Prior	None (coefficients shrink toward 0)
Value	Offensive plus defensive coefficient, per 100 possessions

λ = 3,000 matches Sill's (2010) cross-validated value for three seasons (2,000 for one); it was not retuned for one season. The comparators are specified in full so they can be reproduced. For HAPM and its prior-informed RAPM, the paper gives the quantity each estimates, the windows and the form of the prior (Table 26).

Table 25. RAPM (k-year) settings, k = 2, 3, 4.

Setting	Value
Rows, outcome, columns, value	As RAPM (1-year) (Table 24)
Seasons	$n - k + 1$ to n, in one regression; nothing after season n
Weights	Possessions × .7 per season back
Penalty	Ridge, λ = 3,000
Selected	.7 per season back and λ = 3,000 for every k in every season, chosen out of sample over decays of 1, .7 and .5 per season back, half-lives of 615, 1,230 and 2,460 league games, and λ = 1,500, 3,000, 6,000 and 12,000

Table 26. PI RAPM (prior-informed RAPM) settings.

Setting	Value
Rows, outcome, columns, value	As RAPM (1-year) (Table 24)
Seasons	Season n's stints alone
Shrinkage	Each coefficient shrinks toward a prior mean, not toward 0
Prior mean	The player's rating from one ridge regression (λ = 3,000) on the stints of the three previous seasons, with possessions weighted 1, .3 and .09 from the most recent season back. That rating is multiplied by a factor that rises with his possessions in those seasons, from 0 for a player with none, and a playing-time term that also rises with those possessions is added. Both are fitted out of sample, by season and separately for offense and defense. A player with no earlier seasons gets the playing-time term alone, about −2.9 per 100
Prior spread	Narrower for players with more possessions in the three previous seasons

E. The Absence Test

When a player misses a game, his team loses his value minus the value of the teammates who take his minutes. The test compares that forecast with what happened.

- *Absences:* games missed by a team's top eight players by possessions, each with at least 20 games for the team, 2013-14 to 2025-26; 30,413 player-games.
- *Forecast change:* the team's available players at their usual minutes, with the absent player's minutes given to the teammates who usually take them. Each value comes from the other half of the season, so no value is scored on the games it was fit on.
- *Actual change:* the team's margin, compared within team and half-season, with home court and opponent strength held fixed.
- *Slope:* the actual change divided by the forecast change. 1 is the right size; above 1, the metric underrates the absent players; below 1, it overrates them.

Table 27. Absence-test slopes for time-decay HAPM, by who sits and by his team's strength.

Absences	Slope (90% interval)
All	1.00 [0.88, 1.13]
Who sits	
Top player by possessions	0.91 [0.77, 1.07]
Second	0.87 [0.69, 1.05]
Third to fifth	0.97 [0.83, 1.12]
Sixth to eighth	1.27 [0.75, 1.90]
His team this season	
Good (net rating above +2.5)	1.22 [1.03, 1.42]
Middle	1.09 [0.86, 1.32]
Bad (net rating −1.4 or below)	0.67 [0.47, 0.88]
Minutes in the forecast	
Actual minutes in that game	0.67 [0.59, 0.75]

Slope: the actual change divided by the forecast change (1 = right size). Intervals: 90%, from 2,000 team-season resamples. Who sits: rank by possessions for the team; third to fifth and sixth to eighth count players on the floor for at least 40% and 20% of the team's possessions in their games. Every row but the last forecasts with usual minutes. With each game's actual minutes, the forecast change (−1.70 per absence) is larger than the actual change (−1.14).

F. Isolation Tests

Table 28. How much team context moves each metric's values (0 means not at all).

Metric	Teammates (teammate bias)	Replacements (team-depth bias)	New team	Old team	Movers' R^2
WOWY (on/off)	+.20	−.21	−.36	−.02	.019
aWOWY	+.23	−.30	−.36	+.09	.037
Par ±	+.00	−.06	−.12	−.05	.024
HAPM, one-year	+.00	−.09	−.13	−.04	.032
HAPM, two seasons	−.02	−.05	−.10	−.06	.039
HAPM, time decay	−.03	−.03	−.07	−.03	.043
RAPM (1-year)	+.23	−.10	−.11	−.07	.025
RAPM (2-year)	+.08	−.03	−.03	−.04	.036
RAPM (4-year)	+.02	−.01	−.03	−.01	.040

- Teammates: for 1,359 player-season pairs in which the player stayed on his team, the slope of next season's value on the strength of the teammates he played with (their previous-season RAPM), given his own value the season before.
- Replacements: the same, on the strength of the players who played when he sat.
- New team and old team: for 506 moves to a new team, the slope of the value after the move on the new team's (or the old team's) net rating without him, given his value before the move.
- HAPM, two seasons: this season's and last season's Par ± pooled, last season weighted .5.
- Movers' R^2: the share of next season's game margins explained when only the players who changed teams carry their previous values.
- Slopes are per standard deviation of the context.
- Nine season pairs from 2013-14 to 2025-26, players with 2,000+ possessions in both seasons. The 90% intervals are about ±.02 to ±.08.

References

- Blackport, D. (2014). How Long Does It Take For Three Point Shooting To Stabilize? Nylon Calculus.
- Carleton, R. (2012). Baseball Therapy: It's a Small Sample Size After All. Baseball Prospectus.
- Chetty, R., Friedman, J. and Rockoff, J. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. American Economic Review 104(9).
- De Bacco, C., Wang, Y. and Blei, D. M. (2024). A Causality-Inspired Adjusted Plus-Minus Model for Player Evaluation in Team Sports. *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, PMLR.
- Engelmann, J. (2011). Improved Multi-Year Ranking: Short Explanation. Advanced Stats for the NBA (stats-for-the-nba.appspot.com), archived October 4, 2011.
- Engelmann, J. (2017). Possession-Based Player Performance Analysis in Basketball. In *Handbook of Statistical Methods and Analyses in Sports*, CRC Press.
- Engelmann, J. (2026). NBA Adjusted Plus-Minus: How to Build It, plus Pro Tips and Tricks. Substack.
- EvolvingWild (2018). Revisiting Relative Shot Metrics. Hockey Graphs.
- Fearnhead, P. and Taylor, B. (2011). On Estimating the Ability of NBA Players. Journal of Quantitative Analysis in Sports 7(3).
- Fyffe, I. and Vollman, R. (2002). Improving Plus-Minus. Hockey Research Journal V(I).
- Goldstein, J. (2018a). Calculating Luck-Adjusted Ratings. Nylon Calculus.
- Goldstein, J. (2018b). Introducing Player Impact Plus-Minus. Nylon Calculus.
- Ilardi, S. (2014). The Next Big Thing: Real Plus-Minus. ESPN. Real Plus-Minus was built by Jeremias Engelmann.
- Hvattum, L. M. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports. International Journal of Computer Science in Sport 18(1).
- Jenkins, B. (2026). Deep Reinforcement Learning for NBA Player Valuation: A Temporal Difference Approach with Shapley Attribution. MIT Sloan Sports Analytics Conference.
- Kline, P., Saggio, R. and Sølvsten, M. (2020). Leave-Out Estimation of Variance Components. Econometrica 88(5).
- Macdonald, B. (2012). Adjusted Plus-Minus for NHL Players using Ridge Regression with Goals, Shots, Fenwick, and Corsi. Journal of Quantitative Analysis in Sports 8(3).
- Medvedovsky, K. (2020). NBA Stabilization Rates and the Padding Approach.
- Narsu, K. and Cranjis McBasketball (2020). LEBRON: Introduction. BBall Index.
- Petridis, A. and Pelechrinis, K. (2026). L-RAPM: lineup-level regularized adjusted plus-minus.
- Johnson, D. (2013). HARO, HARD and HART. Hockey Analysis.
- Myers, D. Box Plus/Minus, Version 2.0. Basketball-Reference.
- Medvedovsky, K. and Patton, A. WOWY RAPM and What is DARKO? DARKO.
- Narsu, K. (2017). When Can We Trust a Team's Stats? Nylon Calculus.
- Pemstein, J. and Dolinar, S. (2015). A New Way to Look at Sample Size. FanGraphs.
- RealGM. The Multi-Year WOWY Database.
- Rosenbaum, D. (2004). Measuring How NBA Players Help Their Teams Win. 82games.com.
- Schuckers, M. and Curro, J. (2013). Total Hockey Rating (THoR): A Comprehensive Statistical Rating of National Hockey League Forwards and Defensemen. MIT Sloan Sports Analytics Conference.
- Shax (2026). When Does an NBA Record Start to Mean Something? Substack.
- Sill, J. (2010). Improved NBA Adjusted Plus-Minus Using Regularization and Out-of-Sample Testing. MIT Sloan Sports Analytics Conference.
- Silver, N. (2019). How Our RAPTOR Metric Works. FiveThirtyEight.

- Snarr, T. About EPM and Methodology. Dunks and Threes.
- Tango, T. Evaluating Catchers.
- Tango, T., Lichtman, M. and Dolphin, A. (2007). *The Book: Playing the Percentages in Baseball*. Potomac Books.
- Taylor, B. (2016). *Thinking Basketball*. And Historical Impact: Introducing WOWYR.
- Terner, Z. and Franks, A. (2021). Modeling Player and Team Performance in Basketball. Annual Review of Statistics and Its Application 8.