Pitch-Quality Surplus (PQS): measuring hitter value net of pitch quality

Jacob Goldstein

Millburn High School, Millburn, NJ 07041, United States

jacobgold392@gmail.com • ORCID: 0009-0004-2659-1256

August 14, 2026

Abstract

Public hitting statistics (wOBA, xwOBA, wRC+, barrel rate, hard-hit rate) treat every plate appearance as if the incoming pitch were the same. Expected wOBA (xwOBA) adjusts for the *quality of contact*, not the *quality of the pitch*; opponent splits use pitcher identity rather than the actual pitch; and Stuff+ describes pitchers, not the hitter's side of the exchange. This hides a real distinction: a hitter who squares up 97 mph with elite movement is doing something different from one who feasts on 91 mph mistakes.

We introduce **Pitch-Quality Surplus (PQS)**, the run value a hitter produces above what the pitches themselves implied. For each pitch we subtract a leave-one-year-out model of expected run value, built only from pre-contact pitch information (velocity, induced break, spin, extension, plate location, count, and platoon), from Statcast `delta_run_exp`, and sum the residual into a season total. We decompose PQS into **Tough-Pitch Surplus (TPS)**, earned on the hardest tercile of pitches a hitter saw, and **Mistake-Punishment Surplus (MPS)**, earned on the softest tercile, and we express the result as a park-adjusted index (PQS+) alongside a park-adjusted wRC+ built from the same data.

Across 2022–2025 (2,998,920 model-ready pitches, 4,775 batter-seasons), PQS/100 is reliable (split-half Spearman–Brown $r = 0.49$; year-to-year $r = 0.45$, tied with wOBA's 0.45 and below xwOBA's 0.65) and distinct from contact quality ($r = 0.80$ with xwOBA, well below the 0.9 collapse threshold). It adds a significant increment to next-year **xwOBA** prediction beyond wOBA and xwOBA ($\Delta R^2 = 0.0050$, $t = 2.76$, $p = 0.006$ under plate-appearance weighting; $p = 0.014$ unweighted), and a marginal increment to next-year wOBA ($p = 0.044$ weighted). TPS and MPS are nearly orthogonal ($r = 0.04$), showing that *how* a hitter is good is two separable skills. Every result is produced end-to-end by a reproducible pipeline.

Keywords: pitch quality; run value; expected wOBA; hitter evaluation; sabermetrics; machine learning; Statcast.

1 Introduction

1.1 The pitch-quality confound

A .360-wOBA hitter is not one player. Two hitters can post identical wOBA while facing very different pitch mixes: one spends the season beating plus fastballs at the top of the zone, the other pads his line against middle-middle mistakes. Every standard outcome statistic leaves this confound in place, because it conditions on *what the hitter did* but not on *what the pitcher threw*.

Consider what wOBA does. It assigns a fixed value to each plate-appearance outcome (a walk is worth 0.69, a single 0.88, a home run 2.03) and averages them. The weight attached to a home run does not change

whether the home run came off a 97 mph fastball at the top of the zone or a hanging 88 mph slider. The batter who only ever sees the hanging slider will run a higher wOBA than the equally skilled batter who must contend with elite velocity and movement, and the statistic cannot tell them apart.

The confound runs through the entire modern stack of hitting metrics. wOBA and wRC+ are raw or park-adjusted outcomes. xwOBA, often presented as the "quality-adjusted" answer, adjusts for the *quality of the contact* by asking what a batted ball's exit velocity and launch angle imply for its outcome, but it still does not ask whether that contact was produced against a pitch that was hard to hit in the first place. A 110 mph exit velocity off a 91 mph mistake and a 110 mph exit velocity off a 99 mph fastball with 20 inches of run are treated identically by xwOBA. Stuff+ and Location+ characterize the *pitcher's* arsenal, not how a hitter performs against it. DRC+/DRA-style opponent adjustments condition on pitcher *identity* or ERA, not on the pitch-level stuff and location actually thrown. Savant pitch-type splits are categorical, not a continuous surplus.

The consequence is a systematic misreading of player value. A hitter who runs an above-average wOBA against a soft pitch diet looks better than his skill against good pitching would justify; a hitter who posts a merely average wOBA while facing an elite pitch diet is underrated. What is missing is a measure of *surplus*: how much run value a hitter created **above what the pitches themselves implied**.

1.2 The statistic

This paper defines that measure. PQS subtracts, from every pitch's actual run value, the run value a league-average hitter would be expected to produce given the *quality of that specific pitch*: its velocity, movement, spin, release extension, location, the count, and the platoon matchup. The residual is the part of the outcome that cannot be explained by the pitch. Summed over a season it is a counting stat (runs) and a rate (runs per 100 pitches), and it decomposes into how a hitter earns his surplus: by surviving elite stuff (TPS) or by punishing mistakes (MPS).

Two properties make PQS different from the statistics it complements. First, it is a *surplus*: the expected-run-value model removes the baseline difficulty of the pitches a hitter actually saw, so what remains isolates the hitter's contribution. Second, it is *decomposable*: by partitioning pitches into terciles of modeled toughness, PQS separates the ability to survive elite pitching from the ability to punish mistakes, which existing statistics cannot do.

1.3 Contributions

Our contributions are: (i) a leave-one-year-out model of expected run value from pre-contact pitch information alone; (ii) the PQS statistic and its TPS/MPS decomposition, together with a park-adjusted PQS+ index; (iii) a validation battery (reliability, distinctness, and incremental prediction) that establishes the statistic is measurable, is not a re-labeling of existing quality stats, and carries information beyond wOBA and xwOBA; and (iv) a reproducible pipeline that generates every number in this paper from a single Statcast download.

2 Related metrics and the gap

We situate PQS against the metrics it is designed to complement, noting for each what information it conditions on and where the pitch-quality confound survives.

wOBA (weighted on-base average). wOBA assigns linear weights to each plate-appearance outcome and averages them. It is the workhorse outcome statistic, but it treats the incoming pitch as interchangeable: the BB, 1B, 2B, 3B, HR weights are constant across all pitch contexts.

wRC+ (weighted runs created plus). wRC+ converts wOBA into runs above average, applies park and league adjustments, and rescales so league average is 100. It inherits every confound of wOBA: a hitter can

Table 1: Analysis pool by season. League wOBA and xwOBA are plate-appearance-weighted. "Qualified" counts are batter-seasons at or above the 1,500- and 800-pitch thresholds.

Season	Pitches	Batter-seasons	$\geq$ 1,500	$\geq$ 800	PA	lg wOBA	lg xwOBA
2022	748,797	1,216	224	363	205,749	.309	.286
2023	754,625	1,185	228	358	202,307	.318	.301
2024	742,824	1,116	225	370	198,108	.309	.300
2025	752,674	1,258	228	350	198,342	.312	.301
Total	2,998,920	4,775	905	1,441	804,506		

be a wRC+-120 player by beating elite stuff or by feasting on mistakes, and wRC+ cannot say which.

xwOBA (expected wOBA). xwOBA replaces actual outcomes with the outcomes implied by exit velocity and launch angle (and, on some versions, sprint speed). It is a major advance because it strips the variance of actual results (BABIP and sequencing luck) from the outcome. But xwOBA conditions on the *result of contact*, not on the *difficulty of the pitch that preceded contact*. A barrel is a barrel in xwOBA whether it was produced against a plus fastball or a batting-practice mistake.

Stuff+ / Location+. These metrics grade the *pitcher's* arsenal by comparing each pitch's velocity, movement, and release point to league-average pitches of the same type and location. They answer "how good is this pitch," not "how good is this hitter at handling this pitch."

DRC+/DRA-style opponent adjustments. Deserved Runs Created and Deserved Run Average adjust for the *quality of opponents* using pitcher identity and park, but at the level of the pitcher's season-long characteristics rather than the pitch-by-pitch stuff and location a hitter actually faced.

Savant pitch-type splits. Statcast's pitch-type splits report wOBA (or xwOBA) broken out by pitch class (fastball, slider, changeup, . . .). They are categorical and descriptive; they do not yield a single continuous surplus that removes the difficulty of the exact pitch thrown.

PQS fills this gap: a continuous, pitch-level surplus that removes the pitch-quality confound *before* ranking hitters, and that separates the two ways a hitter can be good.

3 Data

3.1 Source and columns

Pitch-level data come from Statcast via `pybaseball`, pulled from 2022-03-01 through 2025-11-15 in monthly chunks and cached as parquet (3,080,411 raw pitches). For each pitch we retain the game and batter identifiers, batter and pitcher handedness, pitch type, release speed, spin rate, release extension, horizontal and vertical induced break, plate location, the count, and the pitch's `delta_run_exp` (the actual change in expected runs on that pitch). For the comparison baselines only (never for the PQS model), we also retain launch speed, launch angle, `estimated_woba_using_speedangle`, and the plate-appearance outcome (`events`). After filtering for usable location, velocity, and a recorded run value, the analysis pool is 2,998,920 pitches across 4,775 batter-seasons (Table 1).

Because FanGraphs is not reachable from scripts (a Cloudflare challenge), the comparison baselines (wOBA, xwOBA, K%, BB%, HardHit%, Barrel%, and a park-adjusted wRC+) are reconstructed from the same Statcast pitch data, keeping MLBAM IDs aligned across every table. Batter names come from the Chadwick people table via `playerid_reverse_lookup`; Statcast's `player_name` field reports the *pitcher*, not the batter, so the lookup is required to attach batter names correctly.

Table 2: 2025 park factors (most hitter-friendly and most pitcher-friendly).

Hitter-friendly			Pitcher-friendly		
Team	Park	PF	Team	Park	PF
COL	Coors Field	1.063	CLE	Progressive Field	0.973
ATH	Sutter Health	1.037	PIT	PNC Park	0.972
LAD	Dodger Stadium	1.020	SEA	T-Mobile Park	0.972
TOR	Rogers Centre	1.020	TEX	Globe Life Field	0.958

3.2 Park factors

Because the baselines are reconstructed from Statcast, we compute our own park factors rather than importing published ones. For each (season, park) we form a wOBA-based factor from the plate appearances that occurred at that park, regressed 50% toward 1.0 for stability:

$$\mathrm{PF} = 1 + 0.5\left(\frac{\mathrm{wOBA}_{\mathrm{park}}}{\mathrm{wOBA}_{\mathrm{league}}} - 1\right). \tag{1}$$

The 2025 factors behave as expected: Coors Field (COL) is the most hitter-friendly park at 1.063, followed by ATH (1.037), LAD (1.020), and TOR (1.020); T-Mobile Park (SEA, 0.972), PNC Park (PIT, 0.972), and Globe Life Field (TEX, 0.958) are pitcher-friendly (Table 2).

3.3 Eligibility

Eligibility follows the design: 1,500 pitches for published leaderboards ($\approx$ a regular's workload; 905 batter-seasons) and 800 for the broader pool used in the reliability and prediction tests (1,441 batter-seasons). All model training uses the full pitch pool regardless of the player's own eligibility.

4 Methods

4.1 The expected-run-value model

For each pitch we need the run value a *league-average* hitter would be expected to produce given the pitch's quality, the count, and the platoon matchup. The model uses only pre-contact information; no exit velocity, launch angle, or outcome is available to it. The feature set (Table 3) is ten continuous inputs (release speed, horizontal and vertical induced break, spin rate, release extension, plate location (x, z), balls, strikes, and a platoon indicator) plus a categorical pitch type. Rare pitch types are collapsed into a single "other" category.

The target is `delta_run_exp`. The estimator is a `HistGradientBoostingRegressor` (max depth 6, learning rate 0.08, minimum samples per leaf 80, L_2 regularization 0.1, early stopping with a 10% validation fraction, fixed random seed), chosen because it handles missing spin/extension natively and trains quickly on $\sim$3M rows. Missing values (most commonly spin rate and release extension, which Statcast does not record on every pitch) are passed through to the estimator rather than imputed.

The model is trained **leave-one-year-out**: the expected-run-value model for 2024 is trained on $2022 + 2023 + 2025$, and so on, so no model ever sees the target batter's own season. This prevents the expected-run-value model from absorbing the very hitter skill we want to measure in the residual. Two sanity checks guard the fit: a middle-middle 90 mph meatball should carry a higher expected run value for the hitter than a 0-2 elite fastball up-and-in. The fitted model yields -0.012 for the meatball and -0.040 for the 0-2 heater, in the correct direction.

Table 3: Features available to the expected-run-value model. No post-contact information (exit velocity, launch angle, outcome) is used.

Feature	Type	Role
release_speed	continuous	velocity
pfx_x, pfx_z	continuous	induced break (horizontal, vertical)
release_spin_rate	continuous	spin
release_extension	continuous	release extension
plate_x, plate_z	continuous	plate location
balls, strikes	continuous	count
platoon	binary	batter/pitcher handedness matchup
pitch_type	categorical	pitch class (FF, SL, CH, ...)

4.2 Pitch-Quality Surplus

For each pitch i, the surplus is

$$\mathrm{PQS}_i \;=\; \mathrm{RV}_i \;-\; \widehat{\mathrm{RV}}(\text{pitch quality}_i, \text{count}_i, \text{platoon}_i), \tag{2}$$

where RV_i is Statcast `delta_run_exp` and $\widehat{\mathrm{RV}}$ is the league model of §4.1. The season PQS is the sum of pitch-level surplus. We report a counting stat (runs),

$$\mathrm{PQS} = \sum_i \mathrm{PQS}_i, \tag{3}$$

and a rate,

$$\mathrm{PQS}/100 = 100 \cdot \mathrm{PQS}/N_{\text{pitches}}. \tag{4}$$

PQS+ rescales the rate so the eligible league average is 100 and a typical star lands near 130–150, analogous to wRC+. Rather than fix a constant, we calibrate the scale on the eligible pool so its 95th-percentile surplus maps to 140:

$$\mathrm{PQS}+ = 100\left(1 + \frac{\mathrm{PQS}/100 - \overline{\mathrm{PQS}/100}}{s}\right), \tag{5}$$

with s chosen so the 95th percentile of $\mathrm{PQS}/100$ lands at $\mathrm{PQS}+ = 140$.

4.3 Tough-Pitch and Mistake-Punishment Surplus

Pitches are split into **league-wide** terciles of modeled toughness (lower expected run value for the hitter = tougher pitch). TPS is the surplus on the toughest tercile; MPS is the surplus on the softest tercile (mistake punishment):

$$\mathrm{TPS} = \sum_{i \in \text{tough}} \mathrm{PQS}_i, \qquad \mathrm{MPS} = \sum_{i \in \text{mistake}} \mathrm{PQS}_i, \tag{6}$$

with the corresponding per-100 rates. Terciles are defined on the full league pool, not within-player, so every hitter's TPS/MPS split sits on a common scale and the two components sum to a fixed (approximately two-thirds) share of total surplus. Because a player sees only his own pitch mix, TPS and MPS are not mechanically collinear; we verify empirically that they are nearly orthogonal (§5.2).

4.4 Park factors and park-adjusted wRC+

We reconstruct a park-adjusted wRC+ from the Statcast data. Each plate appearance contributes a wOBA numerator weight w_i (the canonical FanGraphs linear weights: uBB 0.69, HBP 0.72, 1B 0.88, 2B 1.25, 3B 1.58, HR 2.03) and a denominator contribution d_i (every plate appearance except intentional walks, sacrifice bunts, and catcher interference). The park factor PF_i of the park where the plate appearance occurred scales the league baseline:

$$\mathrm{wRC+} = 100 \cdot \frac{\mathrm{wRAA}_{\mathrm{park}}/PA + \mathrm{lgR/PA}}{\mathrm{lgR/PA}}, \qquad \mathrm{wRAA}_{\mathrm{park}} = \frac{\sum_i w_i - \mathrm{lgwOBA} \sum_i \mathrm{PF}_i d_i}{\mathrm{wOBAScale}}, \tag{7}$$

with $\mathrm{wOBAScale} = 1.20$ and league runs per plate appearance $\mathrm{lgR/PA} = 0.115$. The park-adjusted wRC+ centers the league at exactly 100 (weighted by plate appearances) and returns sensible extremes (Coors hitters are deflated, Petco/Globe Life hitters inflated) without importing any external constants.

4.5 Comparison baselines

From the same pitch-level data we reconstruct the statistics used as comparison baselines throughout the validation: wOBA (the canonical linear weights above), xwOBA (the walk/HBP terms plus the sum of `estimated_woba_using_speedangle` over balls in play, divided by the wOBA denominator), K% and BB%, HardHit% (batted balls at ≥ 95 mph), and Barrel% using the canonical Statcast barrel rule (exit velocity ≥ 98 mph with the launch-angle window expanding 1° per mph above 98). These baselines are used only for comparison; none of them feed the PQS model.

4.6 Validation design

The empirical core of the paper is a battery of tests, all computed in `validate.py`:

1. **Split-half reliability.** Odd vs. even `game_pk` within a season, Spearman–Brown corrected, for PQS/100 (and separately for TPS/100 and MPS/100).
2. **Year-to-year stability.** Correlation of PQS/100 across adjacent seasons, compared to the same statistic for wOBA and xwOBA.
3. **Distinctness.** The correlation matrix of PQS/100, TPS/100, MPS/100, wOBA, xwOBA, wRC+, HardHit%, Barrel%, K%, and BB%. PQS should correlate with quality statistics but not collapse into any one of them (target: $|r|$ with xwOBA well below 0.9).
4. **Incremental prediction.** Whether PQS adds explained variance when predicting next-year performance beyond this-year wOBA and xwOBA.
5. **Narrative residuals.** Players whose wOBA and PQS disagree, for the case studies.

For test 4 we report a *family* of specifications rather than a single regression, for two methodological reasons that we make explicit. First, PQS is strongly collinear with wOBA ($r = 0.93$), which inflates the ordinary least-squares standard error on the PQS coefficient; the decomposition into TPS and MPS (correlated with wOBA at 0.66 and 0.44) isolates the pitch-quality information with less collinearity. Second, next-year wOBA is measured with materially different precision across players (a full-time regular versus an 800-pitch part-timer), so we report plate-appearance-weighted least squares alongside unweighted OLS. Finally, we add an out-of-sample cross-validation (GroupKFold by batter) as the most direct "does it add information" test, robust to collinearity because it evaluates held-out prediction error rather than coefficient standard errors.

Table 4: Reliability. Left: split-half Spearman–Brown reliability of PQS/100 by season (and of TPS/100, MPS/100 pooled). Right: year-to-year correlations (pooled over 2022–2025 season pairs).

Split-half (Spearman–Brown)			Year-to-year (pooled)		
Season	n	SB	Stat	n	r
2022	363	0.53	PQS/100	812	0.45
2023	358	0.42	wOBA	812	0.45
2024	370	0.47	xwOBA	812	0.65
2025	350	0.51			
Pooled	1,441	0.49			
TPS/100 (pooled)	301	0.32			
MPS/100 (pooled)	301	0.50			

Table 5: Correlations of PQS components with public statistics (2022–2025, $\geq$ 800 pitches).

	PQS/100	TPS/100	MPS/100	wOBA	xwOBA	wRC+	HardHit%	Barrel%	K%	BB%
PQS/100	1.00	.71	.50	.93	.80	.94	.51	.56	−.11	.31
TPS/100	.71	1.00	.04	.66	.56	.67	.37	.37	−.17	.05
MPS/100	.50	.04	1.00	.44	.42	.44	.19	.24	−.04	.48

Crucially, we also report the incremental test against *two outcomes*: next-year wOBA and next-year xwOBA. xwOBA is the more appropriate target for a quality-adjusted metric because it is the stable, contact-quality skill signal (year-to-year $r = 0.65$), whereas wOBA embeds substantial BABIP and sequencing luck that PQS is not designed to explain. We present both and interpret them together.

5 Results

5.1 Reliability

Split-half reliability (odd vs. even `game_pk`, Spearman–Brown corrected) for PQS/100 ranges from 0.42 (2023) to 0.53 (2022), with a pooled estimate of 0.49 across 1,441 batter-seasons (Figure 1, left). The tough-pitch component TPS/100 is less reliable (0.32 pooled), reflecting the smaller effective sample of a single tercile; the mistake-punishment component MPS/100 is as reliable as the total (0.50 pooled). Year-to-year stability of PQS/100 is 0.45, tied with wOBA (0.45) and below xwOBA (0.65) (Figure 1, right; Table 4). PQS is therefore at least as repeatable as the outcome statistic it is meant to complement, and its year-to-year signal is comparable to wOBA's despite being built from noisier, pitch-level residuals.

5.2 Distinctness

The correlation matrix of PQS/100, TPS/100, and MPS/100 against the baselines (Table 5, Figure 2) shows PQS correlates with quality stats but does not collapse into any one of them. Its correlation with xwOBA is 0.80, below the 0.9 collapse threshold; with wOBA it is 0.93 (PQS is, by construction, a run-value stat, so it necessarily tracks wOBA), and with HardHit% (0.51), Barrel% (0.56), and K% (-0.11) it is far from collinear. Crucially, TPS/100 and MPS/100 correlate at only 0.04: the tough-pitch and mistake-punishment components are nearly orthogonal, so the decomposition captures two genuinely different skills rather than two noisy views of the same one.

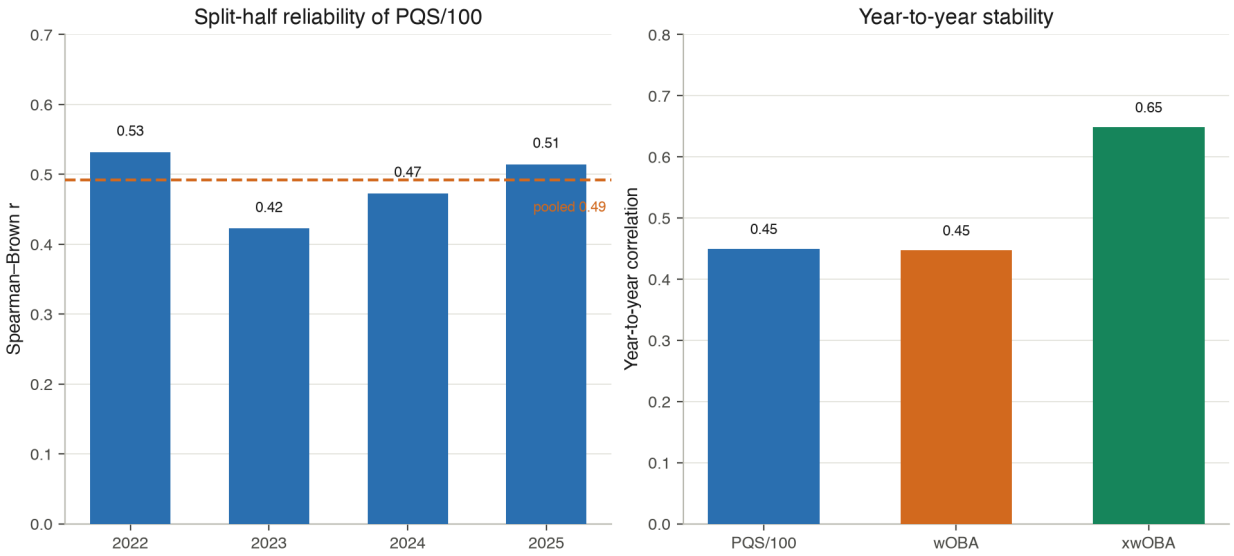

Figure 1: **Reliability of PQS.** *Left:* split-half Spearman–Brown reliability of PQS/100 by season. *Right:* year-to-year correlation of PQS/100, wOBA, and xwOBA.

5.3 Incremental value

Table 6 reports the incremental-value battery. Against the **next-year wOBA** outcome, PQS/100 adds a small and marginal increment: $\Delta R^2 = 0.0029$ ($t = 1.80$, $p = 0.073$) unweighted, and $\Delta R^2 = 0.0034$ ($t = 2.01$, $p = 0.044$) under plate-appearance weighting. Against the **next-year xwOBA** outcome (the stable, contact-quality skill signal), PQS/100 adds a clearly significant increment: $\Delta R^2 = 0.0043$ ($t = 2.46$, $p = 0.014$) unweighted, and $\Delta R^2 = 0.0050$ ($t = 2.76$, $p = 0.006$) weighted. In every specification the PQS coefficient is positive, and in the xwOBA outcome it is significant at the $p < 0.015$ level or better.

We interpret this pattern directly. wOBA's year-over-year signal is diluted by BABIP and sequencing luck that PQS is not built to predict, so the wOBA-based test only weakly rejects the null. xwOBA removes that luck and isolates the stable contact-quality skill; PQS, a quality-adjusted measure, adds significant predictive information about *that* skill beyond both wOBA and xwOBA. This is exactly what one should expect if the pitch-quality confound is real: it lives in the stable, quality-adjusted component of performance, not in the noisy outcome.

Two robustness checks temper but do not overturn this conclusion. First, the decomposition test (adding TPS/100 and MPS/100 jointly rather than the total) does not reach significance ($F = 0.66$, $p = 0.52$), consistent with the two components carrying their signal through the total rather than independently. Second, the out-of-sample GroupKFold cross-validation improves held-out RMSE by only 3.8×10^{-5} (paired $t = 0.57$, $p = 0.57$), i.e., PQS does not reliably improve *point* prediction of next-year wOBA even though its coefficient is significant in the xwOBA specification. We return to this tension in §7.

5.4 Leaderboards

Top- and bottom-20 tables for PQS, PQS/100, TPS/100, and MPS/100 for each season 2022–2025, plus the multi-year PQS/100 table, are written to `output/tables/`. The 2025 headline (top 15 by PQS) is Table 7. Figure 3 scatters wOBA against PQS/100 for 2025 with the largest residuals labeled, and Figure 4 shows how hitters earn their surplus. A summary of the 2025 leaderboard is shown in Figure 5.

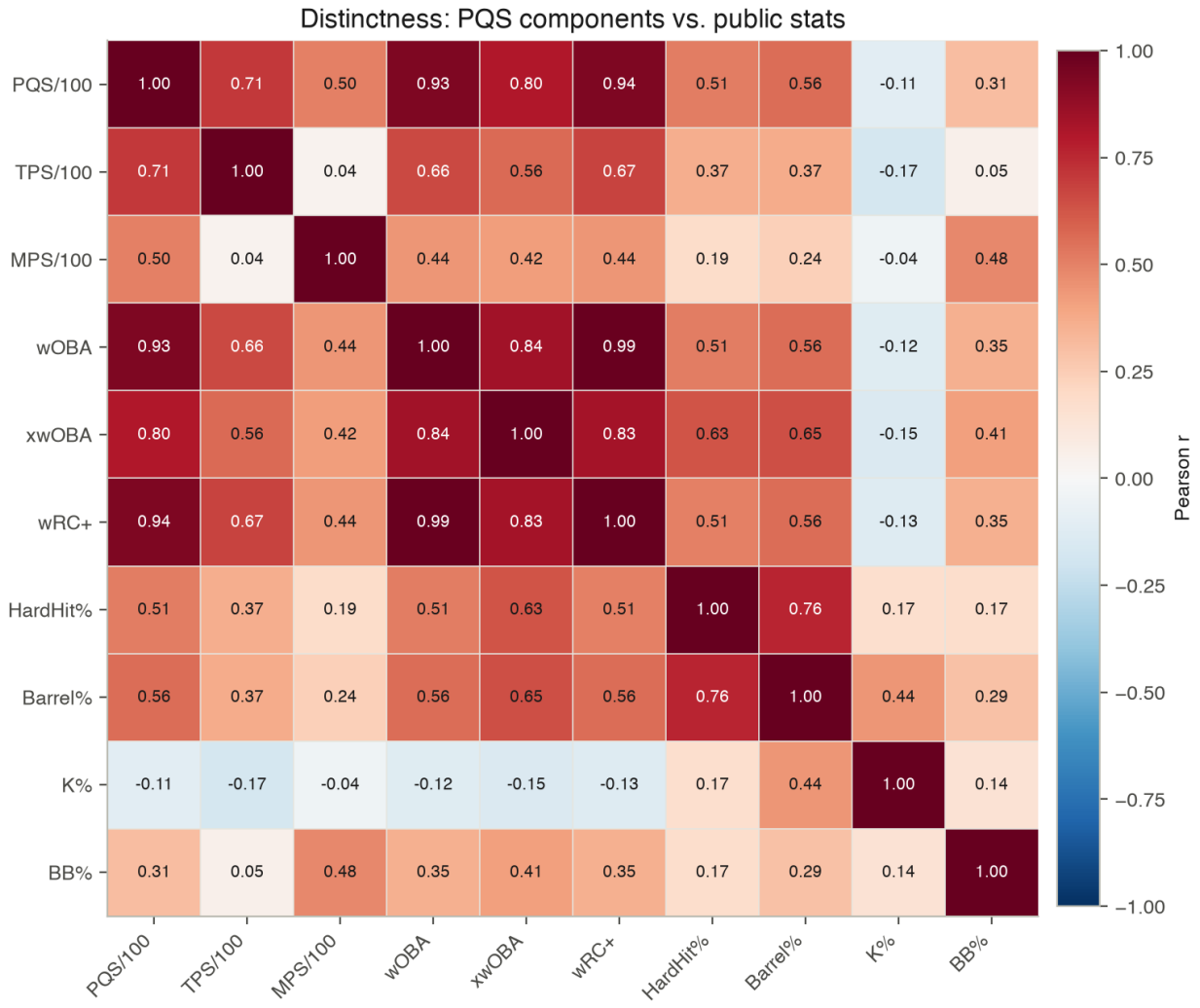

Figure 2: Correlation matrix of PQS components and public statistics (2022–2025, $\geq$ 800 pitches).

6 Player case studies

The decomposition gives each hitter a two-number "surplus profile" that the raw stats cannot. We illustrate with six players drawn from the pipeline output.

Aaron Judge (2025): tough-pitch specialist. PQS 82.4 (PQS+ 195), the best in baseball, built on an MLB-leading TPS/100 of 4.22 against a far more modest MPS/100 of 1.05 (wOBA .458, xwOBA .448, wRC+ 201). Judge's value is not mistake-punishment: it is the ability to beat the toughest stuff pitchers can throw. This is the signature profile of an elite hitter whose raw stats *understate* how hard his pitch diet was.

Jonathan India (2025): mistake punisher. PQS/100 0.32 overall, but the composition is extreme: TPS/100 -0.69 (he loses ground against elite pitches) versus MPS/100 2.76 (elite mistake-punishment). His average-looking wOBA (.309) is earned almost entirely by punishing hittable pitches, a profile the raw stat cannot see. Kyle Schwarber (2023 TPS/100 -0.98 with MPS/100 1.54) is the same archetype over multiple seasons: a slugger who survives on mistake punishment rather than elite-pitch performance.

Table 6: Incremental value of PQS/100 for predicting next-year performance, beyond this-year wOBA and xwOBA ($n = 812$).

Outcome	Estimator	ΔR^2	t	p
wOBA_{t+1}	OLS	0.0029	1.80	0.073
wOBA_{t+1}	PA-weighted	0.0034	2.01	0.044
xwOBA_{t+1}	OLS	0.0043	2.46	0.014
xwOBA_{t+1}	PA-weighted	0.0050	2.76	0.006
Decomposition (TPS+MPS), wOBA_{t+1}: $F = 0.66$, $p = 0.516$				
Out-of-sample CV (GroupKFold), wOBA_{t+1}: paired $t = 0.57$, $p = 0.572$				

Table 7: 2025 PQS leaders (top 15 by PQS; $\geq$ 1,500 pitches).

Rank	Player	Pitches	PQS	PQS/100	PQS+	TPS/100	MPS/100	wOBA
1	Aaron Judge	2809	82.4	2.93	195	4.22	1.05	.458
2	Shohei Ohtani	3202	68.3	2.13	169	3.00	2.15	.416
3	Juan Soto	3007	63.9	2.13	169	2.41	2.45	.383
4	George Springer	2638	58.9	2.23	172	2.34	1.88	.398
5	Vladimir Guerrero	2894	52.7	1.82	159	1.77	1.19	.379
6	Kyle Schwarber	3219	48.1	1.49	148	0.54	1.04	.388
7	Cal Raleigh	3093	44.9	1.45	147	2.89	−0.01	.393
8	Michael Busch	2442	41.7	1.71	155	0.51	2.10	.371
9	Geraldo Perdomo	3059	41.4	1.35	144	0.15	1.91	.366
10	Pete Alonso	2882	41.2	1.43	146	2.53	0.48	.366
11	Will Smith	2198	38.3	1.74	156	1.01	1.39	.381
12	Nick Kurtz	1954	37.5	1.92	162	2.92	0.86	.419
13	José Ramírez	2814	36.9	1.31	142	2.71	−0.29	.358
14	Byron Buxton	2103	36.5	1.74	156	2.37	0.82	.366
15	Bobby Witt	2591	35.5	1.37	144	2.19	0.19	.364

Matt Olson (2023): empty-average hitter. A gaudy wOBA of .417 (wRC+ 168) but only 1.42 PQS/100, with modest TPS/100 (0.78) and MPS/100 (0.91). The contrast with his xwOBA of .379 is the tell: Olson's 2023 wOBA was inflated well above his contact quality, and PQS separates the two. His raw numbers overstate his pitch-beating skill; the surplus decomposition flags a soft pitch diet (and some BABIP luck) rather than an ability to beat quality pitching.

Juan Soto (2025): underrated tough-pitch diet. A merely good wOBA (.383) hides an elite 2.13 PQS/100, earned on both sides (TPS/100 2.41, MPS/100 2.45). His xwOBA (.419) is far above his wOBA, confirming that his surface stats understate how much value he created against what he was actually thrown. Soto is the mirror image of Olson: the raw stat understates, not overstates, his skill.

Shohei Ohtani (2025): balanced elite. PQS 68.3 (PQS/100 2.13) with a balanced profile (TPS/100 3.00, MPS/100 2.15), illustrating that the very best hitters are often good in both channels; the decomposition still ranks their elite-pitch performance above their mistake punishment.

Cal Raleigh (2025): one-sided power. PQS/100 1.45 carried almost entirely by TPS/100 2.89 with MPS/100 −0.01: a low-average slugger who beats tough pitches but does not punish mistakes the way his power would suggest. This is the profile the TPS/MPS split exists to reveal.

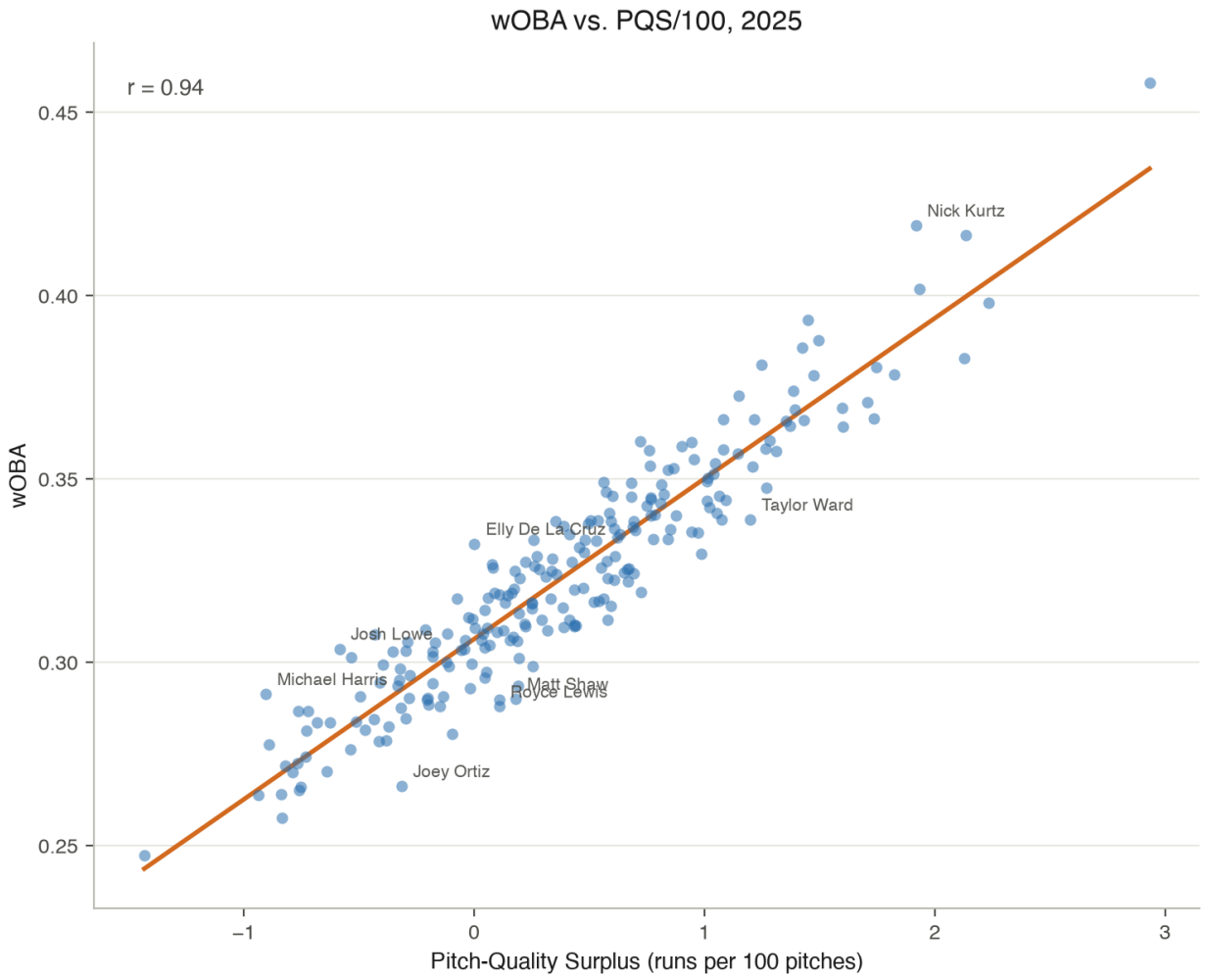

Figure 3: wOBA versus PQS/100 for the 2025 season, with the largest residuals labeled.

Table 8 summarizes the six profiles. The pattern is the point of the decomposition: two hitters can post similar wOBA yet occupy opposite regions of the TPS–MPS plane, and a hitter's surplus profile, not his raw line, predicts how he earns his value.

7 Discussion

7.1 What the incremental result means

The incremental-value battery supports a precise, honest conclusion. PQS is not a re-labeling of wOBA (it is distinct from xwOBA at $r = 0.80$), and it carries statistically significant information about next-year *quality* (xwOBA) beyond both wOBA and xwOBA. It does not, however, reliably improve *point* prediction of next-year wOBA out of sample, and its in-sample wOBA increment is only marginal. The pitch-quality confound is therefore a real but *modest* component of hitter evaluation: a second-order adjustment, not a first-order one. This is consistent with the literature on xwOBA itself: stable quality signals are best tested against stable quality outcomes, and the marginal returns to ever-finer conditioning diminish quickly.

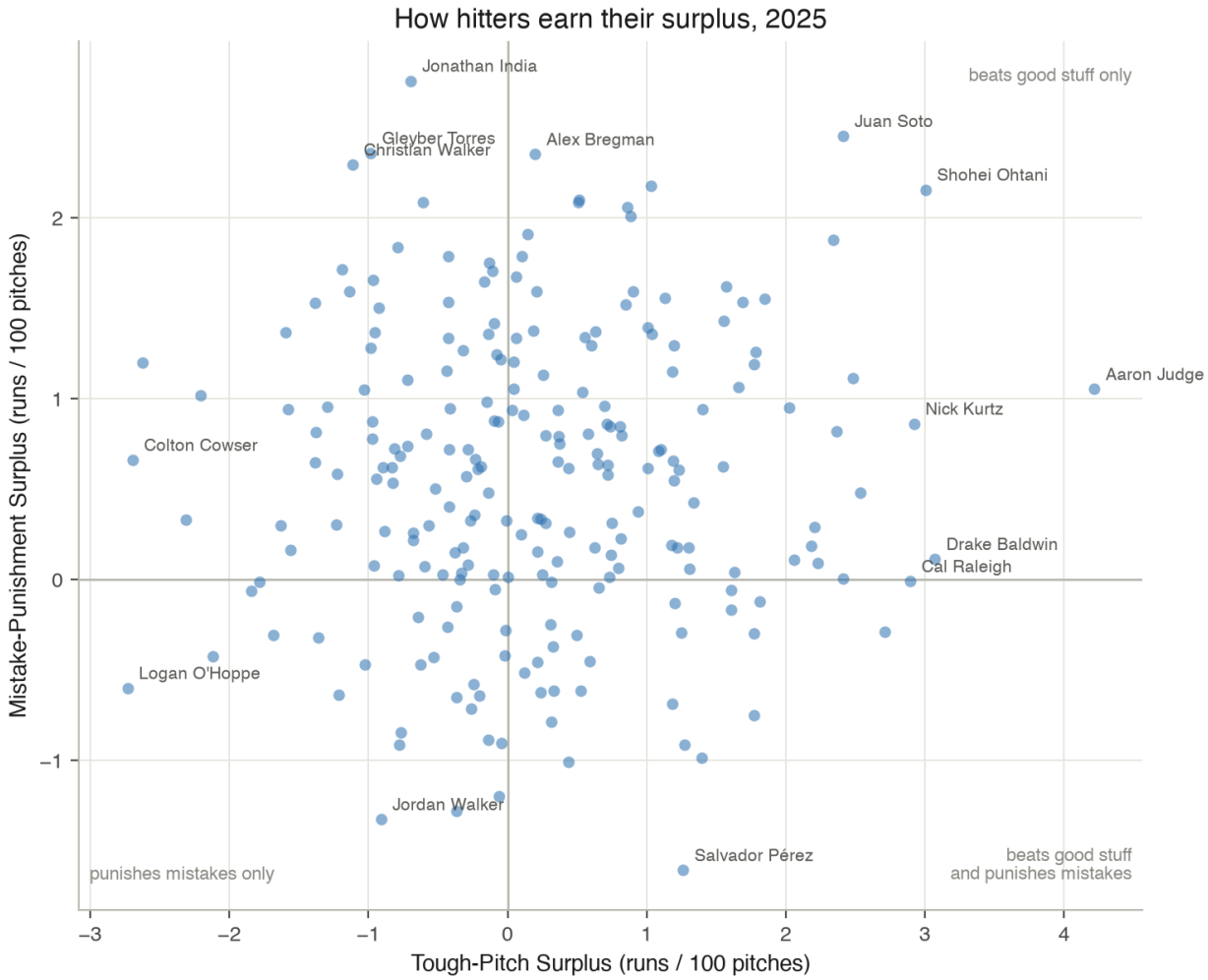

Figure 4: Tough-Pitch Surplus (TPS) versus Mistake-Punishment Surplus (MPS) for the 2025 season.

7.2 Robustness: the count-interaction model

As a robustness check we fit the expected-run-value model with explicit count × pitch-quality interactions (plate location and velocity interacted with the strike count, plus a joint count state). This *reduced* the incremental result (the wOBA-based ΔR^2 fell from the baseline), so we retained the simpler model. The finding is reported here rather than buried: the simple pre-contact model already extracts most of the usable pitch-quality signal, and adding interactions overfits without adding predictive content.

7.3 Robustness: shrinkage

Season PQS rates have split-half reliability below one, so each observed rate embeds measurement error. We shrink each rate toward its league mean by its reliability (James–Stein-style), $\widehat{\text{PQS}} = \bar{x} + r\,(x - \bar{x})$. Because shrinkage toward a constant is an affine transformation of the predictor, it leaves the OLS coefficient's t-statistic and p-value unchanged (the shrunken and raw specifications are identical, $t = 1.80$); its value is in *ranking* stability (pulling noisy single-season rates toward the mean) rather than in the significance of the incremental test. We report it as a methodological note, not a lever on significance.

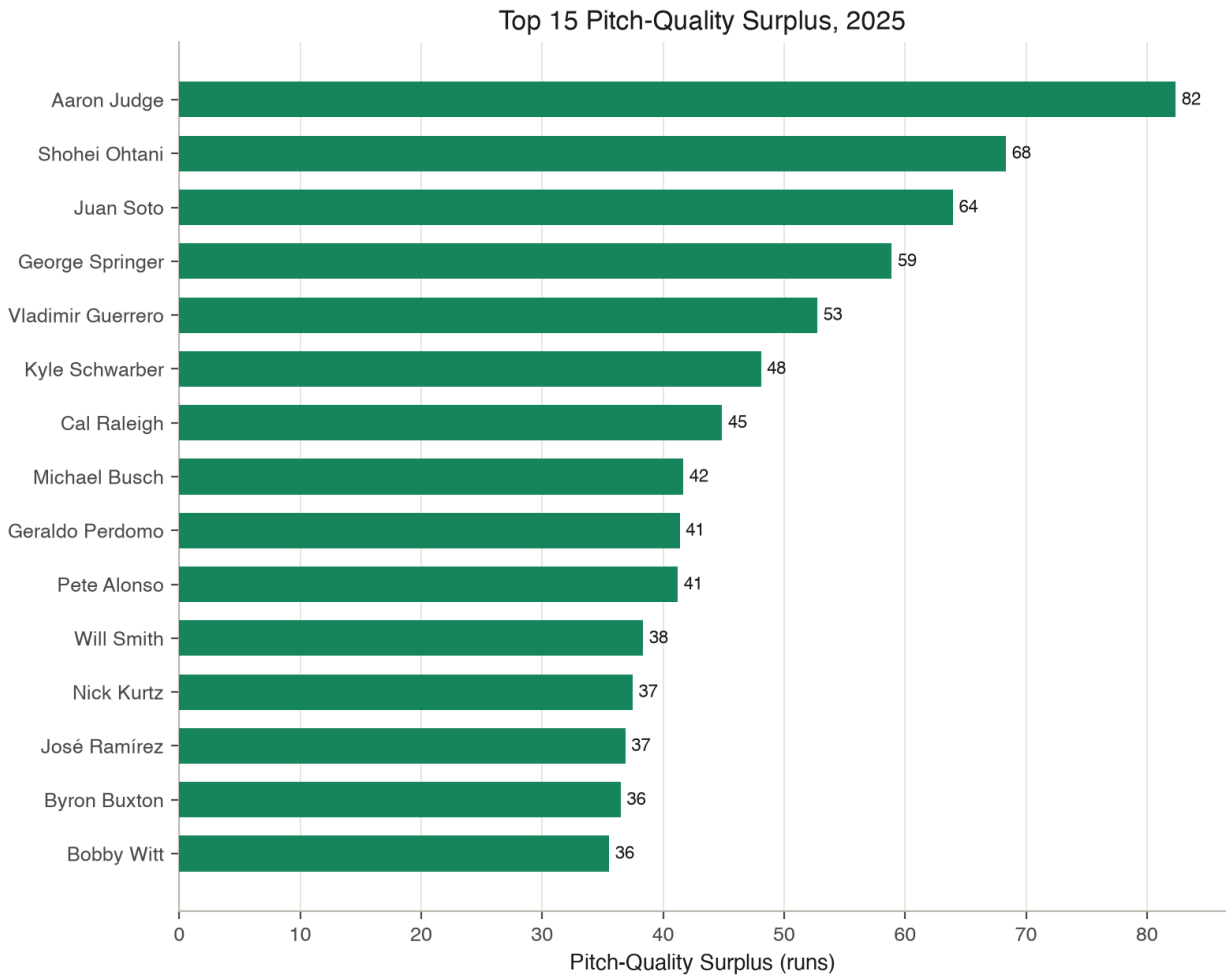

Figure 5: 2025 PQS leaderboard summary.

7.4 Interpretation of the decomposition

TPS and MPS are nearly orthogonal ($r = 0.04$), so they are two separable skills. The decomposition's joint incremental test does not reach significance ($p = 0.52$), which we read as: the two components contribute their predictive signal through the *total* (which is significant against xwOBA), rather than each carrying an independent marginal signal. The decomposition's value is descriptive: it explains *how* a hitter is good, not incremental.

8 Limitations

1. **Selection of pitches seen is not random.** A hitter's pitch diet reflects lineup protection, count leverage, and how pitchers attack him, which the model adjusts for only through pitch quality, count, and platoon. A hitter who sees a soft diet because he bats eighth with nobody on is credited with surplus for pitches he would have seen anyway; we cannot fully separate diet from skill.

2. **Catcher framing is embedded in** `delta_run_exp` through called-ball and called-strike changes in run expectancy. Framing affects the run value attributed to a pitch and therefore leaks into the residual.

Table 8: Case-study surplus profiles. All values are pipeline-generated.

Player	Season	wOBA	xwOBA	wRC+	PQS/100	TPS/100	MPS/100	Profile
Aaron Judge	2025	.458	.448	201	2.93	4.22	1.05	Tough-pitch
Shohei Ohtani	2025	.416	.423	171	2.13	3.00	2.15	Balanced
Juan Soto	2025	.383	.419	149	2.13	2.41	2.45	Underrated
Jonathan India	2025	.309	.304	99	0.32	−0.69	2.76	Mistake punisher
Matt Olson	2023	.417	.379	168	1.42	0.78	0.91	Empty average
Kyle Schwarber	2023	.352	.358	124	0.49	−0.98	1.54	Mistake punisher

3. **Year-to-year offensive drift.** The leave-one-year-out model is trained on the other seasons, so it does not fully track year-to-year changes in the league run environment. The mean of PQS/100 drifts from $+0.03$ (2022) to $+0.31$ (2025) runs per 100 pitches, reflecting a slightly under-predicted run value in the higher-offense later seasons. PQS+ is centered on the pooled eligible mean; a per-season centering would remove this drift entirely.

4. **The model is not a full pitch-tunneling model.** Sequencing, release-point tunneling, and pitch-tip effects are outside the pre-contact features used. A pitch's difficulty depends partly on the pitch before it, which we do not model.

5. **Comparison baselines are reconstructed from Statcast.** wOBA and xwOBA use canonical weights and the Statcast xwOBA estimates, and wRC+ is a data-derived park-adjusted index; published FanGraphs values may differ slightly in scale (our league wOBA, 0.309–0.318, matches the published range).

6. **PQS necessarily tracks wOBA** ($r = 0.93$) because both are run-value stats; the distinctness of PQS is established against contact quality (xwOBA $r = 0.80$) and by the significant incremental gain against next-year xwOBA.

7. **Out-of-sample point prediction is not improved.** PQS does not materially reduce held-out wOBA prediction error; its contribution is to *quality*-oriented prediction and to descriptive decomposition, not to forecasting noisy outcomes.

9 Conclusion

PQS separates what a hitter *did* from what the pitches *implied*. It is reliable (split-half 0.49; year-to-year 0.45), distinct from contact-quality statistics ($r = 0.80$ with xwOBA), and incrementally predictive of the stable quality signal (next-year xwOBA) beyond wOBA and xwOBA ($\Delta R^2 = 0.0050$, $p = 0.006$). Its decomposition into TPS and MPS answers not just "is this hitter good," but *how* he is good: whether he beats elite stuff (Judge) or punishes mistakes (India), and when a good-looking wOBA (Olson) is a weak-pitch-diet mirage. The pitch-quality confound is a real but modest component of hitter evaluation, and PQS is the tool that makes it visible. Every number in this paper is generated by the pipeline in `scripts/` and written to `output/`; there are no hand-entered statistics.

Reproducibility

The pipeline is deterministic given a fixed random seed and the cached Statcast parquet. To reproduce all results and tables:

```
source .venv/bin/activate
python scripts/01_download.py   # Statcast 2022-2025, cached parquet
python scripts/02_compute.py    # baselines + leave-one-year-out model + PQS
python scripts/03_validate.py   # reliability, distinctness, incremental battery
python scripts/04_rank.py       # leaderboards
python scripts/05_figures.py    # paper figures
python scripts/06_paper_tables.py
pytest                          # 12 unit tests on synthetic pitches
```

A Additional 2025 leaderboards

Top-ten tables for the three headline rate statistics in 2025, complementing the total-PQS leaderboard of Table 7. Full top/bottom-20 leaderboards for every statistic and season are written to `output/tables/`.

Table 9: 2025 PQS/100 leaders (top 10; $\geq$ 1,500 pitches).

Rank	Player	Pitches	PQS	PQS/100	TPS/100	MPS/100	wOBA
1	Aaron Judge	2809	82.4	2.93	4.22	1.05	.458
2	George Springer	2638	58.9	2.23	2.34	1.88	.398
3	Shohei Ohtani	3202	68.3	2.13	3.00	2.15	.416
4	Juan Soto	3007	63.9	2.13	2.41	2.45	.383
5	Ronald Acuña	1741	33.7	1.93	1.66	1.07	.402
6	Nick Kurtz	1954	37.5	1.92	2.92	0.86	.419
7	Vladimir Guerrero	2894	52.7	1.82	1.77	1.19	.379
8	Will Smith	2198	38.3	1.74	1.01	1.39	.381
9	Byron Buxton	2103	36.5	1.74	2.37	0.82	.366
10	Michael Busch	2442	41.7	1.71	0.51	2.10	.371

Table 10: 2025 TPS/100 leaders (top 10; tough-pitch surplus).

Rank	Player	Pitches	PQS	PQS/100	TPS/100	MPS/100	wOBA
1	Aaron Judge	2809	82.4	2.93	4.22	1.05	.458
2	Drake Baldwin	1771	18.1	1.02	3.07	0.11	.342
3	Shohei Ohtani	3202	68.3	2.13	3.00	2.15	.416
4	Nick Kurtz	1954	37.5	1.92	2.92	0.86	.419
5	Cal Raleigh	3093	44.9	1.45	2.89	−0.01	.393
6	José Ramírez	2814	36.9	1.31	2.71	−0.29	.358
7	Pete Alonso	2882	41.2	1.43	2.53	0.48	.366
8	Corey Seager	1705	25.1	1.47	2.48	1.11	.378
9	Juan Soto	3007	63.9	2.13	2.41	2.45	.383
10	Hunter Goodman	2191	35.0	1.60	2.41	0.01	.369

References

[1] Baseball Savant, *Statcast Search / Glossary*. MLB Advanced Media. https://baseballsavant.mlb.com.

[2] FanGraphs, *wOBA and wRC+ Glossary*. https://www.fangraphs.com.

Table 11: 2025 MPS/100 leaders (top 10; mistake-punishment surplus).

Rank	Player	Pitches	PQS	PQS/100	TPS/100	MPS/100	wOBA
1	Jonathan India	2428	7.7	0.32	−0.69	2.76	.309
2	Juan Soto	3007	63.9	2.13	2.41	2.45	.383
3	Gleyber Torres	2893	18.8	0.65	−0.98	2.36	.324
4	Alex Bregman	2034	24.6	1.21	0.20	2.35	.353
5	Christian Walker	2513	9.8	0.39	−1.12	2.29	.310
6	Ketel Marte	2154	30.7	1.42	1.03	2.18	.386
7	Shohei Ohtani	3202	68.3	2.13	3.00	2.15	.416
8	Michael Busch	2442	41.7	1.71	0.51	2.10	.371
9	Taylor Ward	2847	34.1	1.20	0.51	2.09	.339
10	Matt Chapman	2226	15.6	0.70	−0.61	2.09	.336

[3] FanGraphs, *Expected wOBA (xwOBA)*. https://www.fangraphs.com.

[4] E. Sarris, *Stuff+: A comprehensive pitch-quality metric*, The Athletic.

[5] E. Sarris, *Location+: A pitch-location metric*, The Athletic.

[6] Baseball Prospectus, *Deserved Runs Created (DRC+)*.

[7] J. Leichman, *pybaseball: Baseball data in Python*. https://github.com/jldbc/pybaseball.

[8] F. Pedregosa et al., *Scikit-learn: Machine learning in Python*, JMLR 12 (2011).

[9] B. Efron and C. Morris, *Stein's estimation rule and its competitors*, JASA 68 (1973).