NOTE: PRELIMINARY DRAFT, SUBJECT TO REVISION

Tom Seaver Versus the Statisticians

Xavier Fünf
xavier_funf@bbcardstats.com

Is *inconsistency* a starting pitcher virtue? As paradoxical as this might sound, there are plausible theoretical reasons to think so. Under the Pythagorean Expectation Formula (PEF), a team's winning percentage can be estimated from the differential in total runs scored and allowed (James, 1983). But where a team concentrates a disproportionately large share of its runs allowed in a modest set of "lost cause" games, it necessarily will surrender a disproportionately smaller share across the remainder—and thus win more of them than PEF predicts. This is exactly the pattern that emerges when a starting pitcher's games are characterized by superior outings punctuated by occasional "blow ups." Based on this insight, theorists have proposed replacing starting-pitcher value estimators based only on run suppression with ones that appropriately credit starter variance, too (Brill & Wyner, 2024). This paper empirically investigates this "Variance Value Added" position. Combining empirically realistic simulations with historical data analyses and Bayesian estimation methods, it reaches a skeptical conclusion. A high level (90th percentile) of starter variance, the analyses show, generates an exceedingly small return in modern baseball: less than 1 win per 162 games for an average team, and even less for a better-than-average one. Moreover, this benefit is almost entirely a consequence of chance fluctuations rather than any persistent pitcher trait in start-to-start variance. There is no reason, then, that a practical evaluator would treat starter variance as a basis for assessing pitchers over and above their ability to suppress opponent scoring.

> *In baseball, my theory is to strive for consistency. . . .*
> —Tom Seaver (Solomon, 1997, p. 969)

> *He said, "Throw strikes and keep 'em off the bases," . . . and I said, "O.K" What else could I say? What else could he say?*
> —Dennis Martinez (Fish, 1987, p. 1773)

1. Introduction

Tom Seaver was a baseball player. He did what he *did*: pitched his teams to victory. He didn't do that by engaging in sports analytics, nor did he generate analytics so others could reproduce what he did on the field.

So Seaver could hardly be blamed if he got the *theory* of starting pitching wrong. Those who *do* theory have rejected the conclusion that consistency is an element of successful starting pitching. On the contrary, it is an impediment to it: holding runs allowed equal, *inconsistent* starting pitching is better.

	Start 1		**Start 2**		**Start 3**		**Start 4**		**Start 5**		**Start 6**		**Totals**	
	IP/RA	*result*	*IP/RA*	*result*	*IP/RA*	*result*	*IP/RA*	*result*	*IP/RA*	*result*	*IP/RA*	*result*	*IP/RA*	*W/L*
Reliable	6 / 3	W: 5 - 4	6 / 3	L: 4 - 5	6 / 3	W: 5 - 4	6 / 3	L: 4 - 5	6 / 3	W: 5 - 4	6 / 3	L: 4 - 5	36/18	3/3
Erratic	6 / 3	W: 5 - 4	8 / 2	W: 4 - 2	7 / 3	W: 5 - 4	6 / 2	W: 4 - 3	7 / 2	W: 5 - 4	2 / 6	L: 4 - 10	36/18	5/1

Table 1. Advantage of inconsistent pitching. Results and totals for two pitchers, Reliable and Erratic, over six starts.

The cases of two pitchers, Mr. Erratic and Mr. Reliable, illustrate (Table 1). They both surrender the same total number of runs over the same total number of innings pitched across six starts. The per-inning rate at which they allowed opponent runs, let's suppose, is exactly equal to the per-inning rate at

which their own teams scored them. But because Erratic concentrated 33% of his runs allowed in one bad outing, he gave his team a greater chance to win in the remainder than did Reliable, who spaced his runs per inning out evenly across all of his appearances: assuming their teams scored consistently with their per-inning rates throughout, Erratic's team will have won five of his six starts but Reliable's only three of his.

Generalizing, inconsistency is a virtue because runs allowed are not uniformly costly. They hurt more in close games but less—to the point of being nearly costless—in games in which a team is far behind. Accordingly, all else equal, a team that concentrates a greater proportion of its runs allowed in fewer games will maximize the chances of winning the remainder, and thereby increase wins on net, relative to a team that spreads out its runs allowed evenly. Call this the Variance Value Added Thesis—or VVAT.

Kicking around for some two decades in sabermetrics circles (Baxamusa, 2006; Gassko, 2006), VVAT was given its most forceful articulation in a recent article by Brill and Wyner (2024). Brill and Wyner make VVAT the basis of a new metric of pitcher value—Grid WAR—that they defend as superior to alternatives, such as FanGraphs's and Baseball Reference's pitcher WAR metrics, precisely because it, unlike them, is sensitive to the contribution that starter variance makes to the prospects of team wins.

But is it really *true* that inconsistency is a pitcher virtue? What do data say about the contribution that higher variance makes to the value of starting pitchers?

The point of this paper is to test VVAT. It will present a number of empirical analyses aimed at determining whether, and if so to what extent and when, higher starter variance generates additional team wins holding runs allowed constant. On the basis of this evidence, it will seek to determine whether a practical evaluator would in fact take starter variance into account in assessing starter pitcher proficiency.

2. The still unfinished biography of VVAT

VVAT is not new (Gassko, 2006; Baxamusa, 2006). It originally surfaced in sabermetrics via a series of on-line exchanges among analysts interested in examining possible deviations from the Pythagorean Expectation formula (James, 1983) (PEF). An extremely influential model that has been adapted to a wide variety of team sports (Chen & Li, 2016; Schatz *et al.*, 2011; Cochran & Blackstock, 2009), PEF states

$$WP = \frac{RS^{\gamma}}{(RS^{\gamma} + RA^{\gamma})}, \tag{1}$$

where "WP" denotes winning percentage, "RS" runs (or points, goals, etc.) scored, and "RA" runs allowed. The value of γ depends on the scale of run scoring and differentials in the sport in question (Kaplan & Rich, 2017). For baseball, the value is usually put at 2, although empirical investigation suggests 1.8 is more accurate (Davenport & Woolner, 1999; Miller, 2007).

Deviations from PEF, it has been shown, are related to the distribution of run differentials per contest (Braunstein, 2010). Gamma presupposes a particular distribution, which is the "ordinary" one for that sport. Teams can exceed their PEF winning percentage via either of two "nonordinary" runs-per-game distributions . One involves a smoother, lower-variance *runs-scored* distribution: this pattern occurs when a team concentrates its offensive production across relatively close games and avoids "wasting" runs in blow-out wins. The other involves a more erratic, higher-variance *runs-allowed* distribution: this is the pattern observed when a team's *opponents* score a disproportionately large share of their runs in

blow-outs, and correspondingly a disproportionately small share in lower-scoring, close games, in which the value of those runs would have been higher.

Obviously, teams don't directly aim to manipulate the distribution of runs to achieve these outcomes. But such results can arise, at least in theory, as a byproduct of the strategic use of team resources. A team that sensibly deploys its use of its best pinch hitters and confines its base-stealing and sacrifice bunts to close games, and avoids using them to run up the score when leading, will end up with lower variance in runs scored; similarly, one that saves its best relief pitchers, and limits its use of fielding substitutions, to situations in which it is trying to preserve a narrow lead—as opposed to wasting their efforts to avert scoring in games that their opponents have well in hand—will end up with a higher variance in runs allowed (Fünf, 2026a).

As the early theorists recognized, *inconsistent starting pitching* furnishes another, albeit less obvious, mechanism for generating higher-than-ordinary runs-allowed variance. As illustrated by the won-loss records of Erratic and Reliable, holding runs constant, a team can expect to win more games when pitchers combine strings of stellar performances with occasional meltdowns than when they allow a steady stream of moderate opposition scoring.

A variety of analysts have over the years poked and prodded at VVAT. They used small samples and tools that were arguably less than ideally suited to the task. Nevertheless, they reached a consensus that starter inconsistency was indeed likely to generate wins in excess of those predicted by PEF (Petti, 2012; Seidman, 2009; Baxamusa, 2006; Gassko, 2006).

The first technically ambitious engagement with VVAT was presented by Lepore (2018). Based on Major League Baseball seasons between 1998 and 2017, Lepore used logistic regression to estimate the probability of a team win (PTW) conditional on how many innings a starter had pitched and how many runs he had allowed upon exiting a game. Lepore calculated the mean PTW for every pitcher in his sample; he also determined the variance in their PTWs across starts. After regressing team winning percentage on pitchers' mean PTWs, he found that the residuals correlated positively with PTW variances. On this basis, he concluded that VVAT was a genuine phenomenon—or at least was "statistically significant"—but that it was likely too small (about 0.5 extra wins per 100 starts for every standard deviation in PTW variance) to be of practical import.

Lepore's finding, however, necessarily rests on model misspecification. As he formulated it, PTW represents the probability of a win in game g by team t given the starting performance of pitcher p. That is, $\mathrm{E}[\mathrm{W}_{tg}|\mathrm{PTW}_{tgp}] = \mathrm{PTW}_{tgp}$. Accordingly, over the course of any number of games started by p (G_{tp}),

$$E\left[\sum_{g=1}^{G_{tp}} W_{tg} \,\middle|\, \left\{\mathrm{PTW}_{tgp}\right\}_{g=1}^{G_{tp}}\right] - \sum_{g=1}^{G_{tp}} \mathrm{PTW}_{tgp} = E\left[\sum_{g=1}^{G_{tp}} W_{tg} \,\middle|\, \left\{\mathrm{PTW}_{tgp}\right\}_{g=1}^{G_{tp}}\right] - \overline{\mathrm{PTW}}_{tp}\, G_{tp} = 0. \tag{2}$$

Team winning percentage in games pitched by p is

$$\mathrm{WP}_{tp} = \frac{\sum_{g=1}^{G_{tp}} W_{tg}}{G_{tp}}. \tag{3}$$

Because

$$E\left[\mathrm{WP}_{tp}G_{tp}\middle|\left\{\mathrm{PTW}_{tgp}\right\}_{g=1}^{G_{tp}}\right]=E\left[\sum_{g=1}^{G_{tp}}W_{tg}\middle|\left\{\mathrm{PTW}_{tgp}\right\}_{g=1}^{G_{tp}}\right]=\overline{\mathrm{PTW}}_{tp}G_{tp}, \tag{4}$$

dividing by G_{tp} results in

$$E\left[\mathrm{WP}_{tp}\middle|\left\{\mathrm{PTW}_{tgp}\right\}_{g=1}^{G_{tp}}\right]=\overline{\mathrm{PTW}}_{tp}. \tag{5}$$

In other words (or just in words), under Lepore's method of calculating starting pitcher value, the expected team winning percentage for any two pitchers who have the *same* mean PTW should be the same—their mean PTWs. How they *arrived* at their respective means over the course of their respective games cannot make any difference.

Assuming Lepore's model for determining PTW was correctly specified, the residuals when team winning percentage is regressed on pitcher mean PTWs should therefore be nothing more than directionless noise. Lepore's finding that those residuals are, in contrast, correlated with PTW variances is thus not evidence that starter variance matters over and above starters' mean PTWs; it is evidence that Lepore's model for deriving starting pitcher value is *not* properly specified: in the manner in which he constructed his regression, $\mathrm{E}[\mathrm{W}_{tg}|\mathrm{PTW}_{tgp}] \neq \mathrm{PTW}_{tgp}$ after all. Variance is spuriously correlated with whatever influence accounts for the misspecification (likely a nonlinearity in how outs and runs map to win probability [see Lepore, p. 23, Fig. 3.2, 2018]). Thus, Lepore's analysis, while innovative and suggestive, fails to demonstrate either that VVAT is true or, if so, how large an impact it has.

Brill and Wyner present the most recent and by far the most rigorous investigation of the effect of starter variance. More data-intensive than Leport's starter value metric, their Grid WAR conveys effectively the same quantity: namely, the marginal expected probability of a team win given the game state (including not just the inning and score but runners on base) when the starter exits. This value is calculated each time a pitcher starts a game.

Conventional pitcher-value metrics—such as FanGraphs's and Baseball Reference's pitcher "Wins Above Replacement" or WAR ones—are based on the *rate* at which a pitcher allows runs per inning or per "game" (9 innings pitched). Because they are indifferent to how those runs are distributed across a pitcher's individual games, these "run suppression" measures are necessarily variance insensitive. Under a run-suppression measure so defined, Erratic and Reliable would end up with the same credited value after their six starts.

Grid WAR differs decisively in this regard. It generates a discrete quantum of value—marginal probability of a team win—every time a pitcher performs. That pitcher's value over a season is formed not by averaging but rather by summing these game-level performances. Reliable and Erratic allowed the same number of runs over the same number of innings. But Grid WAR will credit Erratic with more season value because his runs allowed were distributed across games in a pattern that generated a higher expected total number of team wins. It is precisely this sensitivity to the impact of variance, over and above runs allowed, that for Brill and Wyner makes Grid WAR a better estimator of individual starter value than the conventional run-suppression alternatives.[1]

[1] Brill and Wyner typically use "game-level Grid WAR" or "game-by-game Grid WAR" to refer to individual-outing scores; the sum of those over a season, for them, forms a pitcher's "Grid WAR," a usage that naturally conveys a

Brill and Wyner (pp. 294, 309-310) identify a number of pitchers whose season or career Grid WAR tallies differ dramatically from those pitchers' FanGraphs's WARs. These include Sandy Koufax in 1966 (11.5 team wins to 8.4), and Catfish Hunter and Whitey Ford, whose career Grid WAR totals exceed their FanGraphs marks by 15 and 25 team wins, respectively. Brill and Wyner attribute these differences to the failure of FanGraphs to attend to the extra value associated with these pitchers' tendencies to combine strings of stellar outings with occasional epic "blowups," the pattern associated with high starter variance.

But as theoretically compelling and statistically elegant as Grid WAR is, Brill and Wyner do not supply the defense of VVAT on which its superiority over run-suppression metrics depends. First, they present no validation of Grid WAR as a pitcher-value metric. That is, they do not offer empirical evidence that Grid WAR in fact performs better in assessing the contribution that starting pitchers make to team wins than do variance-insensitive run-suppression measures.

Second, they do not establish that any advantage Grid WAR enjoys in this regard is a consequence of its sensitivity to starter variance. They do show a relationship between variability in Grid WAR scores across games and the difference in Grid WAR's and FanGraphs's starter pitcher evaluations (p. 306, Fig. 12). But this analysis not only begs the question whether FanGraphs undervalues pitchers or instead Grid WAR overvalues them; it also necessarily fails to establish that any relationship between the two metrics is related to Grid WAR's variance sensitivity.

Variances in Grid WAR *can't* affect the probability that a starter's team will win games for the same reason that variances in Lepore's PTWs can't. Like a PTW, a Grid WAR score by construction characterizes the probability of a team win conditional on a pitcher's performance in a particular game. Variability in Grid WAR scores is directionless noise: two pitchers with equal season mean Grid WAR scores necessarily create the same expected probability of a team win whenever they pitch—however they compiled their identical mean scores across appearances. So if Grid WAR generates a different assessment (higher or lower) of starter value than FanGraphs's WAR, per-start variance in Grid WAR scores themselves cannot be the source of that.[2]

Such an advantage—if shown to exist—*could* be a result of Grid WAR's greater variance sensitivity. But to establish that, starter variance would have to be measured in relation to some species of performance *independent* of Grid WAR scores. Brill and Wyner don't offer any such measure of starter variance.

What makes Grid WAR superior to run-suppression estimators like FanGraphs, according to Brill and Wyner, is the latter's inattention to how starting pitchers' *runs allowed* are distributed across games. To empirically test that claim, then, requires a correctly specified measure of variance in *runs allowed* per start. That is what *any* proof of VVAT necessarily depends on: without that, it will not be possible to

---

pitcher's expected-win contribution to his team's season record. Because this paper focuses principally on the distribution of game-level performances, it will for economy use "Grid WAR score" to refer to the single-game marginal expected probability of a team win that Brill and Wyner assign to any given pitching performance, *not* the season-level sum of such performances. The paper will also use "season mean Grid WAR" to refer simply to the mean of those scores across a pitcher's starts. And like Brill and Wyner, the paper will, where the context clearly conveys such a meaning, also use the term "Grid WAR" to refer to the method Brill and Wyner have constructed to measure pitcher value generally.

[2] Indeed, Brill and Wyner shrink Grid WAR scores across multiple seasons to estimate latent pitcher skill (pp. 314-16). Because this procedure treats game-level Grid WAR variance as measurement error, it necessarily presupposes that variance in Grid WAR scores *themselves* is not a meaningful basis for appraising the contribution starter variance makes to pitcher value.

show that a starting pitcher's value should be measured in any currency other than the one that matters under PEF—namely, total runs allowed without regard to their distribution across games.

But even a properly designed specification of starter variance, while necessary, is not *sufficient* to test VVAT. As mentioned, teams will out- or underperform their expected PEF records if their runs-allowed variances deviate sufficiently from the ones associated with the scale of run-scoring in their sport. Such deviations, however, are destined to occur by chance: run-distribution variances are themselves random variables. For differences in per-start variances to be viewed as a component of value over and above differences in starting pitchers' abilities to suppress runs, it must be shown, in addition, that some meaningful portion of the extra wins generated by variance over, say, a season is attributable to a persistent, genuine individual trait and not just random fluctuations in uncontrolled and likely uncontrollable influences that affect pitcher success across outings. A method for disaggregating real between-pitcher differences in variance from completely fortuitous ones also is missing from Brill and Wyner's analysis.[3]

In sum, Brill and Wyner's development of Grid WAR puts, more strongly than do previous studies, the issue whether starter variance really matters. Their article, however, doesn't itself supply the empirical data or analyses necessary to answer that question.

3. Methods and analytical strategy generally

To investigate VVAT, a series of related tests were developed. Their common objective was to answer two questions: first, is VVAT a genuine effect; and second, if so, is it an effect of genuine consequence—one that actually admits of detection and supports practical inferences within the real-world evaluation frames (particularly individual seasons) within which those who use baseball data operate.

These tests, moreover, were to be linked by a common analytical component: the specification of a per-game runs-allowed variance metric. As explained, the putative advantage of a variance-sensitive pitcher evaluator like Lepore's or Brill and Wyner's consists not in the variability of that evaluator's own scores but rather in its sensitivity to how *runs* are distributed across pitcher starts. Each test, then, includes some account of variance measured in those terms.

The first test uses an empirically realistic, data-driven Monte Carlo simulation. The simulation rests on a nonparametric run-per-game variance estimator, which derives from American League/National League games since 1900. Using the estimator to populate exponentially tilted empirical likelihood (ETEL) distributions of mean runs per game and their associated variances, the simulation generates counterfactual records for teams with *equal* runs-allowed *means* but *disparate* runs-allowed per game *variances*. The differences in those records supply one device for quantifying VVAT.

The second test performs a parallel analysis of historical data. Using a starter-variance metric based on expected-outcome residuals adjusted for exposure, it determines how variance in starting-staff variance has influenced the performance of teams over the course of AL/NL history since 1900. It then uses a split-sample analysis to extract the true between-team effect of starter variance from chance fluctuations in the same and to form a Bayesian posterior estimate of the size of the former.

---

[3] As discussed, Brill and Wyner use shrinkage as part of their assessment of whether the proficiency that Grid WAR measures is a durable latent skill. However, they neither specify any measure of variance independent of Grid WAR's own—which is exactly what shrinkage removes—nor empirically disentangle its contribution to Grid WAR scores from that made by starter run suppression. Accordingly, the persistence of the type of skill measured by Grid WAR does not furnish evidence that starter variability is itself a meaningful and persistent between-pitcher trait.

The third test compares variance-sensitive to variance-insensitive starting-pitcher metrics. Holding constant team offensive capacity, fielding proficiency, and relief-staff quality, the analysis determines whether Grid WAR better explains variance in team winning percentages than do two starter runs-saved measures—the one that undergirds FanGraphs's pitcher WAR and another derived from eFIP, an empirically enhanced variant of fielding independent pitching. Grid WAR should explain more variance, if VVAT is true. The sample consisted of individual team-seasons since 1952, the first year for which Grid WAR data are available.

Finally, the fourth test decomposes Grid WAR. Using the post-1951 sample of pitchers (the ones for whom Grid WAR scores are available), it determines the relative contribution of run suppression and per-start runs-allowed variance to each pitcher's season mean Grid WAR score. A split-sample analysis is again used, this time to calculate for individual pitchers a posterior estimate of the size of true, between-pitcher starter variance on team wins per 30 starts.

These tests rests on three data sources. The first is Retrosheet, which was used to compile all team and pitcher data at the season and game level.[4] The second is the Grid WAR shiny app, which was used to obtain individual pitcher Grid WAR outcomes and team aggregations of the same.[5] And the third is FanGraphs.com, which was used to access FanGraphs's starter pitching-runs saved for teams between 1952 and 2025.

These summaries describe the tactics employed; the overall *strategy* was unite and conquer. By examining the issue with a variety of analytically independent tests, the hope was to generate a set of results the overall level of agreement of which would attest to their individual and collective validity.

4. Simulating a VVAT upper bound

4.1. Overview

This test seeks to create an empirically realistic but self-consciously constructed and highly general picture of the effects of runs-allowed per game (RAPG) variance. By using a data-driven simulation, it identifies the characteristics of teams most likely to be benefitted by RAPG variance and by how much.

Because it uses RAPG variance, the test is more favorable to VVAT than one that focuses specifically on starting-pitcher variance. Starter variance benefits will necessarily be attenuated by the quality of a team's relievers. But when RAPG variance generally is manipulated, the results are equivalent to putting the variance of starters and relievers in lock step, thereby conserving all the PEF-deviation benefit associated with concentrating runs allowed in blow-out losses, in which the marginal cost of runs allowed is therefore minimal. This effect is amplified by examining the impact of differences in RAPG variance among teams that vary in *mean* runs allowed and runs scored per game, a manipulation that captures the plausible interaction of RAPG variance with team defensive and offensive capacities generally.

The simulation, it bears emphasis, strikes a completely uncritical stance toward the source of RAPG variance. That is, the simulation takes the LOWESS-derived association between RAPG means and variances as given—without any effort to separate persistent trait-based sources of variance in a team's run-suppression components (its starting and relief pitchers and to a lesser extent its fielders), on

---

[4] Retrosheet owns the copyright in these data sets and permits them to be used free of charge. Interested parties may contact Retrosheet at 20 Sunset Rd., Newark, DE 19711.

[5] https://gridwar.xyz/. I thank Ryan Brill for his generous assistance in answering queries and otherwise enabling efficient use of the data.

the one hand, from chance on the other. Combined with its use of RAPG rather than starting-pitcher variances, this agnosticism toward the contribution chance might have played in generating such variances warrants regarding the simulation results as effectively a VVAT "upper bound," ones consisting of team-record impacts that will inevitably exceed any attributable to true (*i.e.*, chance-corrected) starting pitcher differences.

4.2. The simulation design

4.2.1. RAPG variance estimator. The first key component of the simulation was a LOWESS runs allowed per game (RAPG) variance estimator. RAPG variance will obviously expand or contract in relation to the level of runs scored. The goal of the estimator was thus to form a variance estimate conditional on any specified RAPG.

The estimator was based on the historical RAPG means and variances of every American and National League team (t) over the 1900-2025 seasons (s):

$$\overline{RAPG}_{ts} = \frac{1}{G_{ts}} \sum_{g} RA_{tsg}. \tag{6}$$

This generated a data set of paired observations for every team, ($\overline{RAPG}_{ts}$, $s^2_{RAPG,ts}$).

Run-scoring dynamics have notoriously shifted over the history of major league baseball—from the low-scoring "deadball era" of the early decades of the twentieth century to the offensive explosion of the 1920s and 1930s, from the pitcher-dominant 1960s and early 1970s to the steroid-fueled home run inflation of the 1990s (Schell, 2005). Because these shifts are likely to affect the dispersion of runs per game, the sample was divided into seven run-scoring periods, identified with Bayesian breakpoint analysis (Fearnhead, 2006; Chib, 1998): (1) 1900-1919; (2) 1920-1939; (3) 1940-1961; (4) 1962-1971; (5) 1972-1991; (6) 1992-2007; (7) 2008-2025 (Figure 1).

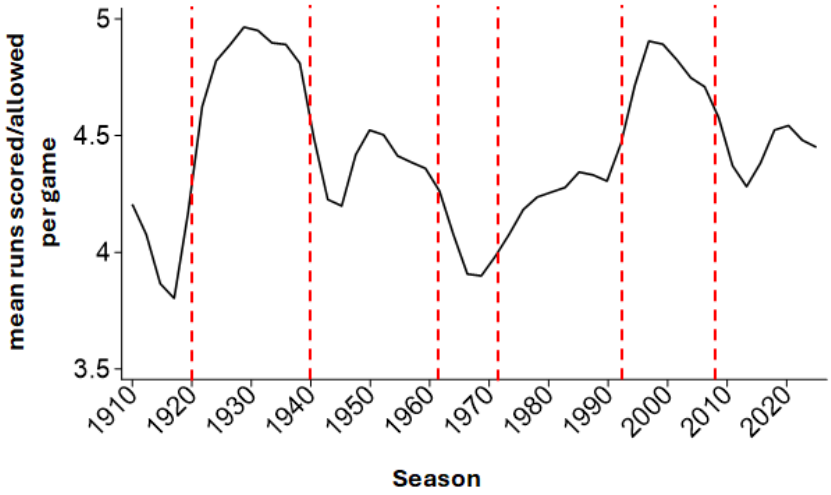


Figure 1. Runs-per-game eras. Periods determined by Bayesian breakpoint analysis (Fearnhead, 2006; Chib, 1998). Breakpoint regions identified by posterior mass of one or more years (1919-1921, 0.98; 1938-1941, 0.98; 1961-1964, 0.78; 1972-1978, 0.74; 1992-1993, 0.98; 2006-2010, 0.94); the year with highest breakpoint posterior was selected as the breakpoint. Trendline formed by polynomial smoothing for presentation.

A separate locally weighted or LOWESS regression (Cleveland, Devlin & Grosse, 1988) was fit to the paired team-season RAPG/variance observations within each of these seven run-scoring periods:

$$\log s^2_{RAPG,ts} = m_p\left(\overline{RAPG}_{ts}\right) + \varepsilon_{ts}. \tag{7}$$

The variances were log transformed so that fitted values would always be positive and multiplicative in relation to RAPG means (Gelman & Hill, 2007, p. 548). Accordingly, within any period, at any specified mean RPG, the predicted variance is

$$\hat{V}_{p,q}(\mu_{RAPG}) = \exp\{\hat{m}_p(\mu_{RAPG}) + \hat{Q}_{p,q}\}, \tag{8}$$

where $\hat{Q}_{p,q}$ is the empirical quantile of the LOWESS residuals within period p.

4.2.2. ETEL distributions. The second key component of the simulation consisted of exponentially tilted empirical likelihood or ETEL distributions of RAPG means and their associated, period-specific variances. ETEL furnishes a nonparametric method for generating distributions that respect empirically derived moments (Schennach, 2007; Imbens, Spady & Johnson, 1998).

Let p_0(r) represent the empirical probability that team t in season s will score or allow r runs in a game. For any target team RPG mean and corresponding period-specific median variance, a tilted distribution can be constructed:

$$p_{\mu_{RAPG,ts},\bar{V}(\mu_{RAPG,ts})}(r) = \frac{p_0(r)\exp(ar + br^2)}{\sum_{j\in R} p_0(j)\exp(aj + bj^2)}. \tag{9}$$

In this formula, $exp(ar + br^2)$ reweights the empirical probability of run total r; the denominator sums over the reweighted empirical probabilities for all run totals $j \in \mathcal{R}$, ensuring that the tilted probabilities sum to one. When a and b are solved for, one can derive weighted distributions that reflect the specified target μ and σ^2:

$$\sum_r r\, p_{\mu_{RAPG,ts},\hat{V}(\mu_{RAPG,ts})}(r) = \mu_{RAPG,ts}, \tag{10}$$

and

$$\sum_r (r - \mu_{RAPG,ts})^2\, p_{\mu_{RAPG,ts},\hat{V}(\mu_{RAPG,ts})}(r) = \hat{V}(\mu_{RAPG,ts}). \tag{11}$$

The simulation is unconcerned with the effect of runs scored per game (RSPG) variance. Nevertheless, the same process was used to construct a parallel set of ETEL RSPGs and variances in anticipation that the simulation would test the impact of varying RAPG variances across different RSPG means, which would all be pegged to their respective median variance levels.

4.2.3. Counterfactual season replays. Using the ETEL distributions, it is possible to play a season for any team at any specified levels of mean RAPG and mean RSPG, and at any specified level of

variance in either. For purposes of comparison, three team mean RAPGs were selected: "superior," "middling," and "poor." These corresponded to RAPG levels at the 25th, 50th, and 75th percentiles for each of the seven run-scoring eras. Two RAPG variance levels were also selected—median (50th percentile) and high (90th)—for each RAPG level.

When crossed, then, these RAPG means and associated variance levels generated six distinct RAPG means and variance targets: (1) superior/median; (2) middling/median; (3) poor/median; (4) superior/high; (5) middling/high; and (6) poor/high. Run-era specific ETEL distributions were generated for each of them.

ETEL *RSPG* distributions—21 of them—were also created. The target RSPG means were set at 5-percentile intervals from the 1st to 99th percentiles for each of the seven run-environment eras. The simulation used the 50th percentile period ETEL variances corresponding to these RSPG means.

Monte Carlo simulations were run for each of the six runs-allowed means/variance teams. The simulation consisted of 10,000 replicate 162-game seasons for each team at each of the 21 percentile levels of the RSPG distribution. Overall, the simulation generated 8,820,000 team-seasons, consisting of 1,428,840,000 games. Wins and losses were assigned on the basis of the runs scored and allowed drawn from the appropriate ETEL distributions.[6]

4.3. Results

The simulation outcomes are reported in Figure 2. They support three principal conclusions.

The *first* is that higher-than-normal variances can be expected to make a positive difference in team records. Across all the run-scoring eras, high-variance teams (those in the 90th percentile of RAPG variance) outperformed median-variance teams.

The *second* is that the benefit was greatest among the weakest teams. The size of the difference in the records of the high-variance and median-variance teams was highest for teams that simultaneously scored substantially fewer than the mean number of runs and allowed substantially more. Teams that simultaneously scored more than the mean number of runs and allowed substantially less got the least benefit from high RAPG variance.

These patterns make sense. Teams that uniformly allow a high number of runs will tend to consistently lose; one that allows the same number of runs on average but with a high degree of volatility will merely be defeated by even larger margins in most of their games but will eke out a higher share of wins in the remainder, in which its pitchers manage to completely shut their opponents down. By the same token, among teams that allow a low number of runs on average, the ones that do so consistently will win consistently, whereas those that allow the same number in a more uneven distribution will merely win more decisively most of the time but lose a higher share of the remainder in which their pitchers blow up.

[6] Ties were broken by selecting a random outcome based on the probability of a team win implied by the team's run-scoring and runs-allowed ETELs: $P(RS_{ETEL} > RA_{ETEL}) / [P(RS_{ETEL} > RA_{ETEL}) + P(RS_{ETEL} < RA_{ETEL})]$.

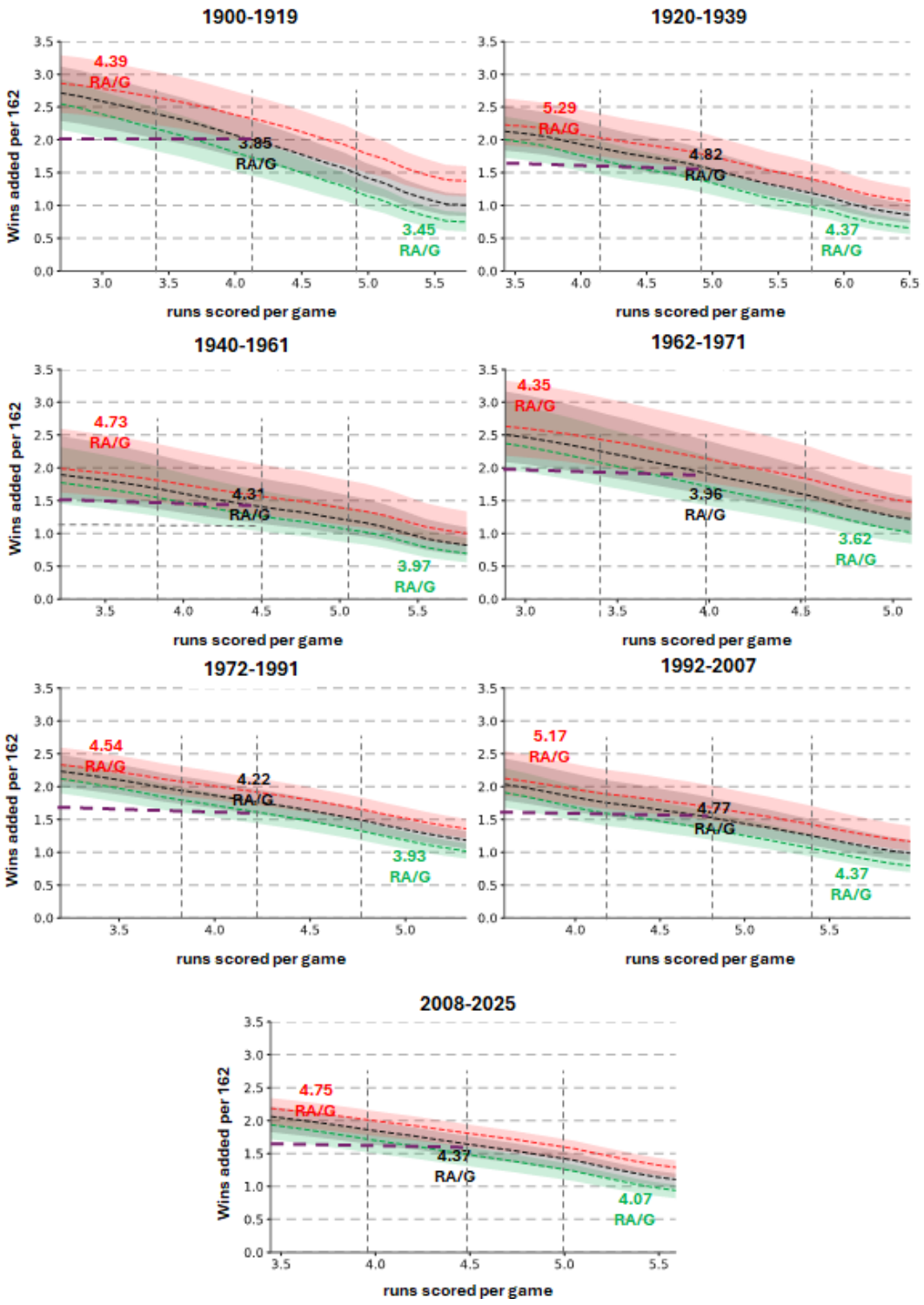


Figure 2. **ETEL simulation: impact of high runs per game variance.** Plotted lines reflect for teams at 25th (green), 50th (black), and 75th (red) percentiles of mean RAPG—over specified mean runs scored per game—the wins-per-162-games impact of being at 90th percentile rather than 50th in RAPG variance. Shaded regions of same color denote 0.95 CIs. Dashed vertical lines are 25th, 50th, and 75th percentiles of mean RPG. Purple dashed horizontal line intersects y-axis at wins added per 162 games for team at 50th percentile RAPG and RSPG means.

The mirror image dynamics apply for run-scoring. Holding total runs allowed constant, teams with anemic offenses will do better with higher RAPG variance: they will merely lose by greater margins than their more consistent counterparts, but will win more often in the remaining games in which their pitchers turn in stellar performances. Offensive powerhouses, in contrast, will tend to win consistently with consistent pitching; volatility will yield even more lopsided wins but at the expense of a greater frequency of losses when their pitchers' abysmal outings erase their teams' normal hitting advantage.

The *third* conclusion is that the benefits of high RAPG variance —that is, variance at the 90th percentile—were modest. The only two periods in which high-variance RAPG teams with mean runs scored and allowed realized gains in the neighborhood of 2 wins per 162 games were the 1900-1919 dead-ball era (2.1, 0.95 CI: 1.8-2.5) and 1961-1971 pitcher-dominance era (1.9, 0.95 CI: 1.6-2.4). Across all other periods, high RAPG-variance teams of average offensive and defensive strength won in the neighborhood 1.5-1.6 more games per season more than teams with median RAPG variances. Teams that combined high variance with high mean RSPGs and low RAPGs gained even less—between 1.0 and 1.5 wins per 162 games.

5. Empirical examination of team starting-staff impacts

5.1. Overview

The ETEL simulation furnished an upper-bound VVAT target. The next analysis uses historical data to assess how much of that potential gain teams likely realized as a result of differences in their starting-staff variance levels, including how much of that gain was a result of genuine propensities of their staffs as opposed to random performance fluctuations that all pitchers can be expected to experience.

5.2. Measuring starter variance

The first task is to form a measure of starter variance. VVAT asserts that variance in run suppression matters over and above runs allowed. An appropriate measure of starter variance, then, must characterize how pitchers comparable in run suppression *disperse runs allowed* across their starts.

The approach to be used here defines variance in terms of *expected-outcome residuals adjusted for exposure*. The method is akin to the one used to define variance in institutional-performance outcomes such as industrial accidents (Ruser, 1991) and is commonplace in insurance rating (Ohlsson & Johansson, 2010, pp. 78-81).

This method consists of three essential steps. First, an event rate is calculated in relation to the pertinent exposure increment (*e.g.*, accidents per worker hour) for each relevant unit (a firm in an industry). Here the relevant events are starter runs allowed, the exposure increment innings pitched in a game (g), and the unit team t in season s:

$$(RA/IP)_{ts}^{SP} = \frac{\sum_{g \in G_{ts}} RA_{tsg}^{SP}}{\sum_{g \in G_{ts}} IP_{tsg}^{SP}} \tag{12}$$

Second, for every observation (a month of production, say, in the case of a hazardous workplace) a residual is computed for the difference between the number of expected and observed events (accidents). That residual here is the difference between the expected and actual number of team starter runs allowed given the number of innings pitched in a particular game:

$$e_{tsg}^{SP} = RA_{tsg}^{SP} - \left(RA/IP\right)_{ts}^{SP} IP_{tsg}^{SP}. \tag{13}$$

And finally, the unit residuals are pooled and used to form a per-observation variance measure for that unit's events (e.g., workplace accidents per month), which here is team runs allowed per start:

$$V_{ts}^{SP} = \frac{1}{G_{ts} - 1} \sum_{g \in G_{ts}} \left(e_{tsg}^{SP}\right)^2. \tag{14}$$

Two other steps were taken to assure the proper performance of this measure in multivariate analyses. The first was to log transform it, a step that accommodates heteroscedasticity and assures a linear slope on the log scale (Rodbard & Frazier, 1975; Harvey, 1976; Davidian & Carroll, 1987, pp. 1083-1084).

The second was to residualize the measure against team starter runs allowed:

$$\ln V_{ts}^{SP} = \alpha_s + b \ln\left(\frac{\sum_{g \in G_{ts}} RA_{tsg}^{SP}}{G_{ts}}\right) + u_{ts}. \tag{15}$$

Residualization prevents the impact of run prevention from passing itself off as the impact of variance, which can be expected to grow disproportionately in relation to increases in runs allowed (Ruser, 1991, p. 336; Eling *et al*., 2018). Once residualized, a measure of starter variance is orthogonal to a team's rate of starter runs allowed and is thus fit for testing exactly what VVAT posits—that starter variance contributes to a team's probability of winning games over and above the number of runs its starters prevent.[7]

5.3. Estimating the contribution of variance to team records

As in the ETEL simulation study, the outcome of interest is the performance of teams with higher variance in runs allowed per game relative to teams of mean variance. Now, however, the data will be based on historical teams, and the focus will be on covariance in runs allowed only by team starting pitchers.

For this purpose a regression model was constructed:

[7] Here *b* is a common slope across all 2,714 team-seasons, identified from within-season differences between teams. Because single seasons have only 16–30 teams, season-by-season slopes are highly variable (10th–90th percentile 0.91–1.51); pooling across seasons reduces this sampling-error noise. Testing confirmed, moreover, that a common slope fits well. The value of *b* (1.22) fits comfortably within the slopes of team-seasons residualizations performed within five starter-usage eras (1.19–1.26; *SE*s 0.04–0.06), which will be introduced in Section 5.3. In addition, when the values were season-standardized (sec. 5.3), the variances formed by pooling all team-seasons correlated 0.9995 with the period-residualized variances and 0.97 with season-by-season residualized variances.

$$
\begin{aligned}
WP_{ts} &= \beta_0 + \beta_{PE} PEF_{ts} + \delta' \mathbf{SP}_{ts} \\
&+ \beta_R zRAPG_{ts} + \beta_S zRSPG_{ts} + \beta_V zSVAR_{ts} \\
&+ zRAPG_{ts} \theta_R' \mathbf{SP}_{ts} + zRSPG_{ts} \theta_S' \mathbf{SP}_{ts} \\
&+ \beta_{RV} zRAPG_{ts} zSVAR_{ts} + \beta_{SV} zRSPG_{ts} zSVAR_{ts} \\
&+ zSVAR_{ts} \phi' \mathbf{SP}_{ts} \\
&+ zRAPG_{ts} zSVAR_{ts} \psi' \mathbf{SP}_{ts} \\
&+ zRSPG_{ts} zSVAR_{ts} \omega' \mathbf{SP}_{ts} + \varepsilon_{ts}, \\
\varepsilon_{ts} &\sim \mathcal{N}\left(0, \sigma^2 \frac{\bar{G}_s}{G_{ts}}\right).
\end{aligned}
\tag{16}
$$

The outcome variable is team winning percentage. The principal predictors are teams' Pythagorean Expectation Formula expected winning percentage (PEF_{ts}) and team starter variance ($zSVAR_{ts}$). As explained, PEF predicts winning percentage based on runs allowed and runs scored (1) without regard to how either is distributed; insofar as VVAT asserts that starter variance contributes to team wins independently of starter runs allowed, the team starter-variance regressor should identify sources of variance that evade prediction based solely on PEF.

The ETEL simulation study found that the impact of RAPG variance varies based on team runs scored and allowed. Accordingly, the regression model included terms for RAPG ($zRAPG_{ts}$) and RSPG ($zRSPG_{ts}$) and their interactions with starter variance. All these variables, including the team-starter variance measure, were season-standardized to remove the impact of shifting run-environments: as run-scoring dynamics change, the standardized variables put on a common scale the impact of between-team differences in these characteristics on team performances (Schell, 1999, 2005).

Variable	Description
WPt_s	winning percentage of team *t* in season *s*
PEF_{ts}	PEF winning percentage of team *t* in season *s*
$zSVAR_{ts}$	season-standardized starter variance for team *t* in season *s*
$zRSPG_{ts}$	season-standardized team *t* runs scored per game in season *s*
$zRAPG_{ts}$	season-standardized team *t* runs allowed per game in season *s*
$\mathbf{SP}_{ts}$	starter-usage era for team *t* in season *s*

Table 2. Team starter-variance model variable notation.

Season-standardization doesn't address, however, shifts in the relative impact of starters over time. Starter usage has evolved over the course of the twentieth and twenty-first centuries, and it stands to reason that the importance of starter variance relative to other influences has evolved along with it.

As with the run-era periods used for the ETEL simulation, starter-usage periods were identified, based on the proportion of total innings pitched by team starters, using Bayesian break-period analysis. Five starter-usage periods were identified: (1) 1900-1910; (2) 1911-1945; (3) 1946-1983; (4) 1984-2013; and (5) 2014-2025. This effect was incorporated into the model with starter-usage variables ($\mathbf{SP}_{ts}$) and associated interactions with the starter variance terms.

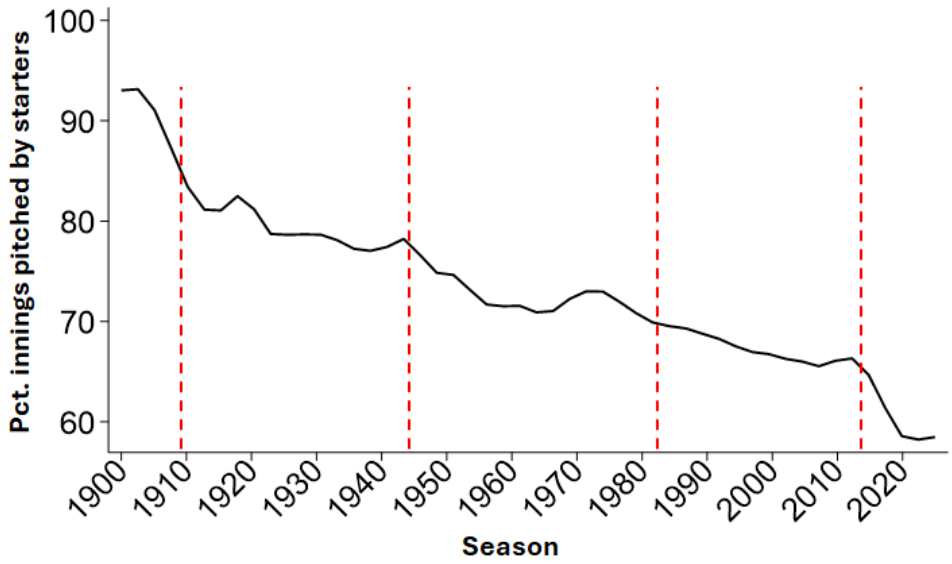


Figure 3. Starting pitcher eras. Eras identified by Bayesian breakpoint analysis (Fearnhead, 2006; Chib, 1998). Where breakpoints were identified by posterior mass comprising multiple years (1945-1947, 95%; 1980-1989, 84%; 2014, 97%), the year with highest breakpoint posterior was selected. Trendline formed with polynomial smoothing for presentation.

The regression was a weighted least squares linear probability model. Weighting was based on games to reflect seasons of differing lengths.[8]

5.4. WLS model output

The regression model results are reported in Table 3. Regressors are entered in steps to separate the explanatory contributions of variables associated with team starter variance from others and to identify which variables reflect inferentially material effects.

Several results from the regression output are worthy of note. One is that team starter variance does register an impact but an extremely small one. By itself, PEF explains 90% of the variance in team winning percentages since 1900; including the team-starter-variance regressors generated only an additional 0.3% (Table 3, step 3).

In addition, the effect is era specific. It is largest in 1900-1910, the first starter-usage period (Table 3, step 3). This result is consistent with the ETEL simulation finding that the impact of RAPG variance was largest in the deadball era. The effect of starter variance did not meaningfully vary over other starter usage periods (Table 3, step 3).

---

[8] Linear probability was selected over binomial logit to reflect the impact of including PEF winning percentage as a predictor. Because PEF explains such a high degree of the variance (a consequence of its use of predictors so proximate to the outcome variable), conditioning on it can be expected to generate residual variance lower than the binomial residual-variance floor (Smith & Faddy, 2019). This problem can be avoided by a linear probability model, because it estimates its variance scale from the data. Moreover, a linear probability model can be expected to generate valid predicted outcomes for probabilities grouped between 0.2 and 0.8 (Chen *et al*., 2017), a range that comfortably brackets the most extreme discrepancies among MLB team winning percentages.

The effect of starter variance also interacts with run scoring. Consistent with the ETEL simulation, the benefits of variance declined as teams scored more runs (Table 3, step 3). Contrary to the ETEL simulation, however, starter variance did not interact with RAPG (Table 3, step 3). Adding period-specific interaction terms for variance and RAPG and for variance and RSPG did not improve model performance (Table 3, step 4).

The practical import of the model results is illustrated in Figure 4. The two panels examine the impact of a high level of variance relative to a mean level separately for the 1900-1910 period, in which the impact of starter variance was historically large, and the 1911-2025 period, in which it was uniform. In both, the impact of high starter variance—90th percentile—is plotted for a mean-RAPG team as its RSPG increases.

	1		*2*		*3*		*4*	
PEF	**1.022**	(125.46)	**0.878**	(17.12)	**0.873**	(17.60)	**0.878**	(17.65)
SP1			0.000	(1.19)	0.000	(0.75)	0.000	(-0.01)
SP2			0.000	(1.74)	0.000	(1.59)	0.000	(0.88)
SP3			**0.001**	(4.36)	**0.001**	(3.77)	**0.001**	(2.20)
SP4			**0.000**	(3.09)	**0.001**	(2.75)	0.000	(1.31)
zRAPG			**-0.007**	(-2.50)	0.002	(0.75)	**-0.006**	(-2.40)
zRSPG			0.002	(0.61)	**-0.007**	(-2.61)	0.002	(0.79)
SP1_x_zRAPG			0.000	(0.18)	**0.006**	(2.76)	-0.001	(-0.48)
SP2_x_zRAPG			**-0.006**	(-4.04)	**0.007**	(3.54)	**-0.007**	(-4.73)
SP3_x_zRAPG			0.001	(0.54)	**0.005**	(2.76)	0.000	(0.20)
SP4_x_zRAPG			-0.002	(-1.28)	**0.005**	(3.17)	-0.002	(-1.68)
SP1_x_zRSPG			**0.005**	(2.31)	0.000	(-0.21)	**0.005**	(2.11)
SP2_x_zRSPG			**0.007**	(3.44)	-0.007	(-4.71)	**0.007**	(3.26)
SP3_x_zRSPG			**0.005**	(2.63)	0.001	(0.37)	**0.005**	(2.50)
SP4_x_zRSPG			**0.005**	(3.11)	-0.002	(-1.51)	**0.005**	(2.93)
zSVAR					**0.004**	(2.64)	**0.004**	(2.73)
SP1_x_SVAR					**0.008**	(3.18)	**0.009**	(3.19)
SP2_x_SVAR					0.001	(0.46)	0.001	(0.48)
SP3_x_SVAR					0.000	(0.16)	0.000	(0.19)
SP4_x_SVAR					0.000	(-0.08)	0.000	(-0.11)
RAPG_x_SVAR					0.000	(-0.03)	0.002	(1.07)
RSPG_x_SVAR					**-0.001**	(-2.51)	0.002	(0.76)
SP1_x_zRAPG_x_zSVAR							-0.006	(-1.71)
SP2_x_zRAPG_x_SVAR							-0.002	(-1.03)
SP3_x_zRAPG_x_zSVAR							-0.004	(-1.50)
SP4_x_zRAPG_x_zSVAR							-0.003	(-1.52)
SP1_x_zRSPG_x_zSVAR							-0.005	(-1.50)
SP2_x_zRSPG_x_zSVAR							-0.002	(-0.90)
SP3_x_zRSPG_x_zSVAR							-0.001	(-0.72)
SP4_x_zRSPG_x_zSVAR							-0.002	(-1.08)
constant	**-0.012**	(-2.84)	**0.060**	(2.34)	**0.063**	(2.53)	**0.060**	(2.42)
N	2,714							
R^2	0.90		0.90		0.91		0.91	
ΔR^2			**0.002**		**0.003**		0.000	
AIC	-12,125		-12,152		-12,233		-12,225	

Table 3. Team starting-staff variance regression model. Outcome variable is team winning percentage. Linear weighted least squares (games played). Raw beta coefficients; clustered standard errors by season; *t*-statistics denoted parenthetically. **Bolded** coefficients and delta R^2s are significant at $p < 0.05$.

For the post-deadball era, the impact of starter variance for a team of average RSPG is 0.75 extra wins per 162 games (0.95 CI: 0.54-0.96). For the teams at the 5th percentile in runs scored, it is 1.32 extra wins per 162 games (0.95 CI: 0.88-1.76), and for those at the 95th, 0.19 extra wins (0.95 CI: -0.30-0.69). These are uniformly lower than the impacts identified in the ETEL simulation: during the corresponding run-scoring eras, teams of mean RSPG and high RAPG variance tended to realize a gain of around 1.3 extra wins per 162 games. The difference is consistent with the smaller impact one would expect for team starter RAPG versus team RAPG simpliciter, since the benefits of the former get diluted by uneven relief-pitcher quality.

More effort and interpretation are necessary to determine the relationship between the WLS-model estimated impact of starter variance for 1900-1910 and the ETEL simulation impact. For a high starter-variance (90th percentile) team with mean RAPG and RSPG, the 1900-1910 model estimates a return of 2.42 additional wins per 162 games (0.95 CI: 1.62-3.23). This point estimate, at least, is larger than the one for an identically situated team under the ETEL for the 1900-1919 run-scoring period: 2.15 additional wins per 162 games (0.95 CI: 1.81-2.47).

The two estimates, however, are not strictly comparable, precisely because the ETEL simulation run-scoring period includes an additional decade. To make the comparison more precise, another ETEL simulation, this time only for 1900-1910, was run. It generated an estimate of 2.20 extra wins per 162 games (0.95 CI: 1.99-2.40).

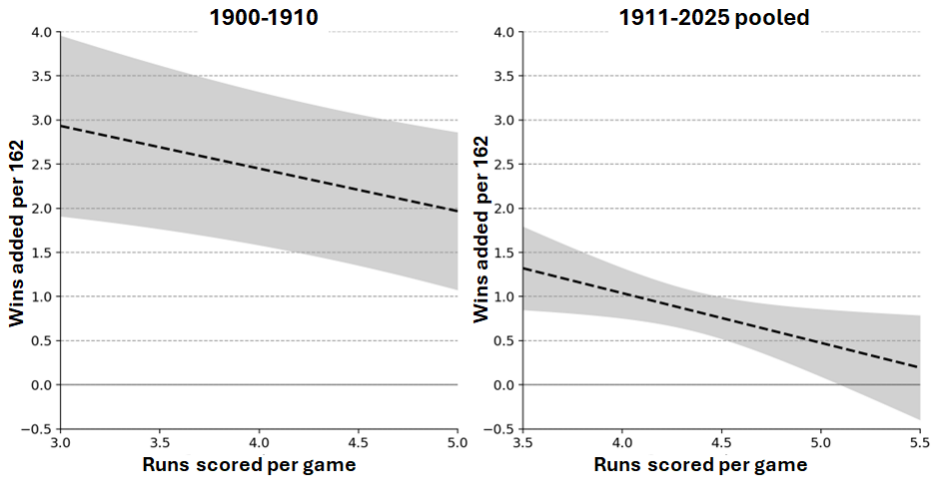


Figure 4. Impacts of high team starting-staff variance. Plotted lines indicate additional wins associated with team starting-staff variance of 90th percentile vs. 50th percentile. Based on team-starter regression model step 3, Table 3. Shaded regions are 0.95 CIs.

The difference, then, between the WLS-model estimate and the ETEL simulation for 1900-1910 period is 0.22 wins per 162 games (0.95 CI: −0.61 to 1.05). The amount is small, but more importantly it is swamped by the imprecision of the estimate, which reflects a 0.95 confidence interval nearly 7.5x its size. It is not possible to say definitively, then, that one estimate is larger or smaller than the other. But insofar as relief-pitcher usage comprised only a small portion of the pitching load in the first decade of the twentieth century (Figure 3), we should not expect a dramatic difference in the ETEL simulation and WLS model estimates.

5.5. Split-sample analysis

Generally. However consequential or inconsequential one might regard the size of the WLS estimates (particularly the ones for modern baseball), they necessarily *overrepresent* the effect of team starter variance for purposes of practical decisionmaking. The reason is that they are innocent of the influence of chance fluctuations unrelated to *true differences* in the variance states of teams' starters.

Imagine, for example, a team general manager is trying to assess the return on obtaining or retaining starters based on their variances. The WLS model would tell her exactly how much of the observed variance of her (or any other) team's starters contributed to team wins over and above the number expected based on run suppression.

But that difference would have no actionable significance unless she could determine how much of it was due to a persistent *characteristic* of the team's pitchers as opposed to their randomly allocated share of the range of uncontrollable influences that affect dispersion of runs across games. Without that information, she would have no idea how much value in the future her team could expect to realize—if any—from assembling a team of starting pitchers selected specifically for their favorable observed variances over some particular stretch of games.

One straightforward way to extract real from chance team starter variance impacts is through a split-sample test. Whereas the WLS model generates a mechanical report of how starter variance happened to work out for a team across a season, the split-sample test uses the teams' starter variances measured in one set of their games to predict how the team would do in another. By comparing the two measures, we can disentangle the respective contributions that the team's true variance state and chance variance fluctuations made to the team's actual record.

Such a test was performed. Alternating games from team seasons were assigned to separate half-season subsamples (h)—"E" for even-numbered games and "O" for odd-. Team starter variances were computed separately within each subsample. The subsample teams were then pooled and analyzed in two separate regressions: in the "OWN" model, subsample team winning percentages were regressed on subsample teams' own PEFs and their own starter variances,

$$WP_{tsh} = \beta_0 + \beta_{PE} PEF_{tsh} + \theta_{OWN}\, zSVAR_{tsh} + \varepsilon_{tsh} \qquad \varepsilon_{tsh} \sim \mathcal{N}\left(0, \sigma^2 \frac{\overline{G_{sh}}}{G_{tsh}}\right); \tag{17a}$$

in the "SWAP" model, team winning percentages were regressed on subsample teams' own PEF but their *opposing subsample counterparts'* starter variance (E's variance assigned to O, and O's variance to E):

$$WP_{tsh} = \beta_0 + \beta_{PE} PEF_{tsh} + \theta_{SWAP}\, zSVAR_{tsh'} + \varepsilon_{tsh} \qquad \varepsilon_{tsh} \sim \mathcal{N}\left(0, \sigma^2 \frac{\overline{G_{sh}}}{G_{tsh}}\right). \tag{17b}$$

A separate model, "COMP," was fit to the complete-season team PEFs and team-starter variances:

$$WP_{ts} = \beta_0 + \beta_{PE} PEF_{ts} + \theta_{COMP} zSVAR_{ts} + \varepsilon_{ts} \qquad \varepsilon_{ts} \sim \mathcal{N}\left(0, \sigma^2 \frac{\overline{G_{s}}}{G_{ts}}\right). \tag{17c}$$

Comparing the OWN and SWAP models, it is possible to disentangle the true from the chance portions of starter-variance impact. Only *true-state* starter variance can influence team winning percentages in the SWAP model; chance fluctuations in starter runs allowed can't affect outcomes in games in which they didn't occur. Accordingly, θ_{SWAP} tells us what the winning percentage impact of teams' persistent, true starter-variance state is. By the same token,

$$\frac{\theta_{OWN} - \theta_{SWAP}}{\theta_{OWN}} \tag{18}$$

tells us what fraction of the E and O subsample teams' starter-variance winning-percentage impacts was a result of chance.

At that point, we can determine *the full-season* impact of teams' true starter-variance states on their records. Over the course of a full season, the true-state effects of the two halves of any team's season will of course combine with one another, while the chance ones will tend to offset and cancel out. As a result, the overall fraction of observed variance that reflects true state will necessarily be larger over a full season than it was in either of the half-season subsamples. That fraction can be derived from the Spearman-Brown formula for calculating split-half reliability,

$$\hat{\rho}_{\text{FULL}} = \frac{2\hat{\rho}_{EO}}{1 + \hat{\rho}_{EO}}, \tag{19}$$

where $\hat{\rho}_{EO}$ is the correlation between teams' E and O starter variances, and $\hat{\rho}_{\text{FULL}}$ the reliability of the full-season measure (Spearman, 1910; Brown, 1910). Multiplying a team's WLS-model estimate of its full-season starter-variance impact by $\hat{\rho}_{\text{FULL}}$ gives the portion attributable to its true variance state.

Once the true and chance elements of starter variance have been separated, moreover, it becomes possible to form Bayesian posterior estimates of the impact of any specified level of starter variance (Smith & Gelfand, 1992). On the season-standardized scale, between-team variance in the true starter-variance state is

$$\text{Cov}\left(zSVAR_{tsE}, zSVAR_{tsO}\right) = \text{Var}\left(t_{ts}\right) = \tau^2, \tag{20}$$

which is the half-season reliability. This quantity is calculated separately for 1900-1910 seasons and 1911-2025 ones. A uniform prior on τ (Gelman, 2006) is set over the range of 0 to τ_{MAX}, where τ_{MAX} is set far above the estimate so it never constrains the estimated posterior:

$$\tau \sim \text{U}(0, \tau_{\max}), \qquad \tau_{\max} = \sqrt{\max(\hat{\rho}_{EO}, 0) + 8\text{SE}(\hat{\rho}_{EO})}. \tag{21}$$

Two million values are drawn from the prior. Each is weighted by its likelihood, the normal density of the E-O half-season correlation given τ^2 and its standard error,

$$\hat{\rho}_{EO} \mid \tau \sim \text{N}\left(\tau^2, \text{SE}(\hat{\rho}_{EO})^2\right), \tag{22}$$

and 400,000 draws are resampled in proportion to those weights. These resampled draws form the posterior. Each draw is converted to full-season reliability with the Spearman–Brown formula:

$$\rho_{\text{FULL}}^{(d)} = \frac{2\left(\tau^{(d)}\right)^2}{1+\left(\tau^{(d)}\right)^2}. \tag{23}$$

This distribution can then be converted to team wins per 162 games:

$$\Delta W_q^{(d)} = 162 \cdot \theta_{COMP}^{(d)} \cdot z_q \cdot \sqrt{\rho_{\text{FULL}}^{(d)}}, \qquad \theta_{COMP}^{(d)} \sim \text{N}\left(\hat{\theta}_{COMP}, \text{SE}(\hat{\theta}_{COMP})^2\right). \tag{24}$$

First, the square roots of $\rho_{\text{FULL}}^{(d)}$ are taken to put the values in the distribution on the same scale as the season-standardized team starter-variances. Next, those values are multiplied by a quantile value based on the team-starter variance level of interest—*e.g.*, 1.28 for the 90th percentile. Each resulting value is then multiplied by a slope drawn randomly from $N(\hat{\theta}_{\text{FULL}}, \text{SE}[\hat{\theta}^2_{\text{FULL}}])$, generating a posterior distribution in winning-percentage impacts. Finally, the values of this distribution are in turn multiplied by 162 to put them on a wins scale. The median is reported as a posterior point estimate, along with the 0.95 highest density interval.

The WLS model showed that the 1900-1910 period produced more starter variance wins than did the remaining seasons. Accordingly, the split-sample test was performed separately for teams in the 1900-1910 seasons and those in the 1911-2025 ones, as were the operations to determine full season impacts and Bayesian posterior estimates.

Outcomes. The main results of the split-season test are reproduced in Table 4. The table reports the slopes for the team-starter variance terms (winning percentage per +1SD).

	1900–1910		*1911–2025*	
θ_{COMP}	0.012	(0.008-0.016)	0.004	(0.003-0.005)
θ_{OWN}	0.014	(0.010-0.018)	0.006	(0.005-0.007)
θ_{SWAP}	0.003	(-0.001-0.007)	0.000	(-0.001-0.001)
$\hat{\rho}_{FULL}$	0.19		0.01	
posterior median of τ^2	0.082	(0.000-0.205)	0.007	(0.000-0.034)

Table 4. Team starter-variance split-sample results. Parentheticals are 0.95 CIs for θ parameter estimates, 0.95 HDI for τ^2 posterior.

The high predominance of chance in the observed results is apparent. For 1900-1910 teams, $\hat{\rho}_{FULL} = 0.19$, signifying that over 80% of the observed starter variance in a full season is due to chance rather than any persistent team state. For the combined half-season teams in the OWN sample, the starter-variance impact, which combines chance and true-state contributions, is 0.014 (0.95 CI: 0.010-0.018) winning percentage points per standard deviation. For the combined half-season teams in the SWAP sample, the impact, now limited to the true-state contribution, is 0.003 (0.95 CI -0.001-0.007). Of the observed variance-impact on team wins during the dead-ball era, then, 79% was due to chance.

The contribution of chance was even larger in the 1911-2025 period. The fraction of observed full-season starter variance attributable to a true state is about 1% ($\hat{\rho}_{FULL} = 0.01$). The undifferentiated chance and true state impact of a +1 SD change in starter variance in the OWN sample is 0.006 (0.005-0.007) winning percentage points; the true-state impact within the SWAP sample is 0.000 (0.95 CI: -0.001-0.001).

It is also worth noting that in both periods the impact of starter variance was *larger* for the combined, half-season OWN sample teams than for the complete-season COMP sample teams. This result addresses the concern that the low reliability of team starter-variance in the tests could be a consequence of splitting the complete season sample into half ones (James, 2008). If that were the source of the low levels of reliability in the split-sample test, the *complete* season impacts would be larger—because the effect of true-state variance on team wins is necessarily diluted more by noise in a half season of games than in a complete one. When in contrast complete-season impacts are smaller than half-season ones, that means that half-season variance impacts are being inflated by chance contributions, which offset each other when combined.

Figure 5 extends the analysis by estimating Bayesian posterior effects. It estimates a wins per 162 effect of 0.92 (0.95 HDI: 0.00-1.59) for teams in the 1900-1910 period—in contrast to an observed effect of 2.47 wins per 162 (0.95 CI: 1.56-3.38). For teams in the 1911-2025 period, the estimated posterior is *0.09* wins per 162 (0.95 HDI: 0.00-0.21), as opposed to an observed effect of 0.77 (0.57-0.97). In the eyes of a practical evaluator, these are the true premiums a team composed of uniformly high-variance rather than mean-variance starters would be expected to net, holding starter run suppression constant.

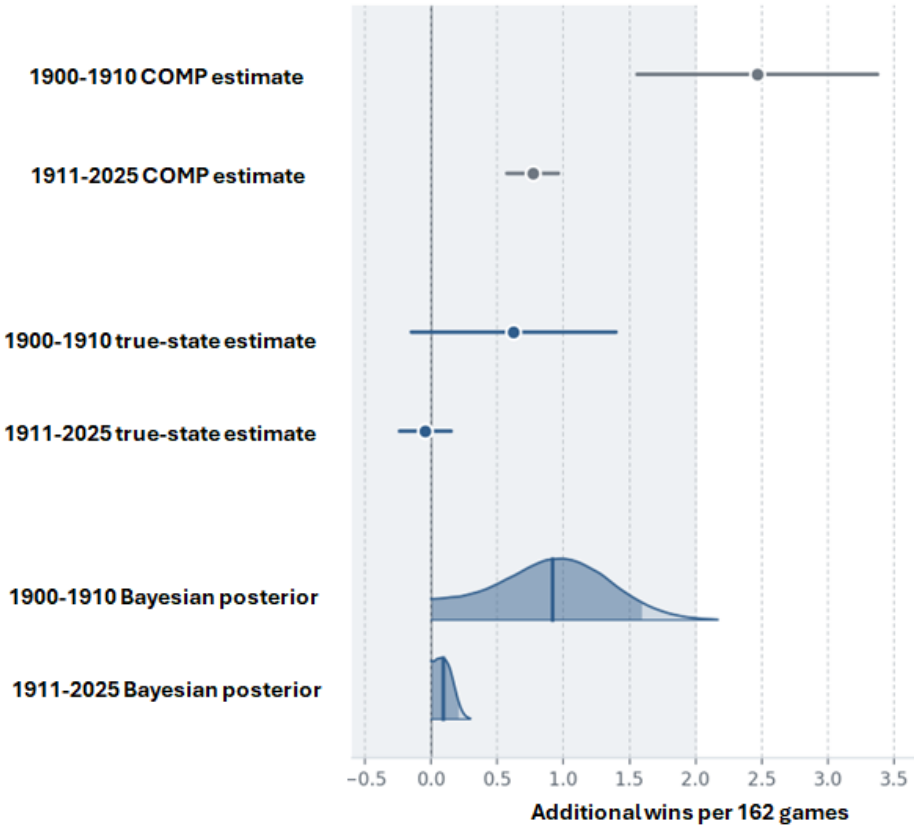


Figure 5. Separating true-state from chance influences of high starter variance on team wins. Additional wins are difference between team at 90th percentile and 50th percentile of starter variance at mean number of runs scored per game over relevant period. True-state and Bayesian estimates based on split-sample tests (Table 4). Brackets for COMP and true-state estimates are 0.95 CI; density distributions for Bayesian posteriors are 0.95 HDI.

6. Dueling metrics

6.1. Overview

The predominant metrics of pitcher value rely on run suppression and ignore how runs are distributed. This is exactly why Brill and Wyner (2024) devised Grid WAR, which they defend as superior on account of its sensitivity to starter variance.

As mentioned, however, Brill and Wyner do not present any evidence to show either that Grid WAR outperforms simple run-suppression measures or that any such difference in performance is attributable to Grid WAR's variance sensitivity. The analysis in this section addresses these matters.

It compares Grid WAR to two run-suppression measures. The first is FanGraphs' pitcher-runs saved metric (PRS), the run-suppression measure to which Brill and Wyner compare Grid WAR.

The other is eFIP PRS. Like FanGraphs's PRS, eFIP PRS is based on fielding independent pitching outcomes—namely, strikeouts, walks, home runs allowed, and hit batters. Unlike FanGraphs's methodology, however, eFIP applies empirically derived, season-specific weights to these outcomes.

The actual contribution of starting pitchers to team value cannot be directly observed. Accordingly, the only way to validate such measures is at the team level—by aggregating the scores of the teams' individual starting pitchers and testing how well that team score predicts team performance. For this purpose, a regression analysis was performed that added FanGraphs's PRS, eFIP PRS, and Grid WAR in turn to a model that otherwise used a common set of predictors for runs scored, relief-pitcher quality, and fielding.[9]

This strategy allows inspection of the respective impacts of the measures to team performance. If Grid WAR enjoys an advantage over the FanGraphs and eFIP runs-saved measures by virtue of Grid WAR's unique sensitivity to starter variance, then one might reasonably expect Grid WAR to add an increment of explanatory power that exceeds those of the respective runs-saved measures.

6.2. Derivation of eFIP

FanGraphs's PRS, the run-saved metric featured in Brill and Wyner (2024), suffers from several defects. First, it applies fixed, ahistorical weights to the components of its fielding-independent pitching (FIP) measure (Slowinski, 2010). It is not obvious that the role of hit batsmen has likely changed in importance over the history of major league baseball, but the relative impacts of strikeouts, home runs allowed, and walks clearly have; a metric that ignores these changes will thus pay a steep toll in explanatory power. Second, FanGraphs adds to its measure of fielding-independent pitching an ERA-mimicking constant (Weinberg, 2017) that results in bias when the site thereafter adjusts it with its multiplicative park factors (Fünf, 2026b).[10] Moreover, FanGraphs isolates the impact of a pitcher from his supporting

[9] The FanGraphs and eFIP PRSs were constructed using the performance of pitchers only when appearing in starting roles.

[10] Both FanGraphs's pitcher-runs saved measure and Grid WAR include park-factor adjustments. Their methods for doing so differ, however, in at least one material respect. For FanGraphs, park factors are used to make a "back end" adjustment: that is, after computing an individual pitcher's runs saved, his total is adjusted upward or downward depending on whether he pitched in a home-team park determined to be hitter-friendly or -unfriendly relative to the league mean (Slowinski, 2012). This adjustment results in a counterfactual estimate of his run-suppression capacity that necessarily biases to some degree the estimate of how his performance contributed to his own team's record in

fielders with fielding-runs saved measures that, over the course of the last quarter century at least, are demonstrably inferior to others (Kahan, 2025).

These concerns motivated the paper's formation of eFIP. This section describes how *e*FIP was formed and used to measure pitching-runs saved.

Conceptually, a team's "pitching runs *allowed*" can be thought of as a notional number of runs that a team's pitchers are responsible for independent of the runs avoided or surrendered by the team's fielders. Pitching-runs allowed can thus be either greater or less than the actual number of runs yielded by the team's pitchers depending on whether the team's fielders *saved* the team runs by superior play or *cost* the team runs by inferior play. Accordingly, for each team *t* in season *s*, one can compute pitching-runs allowed by *adding* to the total team runs allowed the quantity (positive or negative) that represents its respective *fielding runs saved*:

$$PRA_{ts} = RA_{ts} + FRS_{ts}. \tag{25}$$

The fielding-runs saved measure used here was one independently shown to generate the highest variance explained in team runs allowed after controlling for fielding-independent pitching (Kahan, 2025).

eFIP is formed by regressing team pitching runs-allowed per-inning on the teams' per-inning rates of FIP outcomes,

$$\left(PRA / IP\right)_{ts} = \alpha_s + \beta_{1s}\left(\frac{K_{ts}}{IP_{ts}}\right) + \beta_{2s}\left(\frac{BB_{ts}}{IP_{ts}}\right) + \beta_{3s}\left(\frac{HBP_{ts}}{IP_{ts}}\right) + \beta_{4s}\left(\frac{HR_{ts}}{IP_{ts}}\right) + \epsilon_{ts}, \qquad \epsilon_{ts} \sim \mathcal{N}\left(0, \frac{\sigma_s^2}{IP_{ts}}\right). \tag{26}$$

This model was used in turn to determine the *expected pitching-runs allowed per inning*—or eFIPs—for team pitchers, the scores of which were measured separately depending on whether they served in a relief or starter role:

$$eFIP_{ts}^{PR} = \hat{\alpha}_s + \hat{\beta}_{1s}\left(\frac{K_{ts}^{PR}}{IP_{ts}^{PR}}\right) + \hat{\beta}_{2s}\left(\frac{BB_{ts}^{PR}}{IP_{ts}^{PR}}\right) + \hat{\beta}_{3s}\left(\frac{HBP_{ts}^{PR}}{IP_{ts}^{PR}}\right) + \hat{\beta}_{4s}\left(\frac{HR_{ts}^{PR}}{IP_{ts}^{PR}}\right) \quad PR \in \{relievers,\ starters\}. \tag{27}$$

Finally, reliever or starter *e*FIPs can be converted into team reliever or starter season pitching-runs saved. First, the league weighted average *e*FIP for the relevant season is calculated:

games in which he pitched. Brill and Wyner (2024, pp. 298-299), in contrast, make a "front end" park-factor adjustment: that is, the scoring characteristics of every park pitched in are a component of the per-game Grid WAR estimate of how a pitcher's performance affected the incremental probability of a team win. Accordingly, Grid WAR estimates of pitching proficiency will be even more tightly calibrated to estimating a pitcher's value to his team than would an estimator that ignores park factors—not to mention one that, like FanGraphs's, deliberately effaces park-specific effects. This feature of Grid WAR gives it an explanatory-value advantage over FanGraphs's pitcher runs-saved measure, and to a lesser extent over eFIP. Unfortunately, nontransparency in FanGraphs's computation methods made it impossible to remove the park factor adjustments from its published estimates of team starter runs saved. Grid WAR will also enjoy a smaller advantage over eFIP in this regard. Because park-factor effects is likely negligible (Brill & Wyner, 2024, p. 311), the disparate treatment of park factors under the three metrics is unlikely to affect the inferences to be drawn from the analysis results reported here.

$$\overline{eFIP_s} = \frac{\sum_t eFIP_{ts}^{TEAM} IP_{ts}}{\sum_t IP_{ts}}. \quad (28)$$

eFIP times innings pitched equals expected pitching runs allowed. Accordingly, determining pitching runs *saved* for the relevant set of team pitchers requires multiplying their collective *e*FIP (27) by their aggregate innings pitched and subtracting that quantity from the weighted league average *e*FIP multiplied by the same number of innings pitched:

$$eFIP_PRS_{ts}^{PR} = \left(\overline{eFIP}_s - eFIP_{ts}^{PR}\right) IP_{ts}^{PR}. \quad (29)$$

A positive value denotes that the relevant set of team pitchers prevented more runs, and a negative value that they yielded more runs than expected relative to the league average.

As should be clear, the eFIP-derived measure of starting pitcher PRS—like FanGraphs's—is variance insensitive. It appraises pitcher value entirely on the rate at which team starters suppress runs without regard to the way in which runs suppressed are grouped across games.

6.3. Regression model

A regression analysis was constructed to measure the respective team-performance impacts of FanGraphs PRS, eFIP PRS, and Grid WAR as measures of starting pitchers. The values for FanGraphs PRS and eFIP PRS are simply the number of pitching-runs saved derived from those measures divided by team games played.[11] For Grid WAR, the team measure consists of the sum of each team starting pitcher's Grid WARs divided by games played.[12]

First, a benchmark model was established for assessing team performance:

$$WP_{ts} = \beta_0 + \beta_R zRUNS_{ts} + \beta_F zFIELD_{ts} + \beta_{RP} zRELIEF_{ts} + \varepsilon_{ts}. \quad (30a)$$

In this step, the winning percentage of team *t* in season *s* is regressed on its season runs scored divided by games, its fielding runs saved divided by games, and its relief-pitching quality. The fielding-runs-saved measure is the same one used to construct eFIP (Kahan, 2025). The relief-pitching-quality measure is the eFIP-derived pitching-runs saved attributable to the team's relief pitchers divided by reliever innings pitched.[13] All variables are season-standardized to remove the impact of shifting run environments from the estimation of the impact of the specified predictor on team winning percentage.

Second, the respective starter-pitching metrics are entered in turn, each of which is likewise season standardized:

[11] FanGraphs's "pitching runs [saved] above average" measure was used for the analysis.

[12] The division of each metric by games removes effects associated with differences in games played in a season.

[13] Team relief-pitcher runs-saved—whether measured in total over a season or per game over a season—are endogenous to starter proficiency (however measured): worse starters are removed from games earlier and thus create more opportunities for relievers to save runs. Dividing total reliever runs saved by total reliever-innings pitched generates a rate statistic measure of effectiveness, one unaffected by team starter quality.

$$WP_{ts} = \beta_0^{(m)} + \beta_R^{(m)} zRUNS_{ts} + \beta_F^{(m)} zFIELD_{ts} + \beta_{RP}^{(m)} zRELIEF_{ts} + \beta_{SP}^{(m)} zSP_{ts}^{(m)} + \varepsilon_{ts}^{(m)},$$
$$m \in \{zfgRS, zefipRS, zGWAR\}. \quad (30b)$$

The respective impact of each measure can then be determined by assessing how much the addition of each one increases model R^2 relative to the benchmark step.

6.4. Results

As expected, adding the team starting-pitching measures substantially increased variance explained. The impact, however, was uneven across the metrics. The addition of FanGraphs's PRS increased variance explained by seven percentage points, to 0.79. Both Grid WAR and eFIP PRS, however, increased variance explained to 0.83, or 11 percentage points.[14]

	Main Model				Supplemental Model		
	1	*2*	*3*	*4*	*5*	*6*	*7*
z_runs	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**
	(32.00)	(32.89)	(39.27)	(36.66)	(33.23)	(39.63)	(36.64)
z_field	**0.02**	**0.02**	**0.02**	**0.01**	**0.02**	**0.02**	**0.01**
	(20.90)	(23.00)	(24.42)	(9.59)	(22.64)	(24.03)	(9.64)
z_prip	**0.03**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**
	(25.95)	(23.54)	(16.24)	(18.37)	(22.24)	(16.19)	(18.38)
z_fgPRS		**0.02**			**0.02**		
		(24.43)			(24.72)		
z_eFIPRS			**0.03**			**0.03**	
			(27.30)			(27.27)	
z_GWAR				**0.03**			**0.03**
				(28.79)			(28.99)
z_SVAR					**0.003**	**0.003**	0.001
					(3.67)	(4.07)	(1.07)
cons	**0.50**	**0.50**	**0.50**	**0.50**	**0.50**	**0.50**	**0.50**
	(6842.14)	(6824.50)	(6801.57)	(6799.69)	(6822.33)	(6800.54)	(6798.57)
N	1,890						
R^2	0.72	0.79	0.83	0.83	0.79	0.83	0.83
ΔR^2 step 1		**0.07**	**0.11**	**0.11**			
ΔR^2 step2					**0.002**		
ΔR^2 step 3						**0.001**	
ΔR^2 step 4							0.0001

Table 5. Team starter impact model. Sample consists of all AL/NL teams from 1952 to 2025. Outcome variable is team winning percentage. Weighted least squares (games). Raw beta coefficients; clustered standard errors by season; *t*-statistics denoted parenthetically. **Bolded** coefficients and delta R^2s are significant at $p < 0.05$.

This result does not support Brill and Wyner's defense of Grid WAR. It is true that Grid WAR explained more variance than FanGraphs's starter runs-saved measure. But because Grid WAR performed no better than the variance-indifferent eFIP starter runs-saved measure, the superior explanatory power of Grid WAR over FanGraphs's measure cannot be attributed to the former's sensitivity to starter variance.

[14] The addition of the Grid WAR predictor reduced the size of the fielding-runs saved effect (Table 5, step 4). This suggests that Grid WAR, unlike FanGraphs's and eFIP's starter run-saved measures, includes a portion of the differences in team fielding proficiency.

6.5. Supplemental model

The test presented is not perfect. Grid WAR enjoys no explanatory advantage over eFIP, a variance-indifferent measure of pitcher value. But it is at least conceivable that this finding reflects the compensatory role that starter variance makes for a deficiency in Grid WAR's assessment of starter run suppression.

If this is so, adding a starter-variance regressor to the model should break the tie. Because Grid WAR already reflects starter variance, the effect of adding this predictor should be nil for the model that uses Grid WAR as the starting-pitcher measure. But eFIP is variance-indifferent; if it overcomes this defect by superior sensitivity to starter run suppression, the addition of the starter variance predictor should vault the model that includes both into the lead. For comparison, the starter-variance predictor was also added to the model that used FanGraphs's PRS metric.

The results are reflected in Table 5, steps 5-7. As expected, the addition of starter variance added no meaningful explanatory power to the model that uses Grid WAR to assess starter-pitcher value (Table 5, step 7). However, adding that same predictor increased the eFIP PRS model by only 0.1% (Table 5, step 6). This increment seems too small to support the conjecture that the explanatory equivalence of Grid WAR and eFIP stems from offsetting asymmetries in their respective attention to run suppression and starter variance. The better explanation is that they assign near equal weight to run suppression, which is what drives both metrics.

Adding starter variance increased by 0.2% the explanatory power of the model that relied only on FanGraphs's PRS (Table 5, step 5). This effect does not approach the R^2 difference between models that featured Grid WAR, on the one hand, and FanGraphs's starter runs-saved measure, on the other (Table 5,, steps 2, 4). Accordingly, it is clear that Grid WAR's explanatory advantage over FanGraphs's WAR estimator is effectively attributable to Grid WAR's superior sensitivity to starter run suppression, not its unique sensitivity to starter variance.

7. Decomposing Grid WAR

7.1. Overview

The analysis presented in this section decomposes the elements of starter value measured by Grid WAR at the individual-pitcher level. Grid WAR is designed to remedy the neglect of starter variance in conventional run-suppression measures; it can achieve this aim, obviously, only if starter variance makes a meaningful contribution to individual pitcher Grid WAR scores over and above the contribution of starter run suppression. Separating the relative weights that Grid WAR affords these components of starter value can determine whether this is the case.

In addition, decomposing Grid WAR scores in this manner complements the analysis from the last section. The results of that analysis supported the conclusion that Grid WAR's performance relative to the FanGraphs and eFIP starter runs-saved measures was not consequentially affected by Grid WAR's variance insensitivity but arose instead from Grid WAR's fidelity to differences in starter run suppression. If this conclusion is sound, then decomposing Grid WAR at the individual-pitcher level should reveal that Grid WAR's variance sensitivity makes no material contribution to starter Grid WAR scores.

7.2. WLS model

Generally. A weighted least squares regression model was constructed to measure the respective impacts of run suppression and starter variance on Grid WAR.

$$
\begin{aligned}
zGWAR_{is} &= \beta_0 + \beta_R zRAPIP_{is} + \beta_{R2}(zRAPIP_{is})^2 + \delta'\mathbf{SP}_{is} + zRAPIP_{is}\phi'\mathbf{SP}_{is} \\
&+(zRAPIP_{is})^2\psi'\mathbf{SP}_{is} + \beta_V zSVAR_{is} + \beta_{RV} zRAPIP_{is} zSVAR_{is} \\
&+\beta_{R2V}(zRAPIP_{is})^2 zSVAR_{is} + zSVAR_{is}\omega'\mathbf{SP}_{is} + zRAPIP_{is} zSVAR_{is}\xi'\mathbf{SP}_{is} \\
&+(zRAPIP_{is})^2 zSVAR_{is}\zeta'\mathbf{SP}_{is} + \varepsilon_{is}, \quad \varepsilon_{is} \sim \mathcal{N}\left(0, \sigma^2 \frac{\overline{G_s}}{G_{is}}\right)
\end{aligned}
\tag{31}
$$

In the model, *season-mean* Grid WAR score for pitcher *i* in season *s* is estimated on the basis of a series of pertinent explanatory variables. These include rate of runs allowed, measured at the inning level (RAPIP); the square of that rate (RAPIP^2); starter-usage period (Figure 3);[15] and the interactions of the usage period with RAPIP and RAPIP^2. A quadratic form was fit after inspection of the model residuals revealed a nonlinear relationship between season mean Grid WAR scores and the rate of runs allowed.

	1		*2*		*3*		*4*	
zRAPIP	**-0.96**	(-114.60)	**-0.97**	(-121.86)	**-0.97**	(-122.01)	**-0.97**	(-121.89)
zRAPIP^2	**0.07**	(16.71)	**0.07**	(17.67)	**0.07**	(17.60)	**0.07**	(17.52)
1984_2013	-0.02	(-1.16)	-0.02	(-1.15)	-0.02	(-1.16)	-0.02	(-1.15)
2014_2025	-0.02	(-1.59)	-0.02	(-1.74)	-0.02	(-1.67)	-0.02	(-1.71)
zRAPIP_x_1984_2013	0.00	(0.26)	0.00	(0.06)	0.00	(0.06)	0.00	(0.05)
zRAPIP_x_2014_2025	**-0.03**	(-2.25)	**-0.03**	(-2.13)	**-0.03**	(-2.13)	**-0.03**	(-2.08)
zRAPIP^2_x_1984_2013	0.01	(1.11)	0.01	(1.11)	0.01	(1.11)	0.01	(1.11)
zRAPIP^2_x_2014_2025	0.00	(0.23)	0.00	(0.33)	0.00	(0.17)	0.00	(0.25)
zSVAR			**0.11**	(14.81)	**0.11**	(14.74)	**0.11**	(13.34)
zRAPIP_x_zSVAR			**0.02**	(4.29)	**0.01**	(2.38)	**0.01**	(2.29)
zRAPIP^2_x_zSVAR			0.00	(0.26)	0.00	(0.35)	0.00	(0.39)
z_svar_x_1984_2013			0.01	(0.59)	0.01	(0.58)	0.00	(0.45)
z_svar_x_2014_2025			-0.01	(-0.76)	-0.01	(-0.71)	0.00	(-0.19)
zRAPIP_x_zSVAR_x_1984_2013					0.00	(-0.03)	0.00	(-0.04)
zRAPIP_x_zSVAR_x_2014_2025					0.01	(1.42)	0.01	(1.47)
zRAPIP^2_x_zSVAR_x_1984_2013							0.00	(0.05)
zRAPIP^2_x_zSVAR_x_2014_2025							0.00	(-0.80)
constant	-0.01	(-0.88)	-0.01	(-0.65)	-0.01	(-0.65)	-0.01	(-0.65)
N	10,677							
R^2	0.90		0.92		0.92		0.92	
ΔR^2			**0.01**		0.00		0.00	
AIC	6116		4,618		4,619		4,621	

Table 6. Grid WAR decomposition model. Sample consists of all AL/NL pitchers with ≥ 11 starts from 1952 to 2025. Outcome variable is season mean Grid WAR. Weighted least squares (games). Raw beta coefficients; *t*-statistics denoted parenthetically; standard errors clustered by individual pitcher. **Bolded** coefficients and delta R^2s are significant at $p < 0.05$.

The model also comprises terms relating to starter variance. These include the starter-variance level; the *interaction* of starter-variance level with RAPIP and RAPIP^2; and the three-way interaction of those variables with starter-usage period. Individual starter variances for this analysis were formed in the same manner as were the team-starter variances [(12)-(15)] but were computed using individual starters'

[15] Because Grid WAR scores are available only after 1951, only usage periods three through five figure in the analysis.

RAPIPs and per start residuals, not the teams'.[16] Including the interaction of starter variance and RAPIP tests whether (as suggested by the ETEL simulation) the contribution of starter variance depends on rate of runs allowed. Season-mean Grid WAR scores, RAPIP, and starter variance were all season-standardized in order to remove the impact of shifting run environments on estimation of the relevant effects.

The sample was confined to pitchers with 11 or more starts in a given season. For pitchers with low numbers of starts, log variances are ill-behaved and can generate extreme values that distort the regression slopes even with weighting. The standard rule of thumb is a minimum 10 degrees of freedom as the "safe" level for log-transformed variances (Bartlett & Kendall, 1946).

Results. The results are reported in Table 6. The model predictors were entered in steps to isolate the impact of starter variance as well as its associated interaction terms. This procedure established that starter variance does contribute to season mean Grid WAR scores. It also determined, consistently with the ETEL simulation, that the impact is conditional on pitcher run suppression: the higher the rate at which a pitcher allows runs, the greater the contribution a high level of variance made to improving his season mean Grid WAR (Table 6, step 2). These patterns did not vary across the starter usage periods (Table 6, steps 3-4).

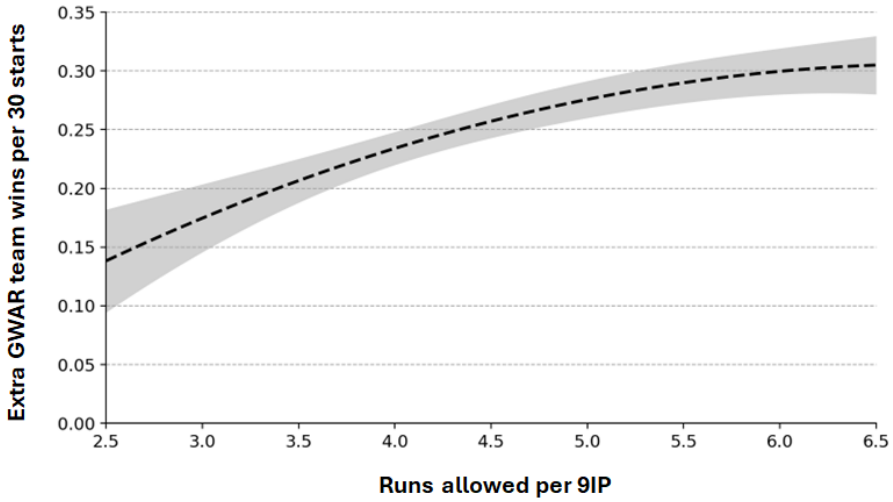


Figure 6. Grid WAR equivalent of extra team wins per 30 starts attributable to pitcher at 90th percentile of starter variance. Dashed line plots over mean runs allowed per 9 innings the incremental number of additional Grid WAR projected team wins (Table 6, step 2) for a pitcher at 90th as opposed to 50th percentile of starter variance over the course of 30 starts. Gray-shaded region is 0.95 CI.

Relative to run suppression, however, the impact of variance on pitcher Grid WARs is extremely small. Pitchers' RAPIP on its own accounts for 90% of the variance in pitchers' season mean Grid WARs. Starter variance added only 1.3% on top of that (Table 6, step 2). In sum, pitcher Grid WAR scores are overwhelmingly attributable to their proficiency in suppressing opponent runs.

This practical import of these effects are illustrated in Figure 6. For a pitcher who allowed 4.5 runs per 9 innings (the period mean), being in the 90th as opposed to the 50th starter-variance percentile generated the Grid WAR equivalent of only 0.26 additional team wins per 30 starts (0.95 CI: 0.24-0.27). For a pitcher who allowed 6.0 runs per 9 innings (95th percentile), the Grid WAR equivalent of being in

[16] In residualizing the variances on log runs per start, a single slope was fit for the entire sample of 1952-2025 starting pitchers.

the 90th percentile would be the Grid WAR equivalent of 0.30 team wins per 30 starts (0.95 CI: 0.28-0.32).

7.3. Split-sample analysis

Generally. Just as in the case of team-starter variance, in the case of individual-pitcher starter variances, a practical evaluator—a GM, say, or a bookmaker—will not take the WLS results at face value. The reason, again, is that the WLS model is completely insensitive to the respective contributions of chance fluctuations and genuine, between-pitcher differences in variance rates. The two must be disentangled before any valid inferences can be drawn about the contribution that *true-variance-state* differences make to pitcher season-mean Grid WAR scores.

Again, the separation of true-state and chance impacts can be achieved with a split-sample test. Here, the starting efforts of pitchers with at least 22 season starts—the minimum necessary to exclude pitchers with ill-behaved log-transformed variances (Bartlett & Kendall, 1946)—were alternately assigned to even-numbered and odd-numbered start subsamples ("E" and "O," respectively). Starter variances, RAPIPs, and sample mean Grid WAR scores were calculated for pitchers in both.

Then, in a procedure equivalent to the one used for the team-level split-sample test, three separate regression models were applied to the combined E and O subsamples. In OWN, mean Grid WAR scores were regressed on pitchers' own subsample RAPIPs and starter variances:

$$\begin{aligned} zGWAR_{ish} &= \beta_0 + \beta_R zRAPIP_{ish} + \beta_{R2}(zRAPIP_{ish})^2 + \theta_{OWN} zSVAR_{ish} \\ &+ \beta_{RV} zRAPIP_{ish} zSVAR_{ish} + \beta_{R2V}(zRAPIP_{ish} zSVAR_{ish})^2 + \varepsilon_{ish} \\ \varepsilon_{ish} &\sim \mathcal{N}\left(0, \sigma^2 \frac{\overline{G_{sh}}}{G_{ish}}\right). \end{aligned} \tag{32a}$$

In SWAP, mean Grid WAR scores were regressed on pitchers' own subsample RAPIPs and their variance levels from the *alternative* subsample:

$$\begin{aligned} zGWAR_{ish} &= \beta_0 + \beta_R zRAPIP_{ish} + \beta_{R2}(zRAPIP_{ish})^2 + \theta_{SWAP} zSVAR_{ish'} \\ &+ \beta_{RV} zRAPIP_{ish} zSVAR_{ish'} + \beta_{R2V}(zRAPIP_{ish} zSVAR_{ish'})^2 + \varepsilon_{ish} \\ \varepsilon_{ish} &\sim \mathcal{N}\left(0, \sigma^2 \frac{\overline{G_{sh}}}{G_{ish}}\right). \end{aligned} \tag{32b}$$

Finally, in COMP, complete-season mean Grid WARs were regressed on each pitcher's complete season RAPIP and complete season starter variances:

$$\begin{aligned} zGWAR_{is} &= \beta_0 + \beta_R zRAPIP_{is} + \beta_{R2}(zRAPIP_{is})^2 + \theta_{COMP} zSVAR_{is} \\ &+ \beta_{RV} zRAPIP_{is} zSVAR_{is} + \beta_{R2V}(zRAPIP_{is} zSVAR_{is})^2 + \varepsilon_{is} \\ \varepsilon_{is} &\sim \mathcal{N}\left(0, \sigma^2 \frac{\overline{G_{s}}}{G_{is}}\right). \end{aligned} \tag{32c}$$

The contributions of pitchers' true-state variances can be determined here in the same manner that the contributions of team true-state starter variances were determined. Because only pitchers' true-state

variances can influence their Grid WAR scores in the SWAP model, θ_{SWAP} represents the impact of true-state variance, disentangled from chance fluctuations, on pitchers' scores. Thus, the fraction of Grid WAR equivalent team wins attributable to chance, as opposed to true differences in individual pitcher variance states is the difference between θ_{OWN} and θ_{SWAP} divided by the former (18). Full-season reliability, $\hat{\rho}_{\mathrm{FULL}}$, can once more be determined by the Spearman-Brown step up for split-half reliability (19).

Bayesian posterior estimates can then be computed in the same fashion as well [(20)-(24)]. The half-season reliability $\hat{\rho}_{EO}$, now formed by the covariance of $z\mathrm{SVAR}_{isE}$ and $z\mathrm{SVAR}_{isO}$, supplies the estimate of τ^2. A uniform prior on τ is set over the range of 0 to τ_{max}. Two million values are drawn from the prior, $\tau \sim \mathrm{U}(0,\ \tau_{max})$, each of which is weighted by its likelihood (22); and 400,000 draws are resampled in proportion to those weights, forming the posterior. After each of the draws is stepped up to full-season reliabilities under the Spearman-Brown formula (23), the distribution can be converted to extra team wins per 30 starts at any desired quantile of starter variance (24).

Results. The results are reported in Table 7. The OWN model estimates that a +1 SD increase in starter variance is worth 0.0067 wins per start (0.95 CI: 0.0064-0.0070). For the SWAP model, the impact is 0.0001 (0.95 CI: -0.0002-0.0004). Accordingly, chance fluctuations make up 98.5% of the estimated impact of starter variance to the team wins associated with season-mean Grid WAR scores. The full-season reliability of true-state variance for a full season is negative, or effectively 0.00.

θ_{COMP}	0.0062	(0.0058-0.0066)
θ_{OWN}	0.0067	(0.0064-0.0070)
θswap	0.0001	(-0.0002-0.0004)
$\widehat{\rho}_{FULL}$	-0.001	
posterior median of τ^2	0.004	(0.000-0.019)

Table 7. Individual starter-variance split-sample results. N = 6,761 (all pitchers with ≥ 22 season starts, 1952-2025). Slope estimates converted from +1 SD impact of starter variance on season-standardized season mean Grid WAR to +1 SD impact of starter variance on raw season mean Grid WAR (wins): θ × the sample median of weighted season SDs of mean season Grid WAR. Parentheticals are 0.95 CIs for θ parameter estimates, 0.95 HDI for τ^2 posterior.

Again, the estimated impact of starter variance in the OWN model, while miniscule, is larger than the estimated impact for the COMP model, which is based on complete season observations rather than half-season ones. Again, this result helps dispel concern that the low reliability for season true-state variance is a consequence of splitting the sample (James, 2008).

The Bayesian estimates of the effect of high true-state variance is illustrated in Figure 7. For a pitcher at the 75th percentile in runs allowed per inning, the COMP Grid WAR benefit of being in the 90th percentile of starter variance (versus the 50th) was equivalent to 0.24 wins per 30 starts (CI: 0.23-0.26); the posterior estimate of the Grid WAR benefit associated with true-state variance for the same pitcher would be 0.020 wins (0.00-0.04). For a pitcher at the 25th percentile in RAPIP, the observed benefit shrinks from 0.20 wins per 30 starts (CI: 0.18-0.22) to 0.018 (0.00-0.05).

These are the best estimates, then, of the Grid WAR return of securing a pitcher of high starter variance. As is evident from their miniscule size, such a pitcher's season-mean Grid WAR—his expected per-start impact on the probability of a team win—is effectively a measure of his ability to prevent opponents from scoring runs.

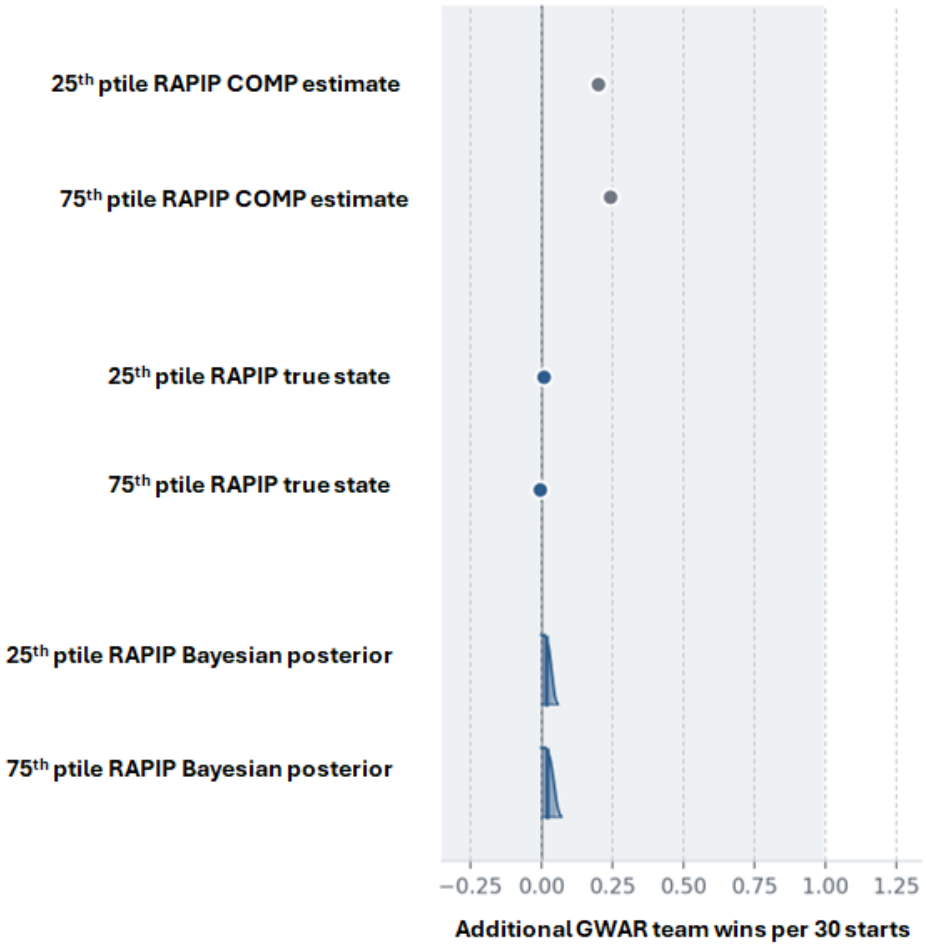


Figure 7. Separating true-state from chance influences of high starter variance on individual pitcher mean Grid WARs. Additional wins are difference between pitchers at 90th percentile and 50th percentile of starter variance at indicated percentile of mean runs allowed per game. True-state and COMP estimates based on split-sample tests (Table 7). Density distributions for Bayesian posteriors are 0.95 HDI.

8. Conclusion

8.1. Taking Stock

The goal of this paper was to assess VVAT—the variance value added thesis, which holds that greater starter variance contributes to team wins over and above starters' proficiency in avoiding runs. VVAT was explored through four independent forms of analysis: first, a simulation that generated counterfactual season performances for teams based on a nonparametric estimation of the relationship between runs allowed per game and runs-allowed variance; second, a historical assessment of the impact of how variance in team starter performance has affected team records since 1900; third, a comparison of the relative power of variance-sensitive and variance-indifferent metrics of starter proficiency to explain variance in team winning percentages; and finally an estimation of the contribution that starter variance makes to the evaluation of starting pitcher effectiveness under the leading variance-sensitive estimator, Grid WAR.

The overall conclusion is straightforward. Whatever effect starter variance contributes to differences in team wins is trivial and a product of random fluctuations, not persistent true differences in starter variance.

The triviality of the starter variance impact is attested to by the simulation and by the second and fourth analyses. The regression model for the impact of starter variance on the performance of historical teams showed that in the deadball era, variance in team starters (and runs allowed per game generally) plausibly accounted for the equivalent of about 2 wins per 162 games for teams at the tail end of the variance distribution—90th percentile and above. Thereafter, even teams with starter variances that extreme typically realized benefits of less than 1 win per 162 over the records predicted by the variance-indifferent Pythagorean Expectation Formula.

The returns were comparably weak for individual pitchers. Those who displayed the highest levels of variance over the last 75 years of MLB history likely generated less than 0.3 wins per 30 starts for their team as measured by Grid WAR (Brill & Wyner 2024).

These results, moreover, were concentrated in the weakest performers. Teams that benefitted most from starter variance were ones with anemic offenses; individual staters that did were the ones that had the highest rates of runs allowed. High scoring teams and high run-suppressing individual pitchers realized little if any benefit from higher variance. Differences of even 1 win per season might be of extreme consequence for teams at the cusp of a post-season berth. But teams of that caliber, the analyses showed, were unlikely to gain even that much from high levels of variance among their starters.

Moreover, these effects are a result of chance fluctuations only. When tests were performed to segregate from sampling error the impact of persistent variance starters, either among team starting staffs or individual pitchers, the posterior predicted benefits of starter variance fell effectively to zero: less than 1/10 of a win per 162 games for teams with starting staffs in the 90th percentile of starter variance, and less than .025 wins per 30 starts for individual pitchers at that level.

Accordingly, no practical evaluator would rely on starter variance in assessing individual pitcher value or team prospects. Apprised of a valid measure of expected runs allowed, there is nothing she can learn from variance that will allow her to form a stronger starting staff or to gauge the value of one pitcher versus another.

As a result, the impact of starter variance furnishes no basis for preferring one estimator of starting-pitcher value over another. Brill and Wyner (2024) identify Grid WAR as assigning pitchers multiple team wins per season—and dozens over a career—relative to conventional WAR measures. Impacts this large are not credible: they exceed the number of wins that variance in runs allowed per game can realistically be expected to deliver in even the best case, even when it results sheerly from chance. Moreover, it is clear that persistent states of starter variance have added nothing to the value of individual pitchers, even under Grid WAR itself. Whatever advantage Grid WAR enjoys over FanGraphs's WAR metric—the alternative estimator examined by Brill and Wyner—is a consequence of Grid WAR's greater sensitivity not to starter variance but to starter run suppression. Indeed, Grid WAR performs no better than a measure of fielding-independent pitching, eFIP, designed to avoid the demonstrated weaknesses of other variance-indifferent measures.

8.2. Tom Seaver and the uses of theory

In a semi-famous essay featuring Dennis Martinez, Earl Weaver, and a sports columnist, literary theorist Stanley Fish (1987) defended the independence of *doing* and *theorizing*, a dichotomy that spans activities as diverse as pitching baseballs and interpreting the Constitution of the United States.

But the sports analytics revolution might well be regarded as proving Fish wrong, particularly in the domain of baseball. Players might not be theorists. But there is no question that *as a result of theorizing*, pitchers, hitters, and fielders now play *differently* from how they did even a quarter century ago, and generate value for their teams more predictably and reliably as a result.

Nevertheless, unlike individual baseball games, seasons, or World Series, the *doing* of sports theorizing doesn't result in the identification of final winners and losers. There is a fundamental difference between sports contests and the empirical study of the same (or of anything else): the former come to a definitive conclusion, whereas the latter never does. Consistent with the logic of scientific discovery, empirical inquiry continues indefinitely, continuously accumulating evidence and formulating tests that result in at most *provisional* best understandings, which persist only until they are overcome by better ones (Popper, 1959).

This paper is part of that process, applied to one of the characteristics that contributes to the value of starting pitching—namely, starter variance. Taking up one position—VVAT—that has deepened understanding and perhaps temporarily altered a prevailing view of what makes starting pitching valuable, the paper has formulated tests, assembled data, and drawn a set of inferences therefrom that suggest an alternative answer. That answer isn't the final word; it can't be because of the nature of what empirically *theorizing* baseball entails. But at least for the time being, the conclusion, it is suggested, is that Tom Seaver's theory—and Dennis Martinez's—is the one statisticians should use to appraise pitching skill: measure consistency in the elements of performance that *suppress runs*—and the rest will take care of itself.

References

Bartlett, M. S., & Kendall, D. G. (1946). The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation. *Supplement to the Journal of the Royal Statistical Society*, *8*(1), 128-138. https://doi.org/10.2307/2983618

Baxamusa, S. (2006, November 6). *Are foolish consistencies the hobgoblins of starting pitchers?* The Hardball Times. https://tht.fangraphs.com/are-foolish-consistencies-the-hobgoblins-of-starting-pitchers/

Braunstein, A. (2010). Consistency and Pythagoras. *Journal of Quantitative Analysis in Sports*, *6*(1). https://doi.org/doi:10.2202/1559-0410.1215

Brill, R. S., & Wyner, A. J. (2024). Introducing Grid WAR: Rethinking WAR for starting pitchers. *Journal of Quantitative Analysis in Sports, 20*(4), 293–329.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322.

Chen, J., & Li, T. (2016). The shrinkage of the Pythagorean exponents. *Journal of Sports Analytics, 2*(1), 37–48. https://doi.org/10.3233/jsa-160017

Chen, K., Cheng, Y., Berkout, O., & Lindhiem, O. (2017). Analyzing proportion scores as outcomes for prevention trials: A statistical primer. *Prevention Science, 18*(3), 312–321. https://doi.org/10.1007/s11121-016-0643-6

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics, 86*(2), 221–241. https://doi.org/10.1016/S0304-4076(97)00115-2

Cleveland, W. S., Devlin, S. J., & Grosse, E. (1988). Regression by local fitting: Methods, properties, and computational algorithms. *Journal of Econometrics, 37*(1), 87–114.

Cochran, J. J., & Blackstock, R. (2009). Pythagoras and the National Hockey League. *Journal of Quantitative Analysis in Sports, 5*(2), 1–13.

Davenport, C., & Woolner, K. (1999, June 30). *Revisiting the Pythagorean theorem: Putting Bill James' Pythagorean theorem to the test*. Baseball Prospectus. https://www.baseballprospectus.com/news/article/342/revisiting-the-pythagorean-theorem-putting-bill-james-pythagorean-theorem-to-the-test/

Davidian, M., & Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association, 82*(400), 1079–1091. https://doi.org/10.2307/2289384

Eling, N., Richard, A. C., Richardson, S., Marioni, J. C., & Vallejos, C. A. (2018). Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Systems, 7*(3), 284–294.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing, 16*(2), 203–213. https://doi.org/10.1007/s11222-006-8450-8

Fish, S. (1987). Dennis Martinez and the Uses of Theory. *The Yale Law Journal*, *96*(8), 1773-1800. https://doi.org/10.2307/796396

Fünf, Xavier, Against Pythagorean Expectation Benchmarking (April 4, 2026). Available at SSRN: http://dx.doi.org/10.2139/ssrn.6520238

Fünf, X. (2026b). *FanGraphs' FIP boner*. BB Cards & Stats. https://bbcardstats.com/fangraphs-fip-boner/

Gassko, D. (2006, April 28). *Same old, same old*. Hardball Times. https://tht.fangraphs.com/same-old-same-old/

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*, 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://books.google.com/books?id=lV3DIdV0F9AC

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica, 44*(3), 461–465. https://doi.org/10.2307/1913974

Imbens, G. W., Spady, R. H., & Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica, 66*(2), 333–357. https://doi.org/10.2307/2998561

James, B. (1983). *The Bill James baseball abstract 1983*. Ballantine Books.

James, B. (2008). Mapping the fog: A response to "Clutch hitting and the Cramer test." *The Baseball Research Journal, 37*, 76–82.

Kahan, D. M. (2025). Fielding before and after baseball's great transformation. *Journal of Sports Analytics, 11*, https://doi.org/10.1177/22150218251349133

Kaplan, E. H., & Rich, C. (2017). Decomposing Pythagoras. *Journal of Quantitative Analysis in Sports, 13*(4), 141–149.

Lepore, J. F. (2018). *Understanding the market value and utility of high-variance starting pitchers* [Unpublished master's thesis]. University of California, Los Angeles.

Miller, S. J. (2007). A derivation of the Pythagorean won-loss formula in baseball. *Chance, 20*(1), 40–48.

Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. Springer.

Petti, B. (2012, March 7). *Does consistent play help a team win?* FanGraphs. https://blogs.fangraphs.com/does-consistent-play-help-a-team-win/

Popper, K. R. (1959). *The logic of scientific discovery*. Basic Books.

Rodbard, D., & Frazier, G. R. (1975). Statistical analysis of radioligand assay data. In *Methods in enzymology* (Vol. 37, pp. 3–22). Academic Press. https://doi.org/10.1016/S0076-6879(75)37003-1

Ruser, J. W. (1991). Workers' compensation and occupational injuries and illnesses. *Journal of Labor Economics, 9*(4), 325–350. https://doi.org/10.1086/298272

Schatz, A., Alamar, B., Barnwell, B., Connelly, B., Farrar, D., Fremeau, B., Gardner, D., Gower, T., Macey, N., & McCormick, S. (2011). *Football Outsiders almanac 2011: The essential guide to the 2011 NFL and college football seasons*. CreateSpace Independent Publishing Platform.

Schell, M. J. (1999). *Baseball's All-time Best Hitters: How Statistics Can Level the Playing Field*. Princeton University Press.

Schell, M. J. (2005). *Baseball's all-time best sluggers: Adjusted batting performance from strikeouts to home runs*. Princeton University Press.

Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics, 35*(2), 634–672.

Seidman, E. (2009, March 3). *Is consistency key?: Adventures in pitching*. Baseball Prospectus. https://www.baseballprospectus.com/news/article/8579/is-consistency-key-adventures-in-pitching/

Slowinski, P. (2010, February 15). *FIP*. FanGraphs. https://library.fangraphs.com/pitching/fip/

Slowinski, P. (2012, March 28). *WAR for pitchers*. FanGraphs. https://library.fangraphs.com/war/calculating-war-pitchers/

Smith, A. F. M., & Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician, 46*(2), 84–88. https://doi.org/10.1080/00031305.1992.10475856

Smith, D. M., & Faddy, M. J. (2019). Mean and variance modeling of under-dispersed and over-dispersed grouped binary data. *Journal of Statistical Software, 90*(8), 1–20.

Solomon, B. (1997). *The baseball timeline: The day-by-day history of baseball, from Valley Forge to the present day*. Avon Books. https://books.google.com/books?id=y_Fb3LgT2OUC

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295.

Weinberg, N. (2017, April 17). *Calculating pitcher WAR, a complete example*. FanGraphs. https://library.fangraphs.com/calculating-pitcher-war-a-complete-example/