A correction to fWAR that takes into account the relationship between starting pitcher quality and length of start

Milutin Gjaja

January 8, 2025

Abstract

I use MLB pitching data from 2022 to 2024 and find a correlation between a starting pitcher's skill, as determined by fWAR, and the average number of innings he pitches per start. Based on this relationship, I claim that fWAR misvalues starting pitchers and propose a correction to the fWAR formula.

1 Introduction

WAR, or Wins Above Replacement, is a widely-used statistic in baseball that summarizes a player's total contribution on the field by a single number. The key concept behind the statistic is that of a "replacement-level player". Replacement-level players are players who are good enough to excel in the minors, but not good enough to get a consistent spot in a Major League team. They are widely available and can be readily signed for cheap. Because of their availability and talent level, they are a good baseline against which MLB players can be measured. Hence, the basic question WAR seeks to answer: "If we lost a player and had to replace him with a replacement-level players, how much value would we lose?" WAR expresses this value in wins, usually on a per-season basis. In addition, WAR allows a comparison between different positions and teams, providing a consistent standard across the entire league.

Unlike most statistics, there is no standard formula for WAR. The three most common variants are fWAR, bWAR, and WARP, which are calculated by the publications FanGraphs, Baseball Reference, and Baseball Prospectus, respectively. This paper focuses on fWAR; I discuss the other two formulas in the conclusion. I will demonstrate in this paper that there is a statistically significant relationship between the quality of a starting pitcher and the number of innings pitched per start. Because fWAR does not take this relationship into account, it consistently misvalues starting pitchers.

1.1 Explanation of fWAR Formula for pitchers

The full fWAR formula for pitchers is as follows[1]. I will discuss each variable in the formula in turn. Full details are provided in Appendix A.

$$\text{fWAR} = \left(\frac{\text{lgFIPR9} - \text{pFIPR9}}{\text{dRPW}} + \text{RL}\right) \cdot \frac{\text{IP}}{9} \cdot \text{Lev} + \text{Lgc}$$

fWAR measures the pitcher's skill, takes the difference with the league average, then adds the replacement level, which is the difference in skill between the league average and the average replacement pitcher. It then scales this quantity to the number of games the pitcher plays and adds a small adjustment to bring the league-wide WAR to a pre-determined number[2]. For relievers, the stat is scaled by an additional factor which seeks to account for the fact that the bullpen has a hierarchy; when a reliever is injured,

[1] WAR uses different formulas for pitchers and fielders because of how different their roles are. All the information below deals exclusively with the pitcher formula.

[2] The accepted standard is that the total league contribution sums up to 1000 WAR. This is divided between pitchers and fielders based on the distribution of league payroll: around 43% of the yearly payroll goes to pitchers, meaning they account for 430 WAR total. The league constant adjusts the sum of all fWAR for pitchers so that they add up to 430.

his replacement slots in at the bottom of the hierarchy, shifting everyone else up.

pFIPR9 is the formula's approximation of the pitcher's skill. It is based on a common stat called Field-Independent Pitching, or FIP, which it then slightly modifies. Because we'll be using pFIPR9 extensively, I explain the calculation in some detail.

$$\text{FIP} = \frac{13\text{HR} + 3(\text{BB} + \text{HBP}) - 2\text{K}}{\text{IP}} + \text{FIPc}$$

We begin with FIP. FIP measures a pitcher's skill independent of the rest of the team's defensive ability. To do this, it only takes into account events which do not involve fielders: home runs (HR), walks (BB), hit-by-pitch (HBP), and strikeouts (K). It scales these to a by-inning basis and adds a constant FIPc so that the league average FIP is the same as the league average ERA, another popular statistic[3].

$$\text{ifFIP} = \frac{13 \times \text{HR} + 3 \times (\text{BB} + \text{HBP}) - 2 \times (\text{K} + \text{IFFB})}{\text{IP}} + \text{ifFIPc}$$

The next step modifies FIP to include infield fly-balls (IFFB). While these involve fielders, they are virtually guaranteed to result in an out, giving them the same value as a strikeout. The ifFIPc calculation is straightfoward, taking the MLB average and scaling it to ERA.

$$\text{pFIPR9} = \frac{\text{ifFIP} + \text{mlbRA9} - \text{mlbERA}}{\text{pPF}/100}$$

pFIPR9 scales ifFIP to Runs Against per 9 Innings (RA9), which is done by adding the difference between the league averages of RA9 and ERA. The Park Factor (PF) is a measure of how beneficial a park is to a pitcher's performance based on factors like altitude, temperature, and dimensions; the higher the park factor, the more hitter-friendly the stadium is. The average PF is 100; most stadiums range between 97 and 103. We divide the whole calculation by PF/100 to get pFIPR9[4].

The rest of the fWAR formula is straightforward. lgFIPR9 is the league average of the statistic we just calculated. fWAR calculates AL and NL averages separately for lgFIPR9 (but not for ifFIPc) because of scheduling differences and the historical absence of the DH in the National League. dRPW, which stands for dynamic Runs Per Win, is used to convert the FIPR9 difference from runs/game to wins/game. RL is the difference between the replacement level and the league average. IP/9, essentially how many full games the pitcher has played, is used to convert the formula into total wins for the season, and Lgc is the adjustment to bring the league-wide fWAR to a pre-determined value. Since we're concerned with starting pitcher fWAR, we can disregard the leverage component, Lev[5].

One note: WAR is, by construction, an approximation. Per FanGraphs, "WAR is not meant to be a perfectly precise indicator of a player's contribution, but rather an estimate of their value to date."[6] Given two players with WAR = 6.1 and 5.9 respectively, we cannot be certain that one is definitively more valuable than the other, but we can be confident that both are around All-Star level and more valuable than a player with WAR = 2.3.

[3] Because pFIPR9 is based on FIP, it has an inverse relationship to skill: the lower the pitcher's pFIPR9, the higher his skill level.

[4] FanGraphs uses a modification of PF for its fWAR calculations, which regresses the regular PF halfway to the mean of 100. This accounts for the fact that a pitcher will play approximately half of his games in his home stadium.

[5] The full formulas for all of these stats can be found in Appendix A.

[6] Slowinski, 2010

2 Methods

For this study, I use pitcher data from 2022 to 2024. A starting year of 2022 is chosen because it is the first year that the universal DH was implemented. Because I am solely interested in starting pitchers, I want to remove relievers from the data set. To do this, I filter by two variables: GS >= 7 and GS/G >= 0.8[7]. The former represents approximately a quarter season's worth of starts. The latter removes starts from pitchers who are usually relievers[8]. After applying the filters, I have n = 510 observations. Data for most stats (HR, BB, HBP, K, IP, G, GS) is taken from Baseball Reference[9], which has a convenient .csv file download option. League averages, IFFB, and fWAR, as well as all bullpen data, is taken from FanGraphs[10]. Financial data to calculate Lgc is taken from Spotrac[11].

fWAR components are not available on FanGraphs. My calculation of fWAR, calc_fWAR, diverges slightly from the official formula. FanGraphs does not provide an overall team-by-team breakdown for players with multiple teams, so I use the league average Park Factor of 100 for them. These players number 36, roughly 7% of my observations. I omit leverage because I'm interested in starting pitchers. As a result of these changes, my Lgc is marginally different from the official calculation, but being a constant, it has no impact on the rest of the study.

3 Results

3.1 Relationship between fWAR and calc_fWAR

I conduct a preliminary linear regression between fWAR and calc_fWAR to ensure that the rest of the study holds for fWAR. The results of this regression are shown in Figure 1.

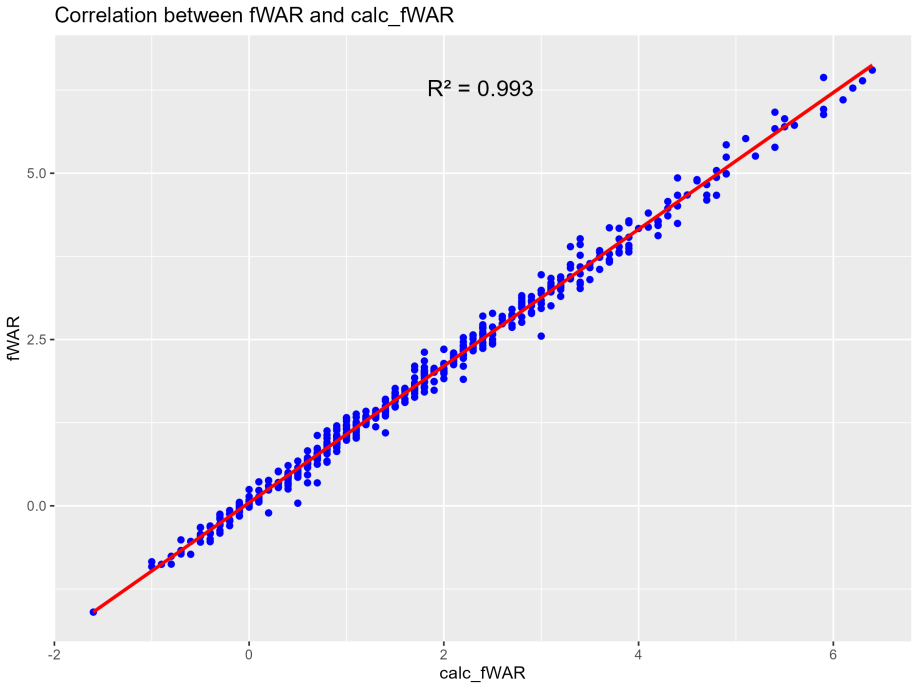


Figure 1: Relationship between fWAR and calc_fWAR

There is an extremely high correlation between calc_fWAR and fWAR ($R^2 > 0.99$, p-value <2.2e-16). There are no major outliers. The average discrepancy between the two is 0.12 and median 0.09, with

[7] G are total games played, GS are total games started. GS <= G for all pitchers.

[8] Occasionally, for reasons of injury or rest management, teams will have "bullpen games" where only relievers are used. These starts are unusual for a couple reasons, the primary being that the "starting" pitcher is only expected to pitch an inning or two before being replaced. It is these types of starts I'm eliminating from the data.

[9] www.baseball-reference.com

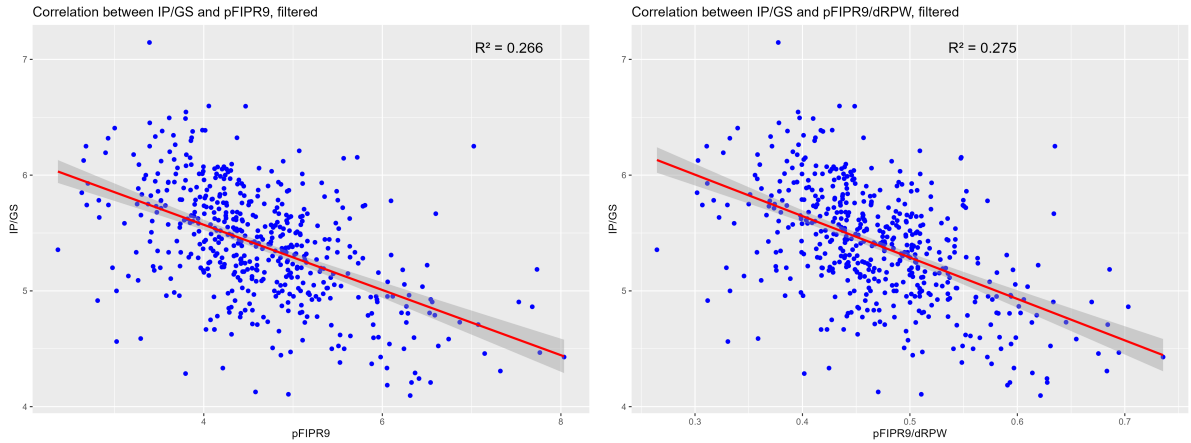
[10] www.fangraphs.com

[11] www.spotrac.com

three quarters of all calc_fWAR values falling within 0.18 of fWAR. As stated in Section 1, WAR is by definition an approximation; thus, this difference is negligeable for evaluating player skill. This evidence points to calc_fWAR being an ideal proxy of fWAR for future calculations.

3.2 Relationship between starting pitcher skill and number of innings pitched per start

fWAR's measure of a pitcher's skill is pFIPR9, or alternatively, pFIPR9/dRPW, its equivalent in wins/game. This quantity correlates strongly with the number of innings pitched per game. I conduct two linear regressions to study the relationship between pitcher skill and average innings pitched per start: IP/GS $\sim$ pFIPR9 and IP/GS $\sim$ pFIPR9/dRPW. The results are summarized in Figure 2 and Table 1.



(a) IP/GS and pFIPR9

(b) IP/GS and pFIPR9/dRPW

Figure 2: Relationship between IP/GS and skill

Measure	Coef.	Est.	Std. Error	t-value	Pr($>\|t\|$)	Res. SE	R^2	p-value
pFIPR9	Intercept	6.70	0.10	68.69	<2e-16	0.42	0.27	< 2.2e-16
	pFIPR9	-0.28	0.02	-13.59	<2e-16			
pFIPR9/dRPW	Intercept	7.08	0.12	57.87	<2e-16	0.42	0.27	< 2.2e-16
	pFIPR9/dRPW	-3.58	0.26	-13.90	<2e-16			

Table 1: Summary of relationship between IP/GS and skill

Both regressions yield similar results: $R^2 \approx 0.27$ and p-values $<$ 2.2e-16 for both. The t-stats for all coefficients are very high, indicating a statistically strong relationship. For every unit of difference of pFIPR9, the pitcher lasts a little under an out (0.33 IP) longer. Since pFIPR9 is scaled to RA9, this could reasonably be understood as follows: for every run per game on average a pitcher gives up, he lasts one out less per game.

To give an idea of the impact of this correlation, I've added three benchmarks to Figure 3, with results summarized in Table 2. Average is the mean of all observations (n=510), replacement-level are players with -0.5 <calc_fWAR< 0.5, and Elite are the top quartile of all observations by calc_fWAR. The green lines represent the average IP/GS for each group, while red lines are average pFIPR9 values. The additional yellow line is the average bullpen pFIPR9.

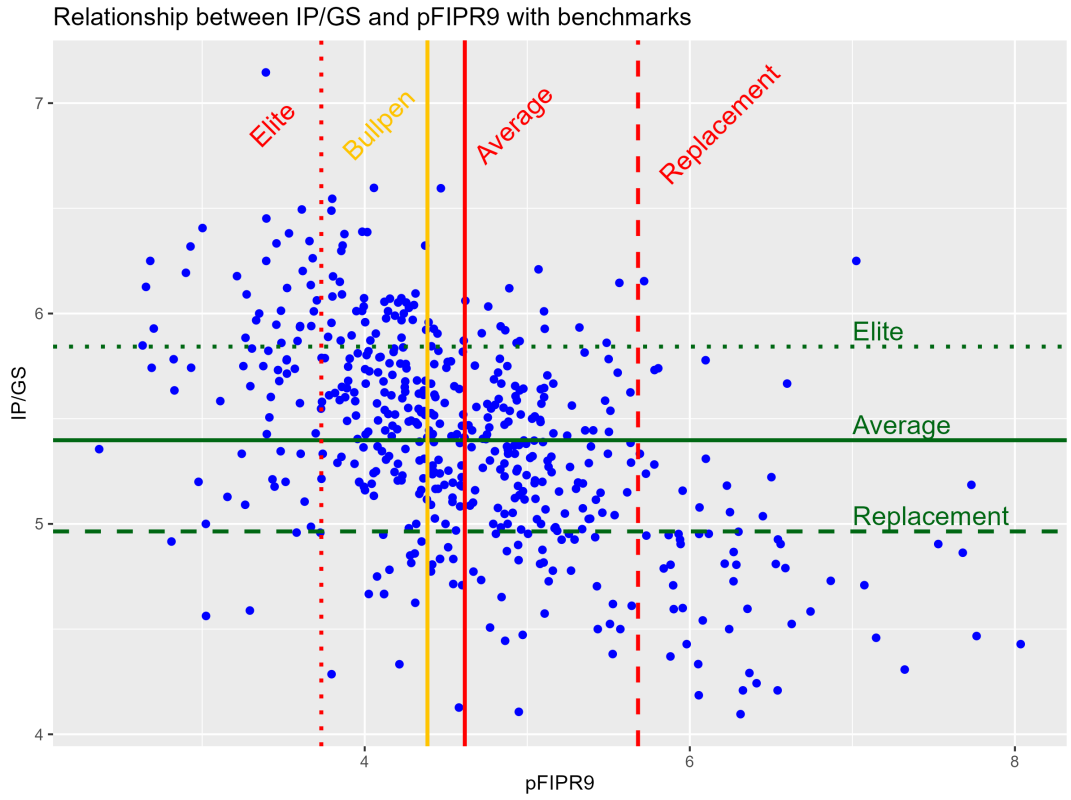


Figure 3: Relationship between IP/GS and pFIPR9 with benchmarks

Benchmark	IP/GS	pFIPR9
Replacement-level	4.96	5.68
Average	5.40	4.62
Elite	5.84	3.73
Bullpen	NA	4.39

Table 2: Summary of benchmarks

4 Results and correction

The relationship between the number of innings pitched per start and pFIPR9 shows a flaw in fWAR's assumption of a 1-for-1 substitution in workload between the pitcher and his replacement-level counterpart. Replacement-level pitchers last, on average, 4.96 innings per start, while the average MLB pitcher lasts 5.40. This creates a gap of about half an inning where the MLB pitcher would still be playing but his replacement would not. This gap doesn't disappear; instead, it goes to the bullpen. Therefore, for the length of this gap in every game, the real loss in skill is the difference between the pitcher and the team bullpen, not the pitcher and his replacement. The values in Table 2 give a quick idea why this matters: the average bullpen is slightly more effective than the average MLB starting pitcher, and far more effective than replacement-level starting pitchers.

While this gap may not amount to much over a single game, it adds up over the course of a season. The average number of games started per season in the dataset is GS ≈ 22.67; for players who play the entire season, this rises to GS ≈ 30. This translates to 10.0 IP or 13.2 IP each season where the bullpen would be pitching instead of the replacement, over a full game. This gap is particularly stark for elite starting pitchers, who last almost a full inning longer than their replacement-level counterparts: here, the difference over the length of a season is 19.9 IP or 26.4 IP, *nearly three full games worth of pitching credit*. This leads to fWAR overvaluing the contribution of a starting pitcher over his replacement, particularly for high-end starters.

To solve this issue, I propose the following correction to the fWAR formula for starting pitchers, called new_fWAR:

$$\text{new\_fWAR} = \left(\frac{\text{lgFIPR9} - \text{pFIPR9}}{\text{dRPW}} + \text{RL}\right) \times \left(\frac{4.96}{9} \cdot \text{GS}\right)$$

$$+ \left(\frac{\text{bpFIPR9} - \text{pFIPR9}}{\text{dRPW}}\right) \times \left(\frac{\frac{\text{IP}}{\text{GS}} - 4.96}{9} \cdot \text{GS}\right) + \text{nLgc}$$

Pitchers gets replacement-level credit for the average replacement-level outing, approximately 4.96 innings. For the rest of the pitcher's outing, the difference in skill is calculated between the pitcher and the team's bullpen, which would be pitching[12]. A new constant nLgc is calculated, as before, for adjustment purposes. The figure of 4.96 can of course be changed to more accurately reflect the replacement-level IP/GS.

Figure 4 shows the difference between the measures in bins of 10%, sorted by calc_fWAR, while Table 3 provides a summary of the difference[13]. Every group which had a positive calc_fWAR has their value reduced. The most drastic change is that for high-end pitchers: there is a difference of 0.55 fWAR for the top 10% of pitchers, from an average of 4.93 to 4.38. For the bottom 70% or so, however, the change is marginal, amounting to 0.2 fWAR or less.

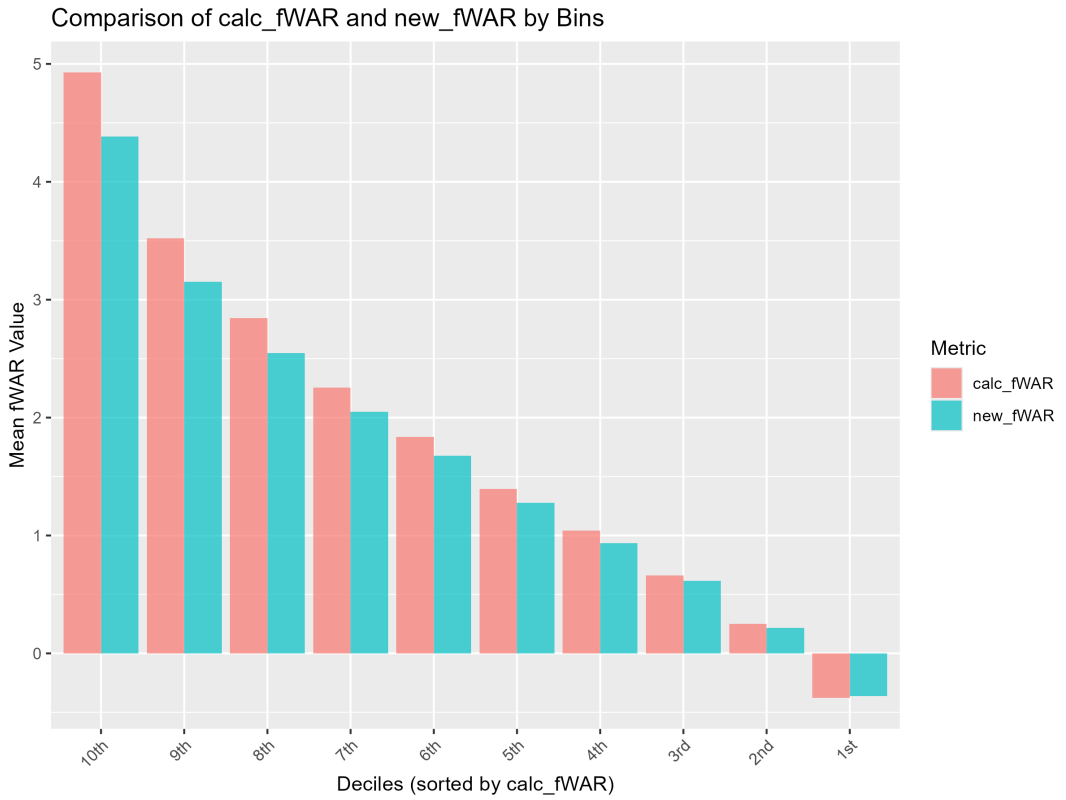


Figure 4: Difference between calc_fWAR and new_fWAR by deciles

[12]This formula only takes GS into account, not G. For pitchers who split their time between the bullpen and the rotation, the correction should be applied to their starts and the original formula to their relief games. For the purposes of this study, the difference is negligible: the mean GS/G in my dataset is 0.98 and the median is 1.

[13]These calculations are run with Lgc, not nLgc. This is mainly done because calculating nLgc would require recalculating new_fWAR for the rest of the league, including leverage and new values for every reliever who starts at least one game. Either way, Lgc represents less than 0.1 fWAR on average and is applied to every pitcher, making the difference between Lgc and nLgc minimal.

Dataset	n=	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All	510	0.00	0.05	0.13	0.21	0.30	1.13
Top 10% of calc_fWAR	51	0.11	0.40	0.51	0.55	0.68	1.13

Table 3: Summary of difference between calc_fWAR and new_fWAR

Figure 5 lists the fifteen largest differences in fWAR value by player and year. This illustrates the impact on the elite level clearly. All players have at least 2.5 calc_fWAR, putting them in the top third of the league. There are three Cy Young winners[14] and seven All-MLB members[15]. The drops in fWAR range from 0.72 to 1.13. The largest difference is over a quarter of the previous calc_fWAR value.

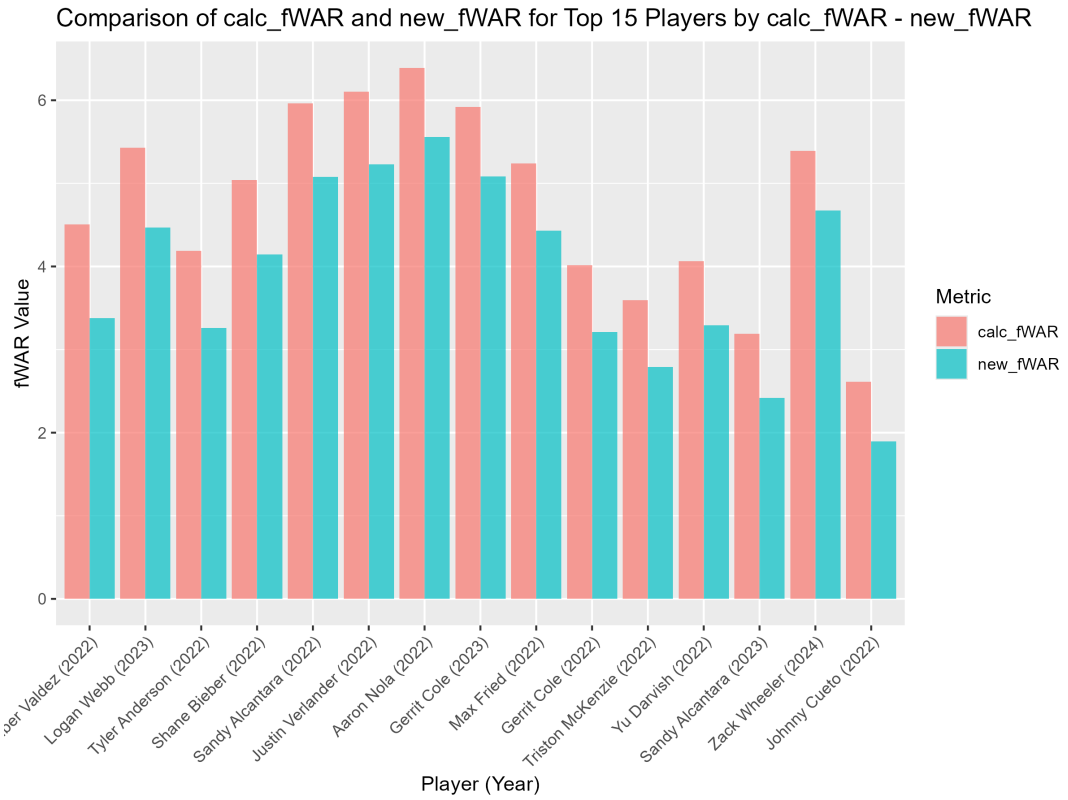


Figure 5: Difference between calc_fWAR and new_fWAR for top 15 by difference

5 Conclusion

Because of the correlation between pitching skill and length of start, fWAR overvalues most starting pitchers in the league. For the average league pitcher, this adjustment is small and comparable to the uncertainty of WAR as a measure. However, the adjustment becomes stark in the case of high-end pitchers, as they tend to last longer and thus have a higher percentage of the workload taken over by the bullpen rather than the replacement player.

I conclude with a few observations. While I conducted this study on fWAR for ease of calculation, this flaw is likely present in other WAR evaluations. Per FanGraphs[16], the main difference in WAR calculations between websites are the base metrics they use to evaluate player skill; the overall philosophy of replacement is consistent. bWAR, for instance, seems to make the same assumption in their calculation of 1-for-1 substitution of the starting pitcher, although it's difficult to be certain because of the opaqueness of the explanation on the Baseball Reference page. A more in-depth study into these metrics would probably elicit a definitive answer as to whether the same flaw as fWAR is present in other

[14] Alcantara '22, Verlander '22, and Cole '23. Cy Young and All-MLB information taken from www.mlb.com.
[15] Alcantara '22, Valdez '22, Verlander '22, Fried '22, Nola '22, Cole '23, and Wheeler '24.
[16] Slowinski, 2012.

WAR calculations and propose a correction.

The main unknown in the correction I propose is the impact of increased usage on the bullpen. In the theoretical WAR world where every starting pitcher is replaced for the season, the bullpen would find itself pitching dozens of additional innings a year. On the one hand, it seems likely this would be taxing in some capacity and decrease the level of the bullpen as a whole. On the other hand, the average bullpen has a higher pFIPR9 than the average starting pitcher to begin with, so even a slight regression would still result in a large talent gap between the bullpen and the replacement-level pitcher.

References

Slowinski, Piper. 02/15/2010. "What Is WAR? — Sabermetrics Library." Sabermetrics Library. https://library.fangraphs.com/misc/war/.

Slowinski, Piper. 03/22/2012. "fWAR, rWAR, and WARP — Sabermetrics Library." Sabermetrics Library. https://library.fangraphs.com/war/differences-fwar-rwar/.

A Appendix: fWAR formula details

A.1 Key

HR = Home runs
BB = Base on balls (walks)
HBP = Hit by pitch
K = Strikeout
IFFB = Infield fly-ball
IP = Innings pitched
$\frac{\text{rIP}}{\text{GS}}$ = Replacement innings pitched per game started (set to 4.96 above)
lg = League (AL or NL)
mlb = MLB (both leagues combined)
pPF = Pitcher's Park Factor (regressed halfway to 100)
G = Games played
GS = Games started
gmLi = Average game leverage when entering game[17]
pPayroll = Percentage of all MLB payroll that goes to pitchers
FIP = Field-independent pitching
ifFIP = Field-independent pitching with infield fly-balls
RA9 = Runs against per 9 innings
ERA = Earned runs against
Lev = Leverage factor
dRPW = Dynamic runs per win

[17]The leverage index, Li, is a measure of how critical a particular moment is in the game. It is calculated by multiplying the likelihood that any possible event should occur (strike, home run, double, etc) by the swing in win expectancy should that event occur, and then taking the sum. The leverage at the top of the first inning is around 0.88; a high-leverage event (bases loaded in the bottom of the ninth, trailing by two for example) is above 2.

A.2 Equations

$$\text{FIP} = \frac{13 \cdot \text{HR} + 3 \cdot (\text{BB} + \text{HBP}) - 2 \cdot \text{K}}{\text{IP}} + \text{FIPc}$$

$$\text{ifFIP} = \frac{13 \cdot \text{HR} + 3 \cdot (\text{BB} + \text{HBP}) - 2 \cdot (\text{K} + \text{IFFB})}{\text{IP}} + \text{ifFIPc}$$

$$\text{ifFIPc} = \text{mlbERA} - \frac{13 \cdot \text{mlbHR} + 3 \cdot (\text{mlbBB} + \text{mlbHBP}) - 2 \cdot (\text{mlbK} + \text{mlbIFFB})}{\text{mlbIP}}$$

$$\text{pFIPR9} = \frac{\text{ifFIP} + \text{mlbRA9} - \text{mlbERA}}{\text{pPF}/100}$$

$$\text{lgFIPR9} = \text{lgifFIP} + \text{mlbRA9} - \text{mlbERA}$$

$$\text{dRPW} = 1.5 \cdot \left(\frac{\left(18 - \frac{\text{IP}}{\text{G}}\right)}{18} \cdot \text{lgFIPR9} + \frac{\text{IP}/\text{G}}{18} \cdot \text{pFIPR9} + 2 \right)$$

$$\text{RL} = 0.03 \cdot \left(1 - \frac{\text{GS}}{\text{G}}\right) + 0.12 \cdot \frac{\text{GS}}{\text{G}}$$

$$\text{Lev} = \frac{1 + \text{gmLi}}{2}$$

$$\text{fWAR\_raw} = \left(\frac{\text{lgFIPR9} - \text{pFIPR9}}{\text{dRPW}} + \text{RL} \right) \cdot \frac{\text{IP}}{9} \cdot \text{Lev}$$

$$\text{Lgc} = \frac{\text{mlbfWAR\_raw} - 1000 \cdot \text{pPayroll}}{\text{mlbIP}} \cdot \text{IP}$$

$$\text{fWAR} = \left(\frac{\text{lgFIPR9} - \text{pFIPR9}}{\text{dRPW}} + \text{RL} \right) \cdot \frac{\text{IP}}{9} \cdot \text{Lev} + \text{Lgc}$$

$$\text{new\_fWAR} = \left(\frac{\text{lgFIPR9} - \text{pFIPR9}}{\text{dRPW}} + \text{RL} \right) \cdot \left(\frac{\frac{\text{rIP}}{\text{GS}} \cdot \text{GS}}{9} \right)$$

$$+ \left(\frac{\text{bpFIPR9} - \text{pFIPR9}}{\text{dRPW}} \right) \cdot \left(\frac{\frac{\text{IP}}{\text{GS}} - \frac{\text{rIP}}{\text{GS}} \cdot \text{GS}}{9} \right) + \text{nLgc}$$