Introducing *Grid WAR*: Rethinking WAR for Starting Pitchers

Ryan Brill, Justin Lipitz, Emma Segerman, Ezra Troy, Adi Wyner

March 1, 2022

Abstract

Traditional methods of computing WAR (wins above replacement) for pitchers are flawed. Specifically, Fangraphs and Baseball Reference compute a pitcher's WAR as a function of his performance averaged over the entire season, which is problematic because they ignore a pitcher's game-by-game variance and ignore the convexity of WAR. Hence we propose a new way to compute WAR for starting pitchers: *Grid WAR* (*gWAR*). The idea is to compute a starter's *gWAR* for each of his individual games, and define a starter's seasonal *gWAR* as the sum of the *gWAR* of each of his games. We find that *gWAR* as a function of the number of runs *R* a pitcher allows during a game is convex, which by Jensen's inequality implies that current implementations of WAR undervalue pitchers who allow few runs (specifically, 0 or 1 run) in many games.

1 Introduction

WAR (wins above replacement) is a fundamental statistic for valuing baseball players, and has recently been proposed to determine arbitration salaries (Perry, 2021). So, it is of utmost importance to use a WAR statistic that accurately captures a player's contribution to his team. However, current popular implementations of WAR for starting pitchers, implemented by Fangraphs (Slowinski, 2012) and Baseball Reference (2011), have flaws. In particular, by computing WAR as a function of average pitcher performance, Baseball Reference and FanGraphs ignore a pitcher's game-by-game variance and ignore the convexity of WAR. Hence in this paper we propose a new way to compute WAR for starting pitchers, *Grid WAR*.

2 Problems with Current Implementations of WAR

2.1 The Problem: Averaging over Pitcher Performance

The primary flaw of traditional methods for computing WAR for pitchers, as implemented by Baseball Reference and Fangraphs, is WAR is calculated as a function of a pitcher's *average* performance. Baseball Reference averages a pitcher's performance over the course of a season via *xRA*, or "expected runs allowed" (Reference, 2011). *xRA* is a function of a pitcher's average number of runs allowed per out. Fangraphs averages a pitcher's performance over the course of a

Table 1: Max Scherzer's performance over six games prior to the 2014 all star break.

game	1	2	3	4	5	6	total
earned runs	0	10	1	2	1	1	15
innings pitched	9	4	6	7	8	7	41

season via $ifFIP$, or "fielding independent pitching (with infield flies)" (Slowinski, 2012). $ifFIP$ is defined by

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

which involves averaging some of a pitcher's statistics over his innings pitched.

2.2 Ignoring Variance

Using a pitcher's *average* performance to calculate his WAR is a subpar way to measure his value on the mound because it ignores the variance in his his game-by-game performance.

To see why ignoring variance is a problem, consider Max Scherzer's six game stretch from June 12, 2014 through the 2014 all star game, shown in table 1 (ESPN, 2014). In Scherzer's six game stretch, he averages 15 runs over 41 innings, or 0.366 runs per inning. So, on average, Scherzer pitches 3.3 runs per complete game. If we look at each of Scherzer's individual games separately, however, we see that he has four dominant performances, one decent game, and one "blowup". Intuitively, the four dominant performances alone are worth more than allowing 3.3 runs in each of six games. On this view, averaging Scherzer's performances significantly devalues his contributions during this six game stretch.

Because "*you can only lose a game once*", it makes more sense to give Scherzer zero credit for his one bad game than to distribute his one poor performance over all his other games via averaging. Hence we should not compute WAR as a function of a pitcher's average performance. Instead, we should compute a pitcher's WAR in each individual game, and compute his season-long WAR as the summation of the WAR of his individual games.

2.3 Ignoring the Convexity of WAR

Additionally, using a pitcher's *average* performance to calculate his WAR is a subpar way to measure his value on the mound because it ignores the convexity of WAR.

Think of a starting pitcher's WAR in a complete game is a function $R \mapsto WAR(R)$ of the number of runs allowed in that game. We expect WAR to be a decreasing function, because allowing more runs in a game should correspond to fewer wins above replacement. Additionally, we expect WAR to be a *convex* function, whose second derivative is positive. In other words, as R increases, we expect the relative impact of allowing an extra run, given by $WAR(R+1) - WAR(R)$, to decrease. Concretely, allowing 2 runs instead of 1 should have a much steeper dropoff in WAR than allowing 7 runs instead of 6. We expect this because "*you can only lose a game once*". If a pitcher allows 6 runs, he has essentially already lost his team the game, so allowing an extra run to make this total 7 shouldn't have a massive difference in a pitcher's WAR during that game. Conversely, if a pitcher allows 1 run, the marginal impact of allowing an extra run to make this total 2 is much larger.

Because we expect *WAR* to be a convex function, Jensen's inequality tells us that averaging a pitcher's performance undervalues his contributions. Specifically, thinking of a pitcher's number of runs allowed in a complete game as a random variable R, Jensen's inequality says

$$WAR(\mathbb{E}[R]) \leq \mathbb{E}[WAR(R)]. \tag{1}$$

Traditional methods for computing WAR are reminiscent of the left side of equation (1) - average a pitcher's performance, and then compute his WAR. In this paper, we devise a WAR metric reminiscent of the right side of equation (1) - compute the WAR of each of a pitcher's individual games, and then average. By equation (1), traditional metrics for computing WAR undervalue the contributions of many starting pitchers. On the other hand, our method allows the convexity of WAR to more accurately value a pitcher's contributions.

3 Defining *Grid WAR* for Starting Pitchers

We wish to create a metric which computes a starting pitcher's WAR for an individual game. The idea is to compute a context-neutral and offense-invariant version of win-probability-added that is derived only from a pitcher's performance.

First, we define a starting pitcher's *Grid WAR* (*gWAR*) for a game in which he exits at the end of an inning. To do so, we create the function $f = f(I,R)$ which, assuming both teams have league-average offenses, computes the probability a team wins a game after giving up R runs through I innings. f is a context-neutral version of win probability, as it depends only on the starter's performance.

To compute a wins *above replacement* metric, we need to compare this context-neutral win-contribution to that of a potential replacement-level pitcher. We use a constant w_{rep} which denotes the probability a team wins a game with a replacement-level starting pitcher, assuming both teams have league-average offenses. We expect $w_{rep} < 0.5$ since replacement-level pitchers are worse than league-average pitchers.

Then, we define a starter's *Grid WAR* during a game in which he gives up R runs through I complete innings as

$$f(I,R) - w_{rep}. \tag{2}$$

We call our metric *Grid WAR* because the function $f = f(I,R)$ is defined on the 2D grid $\{1,...,9\} \times \{1,...,25\}$.

Next, we define a starting pitcher's *Grid WAR* for a game in which he exits midway through an inning. To do so, we create a function $g = g(R|S,O)$ which, assuming both teams have league-average offenses, computes the probability that, starting midway through an inning with $O \in \{0,1,2\}$ outs and base-state

$$S \in \{000,100,010,001,110,101,011,111\},$$

a team scores exactly R runs through the end of the inning. Then we define a starter's *Grid WAR* during a game in which he gives up R runs and leaves midway through inning I with O outs and base-state S as the expected *Grid WAR* at the end of the inning,

$$\sum_{r\geq 0} g(r|S,O) f(I,r+R) - w_{rep}. \tag{3}$$

Finally, we define a starting pitcher's *Grid WAR* for an entire season as the sum of the *Grid WAR* of his individual games. In order to compute *Grid WAR* for each starting pitcher, we need only estimate the grid functions f and g and the constant w_{rep}.

4 Estimating the Grid Functions f and g

In this section, we estimate f, w_{rep}, and g. To do so, we use data scraped from Retrosheet (2021). Our cleaned data is freely available for download on Dropbox (Brill, 2021).

4.1 Estimating f

First, we estimate the function $f = f(I,R)$ which, assuming both teams have league-average offenses, computes the probability a team wins a game after giving up R runs through I complete innings. We estimate f using logistic regression. The response variable is a binary variable indicating whether a pitcher's team won a game after giving up R runs through I innings. We model I and R as fixed effects (i.e., we have separate coefficients for each value of I and R). In order to make f context neutral, we also adjust for home field, National vs. American league, and the year, each as a fixed effect. This process is essentially equivalent to binning, averaging, and smoothing over the variables (I,R) after adjusting for confounders. Additionally, recall that if a home team leads after the top of the 9^{th} inning, then the bottom of the 9^{th} is not played. Therefore, to avoid selection bias, we exclude all 9^{th} inning instances in which a pitcher pitches at home.

In figure 1, we plot the functions $R \mapsto f(I,R)$ for each inning I, for an away-team American League pitcher in 2019. For each inning I, $R \mapsto f(I,R)$ is decreasing. This makes sense: within an inning, if you allow more runs, you are less likely to win the game. Also, $R \mapsto f(I,R)$ is mostly convex. This makes sense: if you have already allowed a high number of runs, there is a lesser relative impact of throwing an additional run. Conversely, if you have allowed few runs thus far, there is a high relative impact of throwing an additional run. Furthermore, for each R, the function $I \mapsto f(I,R)$ is increasing. This makes sense: giving up R runs through I innings is worse than giving up R runs through $I+i$ innings for $i > 0$, because giving up R runs through $I+i$ innings implies you gave up fewer than R runs through I innings, on average.

4.2 The Convexity of *gWAR*

As shown in figure 1, the function $R \mapsto f(I,R)$ is convex for each inning I. Therefore, by Jensen's inequality, we expect that traditional WAR metrics undervalue players relative to *gWAR*.

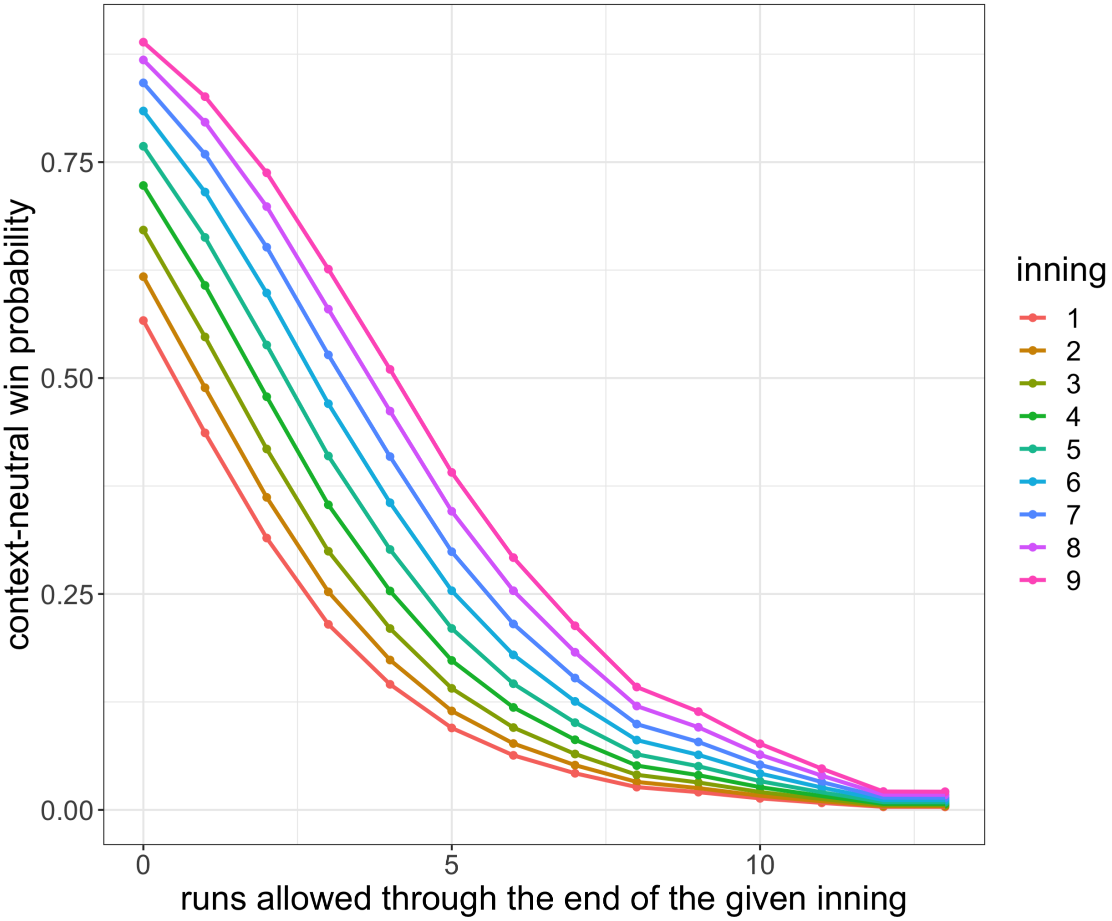
To see why, suppose a starting pitcher allows R runs through $I = i$ innings, where i is a fixed number and R is a random variable. Then by Jensen's inequality,

$$f(i,\mathbb{E}[R]) \leq \mathbb{E}[f(i,R)]. \tag{4}$$

Therefore, supposing a pitcher in each of $j \in \{1,...,n\}$ games allows R_j runs through $I = i$ complete innings, we approximately have

$$f\left(i, \frac{1}{n}\sum_{j=1}^{n} R_j\right) \leq \frac{1}{n}\sum_{j=1}^{n} f(i,R_j). \tag{5}$$

Figure 1: The function $R \mapsto f(I,R)$ for each inning I, for an away-team American League pitcher in 2019.



In other words, for a pitcher who pitches exactly $I = i$ innings in each game, the $gWAR$ of his average number of runs is less than the average $gWAR$ of his individual games. On this view, traditional WAR metrics undervalue the win contributions of many players, especially those of high variance pitchers or pitchers with skewed distributions. In this paper, by computing season-long WAR as the summation of the WAR of his individual games, we allow the convexity of WAR to more accurately describe pitchers' performances.

4.3 Estimating w_{rep}

To compute a wins *above replacement* metric, we need to compare the context-neutral win-contribution to that of a potential replacement-level pitcher. Thus we define a constant w_{rep} which denotes the context-neutral probability a team wins a game with a replacement-level starting pitcher, assuming both teams have a league-average offense. We expect $w_{rep} < 0.5$ since replacement-level pitchers are worse than league-average pitchers.

It is difficult to estimate w_{rep} because it is difficult to compile a list of replacement-level pitchers. According to Fangraphs (2010), *replacement-level* is the "level of production you could get from a player that would cost you nothing but the league minimum salary to acquire.” Since we are not members of an MLB front office, this level of production is difficult to estimate. Ultimately, the value of w_{rep} doesn’t matter too much because we rescale all pitcher’s *Grid WAR* to sum to a fixed amount, to compare our results to those of Fangraphs. So, we arbitrarily set $w_{rep} = 0.41$.

4.4 Estimating g

Now, we estimate the function $g = g(R|S,O)$ which, assuming both teams have league-average offenses, computes the probability that, starting midway through an inning with $O \in \{0,1,2\}$ outs and base-state

$$S \in \{000,100,010,001,110,101,011,111\},$$

a team scores exactly R runs through the end of the inning. We estimate $g(R|S,O)$ using the empirical distribution, for $R \in \{1,...,13\}$. Specifically, we bin and average over the variables (R,S,O), using data from every game from 2010 to 2019. Because g isn’t significantly different across innings, we use data from each of the first eight innings.

In figure 2 we plot the distribution of $g(R|S,O=0)$, with $O=0$ outs, for each base-state S. With no men on base ($S=000$), 0 runs allowed for the rest of the inning is most likely. With bases loaded ($S=111$), 1 run allowed for the rest of the inning is most likely, and there is a fat tail expressing that 2 through 5 runs through the rest of the inning are also reasonable occurences. With men on second and third, 2 runs allowed for the rest of the inning is most likely, but the tail is skinnier than that of bases loaded.

5 Results

We compute the *Grid WAR* (*gWAR*) of each starting pitcher in 2019 using data scraped from Retrosheet (2021). Our cleaned data, consisting of every plate appearance from 1990 to 2020, is freely available for download on Dropbox (Brill, 2021). We acquire the 2019 FanGraphs WAR ($fWAR$) of 58 starting pitchers from Fangraphs (2019). To legitimize comparison between *gWAR* and $fWAR$, we rescale *gWAR* so that the sum of these pitchers’ *gWAR* equals the sum of their $fWAR$. Because their $fWAR$ sums to 205 and their *gWAR* sums to 198, the impact of rescaling is small.

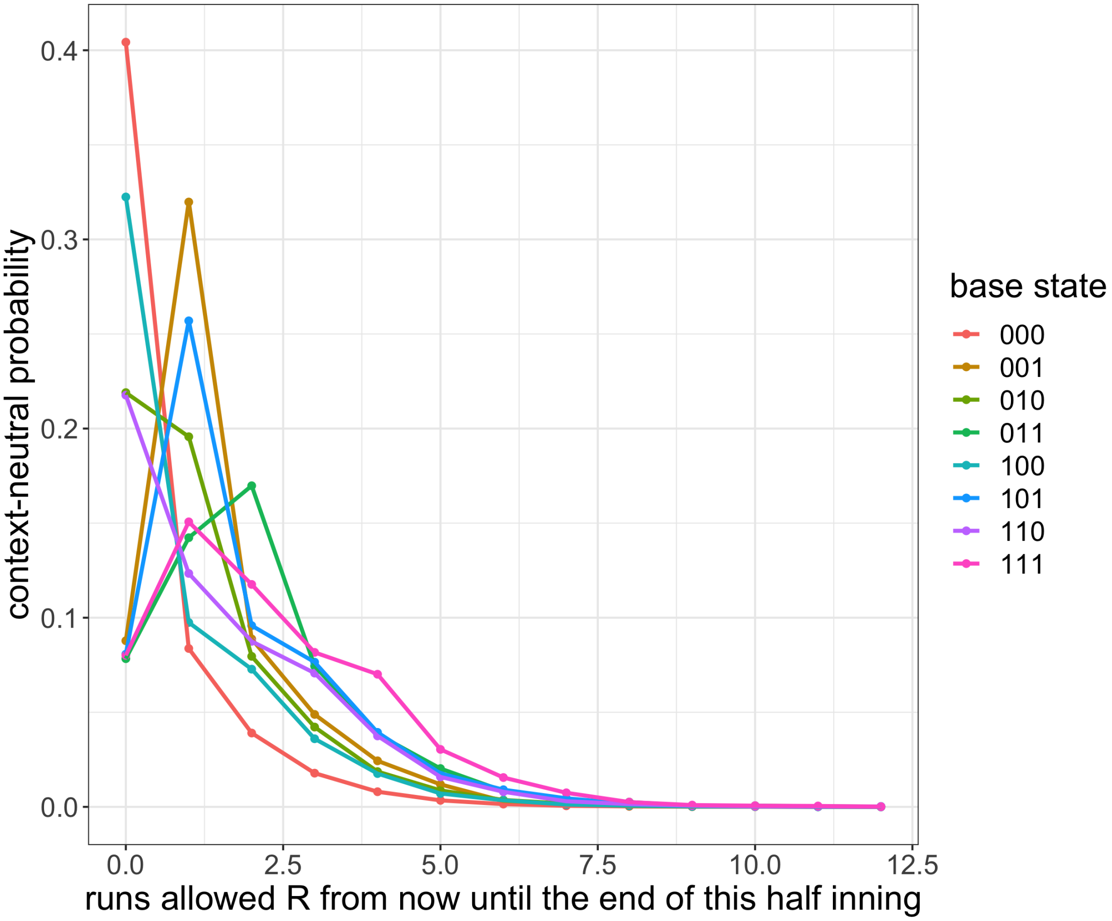
5.1 *Grid WAR* vs. Fangraphs WAR in 2019

In figure 3 we plot *gWAR* vs. $fWAR$ for starting pitchers in 2019. We define the metric `vertical distance` (*vd*), which is a pitcher’s difference in *gWAR* and $fWAR$,

$$vd := gWAR - fWAR. \tag{6}$$

In figure 3, a player’s *vd* is the y-distance from the point $(x,y) = (fWAR, gWAR)$ to the line $y = x$. According to *Grid WAR*, players with large positive *vd* values are undervalued, players with small $|vd|$ values are equally valued, and players with large negative *vd* values are overvalued, relative to FanGraphs. In figure 3, we include the names of the five most undervalued, equally valued, and overvalued starting pitchers, determined via their *vd* scores.

Figure 2: The discrete probability distribution $R \mapsto g(R|S, O = 0)$ for each base-state S.

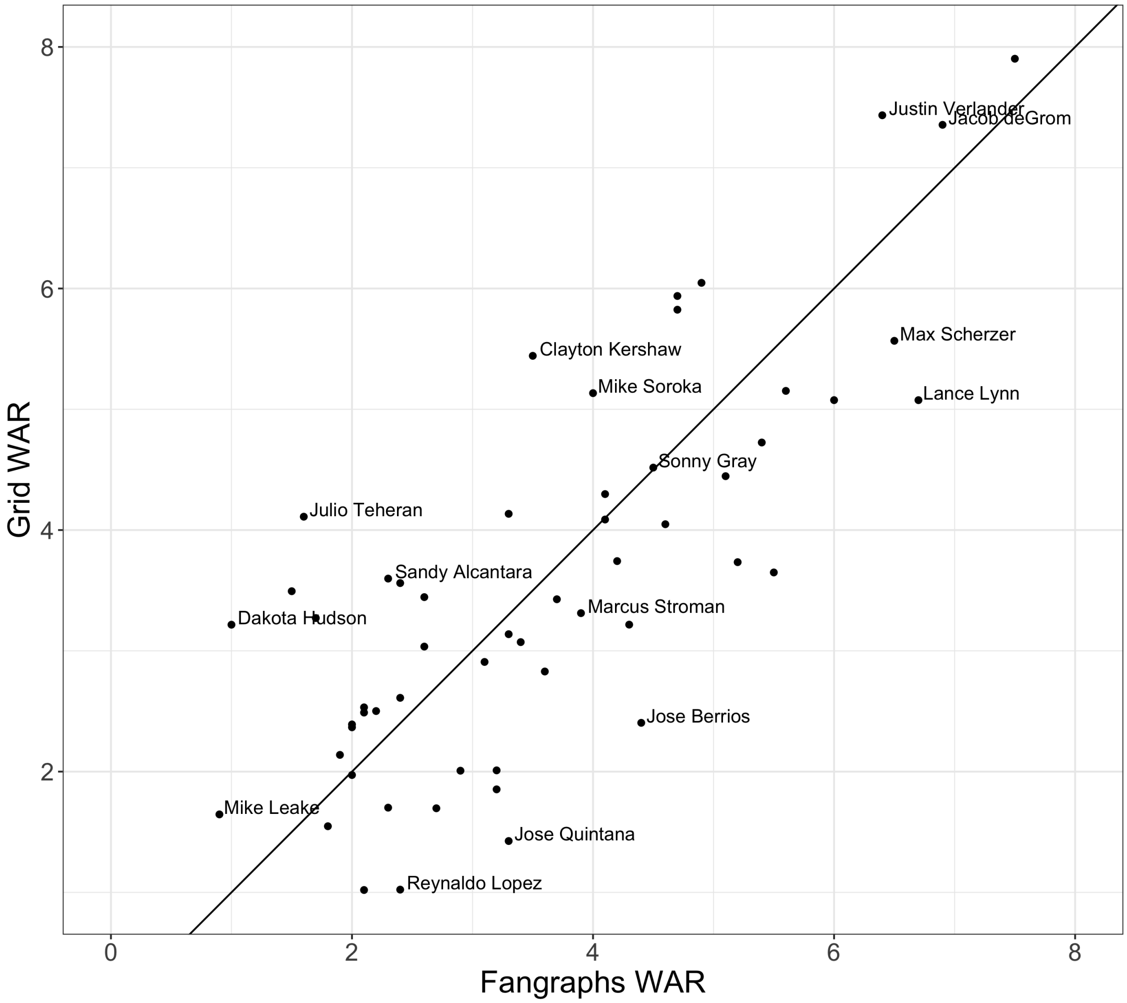


As discussed in section 4.2, by the convexity of *Grid WAR*, we expected most of the points in figure 3 to lie above the $y = x$ line. However, we see that this is not the case. We suspect this is because FanGraphs WAR is a complicated formula which adds many adjustments on top of a pitcher's average performance (Slowinski, 2012). Nevertheless, examining individual player-seasons sheds light on the difference between each of these players' $gWAR$ and $fWAR$ values.

5.2 Overvalued vs. Undervalued Pitchers on Aggregate in 2019

In figure **??**, we bin the 2019 starting pitchers into three categories - overvalued (negative vd), equally valued (low $|vd|$), and undervalued (high vd) - and plot the empirical distribution of runs allowed in a game, for each bin. We see that undervalued pitchers have a high relative proportion of games with 0 and 1 run allowed. This makes sense: averaging a pitcher's performance over all his games dilutes his exceptional games, which undervalues his performance by the convexity of *Grid WAR*. Furthermore, both overvalued and equally values pitchers have a high relative proportion of games with 2 and 3 runs allowed. The difference between these two categories appears to be that equally valued pitchers have a higher relative proportion of games with 0 and 1 run allowed, whereas overvalued pitchers have fatter tails.

Figure 3: *Grid WAR* vs. FanGraphs WAR in 2019. The names of the five most undervalued, equally valued, and overvalued starting pitchers in 2019, according to *gWAR* relative to *fWAR*, are included.



5.3 Noah Syndergaard vs. Mike Fiers in 2019

Previously, we considered the difference between overvalued, equally valued, and undervalued pitchers on aggregate. Now, we examine these differences by considering individual player-seasons, and we see similar trends. We begin with Noah Syndergaard and Mike Fiers.

In 2019, Noah Syndergaard has 4.3 *fWAR* and 2.3 *gWAR*, whereas Mike Fiers has 1.7 *fWAR* and 4.12 *gWAR*. In fact, Mike Fiers is the most undervalued starting pitcher in 2019, according to *Grid WAR*, with a *vd* of 2.4. In true Moneyball fashion, in 2019, at the ripe age of 33, Mike Fiers played for the Oakland Athletics. Upon first glance, noting that Syndergaard is a well-regarded pitcher, we fear that we made a mistake. Nevertheless, examining their 2019 game-by-game statistics allows us to understand their differences in *gWAR* and *fWAR*, and suggests that *Grid WAR* is not mistaken after all.

In figure **??** we plot the histogram of runs allowed in a game for Mike Fiers and Noah Syndergaard in 2019. The number in the bin corresponding to allowing *r* runs in a game refers to the mean number of innings pitched in those games. For example, Mike Fiers allowed 1 run in 10 games, and pitched an average of 6.3 innings in those games. Note that because both pitchers have a similar mean number of innings pitched for each of these runs-allowed bins, there is negigible bias caused by one pitcher leaving a game earlier than the other.

Fiers has a high proportion of games in which he allows 0, 1, or 2 runs. On the other hand, Syndergaard has a high proportion of games in which he allows 2, 3, or 4 runs. Therefore, it makes sense that Fiers' *gWAR* is so much higher than Syndergaards'. Again, by the convexity of *gWAR*, allowing 0 or 1 run is extremely valuable. Nonetheless, it is perplexing why Syndergaards' *fWAR* is so high and Fiers' *fWAR* is so low.

5.4 Lance Lynn vs. Hyun-Jin Ryu in 2019

Now, we consider the individual player-seasons of Lance Lynn and Hyun-Jin Ryu.

In 2019, Lance Lynn has 6.7 *fWAR* and 4.3 *gWAR*, whereas Hyun-Jin Ryu has 4.9 *fWAR* and 6.7 *gWAR* - roles reversed! In fact, Lance Lynn has the third highest *fWAR* of all starting pitchers in 2019. He is also the most overvalued pitcher in 2019, according to *gWAR*, with a *vd* of -2.4. Again, examining their 2019 game-by-game statistics allows us to understand their differences in *gWAR* and *fWAR*.

In figure **??** we plot the histogram of runs allowed in a game for Lynn and Ryu in 2019. Again, the number in the bin corresponding to allowing *r* runs in a game refers to the mean number of innings pitched in those games. For example, Ryu allowed 0 runs in 10 games, and pitched an average of 7.2 innings in those games. Note that because both pitchers have a similar mean number of innings pitched for each of these runs-allowed bins, there is negigible bias caused by one pitcher leaving a game earlier than the other.

We see that Ryu's runs-allowed distribution is concentrated on allowing 0, 1 or 2 runs, whereas Lynn's distribution is concentrated on allowing 1, 2, and 3 runs. The primary difference in their distributions appears to be that Ryu allows 0 runs in 8 more games than Lynn, and Lynn allows 4 runs in 3 more games than Ryu. Again, by the convexity of *gWAR*, all the games in which Ryu allows 0 runs are extremely valuable, and are enough to give him 2 more *gWAR* than Lynn. Also, that Lynn's distribution is concentrated on allowing 1, 2, and 3 runs, as opposed to 3, 4, and 5 runs per game explains why he still has a high *gWAR*, particularly in comparison to Syndergaard from the previous section.

6 Conclusion

Traditional methods of computing WAR are flawed because they compute WAR as a function of a pitcher's *average* performance. Averaging over pitcher performance is a subpar way to value a pitcher's performance because it ignores a pitcher's game-by-game variance and ignores the convexity of WAR. So, in this paper, we devise *Grid WAR*, a new way to compute a starting pitcher's WAR. We compute a pitcher's *gWAR* in each of his individual games, and define his seasonal *gWAR* as the sum of the *gWAR* of his individual games. We compute *gWAR* on a set of starting pitchers in 2019, and compare them to their FanGraphs WAR. Examining the trends of pitchers who are overvalued, equally valued, and undervalued by *gWAR* relative to *fWAR* in 2019, we see that *gWAR* highly values games in which a pitcher allows few runs (0 or 1). This makes sense, becuase by the convexity of WAR, the more runs a pitcher allows, giving up an additional run has less of a marginal impact. Additionally, by examining individual player-seasons in 2019, we see the convexity of WAR again highly value pitchers who allow few runs in many games.

6.1 Future Work

In this paper, we show that current implementations of WAR for starting pitchers are flawed, and we propose a new way to compute WAR for starting pitchers: *Grid WAR*. Our method, however, does not translate to valuing relievers in an obvious way. In particular, relievers enter the game at different times, which makes it difficult to value their context-neutral win contribution. Also, there is no obvous analog of w_{rep} for relievers. Nevertheless, for future work we suggest extending *Grid WAR* to value relief pitchers.

6.2 The Code

Our code is available on github at [LINK].

References

Ryan Brill. Cleaned retrosheet play-by-play data. https://upenn.box.com/v/retrosheet-pa-1990-2000, June 2021.

ESPN. Max scherzer 2014 game log. https://www.espn.com/mlb/player/gamelog/_/id/28976/year/2014/category/pitching, 2014.

Fangraphs. Replacement level. https://library.fangraphs.com/misc/war/replacement-level/, 2010.

Fangraphs. Fangraphs 2019 starting pitcher war leaderboard. https://www.fangraphs.com/leaders.aspx?pos=all&stats=sta&lg=all&qual=y&type=8&season=2019&month=0&season1=2019&ind=0&team=0&rost=0&age=0&filter=&players=0&startdate=&enddate=, 2019.

Dayn Perry. Mlb proposes determining arbitration salaries by using the war statistic, per report. https://www.cbssports.com/mlb/news/mlb-proposes-determining-arbitration-salaries-by-using-the-war-statistic-per-report/, 2021.

Baseball Reference. Pitcher war calculations and details. https://www.baseball-reference.com/about/war_explained_pitch.shtml, 2011.

Retrosheet. Retrosheet play-by-play data files (event files). https://www.retrosheet.org/game.htm, 2021.

Piper Slowinski. War for pitchers. https://library.fangraphs.com/war/calculating-war-pitchers/, 2012.