Wharton University of Pennsylvania SPORTS ANALYTICS Student Research Journal

Introducing *Grid WAR*

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

May 2, 2022

Abstract

Traditional methods of computing WAR (wins above replacement) for pitchers are based on an invalid mathematical foundation. Consequently, these metrics, which produce reasonable values for many pitchers, can be substantially inaccurate for some. Specifically, Fangraphs and Baseball Reference compute a pitcher's WAR as a function of his performance averaged over the entire season. This is wrong because not all runs allowed have the same impact in determining the outcome of a game: for instance, the difference in impact between allowing 1 run in a game instead of 0 is much greater than the difference in impact between allowing 6 runs in a game instead of 5. 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 $gWAR$ highly values games in which a pitcher allows few runs (0 or 1).

Keywords: Wins Above Replacement (WAR), Baseball, Ignoring Variance

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 Reference (2011), have flaws. In particular, by computing WAR as a function of a pitcher's season average performance, these methods ignore a pitcher's game-by-game variance. 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 season via $ifFIP$, or "fielding independent pitching (with

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

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. Another way of thinking about this is

"*not all runs have the same value.*"

For instance, the difference between allowing 10 runs instead of 9 in a game is much smaller than the difference between allowing 1 run instead of 0. On this view, the tenth run allowed in a game has much smaller impact than the first.

Hence we should not compute WAR as a function of a pitcher's average game-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.

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

$$f(I,R) - w_{rep}. \tag{1}$$

We call our metric *Grid WAR* because the function $f = f(I,R)$ is defined on the 2D grid $\{1,...,9\} \times \{0,...,R_{max} = 10\}$. We restrict $R \leq R_{max} = 10$ because there is not much data to estimate $f(I,R)$ for $R > 10$, and because $f(I,10)$ is essentially zero. In particular, for $R > R_{max}$, we set $f(I,R) = f(I,R_{max})$.

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,

$$\sum_{r \geq 0} g(r|S,O) f(I,r+R) - w_{rep}. \tag{2}$$

Finally, we define a starting pitcher's *Grid WAR* for an entire season as the sum of the *Grid WAR* of his individual games.

4 Estimating the Grid Functions f and g

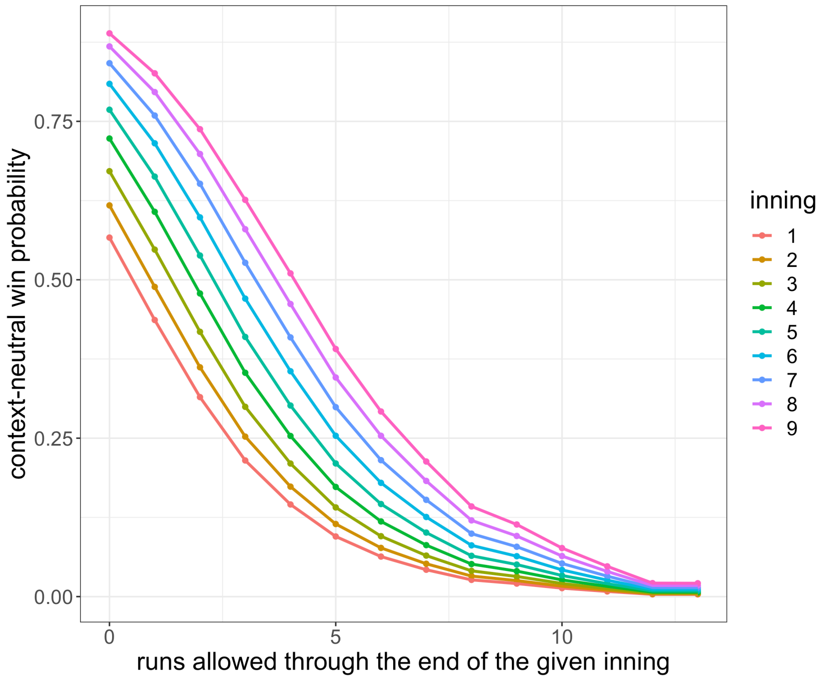
In this section, we estimate f, w_{rep}, and g. To do so, we use data scraped from Retrosheet (2021). Our cleaned data is freely available for download on Dropbox (Brill, 2021).

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

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



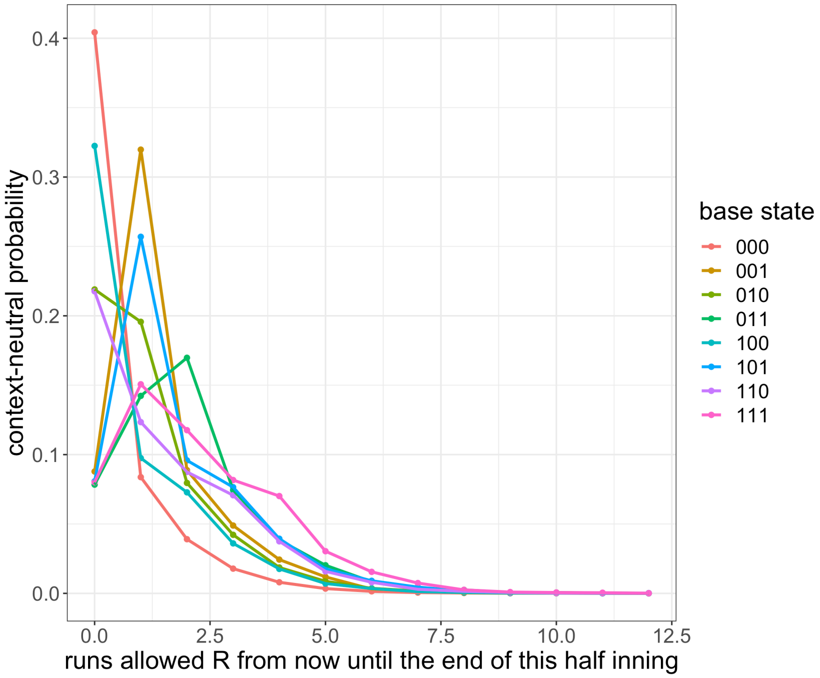
run. Furthermore, for each R, the function $I \mapsto f(I,R)$ is increasing. This makes sense: giving up R runs through I innings is worse than giving up R runs through $I+i$ innings for $i > 0$, because giving up R runs through $I+i$ innings implies you gave up fewer than R runs through I innings, on average.

4.2 Estimating w_{rep}

To compute a wins *above replacement* metric, we need to compare a starting pitcher's 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 $w_{rep} = 0.41$.

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



4.3 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

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

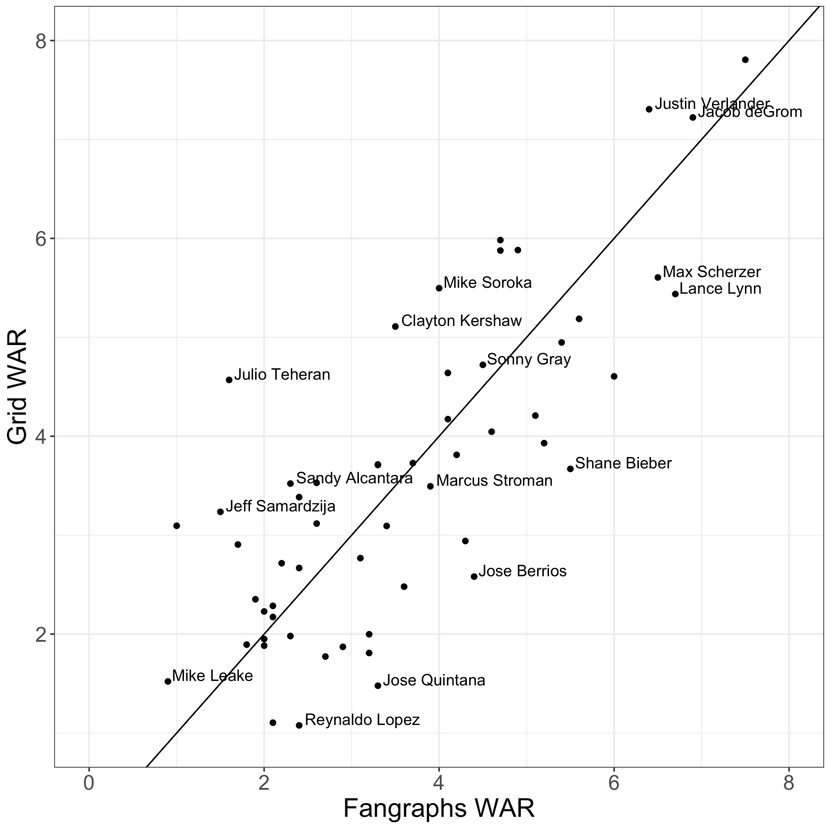
a team scores exactly R runs through the end of the inning. We estimate $g(R|S,O)$ using the empirical distribution, for $R \in \{1,...,13\}$. Specifically, we bin and average over the variables (R,S,O), using data from every game from 2010 to 2019. Because g isn't significantly different across innings,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

Figure 3: *Grid WAR* vs. FanGraphs WAR in 2019. Pitchers above the line $y = x$ are undervalued according to $gWAR$ relative to $fWAR$, and pitchers below the line are overvalued.



sum of these pitchers' $gWAR$ equals the sum of their $fWAR$. By rescaling, we compare the *relative* value of starting pitchers according to $gWAR$ to the *relative* value of starting pitchers according to $fWAR$. In figure 3 we plot $gWAR$ vs. $fWAR$ for starting pitchers in 2019.

5.1 Comparing Players with Similar $fWAR$ and Different $gWAR$ Values in 2019

In figures 4, 5, and 6, we compare pairs of players who have similar $fWAR$ and different $gWAR$ values. In figure 4 we compare Jacob deGrom to Lance Lynn. They have similarly high $fWAR$ (Lynn 6.7, deGrom 6.9), but deGrom has much higher $gWAR$ (deGrom 7.2, Lynn 5.4). In figure 5 we compare Sonny Gray and Jose Berrios. They have similarly moderate $fWAR$ (Gray 4.5, Berrios 4.4), but Gray has much higher $gWAR$ (Gray 4.7, Berrios 2.6). In figure 6 we compare Sandy Alcantara and Reynaldo Lopez. They have similarly low $fWAR$ (Alcantara 2.3, Lopez 2.4), but Alcantara has much higher $gWAR$ (Alcantara 3.5, Lopez 1.1).

In each of these comparisons, we see a similar trend explaining the differences in $gWAR$. Specifically, the pitcher with higher $gWAR$ allows fewer runs in more games, and allows more runs in fewer games. This is depicted graphically in the "Difference" histograms, which show the difference between the histogram on the left and the histogram on the right: the green bars are shifted towards the left (pitchers with higher $gWAR$ allow few runs in more games), and the red bars are shifted towards the right (pitchers with lower $gWAR$

allow more runs in more games). For instance, consider figure 5. Gray pitches 7 more games than Berrios in which he allows 3 runs or fewer, and Berrios pitches 8 more games than Gray in which he allows 4 runs or more. On this view, Gray should absolutely have a higher WAR than Berrios. Similarly, consider figure 6. Alcantara pitches 9 more games than Lopez in which he allows 0, 2, or 4 runs, whereas Lopez pitches 9 more games than Alcantara in which he allows 1, 3, or 5 runs. So, in 9 games, Alcantara pitches exactly 1 run fewer than Lopez! Hence he should have a higher WAR than Lopez. Additionally, consider figure 4. DeGrom pitches 6 more games than Lynn in which he allows exactly 0 runs, and Lynn pitches 7 more games than DeGrom in which he allows 3 runs or more.

Figure 4: Histogram of runs allowed in a game for Jacob deGrom and Lance Lynn in 2019. They have similar $fWAR$ (Lynn 6.7, deGrom 6.9), but deGrom has much higher $gWAR$ (deGrom 7.2, Lynn 5.4). This is because deGrom pitches 6 more games in which he allows exactly 0 runs, and Lynn pitches 7 more games in which he allows 3 runs or more.

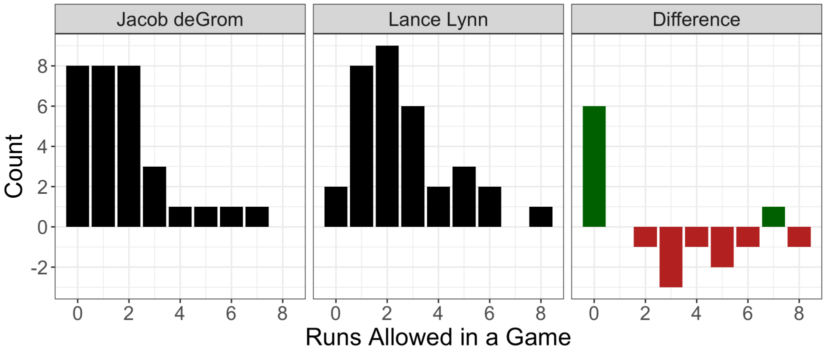


Figure 5: Histogram of runs allowed in a game for Sonny Gray and Jose Berrios in 2019. They have similar $fWAR$ (Gray 4.5, Berrios 4.4), but Gray has much higher $gWAR$ (Gray 4.7, Berrios 2.6). This is because Gray pitches 7 more games in which he allows 3 runs or fewer, and Berrios pitches 8 more games in which he allows 4 runs or more.

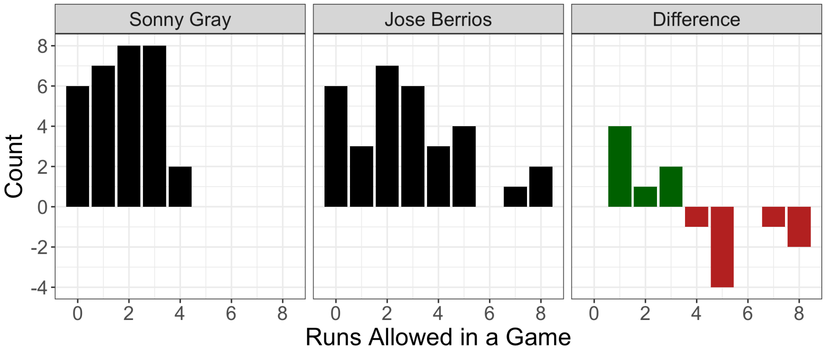
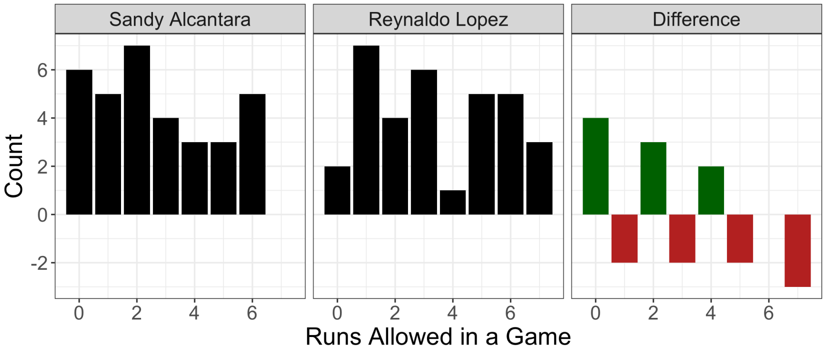


Figure 6: Histogram of runs allowed in a game for Sandy Alcantara and Reynaldo Lopez in 2019. They have similar $fWAR$ (Alcantara 2.3, Lopez 2.4), but Alcantara has much higher $gWAR$ (Alcantara 3.5, Lopez 1.1). This is because Alcantara pitches 9 more games in which he allows 0, 2, or 4 runs, and Lopez pitches 9 more games in which he allows 1, 3, 5, or 7 runs.



5.2 Comparing Undervalued and Overvalued Players 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 $fWAR$,

$$vd := gWAR - fWAR. \tag{3}$$

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 similarly valued, and players with large negative vd values are overvalued, relative to FanGraphs.

In figure 7, we bin the 2019 starting pitchers into two categories - overvalued (negative 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 runs allowed. This makes sense: averaging a pitcher's performance over all his games dilutes his exceptional games, which undervalues his performance. Furthermore, overvalued pitchers have a high relative proportion of games with 2 and 3 runs allowed.

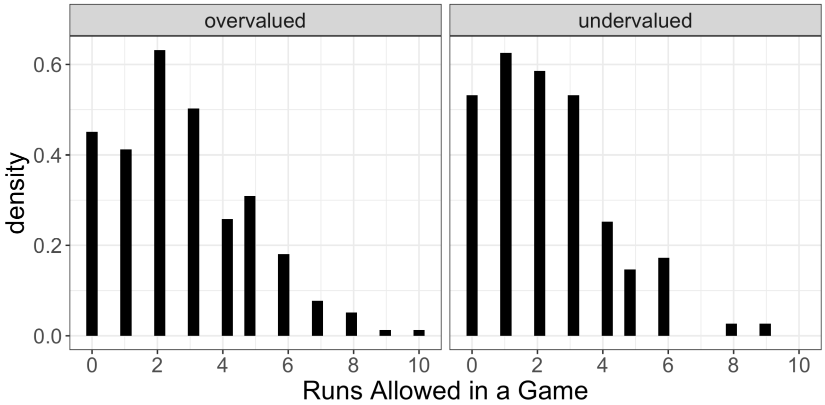
We see a similar trend when we examine the individual player-seasons of undervalued and overvalued pitchers. In figure 8, we examine two of the most undervalued pitchers in 2019 according to $gWAR$ relative to $fWAR$, Clayton Kershaw and Julio Teheran. Teheran is the most undervalued pitcher in 2019, with $gWAR - fWAR = 2.5$, and Kershaw's $gWAR - fWAR = 1.6$. We again notice that both pitchers have a higher relative proportion of games in which they allow 0 or 1 run.

Conversely, in figure 9, we examine two of the most overvalued pitchers in 2019 according to $gWAR$ relative to $fWAR$, Jose Berrios and Jose Quintana. Both Quintana and Berrios have $gWAR - fWAR = -1.8$. We again notice that both pitchers have a lower relative proportion of games in which they allow 0 or 1 run; both run distributions are concentrated around allowing 2 or 3 runs.

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 ignores the idea that "*not all runs have the*

Figure 7: The distribution of runs allowed in a game for overvalued and undervalued starting pitchers in 2019. Undervalued pitchers have a higher proportion of games in which they allow 0 or 1 run, and overvalued pitchers have a higher proportion of games in which they allow 2 or 3 runs.



same value" - for instance, allowing the tenth run of a game has a smaller marginal impact than allowing the first run of a game, because by the time a pitcher has thrown nine runs, the game is essentially already lost. In other words, "*you can only lose a game once.*" 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 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: the more runs a pitcher allows, giving up an additional run has less of a marginal impact.

6.1 The Code & Data

Our code is available on github at `https://github.com/snoopryan123/grid_war`. The data is available on Dropbox at `https://upenn.app.box.com/v/retrosheet-pa-1990-2000`.

Figure 8: Histogram of runs allowed in a game for Clayton Kershaw and Julio Teheran in 2019. Teheran and Kershaw in 2019 are undervalued according to $gWAR$ relative to $fWAR$, as Teheran's $gWAR - fWAR$ is 2.5 and Kershaw's $gWAR - fWAR$ is 1.6. Both pitchers have a high relative proportion of games in which they allow 0 or 1 run.

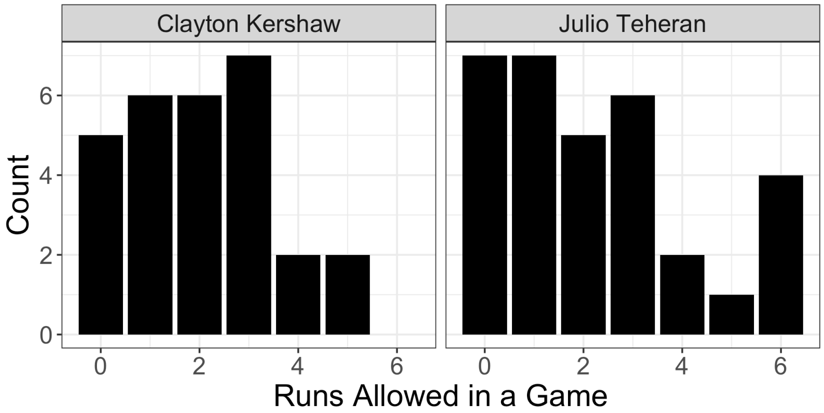
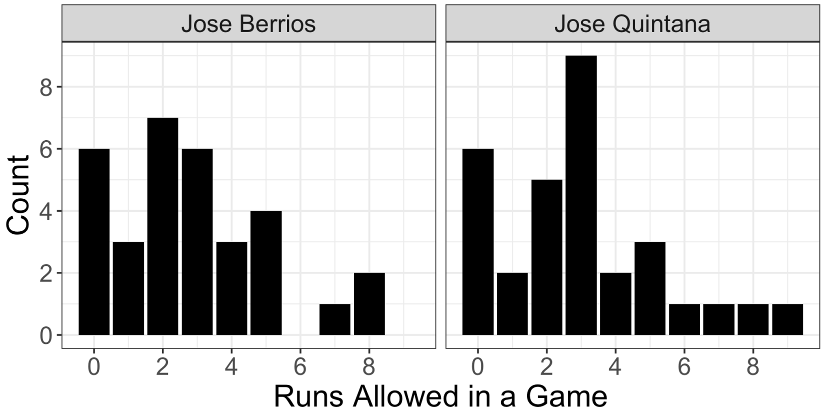


Figure 9: Histogram of runs allowed in a game for Jose Berrios and Jose Quintana in 2019. Berrios and Quintana in 2019 are overvalued according to $gWAR$ relative to $fWAR$. Both pitchers' $gWAR - fWAR$ is -1.8. Both pitchers have a high relative proportion of games in which they allows 2 or 3 runs, and a lower relative proportion of games in which they allow 0 or 1 run.



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.