By the Numbers

Volume 30, Number 1 The Newsletter of the SABR Statistical Analysis Committee March, 2021

Review

Academic Research, 2019-21

Charlie Pavitt

Charlie reviews what he found to be the most interesting contributions from the academic literature since the last issue of BTN.

Nobuyoshi Hirotsu and J. Eric Bickel (2019), Using a Markov decision process to model the value of a sacrifice bunt, Journal of Quantitative Analysis in Sports, Vol. 15 No. 4, pp. 327-344.

This is a potentially significant contribution to the literature on the value of sacrifice hits. The authors examined not only the two standard criteria for evaluation (the probability of scoring and the expected number of runs scored) but additionally the probability of winning by including both batting orders and distinguishing between the top (visiting team at bat) and bottom (home team at bat). Further, they controlled for base-out situation, inning, lead (from 20 runs behind to 20 runs ahead), and batting order position.

There were several simplifying assumptions that must be kept in mind when considering the practical significance of the results. Batting performance was measured by taking the average of each lineup position across the 2013 National League, including only pitchers for the ninth spot. Double plays and home field advantage were ignored. Advancement on hits was accomplished using the often-used D'Esopo/Lefkowitz assumptions:

- Runner on second scores on a single, but otherwise runners never take the extra base on hits;
- Runners do not advance on outs;
- Bunts erase the lead runner (if unsuccessful) or the batter (if successful), and all other runners advance

What is critical about using probability of winning as the criterion of evaluation is that, unlike probability of scoring and expected runs scored, the impact of what occurs in one inning affects what occurs in the next. This can lead to different implications than what has been generally accepted, and that is exactly what the authors found.

They reported in detail what they uncovered for sac bunts by the home team with runner-on-first-and-nobody-out when the probability of a successful bunt ranged from .4 to .6 to .8 to 1.0. To maximize expected runs, nobody should ever bunt with success rates of .4 or .6, and the #9 hitter should bunt with a success rate of .8 or higher. To maximize the probability of scoring, #9 should bunt with a success rate of .4 or higher, #2, #5, and #6 should also bunt with a success rate of .6 or higher, and all lineup positions should bunt with a success rate of .8 or more. None of this should be too surprising, as previous work has rejected the sacrifice for overall run scoring but noted situations in which it increases the odds of scoring at least one run.

Things get complex when we turn to probability of winning. I can't list all of the details here; they are very specific concerning when bunting is and is not a good idea based on batting order position, win probability, inning, and score. The point is that the

In this issue

The previous issue of this publication was June, 2019 (Volume 29, Number 1).

implications of these results are starkly different from those of earlier research. However, it should be kept in mind that employing these tactics would, according to Hirotsu and Bickel's calculations, never increase the probability of winning even as little as one percent.

The study included a number of demonstrations of their estimations, all assuming a .8 success rate. The most interesting of these are:

- A list showing the top ten circumstances in which a bunt increased the probability of winning with multiple base runners, not surprisingly all occurring in the 9th inning;
- The change when baserunner advancement is assumed to be more aggressive (1st to 3rd on all singles, 1st to home on all doubles) and less aggressive (2nd to 3rd on singles) than the D'Esopo/Lefkowitz model;
- The impact when double plays are added to the models (the results barely change).

The authors also claimed that the model can allow for incorporating differences in lineup position for bunting success, probabilistic information for baserunning advancement, and home field advantage probability. Although they did not mention incorporating different hitting ability at the lineup positions, I would think that would be easy – just different figures at the beginning.

Jeff Barden and Alex Vestal (2019), Horizontal competition and interorganizational exchange partner selection: An analysis of Major League Baseball player trades, Strategic Organization, Vol. 17 No. 3, pp. 311-333.

Way back in November 2007, I reported in BTN on a very interesting article written by Jeff Barden and Will Mitchell on the extent to which past relationships among general managers impacted on player exchanges among them. Twelve years later, Barden returned with a different second author, this time exploring the impact of various factors on the number of trades between teams from March 1985 to April 2003: a total data set of 1,636 transactions in all.

Results showed teams were indeed less likely to trade within division and with geographically close teams, supposedly because of the inherent rivalries, and more likely to trade with teams against whom they have played a lot of games, particularly if they have not traded in the past, supposedly because of increased knowledge about one another's players. (Would the two cancel one another out in practice?) The interesting control variables of payroll and performance differences, plus having a common trading partner, increased trading activity alone but washed out when the hypothesized predictors were entered into subsequent models.

John Charles Bradbury (2019), Monitoring and employee shirking: Evidence from MLB umpires, Journal of Sports Economics, Vol. 20 No. 6, pp. 850-872.

Frequent-contributor-to-the-literature John Charles Bradbury used Retrosheet data from 2000 to 2009 to examine the impact of QuesTec on how umpires called balls and strikes. QuesTec was the system installed in 11 ballparks from 2001-2008 that allowed for the evaluation of home plate umpire calls.

Bradbury found that ballparks with QuesTec had fewer called strikes than the ballparks without it, to the tune of .016 per PA or .81 per game on average. This impact was overwhelmed by other factors, most notably a directive to umpires to be more accurate, leading to the called strike rate to increase by two percent between 2000 and 2001 (the year of the directive) and another half-percent in subsequent seasons.

As for the effect of control variables: consistent with past research, there were fewer called strikes for home team batters, which is part of one of the research-supported explanations for home team advantage, crowd noise; yet *more* called strikes due to the attendance/home team batter interaction, which is inconsistent with that explanation. In addition, there was deference for experienced batters and pitchers (consistent with previous literature) and more called strikes for catchers (different from the previous literature).

Jim Downey and Joseph McGarrity (2019), Pressure and the ability to randomize decision-making: The case of the pickoff play in major league baseball, Atlantic Economic Journal, Vol. 47 No. 3, pp. 261-274.

Back in 2015, the authors published an article in this journal in which they researched when pickoff attempts were more versus less likely. This time they examined the *sequence* of pickoff throws, rather than the frequency.

They kept their dataset (from Retrosheet) from the previous study – all pitches thrown in the A.L. between June 9 and June 13, 2010, limited to situations with a runner on first only (sample size of 1,738).

They found that pitchers were pretty good at randomizing their alternation between throws to plate and to first, with one exception: righty pitchers vs. good base stealers (those in the upper third in SB/times on first) in relatively close games (–2 to 2 runs score difference). In that one case, they tended to alternate predictably between pitches to the plate and throws to first.

In addition, the authors hypothesized that when it is more likely for the batter to be successful, there is less of an incentive for a baserunner to try to steal and so less reason to throw to first. Consequently, there were more throws to first with more strikes on the batter, and fewer after the count went to three balls.

Theo Tobel (2020), The art of the brushback, Baseball Research Journal, Vol. 49 No. 2, pp. 41-46.

In this article from our own *Baseball Research Journal,* Theo Tobel looked at brushback pitches (which he defined as those pitches in the batter's box that do not hit the batter), and explored their association with relevant offensive indices in a well-intentioned but flawed study.

Theo began by trying to determine whether brushbacks had an effect on batting averages. Instead of using the entire 2018 season as he should have, Theo took three separate random samples of 15,000 at bats without letting the reader know whether a specific plate appearance could or could not have been included in more than one sample. In any case, I summed the three samples together (despite the fact that some PAs may have been in multiple samples and, if so, would be double counted).

Despite Theo's claims of no evidence, the actual difference between the two (.238 with brushback and .251 without) points to a significant impact.

Theo performed the same analysis with walk rate, and here the difference was obviously important—summing two samples, 18.9 percent with a brushback but only 6.7 percent without.

Turning to better-performed analyses, Theo determined:

- Two-seam fastballs, sinkers, and cutters were more likely to be brushback pitches, and four-seam fastballs and changeups less likely;
- As expected, brushbacks were proportionally more likely when the pitcher was ahead or even in the count. But surprisingly, with three-ball counts, they occurred more often with increased number of strikes.

Charlie Pavitt, chazzq@udel.edu ♦

Back issues

Back issues of "By the Numbers" are available at the SABR website, at http://sabr.org/research/statistical-analysis-research-committee-newsletters, and at editor Phil Birnbaum's website, www.philbirnbaum.com .

The SABR website also features back issues of "Baseball Analyst", the sabermetric publication produced by Bill James from 1981 to 1989. Those issues can be found at http://sabr.org/research/baseball-analyst-archives.

Submissions

Phil Birnbaum, Editor

Submissions to *By the Numbers* are, of course, encouraged. Articles should be concise (though not necessarily short), and pertain to statistical analysis of baseball. Letters to the Editor, original research, opinions, summaries of existing research, criticism, and reviews of other work are all welcome.

Articles should be submitted in electronic form, preferably by e-mail. I can read most word processor formats. If you send charts, please send them in word processor form rather than in spreadsheet. Unless you specify otherwise, I may send your work to others for comment (i.e., informal peer review).

I usually edit for spelling and grammar. If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does.

I will acknowledge all articles upon receipt, and will try, within a reasonable time, to let you know if your submission is accepted.

Send submissions to Phil Birnbaum, at 110phil@gmail.com .

"By the Numbers" notifications

SABR members who have joined the Statistical Analysis Committee will receive e-mail notification of new issues of BTN, as well as other news concerning this publication.

The easiest way to join the committee is to visit http://members.sabr.org, click on "my SABR," then "committees and regionals," then "add new" committee. Add the Statistical Analysis Committee, and you're done. You will be informed when new issues are available to download from the SABR website.

An Exponential Measure of Career Greatness

Randy Klipstein

There have been different statistics created to measure a player's career. One such stat is JAWS, created by Jay Jaffe, which is partially based on the seven best seasons of a career. Here, Randy Klipstein suggests that the number seven seems arbitrary, and discovers a simpler, more elegant measure that corresponds very closely to JAWS, with a structure similar to those already used for other sabermetric purposes.

The invention of advanced statistics such as bWAR, fWAR, WARP, Win Shares, WAA, or whatever, has provided a construct to compare players within a given season and across seasons. To the extent that we agree with the underlying calculations, we now have tools to compare seasonal player results. Everyone has their favorite metric. Some use a simple or weighted average of the various metrics, reasoning that a combination of such statistics will cancel out flaws in the individual methods.

While such statistics are useful in evaluating seasons, it is clear that when attempting to evaluate careers, simply summing lifetime Wins Above Replacement, or any similar statistic, doesn't work as a measure of greatness. For instance, Rick Reuschel amassed about 40% more bWAR than Sandy Koufax. Moving to Wins Above Average helps, but Reuschel is still 20% ahead. The conclusion: peak seasons matter. WAR is a measure of value, and ten units of WAR amassed in one year is as valuable as ten units of WAR amassed over a twenty-year career. But there is an opportunity cost. That cost is paid in roster spots.

We evaluate careers to settle debates (who was the best, who are the top 10, who was the best first baseman in the 1980s, etc.) and to provide an objective measure of which players are worthy of enshrinement in the Baseball Hall of Fame. A breakthrough came when Jay Jaffe developed JAWS, Jaffe Wins Above Replacement Score, in 2004. JAWS is calculated by averaging a player's career bWAR with the sum of the player's seven best (not necessarily consecutive) seasons of bWAR. In this manner, JAWS, in one number, reflects the twin attributes of our immortals: impressive career totals and a sustained peak, the *de facto* standards of the Hall. Indeed, HOF voters are unimpressed with extraordinary lifetime totals without a sustained and lofty peak (think Lou Whitaker and Tommy John). They are similarly unimpressed with an imposing peak without towering lifetime totals (think Dale Murphy and Johan Santana).

The JAWS equation is:

$$JAWS = \frac{\sum_{Peak\ 7\ Seasons} WAR + \sum_{All\ Seasons} WAR}{2}$$

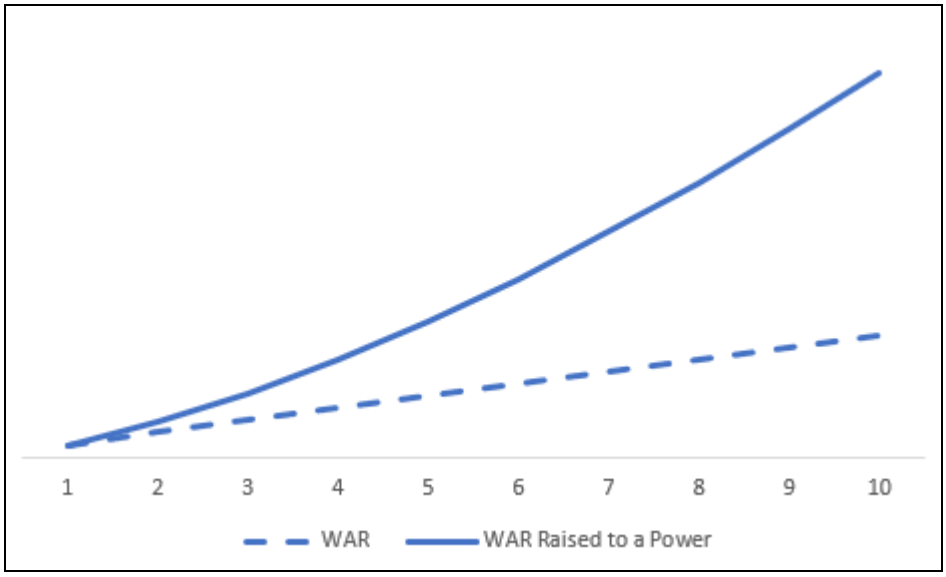
It can be rewritten as:

$$JAWS = \frac{2\sum_{Peak\ 7\ Seasons} WAR + \sum_{Other\ Seasons} WAR}{2}$$

Recognizing that JAWS is only as good as the underlying assumptions in the version of WAR that is used, I always admired JAWS. I had a concern, however, about the arbitrary use of seven seasons. Why not five or ten or any other number? And why are the best seasons counted twice and not three times?

I sought to develop an alternative method for evaluating a player's career, and thus Hall of Fame merit, which is less arbitrary than JAWS. I wondered if a solution might be found with a non-linear equation. What if I simply summed WAR raised to a power for each season of a player's career? This would provide an alternative method to reward players for good and great seasons. What made this look promising to

me is that the career value would increase faster than the actual seasonal WAR figure. In other words, the magnification of seasonal results increases as the accomplishment becomes greater and increasingly rare. I'll call this the exponential method, and here's how it increases relative to plain WAR:



The next decision was what value to use for the exponent. With a nod to Bill James' Pythagorean Theorem, I chose two. Then, I downloaded seasonal data for all MLB players, including WAR, from www.baseball-reference.com, which would make it easy to perform the calculations.

Because squaring a negative number turns it positive, I did not square seasonal WAR figures that were less than zero, but rather used the actual WAR. Initially, I was pleased with the results. The top three were Babe Ruth, Walter Johnson, and Cy Young. Upon further examination, however, I saw that the method indicated that Jose Rijo had a slightly better career than Tommy John and I set the project aside for a while.

Returning to the project, it was obvious that two was too high of an exponent. It distorted great seasons, like Rijo's 9.2 WAR in 1993, beyond reasonableness. So, if two was too high and one would just yield the sum of a player's WAR, I chose the obvious exponent for my next attempt: 1.5. Since raising a negative number to a fractional exponent creates an imaginary number (and imaginary players exist only in fiction), I kept the same logic for seasonal WAR figures below zero. The equation became:

$$\sum_{\substack{All\\Seasons}} WAR^{p}$$

where p = 1.5 when WAR >= 0, otherwise p = 1

I re-ran the numbers for each player and put the resulting "exponential" figure and ranking for each player next to the player's JAWS figure and where they ranked based on JAWS. I was shocked to see that the methods produced almost identical rankings. The top six players were ranked identically (Ruth, Johnson, Young, Bonds, Mays, and Cobb, respectively). Among the JAWS' top 30, no one's exponential ranking was off by more than +/- 3 positions. In JAWS' top 100, only nine players were off by double digits, and five of them were nineteenth century pitchers. The other four were off by ten to thirteen positions.

The disparities grow as you go further down the list. For instance, JAWS ranks Don Baylor as the 1000th greatest player, while the exponential method has him as the 1088th greatest. As you move down the list, or up the normal distribution curve, the data becomes denser. There is actually little difference between Don Baylor and the 88 players ranked immediately ahead of him.

And, by the way, Tommy John now ranked comfortably ahead of Jose Rijo.

Of course, the scales are quite different. Babe Ruth has a JAWS figure of 133 and an exponential figure of 573 (both figures are rounded to whole numbers and are based on Baseball Reference's version of WAR in effect when I downloaded the figures from their website).

I am not sure why the two methods come so close. When compared to the JAWS ranking, the exponential method overrates players with relatively short careers and high peaks, like certain nineteenth century pitchers (Silver King and Guy Hecker, in particular). The exponential method underrates players who had long careers without particularly notable peak seasons, like Cap Anson, Nolan Ryan, Ryne Sandberg, and Tim Raines. Perhaps if I lowered the exponent slightly, the results would match JAWS even more closely, but that was never the goal of this project.

Bill James himself discovered the magic of raising seasonal totals to the power of 1.5. In an article published on his website[1], ranking the Royals' all-time greats, Bill described his reasoning:

> Why 1.5? Why do we raise this [season Win Shares] to the power 1.5?
>
> Because 1.0 isn't enough, and 2.0 is too much. Just my judgment, you know. If you don't raise it to any power at all, then three seasons as an average player is worth as much as an MVP, which is not right. If you Square the Win Shares—raise it to the power 2—then an MVP is worth as much as 9 average players, which doesn't seem right, either. I tried 1.1, 1.2, 1.3, 1.4, 1.5 ... all the way up to 2.0, and 1.5 seemed right. At 1.5, a 30-Win Share season is worth 41% more than two 15-Win Share seasons. That seems reasonable.

The ultimate question is which method is better for evaluating careers. I do not know. And I do not know whether 1.45 or 1.55, or some other number, would be a better exponent to use. I know that this method, with exponent 1.5, produces remarkably similar results to JAWS in terms of ranking players. Both methods are easy to compute, but this exponential method is a bit easier to state and seems less arbitrary.

This project has given me a heightened sense of confidence in using JAWS as an indicator of greatness and Hall of Fame worthiness. Each method, of course, is only as good as the quality of the underlying seasonal calculation.

Here are the top 10 players as ranked by JAWS, the exponential method, using the power of 1.5, and simply summing seasonal WAR (to the power of 1):

Rank	JAWS	Power of 1.5	WAR
1	Babe Ruth	Babe Ruth	Babe Ruth
2	Walter Johnson	Walter Johnson	Cy Young
3	Cy Young	Cy Young	Walter Johnson
4	Barry Bonds	Barry Bonds	Barry Bonds
5	Willie Mays	Willie Mays	Willie Mays
6	Ty Cobb	Ty Cobb	Ty Cobb
7	Roger Clemens	Hank Aaron	Hank Aaron
8	Hank Aaron	Rogers Hornsby	Roger Clemens
9	Rogers Hornsby	Roger Clemens	Tris Speaker
10	Tris Speaker	Tris Speaker	Honus Wagner

[1] "A Galaxy of Royal Stars", September 25, 2020. https://www.billjamesonline.com/a_galaxy_of_royal_stars/

There are 117 players who appear on the top 100 ranking of at least one of the three methods. Thirteen appear only on the WAR top 100:

Paul Molitor	Jim Thome	Rafael Palmeiro	Barry Larkin	Carlos Beltran
Bill Dahlen	Paul Waner	Ted Lyons	Frankie Frisch	
Lou Whitaker	Derek Jeter	Alan Trammell	Scott Rolen	

Eleven players are not on the WAR top 100 but make the other two lists:

Amos Rusie	Charlie Buffinton	Mickey Welch	Wes Ferrell	Carl Hubbell
Tommy Bond	Ed Walsh	Ernie Banks	Robinson Cano	Bob Feller
Al Spalding				

Sam Crawford appears on two of the three lists, missing on the JAWS list. Bobby Grich also makes two out of three, he misses on the Power of 1.5 list. Three players appear only on the JAWS list: Jim Palmer, Duke Snider, and Hal Newhouser. And three players appear only on the Power of 1.5 list: Tony Mullane, Bob Caruthers, and Roy Halladay.

Here's that last paragraph in chart form. A blank cell means the player didn't make that list:

JAWS	Power of 1.5	WAR
	Sam Crawford	Sam Crawford
Bobby Grich		Bobby Grich
Jim Palmer Duke Snider Hal Newhouser		
	Tony Mullane Bob Carruthers Roy Halladay	

Notice that JAWS and the Power-of-1.5 method share 96 of the top 100 players.

For the record, here are the 85 players who appear in the top 100 on all three lists:

Babe Ruth	Alex Rodriguez	Warren Spahn	George Brett	Robin Yount
Walter Johnson	Big 6 Mathewson	Eddie Mathews	Al Kaline	Johnny Mize
Cy Young	Lefty Grove	Bob Gibson	Chipper Jones	Nolan Ryan
Barry Bonds	Mickey Mantle	Hoss Radbourn	Jeff Bagwell	Ed Delahanty
Willie Mays	Mike Schmidt	Cal Ripken	Curt Schilling	Ozzie Smith
Ty Cobb	Tom Seaver	Carl Yastrzemski	Cap Anson	Robin Roberts
Hank Aaron	Rickey Henderson	Roberto Clemente	George Davis	Ken Griffey
Rogers Hornsby	Nap Lajoie	Phil Niekro	Rod Carew	Fergie Jenkins
Roger Clemens	John Clarkson	Bert Blyleven	Mike Mussina	Charlie Gehringer
Tris Speaker	Mel Ott	Jim McCormick	Dan Brouthers	Joe DiMaggio
Honus Wagner	Albert Pujols	Wade Boggs	Pete Rose	Roger Connor
Kid Nichols	Greg Maddux	Steve Carlton	Ron Santo	Bobby Wallace
Ted Williams	Randy Johnson	Eddie Plank	Brooks Robinson	Larry Walker
Stan Musial	Frank Robinson	Pedro Martinez	Arky Vaughan	Frank Thomas
Pete Alexander	Tim Keefe	Adrian Beltre	Johnny Bench	Gary Carter
Eddie Collins	Jimmie Foxx	Gaylord Perry	Tom Glavine	Harry Heilmann
Lou Gehrig	Joe Morgan	Pud Galvin	Reggie Jackson	Luke Appling

Randy Klipstein, rbk65@optonline.net ♦

Career Park Effects for Individual Players

Pete Palmer

Park effects are widely available for ballparks, showing which were more or less favorable to offense. But not all players will show the same effects. Some are poorly suited to the park, and some may not have played well at home for any number of other reasons, including random luck. Here, Pete Palmer calculates career park effects for every player in MLB, describes the methodology, lists leaders, and offers a download of park and player data going back to 1904.

When the original work on park effects was started in the 1970s, I spent hundreds of hours over several years going through American League microfilm to compile stats for 150 or so players. Now, thanks to the wonderful work at Retrosheet by Dave Smith, Tom Ruane, and others, we now have play-by-play data for every game back to 1928 and box scores back to 1904. As a result, I was able to generate data for every player and team in a matter of minutes.

Definitions and theory

Park Factor is equal to runs scored and allowed at home per game divided by the same total in road games. However, variation due to chance is quite large, so the SD for a single season is just about equal to the typical difference between parks.

In baseball, the standard deviation of any sample of runs is approximately equal to the square root of two times the number of runs. That's higher than in soccer or hockey; for those, the SD is only the square root of the number of goals. Baseball is higher because runs come in bunches—you can score more than one run at a time and a play that leads to a run often sets up a situation for more runs.

A typical team whose home games see 720 runs scored (both teams combined), would have an SD about 38 (the square root of 1440). What that means is if you simulated 81 games using typical probabilities for the various events, your total number of runs scored would be between 682 and 758 runs about two-thirds of the time, and between 644 and 796 about ninety-five percent of the time. The road total would be the same.

The SD of the difference between home and road would be about 54 (the square root of 38 squared plus 38 squared). Fifty-four runs out of 720 is exactly 7.5 percent, so the SD of one-year park factor is 7.5 points. That's usually expressed as a park factor of "107.5", in the sense that teams in that park score 107.5% as many runs as in the average park.

(Important note: my park factors are "twice as extreme" as those used at sites such as Baseball Reference. That's because those others provide park effects for players, who play only half their games at home. My SD of 7.5 points is equivalent to 3.75 points on Baseball Reference. That's because a given player might play in the context of a 107.5 park at home, and 100 on the road, for an overall average of 103.75. I'll continue to use the term "park factor" to mean the undiluted number (home/road), and "BPF" to represent the diluted park factor, since "BPF" is what Baseball Reference uses. My "BPF" is not exactly the same as Baseball Reference's, as theirs makes various adjustments, but it should be close.)

If you calculate actual park factors, you'll find the SD is about 11 points. Since park factor is the sum of true park differences and random variation, we can say

```
SD(observed)^2 = SD(random)^2 + SD(true)^2
11^2           = 7.5^2 + SD(true)^2
SD(true)^2     = 11^2  - 7.5^2
SD(true)       = 8
```

That's for a single season. For a decade, SD(random) is reduced by the square root of 10, which brings it down from 7 to 2.4. So, for a ten-year period,

```
SD(expected to observe)^2 = SD(random)^2 + SD(true)^2
SD(expected to observe)^2 = 2.4^2 + 8^2
SD(expected to observe)   = 8.4
```

And, in fact, we do observe an SD of around 8.4 for decade samples.

Teams

For every year in MLB history, I looked at the subsequent 10-year park factor for every team.

Unsurprisingly, Colorado had the highest park factors. From 1993 to 2002, the park averaged 147.4. In 2003, the team started storing their baseballs in a humidor, which prevented the balls from drying out, and the number was reduced to 121. But the past decade has seen it jump back up to 135.

There were only 18 non-Colorado seasons with park factors of 120 or higher. Philadelphia's Baker Bowl is clearly in second place, with 16 of those 18. Wrigley Field in the seventies accounted for the other two.

Every other park in history came in between 119 and 82. The bottom eight slots, and eleven of the lowest twenty, belong to San Diego around the first few years of Petco Park, with park factors in the low 80s.

Home field advantage

Team park factors are based on runs scored by both teams combined. If you were to look only at the home team, you'd find the apparent park factor is higher because home teams hit better with home field advantage, and visiting teams worse.

A typical team plays at a .540 rate at home and .460 on the road, or around 44-37 vs. 37-44, respectively. This means the runs scored rate at home is about 5 percent higher, and the runs allowed rate at home is 5 percent lower. (However, because they don't always bat in the bottom of the ninth, home team batters get about five percent fewer plate appearances, so their own total runs are about the same home as road. But it's the rates that we'll be working with here, so the bottom-of-the-ninth effect won't factor into play.)

So even in an average park (100), if you look only at the home team's runs, they will appear to have a park factor of 110.

Runs vs. OPS

That 110 vs. 90 is the ratio for runs the team produces. When we talk about hitters, we won't be looking at runs, but on their usual batting statistics, like OPS. So the scale will be different.

Since runs are proportional to on-base times slugging, a 10 percent increase in runs is equivalent to a 5 percent increase in both OBP and SLG (since 1.05 * 1.05 is approximately 1.10). It's therefore a 5 percent increase in OPS, since a 5 percent increase in both OBP and SLG is also a 5 percent increase in their sum.

That means that, point-for-point, an OPS park factor is twice as "potent" as a park factor in runs.

Player park distribution

For all players in MLB history, I calculated their career park factor by averaging the parks they played in. The average for all players was 100, with an SD of about 10 points. So two-thirds of players had career home park factors between 90 and 110, which means two-thirds had career OPS home park factors between 95 and 105.

Home field advantage adds another 10 points of park factor for home players, equivalent to 5 points of OPS park factor. So the distribution of player park factors will have a mean of 5 points and a standard deviation of 5 points.

For an average player OPS of .750, one point (percent) of park factor is .0075, so if you're interested in the arithmetic difference (by subtracting rather than dividing), home field advantage is worth about .038 points of OPS, and the standard deviation of park effects is also around .038 points of OPS.

The 5-point SD is only for the differences in environment. There are other sources of variation in actual player performance, such as (a) the player's suitability to the park, such as L/R differences, (b) random luck, and (c) other things.

Players

For every player, I used Retrosheet data to figure his actual home and road batting line for his career. I then divided his home OPS by his road OPS and multiplied by 100.

I will call that "OPS ratio" or "park ratio" to distinguish it from "park factor," which will be the expected number for the park only. Wrigley Field might have an OPS park factor of (say) 110—which is 115 including home field advantage. But a particular Cub might hit better or worse than expected at Wrigley, so his actual park ratio could be significantly higher or lower than 115, for any given year.

(You can download a full listing of player park ratios—and, in fact, most of the other data mentioned in this paper—from the SABR website, in an Excel spreadsheet. It's at https://sabr.box.com/v/palmer-park-factors.)

As expected, most of the players with high park ratios are from Colorado.

The table below shows expected and actual OPS ratios, and actual OPS values home and away, for selected Colorado players. Blackmon has the highest OPS ratio in history, Gonzalez is fourth, and Story is fifth.

Walker and Helton each had Hall of Fame-caliber stats on the road, but Walker had to wait until his tenth year to get in. Helton is running at about a quarter of the vote, and it doesn't look like he will make it. Both are at 34 wins above average on my total player rating list, well into Hall of Fame territory.

	career PF	expected OPS ratio(includes HFA)	actual OPS ratio	OPS home	OPS road
Nolan Arenado	130	120	124	.999	.802
Charlie Blackmon	130	124	134	1.000	.744
Todd Helton	128	119	123	1.054	.857
Neifi Perez	128	119	118	.730	.619
Dante Bichette	128	119	129	.942	.733
Carlos Gonzalez	127	119	132	.963	.726
DJ LeMahieu	127	119	123	.857	.698
Trevor Story	127	119	132	1.000	.757
Vinnie Castillo	125	118	116	.860	.741
Larry Walker	125	118	123	1.072	.867

Chuck Klein was probably the only other player to have his HOF chances impacted by playing in a good hitting park. He had a 126 OPS ratio, mostly because of Baker Bowl (133), but with some contribution from Wrigley. He got minimal support from the writers (less than a quarter of the vote), but was finally selected by the Veterans' Committee in 1980. At 21 wins above average, he is a reasonable choice.

You might think Mel Ott benefited greatly from playing in the Polo Grounds, since he hit 323 of his 511 homers there. But that park, although good for homers, was not good for scoring runs. Ott had a .980 OPS at home and .918 on the road for a 107 ratio, only slightly above average.

Most of the great players were not particularly helped or hurt by their home park. People remember Henry Aaron playing in homer-happy Atlanta, but he was in Milwaukee for many years. His overall OPS ratio was a below-average 102 (110 in Atlanta and 97 in Milwaukee).

Here's a selection of notable players in MLB history.[1] In the chart below, TPR is my "total player wins". Lajoie and Wagner are since 1904 only.

	career value (TPR)	career park factor	expected OPS ratio(includes HFA)	actual park OPS ratio	OPS home	OPS road
Barry Bonds	130	96	103	103	1.070	1.040
Babe Ruth	129	95	103	103	1.181	1.148
Nap Lajoie	95	100	105	115	.855	.743
Ted Williams	86	113	112	106	1.150	1.082
Rogers Hornsby	86	97	103	103	1.026	.996
Ty Cobb	96	102	106	101	.949	.941
Willie Mays	84	98	104	102	.956	.934
Henry Aaron	83	102	106	102	.939	.925
Tris Speaker	83	104	107	114	.990	.869
Honus Wagner	82	103	107	105	.866	.821

Fenway Park was favorable to hitters, but less so to left-handed hitters like Ted Williams. The park factors in the chart are based on performance by both lefty and righty batters, which is why Williams shows a lower park ratio than the total numbers suggest.

There are other parks that seem to have significantly different effects between lefties and righties.

LH/RH splits for parks

To look at handedness, I took each team-season and compared the OPS ratio for left-handed batters to the OPS ratio of right-handers. The only decade periods where righties had a home/road OPS at least 10 percent higher then lefties were all the Fenway 10-year periods contained in 1937-54. Fenway gradually moved towards parity, and since 1970 has been neutral between lefties and righties.

Tris Speaker played in League Park in Cleveland for most of his career, which was very favorable to lefties, with its 340-foot right field power alley. He had a 109 ratio in Boston and 120 in Cleveland. There were six parks that an OPS ratio of 110 or more for a long period: Sportsman's Park in St. Louis (Browns 1913-42, Cardinals 1953-65), Baker Bowl for the 1925-37 Phillies, League Park for the 1926-37 Indians, Yankee Stadium 1948-84 (including two years at Shea Stadium while Yankee Stadium underwent a makeover) and the two Montreal parks. All but Montreal had a large difference between the power alleys in left and right. Sportsman's Park was also used by the Cardinals for 1920-42 and beyond, and although left-hand friendly at 105, it was not quite as much so as it was as an AL park.

It is difficult to assign left- and right-handed park factors, however. Fenway Park, while favoring righties, has had several left-handed hitters who did well there, such as Carl Yastrzemski and Wade Boggs. The outlier is Fred Lynn, who, while with Boston, had an OPS ratio of 132—the highest in league history for any AL player (minimum 500 games with a team), although he did that in the period when the park was balanced. (Lynn's career ratio wound up at only 119, due to his seasons in other parks.)

Yankee Stadium is a confusing park because although it does have a short right field foul pole, the park opens quickly in right and had a huge left field until the remodeling in 1976. In my original research, I never found a Yankee who hit especially well there. Yogi Berra had a park ratio of 107, Tommy Henrich 106, Earl Combs and Bill Dickey 104, Charlie Keller 100 and Roger Maris 97. Mickey Mantle, a switch-hitter who usually batted left-handed, was 104. Thanks to Retrosheet, I was now able to identify Bobby Murcer, at 119, as the only lefty with a high home advantage.

Lou Gehrig had difficulty at home, with a ratio of 96, where 105 is average. He had a lifetime batting average of .351 in road games, with an OPS of 1.105. In 1930, in only 78 road games, he had 117 RBI; he had 98 road RBI in two other years.

[1] My Excel spreadsheet of all players and teams (1904-2019) can be downloaded from the SABR website at https://sabr.box.com/v/palmer-park-factors .

Babe Ruth had unbelievable stats in the Polo Grounds, including a 1.516 OPS in 1920. But Fenway Park was a terrible home run park. When Ruth led the league in HR in 1918 while playing outfield half time, all 11 of his homers were on the road, which set an all-time record for the American League. In 1919, there were 13 homers hit by both teams in Fenway and Ruth had 9 of them—plus another 20 on the road. It should have been no surprise when he topped 50 in 1920 for New York.

Righties are definitely at a disadvantage in Yankee Stadium. The only righties at average (105) or better (minimum 3500 at-bats) are Willie Randolph at 106 and Phil Rizzuto at 105. Dave Winfield had a park ratio of 97, Elston Howard 95, Bill Skowron 93, Joe DiMaggio 92, and Joe Gordon 91 ... and, then, at the bottom, there's Gil McDougald.

McDougald was perhaps the hitter most unsuited to his park in MLB history. His OPS park ratio of 80 is the lowest of any player with 3500 at-bats. It's so low that there's nobody even close. The next two lowest come in at 88—Johnny Logan of the Milwaukee Braves and Jim Ray Hart of the San Francisco Giants—and there's nobody else below 90.

From Baseball Reference, here are McDougald's home/road splits, adjusted to 162 games each:

	AB	R	H	2B	3B	HR	RBI	BB	K	avg	obp	slg	OPS	tOPS+
Home	542	68	138	16	6	6	59	61	70	.255	.333	.348	.680	78
Road	591	100	175	29	5	20	80	75	81	.296	.379	.469	.847	121

McDougald hit homers more than three times as often on the road as he did in Yankee Stadium. Still, he received MVP votes in five of his ten seasons with New York, and was also a five-time All-Star.

Pete Palmer, petepalmr@aol.com ♦

Declaring WAR: An Analysis of "Wins Above Replacement" and Its Implementations

Charlie Pavitt[1]

"Wins Above Replacement" has become one of the most commonly cited sabermetric statistics. But there are various versions of WAR, and what does "replacement" really mean, anyway? Here, Charlie Pavitt compares four different WAR statistics and their underlying explicit or implicit definition of "replacement value."

WAR (Wins Above Replacement) has apparently become one of the most recent sabermetric concepts to permeate today's media environment. But the media are vague in their usage. When they say that Joe Shlabotnik (if you don't know who he was, you youngster, look him up) has a WAR of 2.3, I have always wondered which version they mean. This is because there are multiple versions of WAR—differing in which components they include, how they measure each of those components, the relative weights of the components, and which replacement level is presumed.

Although versions differ, they are all based on the concept that they wish to evaluate performance compared to a hypothetical replacement player. Indeed, Tom Tango refers to the concept of WAR as a "framework," and each version of WAR as an "implementation" of that framework.[2] The common concept (framework) means the different versions (implementations) of WAR have much in common, which is what will make it possible to focus on their specific differences.

Past research has concluded that the versions are highly intercorrelated, but even so, some could be calibrated to give consistently higher or lower WAR figures than others. Further, even if their averages are close, some might still rate certain types of players systematically higher, counterbalancing that with types of players just as systematically lower, with their differences cancelling out.

Just a quick glance shows at least some players valued far differently across WAR versions. Bryce Harper always appears comfortable being an outlier, but I assume he would be far happier with his 2018 openWAR of 7.1 than his bWAR of 1.3 ... and both are significantly different from his fWAR of 3.5 and his WARP of 3.4.

Version	Developed By	Bryce Harper, 2018
bWAR	Baseball Reference	1.3
fWAR	Fangraphs	3.5
openWAR	Paper by Baumer et al.	7.1
WARP	Baseball Prospectus	3.4

It was this Bryce Harper discrepancy in particular that made me curious about the inner statistical workings of each version and motivated me to begin doing this study.[3]

For position players (we use this term to refer to those not pitching; yes, that is also a position, but we can't think of a better concise term), to keep the playing field level, we will be dissecting them in terms of the three components they all have in common—hitting, baserunning, and fielding.

Pitchers' WARs are generally constructed from one index only and will earn a simpler analysis.

[1] With plenty of help from Phil Birnbaum.

[2] See, for instance, http://tangotiger.com/index.php/site/comments/how-is-war-perfect-in-its-framework

[3] For a less-detailed comparison of fWAR, bWAR, and WARP (but not openWAR), see Slowinski (2012); Wikipedia (the font of all useful knowledge) also has a review.

1. Position players—the four WARs

1.1 fWAR (Fangraphs)

The Fangraphs version of WAR (fWAR) is described at Slowinski (2012a), with Weinberg (2014) running through an example, and with links to exact details.

It measures offense by wRAA (their version of Tom Tango's wOBA – adjusted for park, league, and number of plate appearances). Fielding is measured by Mitchel Lichtman's Ultimate Zone Rating (originally based on Sports Info Solutions data. According to Slowinski (2012b), catchers are only evaluated by their Stolen Base Runs and Runs Saved on Passed Pitches, but an examination of the fielding ratings of catchers known to be good and bad pitch framers implies that this skill now has a significant impact on these ratings.[4] For baserunning, they use their own BsR.

Preliminary figures are adjusted for position and league, and then finally to league totals, to get the sum equal to the given year's number of runs representing replacement level.

Where is replacement level? FanGraphs and Baseball Reference agreed to jointly define it as a total of 1000 WAR units per 2430 games (a 162-game schedule across the 30 teams). This figure is presumably pro-rated for strike (and, we presume, coronavirus) seasons.

That means the average team totals 33.3 WAR. Subtracting that from 81-81 means this method pegs a replacement level team at 48-114.

FanGraphs decided to allot 57 percent of the total to position players, with the remaining 43 percent to pitchers—revealing their belief about relative importance. That means 570 WAR will be distributed among position players.

For each player, his share of the 570 WAR is calculated by adding six numbers: batting, fielding, baserunning, positional adjustment, league adjustment, and replacement runs.

Batting, fielding, baserunning

Start by calculating the player's total runs by adding his batting, fielding, and baserunning contributions. So far, these are denominated in Runs Above Average (because of the metrics that FanGraphs chose).

Convert that to Wins Above Average by estimating how many marginal runs make up a marginal win. To do that, use Tom Tango's quick and dirty formula:

```
RPW = 9 * (MLB Runs Scored / MLB Innings Pitched) * 1.5 + 3
```

Divide every player's RAA by RPW to give his WAA:

```
WAA = RAA / RPW
```

Positional adjustment

Now, adjust for his position, based on innings played at that position. The following lists fWAR's positional adjustments in runs per 162 defensive games:

C	SS	2B	CF	3B	RF	LF	1B	DH
+12.5	+7.5	+2.5	+2.5	+2.5	−7.5	−7.5	−12.5	−17.5

Note both that these sum to zero for the eight fielding positions—which I assume is done to maintain the adjustments centered at league average—and that DHs are given a huge penalty on top of that. I suppose that this makes some sense in terms of fielding responsibility, but

[4] Thanks to Tom Tango for pointing this out.

then why not give pitchers a +17.5 score to keep overall mean at zero? As Tom Tango reminded me in an email, there is no positional adjustment for pitchers because they are treated as a separate population[5].

Again, divide these figures by Tango's RPW estimate to convert runs to wins.

League adjustment

The player's league adjustment results from normalizing the league total to actual league runs scored. It is calculated as follows:

1. Sum league run totals for batting, fielding, baserunning, and positional adjustments.
2. Subtract that from the actual league total runs, giving the total discrepancy for the league.
3. Calculate the player's discrepancy share as the same as his share of league PA.
4. Divide by Tango's RPW to convert to wins.

Replacement adjustment

As mentioned, the five values we have so far are denominated in wins above average. The last adjustment is the difference between wins above average and wins above replacement.

Since the player totals so far are denominated against average, they must all sum to the average, zero. But replacement level for the league is 570 wins. So, distribute those 570 wins among the players in proportion to their plate appearances:

```
Replacement Wins = 570 * (player's PA) / (league PA)
```

Summing

Now we have all six figures:

1. Batting wins above average (wRAA)
2. Fielding wins above average (UZR)
3. Baserunning wins above average (BsR)
4. Positional adjustment (as per chart)
5. League adjustment (to get the league to sum to zero)
6. Replacement adjustment to batting wins (to get the league to sum to 570)

Adding those figures gives the player's final fWAR, as calculated by FanGraphs.

We will not describe the other methods in this level of detail, as most of the calculations are similar.

1.2 bWAR (Baseball Reference)

The Baseball Reference version of WAR is usually known as bWAR, but alternatively as rWAR. I provide a description here, but full details can be found at their website.[6]

Like fWAR, bWAR rates offense by wRAA, but with a number of revisions relevant to issues such as times reaching base on errors, differences between infield and outfield singles, differences between strikeouts and other types of outs, adjustment for league, and more.

[5] When asked about the overall logic behind the adjustments, Tom replied that they are convenient, help maintain equality across leagues, that pinch-hitters should also be included (same as DHs, I presume), and that any discrepancies are compensated for at the end.

[6] https://www.baseball-reference.com/about/war_explained_position.shtml

The index for fielding used to be Sean Smith's TotalZone, but in 2012 it was changed to STATS's Defensive Runs Saved. Baserunning is measured several components: SB/CS, events during plate appearances (passed balls, wild pitches, balks, successful pickoffs, errors on attempted pickoffs, and defensive indifference), and an intricate evaluation of advancement on batted balls. There is an additional component for grounding into double plays. There are adjustments for position, and an adjustment to convert from average to replacement level.

The positional adjustments for bWAR are:

C	SS	2B	CF	3B	RF	LF	1B	DH
+9	+7.5	+3	+2.5	+2	−7	−7	−9.5	−15

These are fairly close to the FanGraphs adjustments; they may appear to be a bit less extreme, but part of that is because they're calibrated to 150 defensive games rather than 162. Again, these sum to almost zero (+.5) without the DH. There is a formula providing a player-specific adjustment for pitchers.

The total of 1000 WAR legislated by agreement with FanGraphs is in this case 59 percent to position players and 41 percent to pitchers (based on the proportion of free agent salaries going to each over 2014 through 2017). The split between leagues has changed over time, but as of 2018, 52.5 percent went to the AL and 47.5 percent to the NL, as measured by the results of interleague play and player performance when changing leagues. The agreed-on replacement level team figure of .294 corresponds to 20.5 runs per 600 PA, so a player's runs above replacement can be calculated by the difference between his run total and 20.5, adjusted for playing time. (This is almost a rule of thumb in any version of WAR—that the average full-time player is about +20 runs, or +2 wins above replacement.)

The runs indicated by each component are summed and then converted to wins. The method has changed over the years; in general, it is based on PythagenPat[7]. Some fine-tuning is likely needed to make sure that the sum across players equals 1000.

1.3 WARP (Baseball Prospectus)

Baseball Prospectus's "Wins Above Replacement Player" (WARP) is based on proprietary methods, and as a consequence BP is not forthcoming with details. In general, it uses the batting, fielding, and baserunning indices that are current for their evaluations, and as such will have changed over time. Rob McQuown, Harry Pavlidis, and Jonathan Judge (2015) included as components Batting, Fielding, and Baserunning Runs Above Average, arm ratings for fielders, and a positional adjustment. We shall be using their Value Over Replacement Player (VORP), Fielding Runs Above Average (FRAA), and Baserunning Runs (BRR) to represent them.

VORP is summarized at Wikipedia. James Click (2011), a BPer before climbing the front-office hierarchy to Astros GM, described the fundamentals for FRAA, which in contrast with the details has probably not changed much over time. It compares the number of plays (assists and putouts as measured by conventional metrics) with a league average adjusted for ballpark, number of balls in play, team pitcher handedness and groundball/fly ball ratio, and the number of double play opportunities successfully completed. The difference between the two is translated into runs saved or given up compared to average with positional adjustments included.

As with the other WARs, a sum of these provides a total run count measured against average, so it requires a recalibration to replacement level and then a division by 10 to convert runs to wins.

Catchers are evaluated for framing. Uniquely to WARP, catchers are also adjusted for throwing (both success in nailing attempted steals and tendencies to attempt steals) and blocking.

1.4 openWAR

Ben Baumer, Shane Jensen, and Gregory Matthews (2015; Baumer & Matthews, 2014[8]) have proposed an interesting attempt at interpreting WAR. They named it openWAR because, unlike the others, they have worked to make it accessible, including making their code open-access.

[7] See https://www.baseball-reference.com/about/war_explained_runs_to_wins.shtml for details.

[8] http://arxiv.org/abs/1312.7158

In one huge difference from the other methods, openWAR computations are based on changes in run expectancy and so are the only major WAR method which is context-dependent. In other words, a home run with bases loaded will count a lot more positively, and a strikeout with based loaded negatively, than either with bases empty. As such, comparing it to the other major WAR methods is to some extent apples versus oranges, as those are intentionally designed to be context-free. OpenWAR does not include a fixed positional adjustment as do the other three measures, but Greg Matthews confirmed in an e-mail that position serves as a control variable in the computations for the batting and fielding components.

Responsibility for fielding is shared between pitcher and fielder based on batted ball location (but not velocity), with the play probability for that location depending on likelihood of different fielders making the play and adjusted for ballpark.

Originally, the authors made a serious error in their conception of this index that invalidated their fielding measure (see Michael Wenz (2016), who highlighted the problem). For some reason, the easier a play to make, the more credit openWAR gave the fielders, and the less credit to the pitcher. This meant the fielder was given major responsibility for catching a popup or lazy fly ball that becomes a hit only 10 percent of the time, but little credit for making a great catch on a screaming liner in the gap that becomes an extra base hit more than 90 percent of the time. As a consequence, early fielding RAA was significantly misleading. Gregory Matthews (2016) acknowledged the issue at the time, and, in a recent email exchange, told me the calculation was fixed shortly thereafter.

Baserunners are given credit for bases advanced on hits relative to average, players are compared to baserunning averages for their position, and hitters lose run value for what is credited to baserunning.

Lastly, the sum of the three components is then re-calibrated from relative to average into replacement level terms and divided by the standard runs-to-wins figure of 10.

There is no intrinsic sum that seasonal WAR across all players is expected to equal, and as a consequence no enforced proportions of WAR divided between position players and pitchers.

2. Replacement level

The methods differ in their respective definitions of "replacement level." That means some will consistently assign WAR values to players that are characteristically higher than others. This results in different actual levels for what winning percentage constitutes replacement level, a subject that has been debated even before WAR became a thing.

There are two parts to the question. First, what should be the conceptual definition of replacement level? Should it be players freely available on waivers? Players available in the minor leagues? Players who aren't regulars on their current team? Players who would never be regulars even on expansion teams? One widely-accepted definition, originally from Bill James (1979 Baseball Abstract, pp. 84-85) and later adopted by Tom Tango, is "freely available talent," a definition which is not precise but has the advantage of being intuitive.

Second, once the definition is settled, there's still the question of how to estimate the actual level of play that would result.

The highest bottom-line figure of which I am aware was Brock Hanke's (1998) estimate of .350. After disagreeing fairly substantially in their initial formulations, Baseball Reference (originally at .320) and FanGraphs (originally .265, the lowest I've seen) agreed that approximately 1000 WAR should be available in a given season. This translates to a .294 winning average for a team of players at replacement level, almost the same as Tom Tango's definition of .292, or (after rounding) 48 wins. The sum of all players' WARs for a team in a given season should be approximately equal to the number of wins over 48 that the team achieved.

In an email exchange, Tom explained the reasoning behind a replacement level just under .300. In short, if you set up a regression equation relating number of expected wins with required salary for players capable of achieving that number, the absolute bottom of the relationship, MLB's minimum salary, predicts that .300-ish figure. In other words, replacement level is equivalent to the best player available at minimum salary.

The population of minor league free agents would be considered the pool from which replacement players would be obtained. Dave Cameron (2003) presented some evidence that this figure probably works well. First, examining 24 position players who were transferred between teams via waiver deal or minor league contract over the previous winter, he noted that, as a group, they had accumulated exactly

zero WAR over the previous two seasons. Second, of 628 players with at least 6000 career PA at that time, the only one below zero was Alfredo Griffin, at –0.08 per year. In other words, every veteran player was either at replacement level (Griffin) or above it.

The Baseball Prospectus position concerning replacement level may have changed over time. In 2002, Keith Woolner calculated the difference between the performance every starter, versus the performance of every other player at the same position on the same team, for the period 1893-1998. In general, the backups were 80 percent as productive as the starters, with the exception of catchers (85%) and first basemen (75%).[9] As I noted above, Woolner's approach implies that replacement level is the performance of all substitutes, and Baseball Prospectus has explicitly accepted that definition.

Here is how Russell Carleton (2013) slightly edited definition in the context of BP's WARP: "A replacement player is just the per plate appearance (or per inning) mathematical (weighted) average performance of all backup center fielders [the position Russell used as an example], multiplied by the number of plate appearances (or innings) that Trout [another example] or any other player whose value we want to assess played," with "all" defined as "all the bench/minors/scrap heap center fielders out there."

The second part of the definition strikes me as a little off; I would have put it as the number of PAs/innings that these fictional players would have played if each were typical of the 30 "true" starting center fielders, rather than Trout or any specific one of the 30. In any case, as WARP includes all substitutes and not only "freely available talent," it should have a higher replacement level and so lower overall WAR figures than either bWAR or fWAR. This would be consistent with BP's claim[10] that a replacement level team should win "a little over 50 games" compared with the bWAR/fWAR/Tango consensus of 48.

Interestingly, in a later post (2017), Carleton noted that replacements for LF (and perhaps other position players) are sometimes infielders who are better hitters than the team's fourth OF, which implies that replacement level for WARP might actually be a bit higher yet. This would appear not to be an issue with the other WARs because, except for pitchers, they do not compute replacement level relative to specific position.

Finally, if using openWAR, there is still another definition. OpenWAR assumes 13 position players and 12 pitchers per team, which, when multiplied by the 30 teams, means 390 position players and 360 pitchers. Computationally, the 390 position players with and 360 pitchers who faced the most plate appearances are considered above replacement; the average of all others who appeared in games form the definition of replacement level.

This approach is similar to that used by WARP—looking at actual levels of non-regular players—but openWAR's threshold implies a lower level of replacement player, meaning openWAR will give higher values than WARP.

Between openWAR and bWAR/fWAR, the difference between scores depends on the relationship between "MLB position players (pitchers) beyond the first 390 (360)" and "freely available talent." As we will see, openWAR seems to posit a much lower replacement level, and thus significantly higher WAR values for individual players.

According to Baumer et al., for 2012-13, bWAR and fWAR each correlated .88 with openWAR, and .91 with one another. I find the former of these correlations surprisingly high given that the first two methods are context-free and the third context-dependent. I know of no previous tests of association in which WARP was included.

It should be noted that regardless of whichever (reasonable) definition of replacement value is used, it's generally accepted, based on empirical data, that replacement-level hitters are generally of near-average skill in fielding and baserunning. It's only in hitting that they fall significantly below league average.

The WARs here are in line with that finding. Those that convert individual categories from average to replacement level (as opposed to adding the categories before converting) apply the conversion only to batting.

[9] As denominated by the stat "equivalent average" (EqA).

[10] https://legacy.baseballprospectus.com/glossary/index.php?search=warp

3. Is context desirable?

The general idea of WAR has been criticized by Bill James, among others, for being context-independent; in other words, ignoring when in the game different events take place. For this reason, it should not be interpreted literally, in the sense of describing exactly how many wins a player contributed to his team in the relevant season.

This is a fair criticism to an extent, particularly in the case of relief pitchers whose contributions to winning are greatly dependent on when they are generally used, and for this reason Tom Tango has suggested including leverage weightings. As just described, openWAR has taken this criticism to heart. However, I would argue that the very point of WAR is to provide values that can be used for roster design planning—for example, deciding whether or not to invest in or bid for a given player, both in fantasy leagues and (I suppose in a more sophisticated form) in real-life team operations. If so, then you want a context-free metric, as contexts change from year to year, rather than an index that measures true single-season performance that has been affected by too many influences to list here. Baumer et al. understood this, noting that they were proposing a performance measure whereas the other WARs were tools for player projection.

There is, however, another sense in which all of the WARs, or more specifically their presumed replacement level figures, are contextual. The bWAR/fWAR/Tango consensus presumes a context in which both a team's offense (batting plus baserunning) and defense (pitching plus fielding) are both performing at replacement level. Let us assume that replacement-level players are 24 percent less productive than average, a figure a bit below Woolner's. Under this assumption, performing a technical Pythagorean Equation calculation (with the 1.83 exponent) for a team that scores 76 percent as many runs as average, and gives up 124 percent as many as average, will result in a winning percentage of .290.

However, if instead we presume a context in which a replacement-level offense is paired with a league-average defense, the Pythagorean prediction is .377; a league-average offense with a replacement-level defense gives us .403. For this reason, Tom Tango sometimes referred to a replacement level of .380 for batters and starting pitchers. Figures in the vicinity of these estimates will come up again later in this paper. Tango (2007) also claimed a figure of .470 for relievers, with the difference due to the better overall performance of relievers as compared to starters implying that a higher standard is needed for a replacement-level reliever than for a replacement-level starter.

In chart form:

```
.380 – team of replacement-level hitters with average starters and relievers
.380 – team of replacement-level starters with average hitters and relievers
.470 – team of replacement-level relievers with average hitters and starters
.290 – team of replacement level hitters, starters, and relievers
```

4. Summary of differences—WAR and position player WAR

	who developed it	hitting stat	fielding stat	baserunning stat
fWAR	Fangraphs	wRAA	UZR	BsR
bWAR	Baseball-Reference	wRAA	DRS	their calculation
WARP	Baseball Prospectus	VORP	FRAA	their calculation
openWAR	academic paper by Baumer et al.	change in run expectancy	their calculation	their calculation

	how replacement level is determined	replacement level in practice	batter/pitcher split
fWAR	fixed at .294	.294	57/43
bWAR	fixed at .294	.294	based on free-agent salaries. In 2018, 59/41
WARP	performance of backups at position	.375*	no specific split
openWAR	hitters after top 13 per team	.300*	no specific split

* estimated later in this paper.

	catcher framing?	runs to wins	notable features
fWAR	yes	Tango formula	.
bWAR	no	PythagenPat	league split based on estimate of relative league strength
WARP	yes	10 runs per win	catcher fielding includes SB/CS
openWAR	no	10 runs per win	context (clutch/clustering) dependent for hitting

5. Batting WARs compared empirically

For this analysis, I used the 2018 season. I obtained the relevant hitting, fielding, and baserunning indices that were available for each version of WAR, either at the relevant websites, or, in the case of WARP, from the 2019 *Baseball Prospectus* book. I do not not include any of the adjustments mentioned above—just the published numbers for the three basic stats for each method.

The sample consists of 214 position players who amassed at least 400 plate appearances in 2018. To be honest, I established that high threshold because I did not have it in me to do all of the time-intensive data-entry-by-hand work that including more than that would have required.

Table 1 presents a summary of the overall WAR values for the four metrics.

Table 1 – Position players, overall WAR (2018, minimum 400 PA)

	mean	SD
fWAR	2.34	2.00
bWAR	2.46	2.08
openWAR	2.97	1.98
WARP	2.27	1.77

Q – Do the different WARs differ systematically in what they consider an average 400+ PA hitter?

A – Differences in what's considered average will likely appear as different overall means for the groups of players.

Given that they agreed to use the same replacement level and base their basic indices on wRAA, it is not surprising that fWAR and bWAR means are pretty close to one another. Importantly, the mean for WARP is not much different from those two. Among these three metrics, what differences do exist are not too meaningful in practical terms.

The fact that WARP is fairly close to fWAR and bWAR may imply a closer association between WARP's definition and bWAR/fWAR's operationalization of "freely available talent" than I originally expected. The relationship of what "replacement level" means in practice among the various WARs will be addressed later in the paper.

The differences among those three are dwarfed by their variation with openWAR. As described earlier, the fact that openWARs tend to be higher than WARPs for the same player is to be expected, given that openWAR's conception of replacement level as "average performance for position players beyond the 13 most active" means a lower replacement level than WARP's definition of "all position players beyond starters."

Q – Do the different WARs differ in their spread of scores among these 400+ PA hitters?

A – The four WARs' standard deviations are fairly close to one another, indicating that there is little difference among the four in the spread of scores across players. Keep in mind, however, that capping the sample at 400 PAs rather than including all players means the low end of WAR scores for each is underrepresented in the analysis. This means that the actual standard deviation for each is probably a good bit higher, and it is possible that larger differences among them would then appear.

Q – How highly do the different WARs correlate with one another? In other words, independently of differences in calibration as reflected in means, are players rank-ordered approximately the same across them?

A – Correlations are provided in Table 2.

Table 2 – Correlations between pairs of WARs

	fWAR	bWAR	openWAR
bWAR	.931		
openWAR	.850	.793	.793
WARP	.865	.840	.840

The correlation between fWAR and bWAR is quite high and a bit larger than that reported by Baumer et al.

Correlations involving openWAR are consistently the lowest, and somewhat smaller than those reported in Baumer et al. Again keep in mind that the sampling criterion (400+ PAs) will decrease correlations for the same reason as it decreases standard deviations, so there is a good chance that these would be closer to Baumer et al.'s report with the full population of players.

The correlations involving WARP fall in between the others.

Q - Even if overall means are the same, are there noticeable differences in the WAR figures assigned to individual players across the four WARs?

A – Table 3 reports the means for the difference scores between all combinations of the four indices in absolute values. (Using "raw" mean differences would result in scores higher and lower than zero canceling one another out.)

Table 3 – Mean absolute differences between pairs of WARs

	fWAR	bWAR	openWAR
bWAR	.605		
openWAR	1.009	1.090	
WARP	.810	.866	1.170

The three combinations including openWAR are greater than 1, which means that the "typical" position player in these data sets will be assigned WAR values more than 1 win apart in these cases. This could indicate different relative measurements, or weightings between openWAR and the others, for batting, fielding, and baserunning. It is more likely, however, that it is a product of the same difference in conception of "replacement level" that led its overall mean WARs to be higher than the others, which will be examined later.

Q – Do the WARs give equivalent weightings to batting, fielding, and baserunning in their computation of WAR scores?

A – In one sense, the weightings have to be equivalent, in the sense that overall WAR is the equally-weighted sum of batting WAR, fielding WAR, and baserunning WAR. In the more interesting sense, there might be different spreads within the three measures among different versions. In other words, the best fielders could be (for instance) +2 in some versions, but only +1 in others. That would effectively mean the former method would find fielding more important than the latter.

The relative size of the standard deviations across the three components gives are given in Table 4, and give an indication of the relative importance that each of the four WARs gives each. Note that they are pretty close to one another for batting and baserunning. For fielding, as a consequence of their flawed method noted above, openWAR is far smaller than the others.

Table 4 – Relative proportions of SDs of components

	hitting	fielding	baserunning
fWAR	59	29	12
bWAR	58	35	8
openWAR	73	14	14
WARP	65	25	10

The relative weightings (thanks to Phil Birnbaum for suggesting this analysis) for the four are listed below. As an example: suppose a player were at the same point on the bell curve (say, +1 SD) in all three of hitting, fielding, and baserunning. In that case, his total fWAR would be 59% attributable to his hitting, 29% to his fielding, and 12% to his baserunning.

Q – To what degree does each component contribute to the differences among the WARs?

A – Phil computed the standard deviations of the difference between different WARs for the same player, which are shown in Table 5.

The standard deviations of the differences for hitting and fielding are very close to one another for the six combinations across hitting and fielding, with almost all in the range of 8 or 9. The exception is the fWAR/bWAR comparison for hitting, which reflects their use of the same metric (wRAA) and is almost certainly responsible for the difference score SD for their total WARs being the smallest.

Differences for baserunning are far smaller, about a quarter of the size of those for hitting and baserunning. This gibes with the relative size of the SDs across players within each WAR being much smaller, and reflects the fact that baserunning has only a small impact on total WAR scores as a whole.

Table 5 – SDs of same-player differences between pairs of WARs

	overall WAR	hitting	baserunning	fielding
fWAR/bWAR	7.6	4.3	2.3	8.9
fWAR/openWAR	10.9	8.9	3.3	8.2
fWAR/WARP	10.1	7.2	2.4	9.0
bWAR/openWAR	13.1	8.8	2.9	8.8
bWAR/WARP	11.2	8.5	2.0	8.5
openWAR/WARP	10.9	9.6	2.9	6.8

In contrast, the fact that those for fielding are as high as those for hitting in this analysis is inconsistent with the Table 3 results, where the SDs across players are far higher for batting than for fielding. The juxtaposition between the Table 3 and 4 results for batting versus fielding suggest that the latter may have a relatively large impact on the differences among WAR measures.

Looking at fWAR vs. bWAR – which, I believe, are the two most frequently compared in the literature and in discussions – we see the differences in fielding are higher than the differences in WAR overall! So in one sense, differences in fielding evaluation comprise much more than just half the difference in WAR. It might not be far wrong to say, at least for comparing fWAR vs. bWAR, that fielding is almost all of it![11]

The batting component

The following table lists basic indices for the batting component. Note that the means and SDs are in runs, rather than wins; dividing by 10 will make them roughly comparable to the previous table. This time, I've put the correlations in the same table as the mean and SD.

Table 6 – Batting components

			correlation with:		
	mean (runs)	SD (runs)	fWAR	bWAR	openWAR
fWAR	6.1	17.0			
bWAR	6.0	16.0	.968		
openWAR	5.1	17.1	.920	.859	
WARP	23.8	18.3	.864	.887	.855

Q – Do the different WARs differ systematically in their hitting figures for given players?

A – WARP calibrates its hitting component far higher than do the other three. That's because their figure is based on a comparison to replacement level, whereas the others are relative to average. It looks like you need to subtract 17 or 18 runs from WARP figures to put the four on analogous scales. (That's consistent with other research comparing replacement value with average, and with B-R's estimate of 20.5 runs per 600 PA.)

Keep in mind, however, that overall WARP figures are not far out of line with the others, because the Baseball Prospectus people adjust their hitting component to compensate for this difference in their final WARPs. (More accurately, the other three measures adjust their runs-above-average numbers to runs-above-replacement numbers, which is what WAR requires.)

Also note that the standard deviations are fairly close to one another, which is likely part of the reason that the relative weightings for hitting are fairly close among the three.

[11] Normally, you'd expect the SD of total WAR to be substantially higher than the SDs of the components—specifically, close to the square root of the sum of squares of the three. It's not as high as that, especially in the bWAR/fWAR case. That could be because fWAR and bWAR are not strictly the sums of the components; the sums are adjusted to bring the league total to 1000 WAR exactly.

Q – How highly do the hitting components of the different WARs correlate with one another?

A – The high correlations indicate that the four WARs do evaluate players similarly to each other. Again, bWAR and fWAR were the most similar and openWAR the most dissimilar. Given that bWAR and fWAR are both based on Tom Tango's wRAA, I would have been amazed if the two correlations were very different. I am surprised that they do correlate a bit differently with WARP.

The fielding component

Table 7 gives the basic indices for fielding, again in runs:

Table 7 – Fielding components

	mean (runs)	SD (runs)	correlation with: fWAR	bWAR	openWAR
fWAR	-1.1	8.5			
bWAR	0.6	9.6	.523		
openWAR	0.2	3.2	.290	.407	
WARP	0.4	7.2	.350	.516	.316

Q – Do the different WARs differ systematically in their zero level for fielding?

A – fWAR's fielding component is noticeably lower than the others, which are quite similar to one another.

Q – Do the different WARs differ systematically in their spread of scores among players?

A – For some reason that cries out for an explanation, the openWAR scores do not have nearly as much variance among fielders as the other three.

Q – How highly do the fielding components of the different WARs correlate with one another?

A – Steve Slowinski (2012) claimed that differences among WAR measures are primarily due to distinctions in fielding measurement; in the case of the 2018 season (at least) he may be right, as fielding correlations were easily the lowest across measurements for the three determining factors. It follows that the different WARs rank-order players' fielding performance very differently from one another, with bWAR the closest to the others and openWAR the greatest deviant overall.

The baserunning component

And, finally, baserunning:

Table 8 – Baserunning components

	mean (runs)	SD (runs)	correlation with: fWAR	bWAR	openWAR
fWAR	0.15	3.5			
bWAR	0.16	2.2	.754		
openWAR	0.18	3.2	.522	.474	
WARP	0.05	2.7	.723	.672	.522

Q – Do the different WARs differ systematically in their baserunning figures for different players?

A – Differences are minimal.

Q – How highly do the baserunning components of the different WARs correlate with one another?

A – The figures here are intermediate in size between hitting and fielding, indicating some divergence in relative standing among players. One more time, openWAR is the outlier.

6. Back to Bryce Harper

So now we can examine the outlier who originally interested me in this project; Bryce Harper:

Table 9 – Bryce Harper's 2018 WARs

	fWAR		bWAR		openWAR		WARP	
	runs	z	runs	z	runs	z	runs	z
batting	+25	+1.2	+29.9	+1.4	+40.3	+2.1	+52.2	+1.6
+ baserunning	− 1	−0.5	+ 0.4	+0.1	− 2.7	−0.9	− 3.2	−1.2
+ fielding	−24	−2.6	−18.1	−2.0	− 0.4	−0.2	−12.2	−1.8
=								
total runs	0		+12.2		+37.2		+36.8	
+ replacement	+20		+20.0		+20.0		(included)	
=								
runs above replacement	+20		+32.2		+57.2		+36.8	
÷10 =								
wins above replacement	+2.0		+3.2		+5.7		+3.7	
+ adjustments (inferred)	−0.5		+0.3		+1.4		−0.3	
=								
Official WAR	+1.5		+3.5		+7.1		+3.4	

Adding the three main components, plus 20 runs to adjust to replacement level batting (except WARP, where the adjustment is already there), gets us close to the "official" WAR values as published. Any remaining differences are from subsequent adjustments—park, position, league, normalization, etc., which we did not attempt to reproduce. Those are listed in the table as "inferred," as they are necessary to get the totals to add up to the published values.

• openWAR gave Bryce the highest WAR figure by far (7.1), sixth best among position players. As hitting is such a predominant component of openWAR, this is mostly due to the fact that openWAR assigns Bryce the highest hitting estimate relative to the other three WARs; the resulting z-score is ninth best among position players. The fielding component is analogously the least negative of the four. Baserunning brought Harper down a peg, but at most it just counterbalanced the fielding ratings. Also keep in mind that openWAR is calibrated more than half a win higher than the other three.

• WARP's figures for hitting and fielding were far worse than those for openWAR, resulting in a lower overall WAR (3.4).

• fWAR's hitting and fielding figures were worse yet. Nonetheless, Harper's fWAR ended up almost the same (3.5) as his WARP, perhaps because fWAR is calibrated a bit higher (average fWAR for the dataset was 2.34, versus WARP at 2.27) and weighs batting a bit more than fielding in determining the total.

• bWAR had the worst batting and fielding figures of all, and although the baserunning figure was not as bad as WARP's and openWAR's, it hardly made up for the other differences. As a consequence, Harper's overall bWAR (1.5) was the lowest in the group.

It still seems to me that the differences among the four are unreasonably extreme. Even so, looking only at fWAR and bWAR, Harper's discrepancy (2.0) between the two was not the largest in the data set; Nick Ahmed came in at 2.6, Willson Contreras scored 2.2, and Charlie Blackmon tied Harper at 2.0.

In all four of these cases, large differences in fielding ratings were responsible for the variation, consistent with our expectations from our earlier discussion surrounding Table 5.

7. Pitchers

7.1 fWAR (Fangraphs)

Steve Slowinski (2012c) has supplied a basic primer to fWAR for pitchers, with links to a relevant technical discussion by Dave Cameron; also see Weinberg, 2017a for an example. The computation process begins with a revised version of FIP (FIPR) including infield popups treated as if they were strikeouts in both the pitcher-specific and league-general components.[12] This is then scaled to runs per nine innings rather than ERA by adding the average unearned runs per game, the sum referred to as "runs above average per nine innings." This figure is then adjusted for ballpark and subtracted from league RAA.

Next, to convert runs to wins, instead of using a fixed ratio of 10 runs, or even a formula for the league run environment, FanGraphs calculates a unique Runs Per Win value (Slowinski called this "Dynamic Runs Per Win") for the specific pitcher. This is due to the fact that if a pitcher gives up many fewer runs than average, each run is that much more valuable in creating a win.

Now it is time to consider replacement level. FanGraphs defines the difference between average and replacement value by the following formula:

$$replacement\,value = average - \left[0.03\left(1 - \frac{GS}{G} \right) + 0.12\frac{GS}{G} \right]$$

which implies that the difference between replacement player and league average is .12 WAR per nine innings for pitchers who only start, .03 WAR per nine innings for those who only relieve, and something in between for those doing some of both. That's likely consistent with Tom Tango's (2007) estimated replacement levels for starters (.380) and for relievers (.470).[13]

Adding this replacement value to the previously computed wins per inning above average gives you a wins per game above replacement, which now needs to be multiplied by the pitcher's average number of innings per game. Relievers are additionally adjusted for leverage by

[12] FIP is the abbreviation for "Fielding Independent Pitching," an often-used pitching evaluation measure inspired by Voros McCracken's research, proposed by Tom Tango, and well-described in Woolner and Perry, 2006. It is the most prominent evaluation metric relying solely on strikeouts, walks plus hit by pitches, and homers, aka "the three true outcomes," under the assumption that, unlike batted balls in play, they are unaffected by the quality of the relevant team's fielding.

[13] The .12 figure is also in line with the equivalent definition for batters. In 2018, there were around 38.3 plate appearances per nine innings. That means that .12 wins per game equals 1.9 wins per 600 PA, very close to the standard 2.0 figure.

$$1 + \frac{their\ average\ leverage\ index}{2}$$

with the formula regressing the index halfway toward average to account for reliever "chaining" (to be discussed later). The resulting figure is like WAR, but needs one final adjustment to calibrate it so that the final sum for all pitchers equals the 430 (43 percent of 1000) FanGraphs allots to pitchers. The adjustment figure is unique for each season and multiplied by innings pitched before being added in.

FanGraphs does not appear to include either pitcher batting or fielding into their WAR figure.

7.2 bWAR (Baseball Reference)

The main way in which bWAR differs from fWAR is that it evaluates pitchers by actual runs allowed. fWAR, on the other hand, evaluates pitchers by their three true outcomes. That means bWAR winds up considering what happens to balls in play (a single is more damaging than a ground ball), and also timing of events (a double with runners on is more damaging than one with the bases empty).[14]

The details:

The Baseball Reference version of pitcher WAR[15] begins with two simple measures; runs allowed and innings pitched. It is compared with what they call xRA, an estimate of opposition offense computed by weighting the average number of runs scored by each opposing team by the number of innings the relevant pitcher has versus each of those teams. It then takes on the following series of adjustments:

1 – Team defense measured by Defensive Runs Saved as follows:

(Pitcher's batted balls in play) / (team batted balls in play) * (team DRS)

Phil Birnbaum (2016), among others, has criticized this adjustment as it incorrectly assumes that team defensive performance is evenly distributed among pitchers, instead of assuming that pitchers with fewer runs allowed likely got more plays from their fielders. This counts as a weakness, one not shared by FanGraphs as it bases its method on FIP.

2 – Starter/reliever RA differences, which they consider to be nonexistent before 1960, worth 0.0583 runs per game from 1960 until 1973, and 0.1125 runs per game afterward.

3 – Ballpark, individualized to include only those the pitcher appeared in, which on occasion can differ considerably from their team's as a whole.

B-R claims that pitcher fielding is included in the computation of pitcher WAR, but I cannot find any reference to that at this website, only to team fielding as a whole. Pitchers' contributions as hitters and baserunners are included. Details on both these issues are provided at the B-R website.[16]

To compute replacement level on pitching itself, B-R takes the league average runs allowed per out, and adds 18.7 percent. Why 18.7 percent? In their words, it's "an empirical factor that makes the final result mostly closely align the sum of all player replacement runs to the desired league total." That makes 1 WAR work out to about 13.2 runs per 600 PA, as compared to approximately 20 runs per 600 PA for batters.

The computation goes as follows:

---

[14] The FanGraphs site allows you to see how much fWAR would have to be adjusted to make it more like bWAR – although they don't explicitly state it as such, and the correspondence is not perfect. Navigate to a pitcher page (for instance, Clayton Kershaw: https://www.fangraphs.com/players/clayton-kershaw/2036/stats?position=P). The bottom table, "value," allows you to account for the results of balls in play (column "BIP-wins") and the effects of event timing (column "LOB-wins"). Add those to total fWAR to get something closer to bWAR. (Thanks to Guy Molyneux for alerting me to this feature.)

[15] Baseball-Reference describes their method in full at https://www.baseball-reference.com/about/war_explained_pitch.shtml.

[16] See https://www.baseball-reference.com/about/war_explained_position.shtml

Step 1 – Multiply the league runs per out (R/O) by 18.7 percent. That's the difference between average and replacement.
Step 2 – Add that to the pitcher's R/O above average.
Step 3 – Multiply this result by the pitcher's number of outs recorded.

Runs are converted to wins based on PythagenPat[17], as for position players above. There are further adjustments for reliever leverage index and to make sure total WAR for the season sums to 1000. Baseball Reference also does not appear to include pitcher batting or fielding in their WAR figure.

7.3 WARP (Baseball Prospectus)

Baseball Prospectus's WARP for pitchers is generally based on their current pitching effectiveness index,[18] plus the pitcher's position player WARP (this latter of which includes their batting and fielding performance). BP assumes replacement level is the average performance of starters outside the five most used starters per team. Based on 1978 through 2001, Keith Woolner (2002) computed the following regression equation representing the replacement starters' Run Average (RA):

starter replacement RA/9 = (1.37 * league-average starter RA) – 0.66

Keith also determined that between 1904 and 2001, a team's top five starting pitchers consistently began between 78 and 84 percent of its games, even during the years with supposedly four-man or even three-man rotations, with the exception of the 1981 and 1994 strike years (when it was about 85 percent). Thus assuming that the top 80 percent of pitchers by usage are "regulars" and the bottom 20 percent are "substitutes," Keith separated the most active relievers who, as a set, accumulated 80 percent of the relief innings from the rest, and computed the following regression equation for calculating an annual replacement level:

reliever replacement RA/9 = (1.70 * league-average reliever RA) - 2.27

I do not know whether Baseball Prospectus still uses these equations. I am unsure whether pitcher batting or fielding are included, although my guess is that they are not.

7.4 openWAR

In openWAR, pitchers are given total credit/blame for the run value of the Three True Outcomes (K, BB, HR) and analogous events (HBP). They also get a degree of responsibility for balls in play (the calculation of how balls in play are split between fielders and pitchers was described earlier in this paper). Run values are then adjusted for ballpark and standard platoon advantage.

As with hitters, the authors computed openWARs for samples of plate appearances for pitchers, and noted that as a set pitchers were more consistent than hitters. openWAR is the only one of the four that I know includes pitcher batting and fielding RAA in the final figure.

[17] see https://www.baseball-reference.com/about/war_explained_runs_to_wins.shtml

[18] Their current index as I write this is "Deserved Run Average," which evaluates the pitcher in changes in run expectancy for each batter. In effect, it's the equivalent of RA/9, but adjusted for context (park, opposing batter, handedness, etc.), with runs split proportionately between pitchers based on probability. For instance, consider a reliever who comes in with no outs and the bases loaded, and gives up one run in completing the inning. That run will be mostly charged to the pitcher who put him on base, while the reliever will get substantial credit for not allowing the other two runners to score. See https://www.baseballprospectus.com/news/article/26195/prospectus-feature-introducing-deserved-run-average-draand-all-its-friends/

8. Summary of differences—pitcher WAR

	who uses it	pitching stat	based on
fWAR	Fangraphs	FIPR	three true outcomes
bWAR	Baseball-Reference	wRAA	runs allowed
WARP	Baseball Prospectus	VORP	run expectation changes for each PA (effectively runs allowed, but with different handling for inherited runners)
openWAR	academic paper by Baumer et al.	their calculation	three true outcomes, but pitcher given some credit for BIP

	affected by BABIP/timing?	pitcher batting, fielding included?	starter/reliever adjustment?
fWAR	no	no	yes
bWAR	yes	no	yes
WARP	yes	no	yes
openWAR	yes	yes	no

9. Pitching WARs compared empirically

The results here used the same 2018 database as was used for position players. The sample size was the 141 pitchers with more than 100 innings pitched that season. See Table 10 for the means, SDs, and correlations:

Table 10 – Pitcher WAR (2018, minimum 100 IP)

	mean	SD	correlation with: fWAR	bWAR	openWAR
fWAR	2.2	1.7			
bWAR	2.1	2.1	.843		
openWAR	2.1	1.8	.903	.880	
WARP	1.9	2.2	.736	.856	.814

Q – Do the different WARs differ systematically in their means?

A – Means for fWAR, bWAR, and openWAR are all extremely close to one another, so much so that I did not bother to perform t-test comparisons. Those for WARP are a bit lower, but not significantly so.

Q – Do the different WARs differ systematically in their spread of scores among players?

A – The standard deviations for fWAR and openWAR are somewhat smaller than those for bWAR and WARP, indicating some difference among the four in the spread of scores across pitchers. One more time, limiting the sample to 100+ IP results in an underestimate of the overall variation.

Q – How highly do the different WARs correlate with one another; in other words, independently of differences in calibration as reflected in means, are players rank-ordered approximately the same across them?

A – fWAR and bWAR again correlate highly, and, in contrast with position players, openWAR correlates well with both. The relative outlier is WARP, although it correlates very well with bWAR.

Q - Even if overall means are the same, are there noticeable differences in the WAR figures assigned to individual players across the four WARs?

A – Table 11 shows that the average (absolute) difference between WARP and both openWAR and bWAR is higher than 1.0, which means that the typical player WARs will in these cases be 1 win apart. There were a few pitchers who were 2 wins higher on two of the methods than the other two. One of those was Jacob DeGrom, but as an outlier (his WARP and openWAR were was about 8, his bWAR and fWAR about 9.5), his is a bad example.

Table 11 – Mean absolute differences between pairs of WARs

	fWAR	bWAR	openWAR
bWAR	.851		
openWAR	.716	.709	
WARP	.942	1.231	1.054

Lance Lynn is a much better case to study; the highest two WARPs are more than two full wins higher than the other two.

Table 12 – Lance Lynn, 2018

	WAR
bWAR	0.4
fWAR	2.8
openWAR	2.1
WARP	0.3

fWAR is based on the relatively context-free FIP with a ballpark adjustment, implying that its figure represents what Lynn "deserved" given his actual pitching performance rather than luck-influenced outcomes. In using run expectancy values rather than actual outcomes with ballpark and handedness adjustments, openWAR analogously tries to measure what is "deserved." In contrast, bWAR uses runs allowed, and so is heavily luck-influenced, while WARP adjusts for everything under the sun. The implication is that Lynn's pitching quality was actually at or above average but bad luck and/or context gave him worse outcomes than his performance "deserved."

10. Empirical estimates of WARP and openWAR replacement levels[19]

bWAR, fWAR

We know the explicit replacement levels for bWAR and fWAR, because they are intentionally calibrated as totaling 1000 WAR across 2430 games, so that a team of players at replacement level would be predicted to end a season with a .294 winning percentage, and a replacement-level offense paired with an average pitching staff would perform at about a .380 level.

However, there is no such explicit replacement level for WARP and openWAR, given that their conceptual definitions are not clear. So, we need to estimate them from available data. I used 1970-2016 data from Baseball Reference, with leagues distinguished from one another.

WARP

WARP replacement level, according to the Baseball Prospectus folks, is equivalent to the average performance of backup players, those beyond the 8 (NL) or 9 (AL) regulars at each position.[20] For each season, mean performance levels were computed for each season for all position players beyond those 8/9.

For WARP, composite replacement-player batting was as follows:

	AB	H	2B	3B	HR	BB	K	avg	obp	slg	RC/G
WARP	538	129	23	3	11	48	105	.240	.295	.355	3.60

A team of such players would produce 3.60 runs created per game. More importantly for our analysis, we can calculate an implied replacement level for this batting line, using Pythagoras. An American League team of such players would be predicted to play at a .368 clip, while National League replacement players would perform at .382.[21]

Both figures are close to the .380 for a replacement level offense paired with an average pitching staff as implied by the .294 bWAR/fWAR/Tango all-replacement-level-team consensus, but are inconsistent with what we would expect given both the difference in conceptual definition (all substitutes versus freely available talent).

Incidentally, the SD of implied winning percentage over the 47 seasons in the study was .016 for the AL and .026 for the NL. The overall difference between the leagues (.368 vs. .382) is significant by two-tailed equal-variance-assumed t-test at a probability level of .005. Diagramming the year-by-year figures revealed no trends for changes over time.

openWAR

For openWAR, Baumer et al. consider the main bench players to be above replacement level, so their definition starts after 13 players, rather than 8/9.[22] So openWAR implies a lower replacement level, as can be seen in their composite batting line:

	AB	H	2B	3B	HR	BB	K	avg	obp	slg	RC/G
openWAR	540	121	22	3	9	46	115	.224	.278	.326	3.05

That translates to 3.05 runs created per game, and Pythagorean predictions of .301 for the AL and .299 for the NL.

[19] Data analysis for this section was provided by Phil Birnbaum.

[20] We will refer to limits in this section as "8 per team" or "9 per team," even though the limit is computed per league rather than per team (so that some teams can have 10+ "regulars" and some might have as few as 7).

[21] For DH seasons, the National League winning percentages are computed (via Pythagoras) using runs allowed from that year's American League. This allows NL and AL replacement players to be compared to each other on the same scale. For 1970-1972, both leagues' runs allowed were multiplied by 9/8 in order to allow direct comparisons to 1973 and later.

[22] As described in the earlier footnote, players are pooled within league and not by team. For openWAR, however, all positions go into the same pool, so in a 12-team league, the pool of 156 above-replacement hitters might, for instance, include more than 12 shortstops or fewer than 12 shortstops.

The overall mean of .300 is substantially lower than the .380 one would expect from a replacement-level-offense/average pitching staff team. The lower figure makes sense for a group of players beyond both starters and what I will call "first line backups", i.e. a second catcher, fourth outfielder, and fifth infielder, and explains why the average openWAR values were so much higher than for bWAR and fWAR.

For openWAR, the SDs of seasonal offensive winning percentage were .027 for the AL and .031 for the NL. Again, there were no discernible changes over time in these figures.

First-line backups

The difference between WARP and openWAR is the first-line backups, the 9th/10th through 13th players, who are included in the replacement-level pool in WARP, but are excluded from openWAR. Here's their composite batting line:

	AB	H	2B	3B	HR	BB	K	avg	obp	slg	RC/G
1st line bench	537	133	24	3	12	50	100	.248	.305	.371	3.92

This team computes to a Pythagorean projection of .408 (SD = .022) in the AL and .420 (SD = .031) in the NL. The league difference is significantly different at a probability level of .025.

In short, the expected performance of first line backups was equivalent to an expansion team, and a little higher than the theoretical .380 for a replacement level offense paired with an average pitching staff.

Adjusting for luck

Here's an interesting possibility that might explain the discrepancy.

In a given year, there is a bias such that players who happen to be having unusually bad seasons due to injury or bad luck end up with lower PA, winding up disproportionately in the replacement pool despite not really being replacement-level players, in the usual sense. If so, then Phil Birnbaum's method (2005) for estimating talent figures, based on observed performance with luck removed, should yield much higher Pythagorean estimates.

For openWAR, Phil's method gives .410 for the A.L. and .405 for the N.L., which on the face of it supports openWAR methodology. For WARP, the same analysis done gives even higher numbers; .429 for both leagues. That it's higher than openWAR is, again, what would be expected considering WARP includes first line backups while openWAR does not.

The first line backups projected even higher, at .446 for both leagues, which is again what would be expected given that they exclude the Quad-A players included in openWAR and WARP.

In addition, with the luck component removed, openWAR's difference from WARP and first line backups—and probably fWAR and bWAR—shrinks.

11. Conclusion

I started this project out of curiosity about the stark differences in Bryce Harper's WAR values with different versions. Although I still do not understand that completely, we have learned a lot about how each version ticks.

Beginning with position players, bWAR and fWAR are pretty close to interchangeable, with the exception of occasional cases in which they are based on radically divergent fielding numbers. WARP is calibrated abut the same but will rank-order players somewhat differently.

The most important differences among those three WARs are due to discrepancies in fielding evaluations. Interestingly, UZR (as used by fWAR) and DRS (as used by bWAR) are based on the same raw data, from Baseball Info Solutions, but use different algorithms to calculate defensive value. On the other hand, FRAA (as used by Baseball Prospectus) is based on conventional indices (putouts, assists, double plays)

massaged for ballpark, balls in play, and pitcher handedness and fly ball/ground ball tendencies. Thus, it should be no surprise that the three intercorrelate only in the .4 to .5 range.

Mike Provenzano (2020) estimated that fielding is half of WAR. That's perhaps a bit of an overestimate of how much fielding contributes to a particular implementation of WAR, but a significant underestimate of how much fielding contributes to *differences* between WARs. As we have seen, the relative size of standard deviations across players within each measure implies that it is between a quarter and a third of total WAR. But the relative size of the SDs of the differences between WARs, rather than the values themselves, suggests it's anywhere from half the difference to almost all of the difference (in the case of fWAR/bWAR).

Having said this, and despite the differences, I trust bWAR and fWAR's fielding components over WARP and openWAR. openWAR's has a drastically lower standard deviation, which lowers my confidence in its figures, and my subjective impression is that those writing commentaries for individual players in the Baseball Prospectus annuals appear to rely on scouting judgments over their FRAA figures more often than not.

As for openWAR on the whole, it is the most distinct of the four versions, and as it is based on run expectancy data (and therefore heavily influenced by base-out situations), it really is a different animal.

In any case, although I would not expect as much for most baseball broadcasts, I do wish that on specialized programs, such as what Brian Kenny does on the MLB Network, when WAR is brought up, it was made clear which version they were using.

Turning to pitchers: Although the four WARs are calibrated closely to one another and intercorrelate fairly highly despite their differences in basic data, again they result in noticeable discrepancies in some cases, as described in the example of Lance Lynn. In this case I strongly prefer fWAR, as its basis (FIP) will do a better job of predicting next year's performance more accurately than bWAR's use of runs allowed, which as I mentioned earlier I believe to be the major reason for calculating WARs in the first place. It also escapes openWAR's problem with fielding evaluation that I described earlier.

In principle, I should prefer WARP even more because Baseball Prospectus is using DRA as its main pitching stat—DRA is founded on individual plate appearances adjusted for myriad contextual factors, and so should be a better predictor still. However, it's also the case that DRA is based on run expectancies and, as I argued earlier, I believe that WAR is of most value when it is context-free. So I find WARP less attractive.

This study was also instructive concerning the empirical definition of "replacement level." Since Bill James's work, it has generally been defined conceptually as "freely available talent," and set at just below .300; .294 for both bWAR and fWAR and .292 by Tom Tango. These translate to about .380 for a team with a replacement level offense paired with an average pitching staff.

openWAR operationalizes replacement level as the average for all position players beyond the 13 a team intends for its roster. An examination of the performance of those players (in terms of plate appearances) results in a Pythagorean prediction of .300 across leagues, which is much lower than the .380 implied by the bWAR/fWAR/Tango consensus. The implication is that what is generally viewed as replacement level performs better than the average roster substitute for an injured "first line backup" (second catcher, fourth outfielder, or fifth infielder) or, equivalently, a replacement for a first line backup when that backup is pressed into regular service due to a starter's injury. This is pretty much what Bill James originally meant by "freely available talent." In contrast, WARP's empirical definition, which is the predicted performance of all backups, resulted not surprisingly in a far higher Pythagorean prediction of .376 across leagues, which is close to the .380 but still lower than the .400 that expansion teams have averaged over the years. An additional analysis for just the first line backups yielded slightly better than expansion team (.413) performance. Finally, the National League figures for WARP and the first line backups, but not for openWAR, were noticeably higher than those for the American League. This may be due to the fact that the most offensively productive backup in the A.L. has in a sense been lost to the designated hitter role whereas that for the N.L. remains on the bench. This fact is irrelevant to openWAR as it only includes Quad-A talent.

Given Tom Tango's defense of the just-below-.300 figure and the fact that it has been shown to distinguish 6000-PA veterans (almost all above it) from waiver transfer/minor league contract players (averaging it), all of this seems to mean that the standard replacement level figure represents the performance of all position players beyond the intended starting 8 or 9. This is basically WARP's conception of replacement level, and as is, as it should be, higher than openWAR's and so higher than "freely available talent." Given that the average WARP figures were only slightly lower than bWAR's and fWAR's in practical terms, the implication is that their in-practice replacement levels could in actuality be equivalent to WARP's predicted performance for all backups, and as such higher than the original sense of "freely available talent." I would hesitate to claim that as fact without including all position players in the analysis.

Finally, Phil's method for estimating luck-free figures brings openWAR in line with expansion team performance, WARP's a bit higher than that, and first-line backups a little higher still. The openWAR figures are the least affected by this estimate, likely because the impact of context (for example, the base-out situation which batters face) is so luck-dependent in a given season. One should be careful with the

implication here that expansion teams generally have replacement level offenses and average pitching staffs. It is more likely that they have offenses and pitching staffs that are both a bit better than replacement level.

Having said all of this, we need to acknowledge the "chaining" problem alluded to earlier. It was, to the best of my knowledge, first discussed by Brock Hanke (1998). Using a version of Brandon Heipp's (2020) interpretation of the problem while using Hanke's example is instructive. Your first baseman is injured at the very beginning of the season, and you are going to lose 500 of his plate appearances. A team of players at his level would be projected to play .600 ball, so we consider that to be his level of expected performance. As a consequence, you have to add a "Quad-A" player to your roster who is at replacement level. As a team full of players at his level of performance would be expected to play at .300, it looks like you have lost .300 in the exchange. But that is only the case if the replacement level player is going to get all of those 600 PAs. Rather, your intended backup, a .450 batter, will be getting the 500 PAs instead. Now, if the intended starter was in the lineup, the intended backup would get getting 150 PAs both substituting for the intended starter and pinch hitting; these 150 will now go to the replacement level player. With a healthy starters, you would get

(500 PAs at .600) + (150 PAs at .450)

If you statistically replace the injured regular's PAs with those for the Quad-A call-up, you would get

(500 PAs at .300) + (150 PAs at .450)

which would be a loss of .300 for those 500 PAs. In truth, however, your first line backup would get the injured regular's PAs, giving you

(500 PAs at .450) + (150 PAs at .300)

which is a loss of .150 for 650 PAs, quite a bit less of a disaster. In a personal communication, Tom Tango acknowledged this problem, although he claimed that the simpler approach is sufficient for most purposes.

For relievers, if a closer is hurt, then the chain can go through several pitchers before getting to a call-up. Although I would continue to argue for a context-free treatment of WAR, the chaining issue does imply that replacement level and thus WAR will differ among individual cases.

12. References

Baumer, Benjamin S., Shane T. Jensen, and Gregory J. Matthews (2013). openWAR: An open source system for evaluating overall player performance in major league baseball. *Journal of Quantitative Analysis in Sports,* Vol. 11 No. 2, pages 69-84.

Baumer, Benjamin S. and Gregory J. Matthews (2014). There is no avoiding WAR. *Chance,* Vol. 27 No. 3, pages 41-44.

Birnbaum, Phil (2005). Which great teams were just lucky? *Baseball Research Journal,* No. 34, pages 60-68.

Birnbaum, Phil (2016). How should we evaluate Detroit's defense behind Verlander? http://blog.philbirnbaum.com/2016/11/how-should-we-evaluate-detroits-defense.html

Cameron, Dave (2003). Unifying replacement level. https://blogs.fangraphs.com/unifying-replacement-level/

Carleton, Russell A. (2013). Daddy, what's replacement level? https://www.baseballprospectus.com/news/article/19773/bp-unfiltered-daddy-whats-replacement-level/

Carleton, Russell A. (2017). The disappearing left fielder. https://www.baseballprospectus.com/news/article/31686/baseball-therapy-the-disappearing-left-fielder/

Click, James (2011). Evaluating defense. In Lindbergh, Ben (Ed.), *Best of Baseball Prospectus 1996-2011* (pages 206-208). Self published.

Hanke, Brock J. (1998). WAR redeclared. In Don Malcolm, Brock J, Hanke, Ken Adams, and G. Jay Walker (Eds.), *The 1998 Big Bad Baseball Annual* (pages 495-503). Indianapolis: Masters Press.

Heipp, Brandon (2020). Tripod: Baselines. http://walksaber.blogspot.com/2020/06/tripod-baselines.html

Matthews, Gregory J. (2016). A response to "Declaring openWAR". https://www.baseballprospectus.com/news/article/28272/prospectus-feature-a-response-to-declaring-openwar/

Provenzano, Mike (2020). What we can glean from Statcast's new infield defense metric. https://www.beyondtheboxscore.com/2020/1/9/21057644/mlb-statcast-defensive-metrics-outs-above-average-galvis-bogaerts-simmons-arenado-chapman

Slowinski, Steve (2012). fWAR, rWAR, and WARP. https://library.fangraphs.com/war/differences-fwar-rwar/

Slowinski, Steve (2012a). WAR for position players. https://library.fangraphs.com/war/war-position-players/

Slowinski, Steve (2012b). Catcher defense. https://library.fangraphs.com/defense/catcher-defense/

Slowinski, Steve (2012c). WAR for pitchers. https://library.fangraphs.com/war/calculating-war-pitchers/

Tango, Tom (2007). The replacement pitchers. http://www.insidethebook.com/ee/index.php/site/comments/the_replacement_pitchers/

Turkenkopf, Dan, Harry Pavlidis and Jonathan Judge (2015). Introducing Deserved Run Average (DRA) and all its friends. https://www.baseballprospectus.com/news/article/26195/prospectus-feature-introducing-deserved-run-average-draand-all-its-friends/

Weinberg, Neil (2014). Calculating position player WAR, a complete example. https://library.fangraphs.com/calculating-position-player-war-a-complete-example/

Weinberg, Neil (2017). Calculating pitcher WAR, a complete example. https://library.fangraphs.com/calculating-pitcher-war-a-complete-example/

Wenz, Michael (2016). Declaring openWAR. https://www.baseballprospectus.com/news/article/28142/caught-looking-declaring-openwar/

Woolner, Keith (2002). Understanding and measuring replacement level. In Joe Sheehan (Ed.), *Baseball Prospectus 2002* (pages 455-466). Washington, DC: Brasseys.

Woolner, Keith and Dayn Perry (2006). Why are pitchers so unpredictable? In Jonah Keri (Ed.), *Baseball Between the Numbers* (pages 48-57). New York, NY: Basic Books.

Charlie Pavitt, chazzq@udel.edu ♦