By the Numbers

Volume 30, Number 2 The Newsletter of the SABR Statistical Analysis Committee November, 2021

Review

Academic Research: Studies from 2021

Charlie Pavitt

Charlie reviews more recent studies from the academic literature.

Paulsen, Richard J. (2021), New evidence in the study of shirking in major league baseball, Journal of Sport Management, Vol. 35 No. 4, pp. 285-294.

This is one of the better of the many studies published over the past forty years asking whether player performance is better or worse in their final year before free agency. Early work was plagued by repeated problems: among others, the failure to consider characteristic performance trajectories that tend to increase relatively quickly, peak at 26 or 27, and then ease down afterward; the failure to compare players with expiring contracts to those without them to see if they differ systematically; and, as pointed out by our editor Phil in the past, the fact that careers often end after years in which the player has randomly underperformed compared to expectation, which results in a bias toward unfairly pessimistic predictions for the final years of good but aging players.

Paulsen's work, previously presented at the 2019 MIT Sloan Sports Analytics Conference, deals with these issues better than have most other efforts. It was based on 535 non-pitchers from 2010-2017, comprising 1068 contracts of various lengths and 1829 contract-seasons. Predictors included age, age squared (to provide for the normal curvilinear career trajectory), and player as a fixed effect.

Paulsen uncovered evidence for a bit of shirking, with an average reduction of 0.08 offensive bWAR for each additional remaining year under contract.

As with other studies, performance was found to be highest in the final year, as compared to previous seasons on multi-year contracts. Results were similar when restricted to free-agent eligible players. Whether or not players changed teams had little effect. There was no evidence of shirking for pitchers and for non-pitcher fielding bWAR.

Jane, Wen-Jhan (in press), Choking or excelling under pressure: Evidence of the causal effect of audience size on performance, Bulletin of Economic Research.

Using 2015–2018 Retrosheet performance data and attendance figures from mlb.com, along with various control variables, Wen-Jhan Jane examined the influence of the latter on the former. Overall, using a metric that I believe is hits divided by plate appearances, Jane found the average performance for both home and away teams to describe an inverted-U function across five attendance categories (less than 10K, 10K–20K, 20K–30K, 30K–40 K, and more than 40K). Home team players peaked in the 30K–40K range whereas away team players did so between 20K–30K. Although present in every inning, the effect for the away team was strongest in the 9th and later innings, with the peak now between 10K–20K. However, there was evidence that "star" players, defined as those who had been All-Stars the previous season, actually improved as attendance rose. Jane's study also revealed more support for home field advantage by means of higher figures on the H/PA metric.

In this issue

The previous issue of this publication was March, 2021 (Volume 30, Number 1).

Black, Dirk E., and Marshall D. Vance (2021), Do first impressions last? The impact of initial assessments and subsequent performance on promotion decisions, Management Science, Vol. 67 No. 7, pp. 4556-4576.

This is an examination of the relationship between draft position and promotion across levels for minor league pitchers between 1987 and 2013. The study controlled for the difference between current season and league average FIP for both the current and previous seasons, along with such factors as age, college, years of past experience, and whether repeating a level.

Current season performance was the strongest predictor of next season promotion, with draft position coming in second strongest, and prior season performance also a predictor. Not surprisingly, given the relatively small amount of available data, initial draft position was a stronger predictor than performance after the first season, but became less so annually up to year four, after which promotion became less predictable overall. Prediction of next-season promotion using both draft position and current year performance improved from Rookie classification up to AAA, but was less accurate from AAA to majors. Results were pretty much the same for starters as they were for relievers.

Fesselmeyer, Eric (in press), The impact of temperature on labor quality: Umpire accuracy in Major League Baseball, Southern Economic Journal.

PITCHf/x and its successors have uncovered good evidence for bias in umpire pitch calling in several areas, most notably what I call the "count compensation bias" in which umps provide a bit of help to whichever participant (pitcher or batter) is disfavored in the count. Studies have also found bias in different called strike zones for right-handed versus left-handed batters, favoritism to veteran and star-level players, and home field advantage that increases with larger crowds (which is one of the most supported explanations for the advantage).

There have also been studies with either slight or contradictory relevant evidence, including of all things air pollution; this one, based on temperature, is close to that one in spirit.

Fesselmeyer found umpires' best call accuracy came with a heat index in the 80s, at 86.8 percent. As with other pitch calling biases, the impact was small; at a heat index of less than 70 degrees, the figure was 86.6, and at more than 110 degrees it was 86.2 percent.

Charlie Pavitt, chazzq@udel.edu ♦

Revisiting cWAR: Analyzing the Cape Cod League as a Path to the Majors

Humbert Kilanowski

The author updates a previous analysis of performance in the summer Cape Cod League, with several additional seasons covered. Among other advantages, the additonal data allows a comparison of how well various sabermetric statistics correlate to MLB draft position.

One of the many paths that baseball players take to the Major Leagues is the collegiate route. Because conferences and schedules vary so much, comparing players across the nation becomes highly difficult. Summer leagues, however, provide an environment in which college players can perform at a more standardized level of competition, against players of similar skill.

One of the most prestigious summer leagues is the Cape Cod Baseball League (CCBL), based in the coastal resort area of Massachusetts. The league attracts some of the strongest collegiate players, many of who play under the observation of MLB scouts. In a previous study,[1] we developed a version of Wins Above Replacement, called cWAR, that measures a player's total contributions in this league, relative to the average temporary player as replacement level, in order to rank each player for the Major League draft.

That study examined only the 2019 season; however, we have since been able to collect data that allows us to analyze several additional seasons (2012-2018). This enables us to compare the metric across time and identify trends, such as which seasons were higher- or lower-scoring. It also provides a larger data pool that not only makes the results more reliable, but allows us to look at other issues, such as clutch hitting and draft position.

A Retrospective Study

First, we note that offense has varied considerably in the eight years of our data. In 2012, the league changed the composition of the baseball, and offensive numbers increased -- so much so that the ball had to be changed back in the following year. Thus, followers of the league refer to 2012 as the "juiced-ball" season.[2]

Our analysis confirms this; the first column of Table 1 shows 2012 provided the most runs, although the two most recent seasons (2018-19) approached the values from that year, mirroring increased offense in the Major Leagues. The lowest-scoring season in the sample, on the other hand, was 2015, a year in which several current Major League sluggers played: Bobby Dalbec, Pete Alonso, and Nick Senzel.

Table 1 – Variation in Cape Cod League offense, by year

	Runs per inning	Run value of HR
2012	.567	2.107
2013	.469	2.193
2014	.535	2.040
2015	.440	2.176
2016	.466	2.175
2017	.517	2.159
2018	.540	2.016
2019	.540	2.085

Normalizing Run Expectancy

Run expectancy tables show the average number of runs expected to score in the remainder of an inning, and are calculated through observation from play-by-play data. In our initial model, we had to modify some of those values from those initially calculated empirically, because some of the rarer base-out states had average run expectancies that were out of order or otherwise deviated from the true average. The most glaring example was that the bases-loaded, no-out state had a lower value than runners on second and third with no outs, suggesting that a walk in this case actually has negative value. That

[1] Humbert Kilanowski, "cWAR: Modifying Wins Above Replacement with the Cape Cod Baseball League," *Baseball Research Journal* 49.1 (2020), 99-105.

[2] Chris Thoms, personal correspondence. I am deeply grateful for Chris' work in compiling the play-by-play data in a format that I have been able to analyze using R.

was because of a few “big innings” in which the team at bat scored eight or nine runs after reaching "second-and-third-no-outs" ended up skewing the average to be too high.

To remedy this problem in the previous study, we calculated a buffered average, inserting 50 instances of each state from the 2019 MLB run expectancy matrix. (The figure 50 was chosen since it was close to the frequency of the rarest state from the Cape League data.) Since the Major Leagues had a higher rate of run scoring than the Cape League, we had to multiply each MLB value by a scaling constant, so that the base state (bases empty, no outs) remains unchanged.

Now that we have more seasons from the Cape, however, no scaling constant is needed, as we can instead insert 50 instances of the average state from eight years of Cape League data. The result is that the Cape League contains a self-correcting mechanism that can be applied to each year, without resorting to importing data from another league or level.

The updated run expectancies are shown in Table 2.

The updated values do change the 2019 run expectancy values from the previous study, and also bring about a small change to the WAR values calculated for that year. The 2019 league leader in WAR, Nick Gonzales, ends up with a slightly lower wOBA (.492) and WAR (3.02) than previously calculated -- still, he he remains the leader in single-season WAR for the entire 2012-2019 period. A list of leaders in batting WAR for the period appears in Table 3.

Table 2 – Run Expectancy, Cape Cod League, 2012-2019

	0 outs	1 out	2 outs
Bases Empty	.509	.255	.089
1st only	.923	.533	.214
2nd only	1.195	.689	.310
3rd only	1.461	1.030	.390
1st and 2nd	1.559	.970	.418
1st and 3rd	1.735	1.240	.529
2nd and 3rd	2.038	1.411	.597
Bases Loaded	2.269	1.646	.764

Does Clutch Performance Matter?

In determining which batter is most valuable to his team, one question that often arises is how the player performs in the clutch: with the game on the line, is this the hitter that you want at the plate more than anyone else? The algorithm for calculating WAR does not measure this ability, since it is based on wOBA, a statistic designed to be context-neutral, as every event of the same type is given the same value: a single to lead off an inning in a blowout game, for example, counts the same as a walk-off single with the bases loaded and two outs, down by one run.

We therefore want to account for clutch performance; if our metric is to determine a player's value, the highest-rated player should be the one that a team wants at the plate, with the game on the line, more than anybody else. Since our WAR metric is based on the context-neutral wOBA, we would have to adjust our formula to determine this. Yet while we tested several ideas, none was conclusive in measuring clutch ability. Our first idea, to multiply each change in run value by the leverage index at each moment in the game (similar to a discretized Riemann-Stieltjes integral from measure theory) proved impractical, since leverage depends on the run environment and does not have a standard method of being calculated. Next, we tried to add a quadratic term, RE24 / PA^2, to each player's weighted runs created total, but this only lowered the baseline by about 0.04 WAR. Finally, we replaced wRC with RE24 completely as the basis for batters' WAR. This increases the correlation between each team's WAR and win total in four of eight seasons (from an overall season average of .764, to .803). However, team RE24 is almost identical to team runs scored, so it is not clear that the team correlation would be applicable to players.

Table 3 – Single-season WAR leaders, 2012-2019

		Team	wOBA	WAR
2019	Nick Gonzales	Cotuit	.492	3.02
2015	Nick Senzel	Brewster	.470	2.95
2012	Conrad Gregor	Orleans	.460	2.88
2012	Tyler Horan	Wareham	.481	2.85
2012	Kyle Schwarber	Wareham	.447	2.69
2018	Matthew Barefoot	Hyannis	.474	2.65
2015	Nick Solak	Bourne	.414	2.55
2015	Bobby Dalbec	Orleans	.520	2.51
2012	Tony Kemp	Cotuit	.487	2.44
2012	Phil Ervin	Harwich	.459	2.43

However, we are able to look at the effect of individual player clutch by correlating with eventual draft position. As we will see, it turns out that RE24 has a slightly weaker correlation with draft position than wRC does, suggesting that there is no significant improvement to the

model when we account for clutch performance. It is better, therefore, and easier, to operate only with a context-neutral metric, as ability in the clutch appears not to measurably improve the predictive accuracy of cWAR.

Pitching: FIP vs. wOBA Allowed

Our model stands to make the largest change with regard to how pitching performance is measured. The original cWAR model largely followed the algorithm of FanGraphs, converting fielding-independent pitching statistics (FIP) into wins per game above average (WPGAA) by a linear transformation, counting infield popups as equivalent to strikeouts. This resulted in a much lower total WAR across the league for pitchers (12.5) than for batters (58.6).

Since WAR for batters is based on each hitter's wOBA, we considered whether using wOBA allowed as the basis for a pitcher's WAR would place pitchers and hitters on a more equal plane. We thus computed each pitcher's wOBA allowed, converted it to wRC, and used wRC per batter faced as the rate statistic for calculating the replacement level baseline, much as batter's WAR uses wRC per plate appearance. By dividing by the league's number of runs per win for the season, we then obtain a pitching metric, which we call WAR2, in units of wins.

By changing the pitching metric from the FIP-based WAR1 to the wOBA-based WAR2, the league total WAR for pitchers increased from 12.5 to 19.6.

For individual pitchers, some see an increase in WAR and some a decrease. However, the leaders in pitching WAR from the 2019 season match up better to the end-of-season awards than before, and the winner of the Closer of the Year award, Zachary Brzykcy, sees his WAR improve tremendously from 0.07 to 0.48 (because the only two runs he allowed were solo home runs, which are the worst possible outcome for FIP).

Quantitatively, the WAR2 model, when added to the clutch version of batters' WAR to provide a team rating, correlates better with the teams' actual win totals in seven out of eight seasons in the sample. The average single-season correlation jumps from .80 to .86.

Table 4 -- Top 10 pitcher seasons by cWAR2, 2012-2019

		Team	cWAR2	cWAR1
2012	Sean Manaea	Hyannis	2.60	1.68
2016	Jeffery Passantino	Falmouth	2.56	1.22
2015	Mitchell Jordan	Orleans	2.48	1.41
2016	Hunter Williams	Harwich	1.94	0.69
2016	Brady Puckett	Falmouth	1.89	0.61
2016	Zacary Lowther	Brewster	1.79	1.26
2016	B. J. Myers	Falmouth	1.79	0.59
2015	Ricky Thomas	Yar.-Den.	1.77	0.80
2016	Peter Solomon	Harwich	1.74	0.51
2013	John Means	Falmouth	1.69	0.64

A list of the best pitchers by WAR2 from 2012 to 2019 appears in Table 4; note that the highest-rated pitcher by both metrics, Sean Manaea (now with the Oakland A's) succeeded on the mound during the "juiced-ball" season of 2012, with 2.60 WAR, more than double the total of the second-place pitcher (Jarrett Arakawa, 1.24) that year.

Correlation to the 2020 MLB Draft

Another correlation that we can measure, one that has some predictive value, compares a player's WAR to his position in the Major League draft the following year. While performance on the Cape is certainly not the only indicator of a player's skill—he may have a significant improvement or decline during the following college season, or may see limited action in the summer—it does provide a more standardized measure of what the player contributes on the field. This was observed in the case of Nick Gonzales, who produced impressive college offensive totals at New Mexico State. However, the fact that Gonzales performed at high altitude, and against only mid-major conference competition, led some scouts to doubt Gonzales' college statistics truly indicated his hitting ability. His MVP performance on the Cape in 2019 removed those doubts, and ESPN commentators noted this as a deciding factor for making him the seventh overall pick, the highest of anyone who played a full season on the Cape that summer.

Moreover, the circumstances of the 2020 draft provided a unique situation to analyze. The college season was cut short, meaning that there was a dearth of recent data on which teams could base their picks. The draft lasted only five rounds, for a total of 160 players, of whom 72 had played on the Cape during 2018 or 2019. In order to examine the immediate effect that a player's Cape League season had on his draft position, we limit our sample to the players who played only in 2019, or who saw more playing time (with the plate appearance or batter faced as the unit of action) in 2019 than in any previous season.

We compared each player's WAR, as well as other metrics, to his position in the 2020 draft. A separate set of Pearson correlation coefficients was computed for both batters and pitchers, and the results appear in Tables 5 and 6. For the purposes of this study, one two-way player, Casey Schmitt, was counted as a batter rather than as a pitcher, since he saw more action at the plate than on the mound.

For the 28 batters in the sample, each statistic shows a moderate correlation between performance and draft position. (The correlations are negative because a lower number draft position correlates with a higher value of the statistic.) The correlations are lower than we might expect (with r-squared values near 10%) because two months constitutes a small sample size; also, a player's performance on the Cape is only one of many factors for scouts to consider.

Of the metrics in the table, the counting statistics (WAR, wRC, and RE24) performed better than the rate statistics (wOBA and wRC+), with wRC at slightly the best.[3] All four statistics showed a linear relationship that was significant at $p \leq 0.10$, but not at 0.05. The fit for WAR might improve if we take fielding into account; we have already added baserunning.

Table 5 -- Correlations with draft position, batting

Metric	Correlation	p-value
cWAR	-0.3081	0.055
wOBA	-0.3001	0.060
wRC+	-0.2518	0.098
wRC	-0.3160	0.051
cWAR (w/clutch)	-0.3084	0.055
RE24	-0.3100	0.054

The sample of pitchers also came to 28 players, and we tested both WAR formulas along with three rate statistics (wOBA allowed, league-adjusted FIP-, and ERA), and two other counting statistics (wRC allowed and RE24). It appears that ERA, which scouts are likely to use, has the strongest correlation; however, one pitcher (Shane Drohan) pitched a small sample of innings and ended up with an ERA of 18.90, or 4.39 standard deviations above the mean. (His other rate statistics were not as extreme; in fact, his FIP- of 110 was only the seventh highest in the sample and his WAR was only 1.61 standard deviations below the mean, not enough to be an outlier.) When this influential observation is removed, ERA only has a correlation of 0.1503, the weakest of any metric. Part of the reason is that ERA is a rate stat, which does not carry any information about playing time.

Table 6 -- Correlations with draft position, pitching

Metric	Correlation	p-value
cWAR1	-0.1645	0.201
cWAR2	-0.2759	0.078
wOBA allowed	0.2097	0.142
FIP-	0.1795	0.180
wRC allowed	0.2480	0.102
RE24	0.1848	0.173
ERA	0.2768	0.077
ERA (w/o outlier)	0.1503	0.223

On the other hand, the modified, wOBA-based formula, WAR2, clearly had the strongest result, with the largest absolute correlation (0.2759), and the only significant linear relationship at the $p \leq 0.10$ level. Interestingly, the correlations for wRC and wOBA were stronger than that for league-adjusted FIP, matching the trend in the two WAR metrics, and suggesting that wOBA may be a better match to the scouts' judgements than FIP is.

If there is a correlation between WAR and draft position, does there also exist a causal relationship between them? Or in other words, did a player's sabermetric statistics have an effect on the draft? We can't really tell without comparing these correlations to those using more traditional statistics.

However, we can consider an example. The final pick of the draft, the 160th overall, went to Houston. The Astros had the choice between Darren Baker of the University of California at Berkeley, who was near the top of the league with a .342 batting average (and moreover is the son of Astros manager Dusty Baker), and Shay Whitcomb of the Division II school, the University of California at San Diego, who had a lower batting average (.303) but ranked in the top ten in WAR (1.53 to Baker's 0.84), due to his higher totals in walks and extra-base hits; Whitcomb also had a higher wOBA (.434 over .368) for the same reasons. The Astros put the classical statistics (and nepotism) aside and chose Whitcomb with the last pick, suggesting some importance of the logic behind the WAR metric. Thus, while a player's performance on the Cape in one summer is certainly not the only factor, it can be used to rank him for the draft, or to tentatively predict his position in the draft the following year.

[3] It turns out that wRC was the metric which we used to conclude that Babe Ruth's 1921 season was the best ever by a hitter in the Major Leagues.

Future Considerations

Although we have made some improvements to the WAR model, some factors are not included. For example, since the accurate hit location data necessary to calculate fielding metrics such as Defensive Runs Saved are not available as yet for the Cape League, we have not taken fielding into account, although we could use an older metric like Range Factor and convert it to runs. We also have not made any positional adjustments, which are used in the WAR algorithms for the Major Leagues, partially because players' positions are more fluid in college (Spencer Torkelson played first base in college, but the Tigers drafted him as a third baseman). If we had added positional adjustments, Nick Gonzales' league-leading WAR for the eight-year period would have been even higher, since his primary position is second base.

Additionally, while wOBA has proven to be a better fit than FIP for pitching, some sabermetric sources (notably FanGraphs' "fWAR" staistic) use FIP, as it provides a way to measure a pitcher's performance independently of his team's defense. FIP is different from wOBA in that it rewards pitchers for striking out batters, but does not account for hard-hit balls that result in hits that are not home runs. Perhaps it can be tweaked from the FanGraphs formula to provide a better fit in the context of the Cape League.

The Baseball Reference version, "bWAR," uses actual runs allowed rather than FIP. The wOBA version we use here is probably closer to bWAR's actual runs than fWAR's FIP-estimated runs.[4]

Yet some of the more interesting changes to the model may have to be made as a result of changes to the Cape League and MLB Draft. The CCBL had already decided to shorten the regular season to 40 games after the 2019 season, and in 2021, the season started later and did not include an all-star game. This means that we would need to use a new baseline; fortunately, we could track which players were on temporary contracts as the season progressed, without having to analyze who fit certain criteria retroactively.

Major League Baseball has also decided to go with a shorter draft in the future, and hold the draft in July (while the Cape League is playing) instead of June (usually the week before the Cape season starts). This is to fit the restructuring of the minor leagues, as well as the repurposing of some minor league teams to serve as summer league teams. A player could then play one summer in the Appalachian League, then a summer in the Cape Cod Baseball League, then a partial summer in the new MLB Draft League before being drafted and beginning a professional career in the Minors later that summer. An interesting study would then be to track a player's trajectory throughout these high-profile summer leagues in college, compared to each league's replacement-level baseline.

The possibility of further revisions, as well as the new structure that the Major Leagues have instituted this year for the pipeline that carries players from college to the professional ranks, assures that this project will continue for years to come.

Fr. Humbert Kilanowski, hkilanow@providence.edu ♦

[4] Charlie Pavitt notes (in the previous issue of *By the Numbers*) that each leading WAR algorithm for the Majors has its own method for calculating pitchers' WAR, such as FIP or runs allowed per nine innings as the base statistic.

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.