# What is the effect of optimized batting order on run production based on player-type matchups and recent performance?

Thesis Proposal

Brandon Zink

# 1. Background and Rationale

## 1.1 Background

Baseball is the most storied and traditional sport in the United States, having been played professionally for 149 years (Major League Baseball). With that, comes an inability to change and adjust, with claims such as "it's always been played that way".

With the popularization of the book and movie *Moneyball* regarding the 2002 Oakland A's, advanced statistics caught on in baseball through most every team (Moneyball). Teams have taken different approaches to advanced statistics (often called Sabermetrics), from the Tampa Bays Rays using the opener, to the Cubs batting Anthony Rizzo (a traditional 3 or 4 hitter) leadoff.

In the MLB, the league averages for each position tend to coincide with what those familiar with baseball would call a "traditional" lineup as you are taught since little league. Below are the averages for each spot in the batting order in the AL in 2018 (to avoid the batting statistics of the pitcher):

| **Order** | **BA** | **ISO** | **BB%** | **K%** | **wRC+** | **Player Type** |
|---|---|---|---|---|---|---|
| 1 | .259 | .159 | 8.9% | 19.6% | 105 | Low strikeout, low power hitter |
| 2 | .262 | .167 | 9.0% | 20.2% | 109 | Contact, low power hitter |
| 3 | .266 | .195 | 9.7% | 20% | 118 | Best all around hitter |
| 4 | .258 | .207 | 8.4% | 22.2% | 113 | High power slugger |
| 5 | .244 | .183 | 8.2% | 22.4% | 100 | Generic, league average hitter |
| 6 | .248 | .157 | 8.4% | 21.9% | 97 | Fourth worst hitter |
| 7 | .235 | .151 | 7.7% | 23.6% | 87 | Third worst hitter |
| 8 | .237 | .144 | 7.2% | 23.3% | 85 | Second worst hitter |
| 9 | .227 | .126 | 6.4% | 24.9% | 74 | Worst hitter |

Not only does this go to show that the MLB (or at least the AL) tends to set their batting order in a "traditional" sense, but those statistics will also serve as a baseline batting order (at least for the AL simulations).

Adding a pitcher into the mix creates a unique challenge as well since pitchers are so much worse, to the point of almost being a free out. Below are the statistics for pitchers batting in the 2018 season:

| **Pos** | **BA** | **ISO** | **BB%** | **K%** | **wRC+** |
|---|---|---|---|---|---|
| P | .115 | .033 | 2.9% | 42.2% | -25 |

Even the best pitchers are bad relative to your average MLB batter. The best hitting pitcher in the MLB in 2018 was Clayton Kershaw with a wRC+ of 82, however there were only 3 pitchers with a wRC+ above 50 (FanGraphs).

## 1.2 Literature Review

There is only one publicly available extensive piece of literature on lineup optimization, and that is from *The Book: Playing The Percentages In Baseball*. Just about any other piece that you find will heavily reference or reiterate the same findings as *The Book*. *The Book* takes an averages based approach to lineup optimization, that is, using tables of base/out states, plate appearance averages, and various other averages categorized by position in the batting order to complete their analysis. Their major finding is the following:

> "Your three best hitters should bat somewhere in the #1, #2, and #4 spots. Your fourth- and fifth-best hitters should occupy the #3 and #5 spots. The #1 and #2 slots will have the players with more walk than those in the #4 and #5 slots. From slot #6 through #9, put the players in descending order of quality." (pg. 134)

And for the NL only, they state the following:

> "The second leadoff hitter theory exists. You can put your pitcher in the eighth slot and gain a couple of extra runs per year." (pg. 151)

*The Book* also offers advice about where to place your base stealers, when to worry about strikeouts, and at what spot in the batting order are double-plays most costly.

There are two major factors in which *The Book* falls short, however. *The Book* uses an averages based approach, but the averages don't really hold when you consider that they have altered their lineup. If the average MLB leadoff hitter is a fast guy who doesn't get on base a whole lot, but you instead put your best on base machine at the top of the order, than that changes the probability of having a runner in scoring position for the two, three, and four hitters, which they do not address. They should have run more iterations, or at least a second iteration, to see how a newly constructed lineup would change the probabilities (or run a simulation with it). The second is that team composition don't always fit this model. Not every MLB team has a high OBP guy with low power, base stealer, and multiple above average hitters who can hit for power. These number will change as your teams change, and that should be taken into account.

## 1.3 Key Terms

National League (NL)/American League (AL): the two leagues in MLB baseball. The key difference is that in the NL, your pitcher comes to bat unless they are pinch hit for (in which case they are done pitching that game). In the AL, each team is allowed a designated hitter (DH) that bats in the spot of the pitcher.

Batting Average (BA/AVG): the percent of the time that a runner reaches base out of all of their at-bats (AB). Note that an at-bat is different from a plate appearance (PA). An at-bat is all of a players plate appearances minus their base on balls/walks (BB), hit-by-pitch (HBP), sacrifice fly or sacrifice bunt (moving a runner on base to a further base or scoring him for the cost of an out), and catcher interferences. The 2018 MLB average for AVG is .248.

On-Base Percentage (OBP): the percent of plate appearances in which a player reaches base by hit, walk, or hit by pitch. Note that this is percent of plate appearances as opposed to percent of at-bats. The 2018 MLB average for OBP is .318.

Slugging Percentage (SLG): a players batting average multiplied the number of bases they get for each hit. For example, if a player goes 3 for 6 (3 hits in 6 at bats) with a single, double, and home run, then his slugging percentage for that game would be (1+2+4)/6 = 1.167, with the one, two, and four representing the number of bases from each hit. The 2018 MLB average for SLG is .409.

Isolated Power (ISO): a players SLG minus a players AVG. ISO is supposed to represent a players raw power, or his ability to hit for extra bases. The 2018 MLB average for ISO is .161.

Base on Balls Percentage (BB%): the percent of plate appearances in which a player reaches base due to a base on balls (usually called a walk) caused by reaching 4 balls in an plate appearance. The 2018 MLB average for BB% is 8.5%.

Strikeout Percentage (K%): the percent of plate appearances in which a player strikes out by reaching three strikes. The 2018 MLB average for K% is 22.3%.

Caught Stealing Percentage (CS%): the percent of stolen base attempts in which a player is thrown out. When looking at a stat line, this is CS/(SB+CS). The 2018 MLB average for CS% is 27.9% (i.e. 72.1% of stolen base attempts are successful).

Weighted Runs Created Plus (wRC+): a normalized version of wOBA about 100 (100 is average, 200 is twice as good as average). Both wRC+ and wOBA measure offensive talent in terms of runs created. By giving each event of a plate appearance (1B, 2B, 3B, HR, uBB, HBP) an expected runs created (as found by taking the MLB average) and dividing by the number of plate appearances, you end up with a stat that measures a players offensive contribution independent of context or quality of the rest of the team. wOBA uses a similar scale to OBP, while wRC+ is centered about 100 (give or take depending on the league year). The 2018 MLB average for wOBA is .315 and wRC+ is 97.

More information, including the equations that are used to calculate them, about any of these statistics can be found in the glossary section of FanGraphs.

## 1.4 Thesis Question

So that brings it around to the big question:

What is the effect of optimized batting order on run production based on player-type matchups and recent performance?

This is broken into a few distinct parts, and each will be detailed below. A player modeling algorithm that takes into account recent performance (such as hot and cold streaks and regression and progression) and sample size (number of games played). Next, generating player types (ex. high strikeout, high walk pitcher) and player-type matchups (what statistics are impacted by specific player type matchups). Lastly, using all of that information and a simulation to find which spots in the batting order are most important to run production and which skills (ex. avoiding strikeouts, walking) are most/least beneficial in which specific spot in the order.

# 2. Methodology

## 2.1 Methods

### 2.1.1 Recent Performance

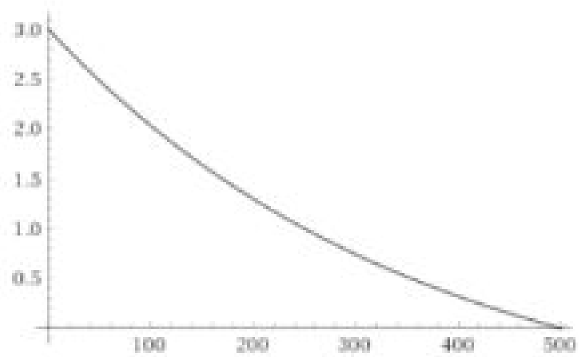
In order to account for recent performance (progression/regression) and hot and cold streaks, I am using a player modeling method that weights statistics from more recent games heavier than those of older games.

To do this, I am going to take player statistics from the past 500 games. I chose games instead of plate appearances to account for platoon and bench/role players, where getting up to 500+ PAs might take 5 or more years and talent level may have of changed greatly over that time.

In order to weight the statistics from each game, I will use the following formula:

$$\sqrt[500]{4}^{-x+500} - 1$$

Which, when graphed, looks like the following:




Where the x-axis is the number of games back (1 would be the most recent game played) and the y-axis is the weight for that given game. This model may be adjusted as the project continues.

This will be done for each individual statistic (AVG, BB%, K%, etc.).

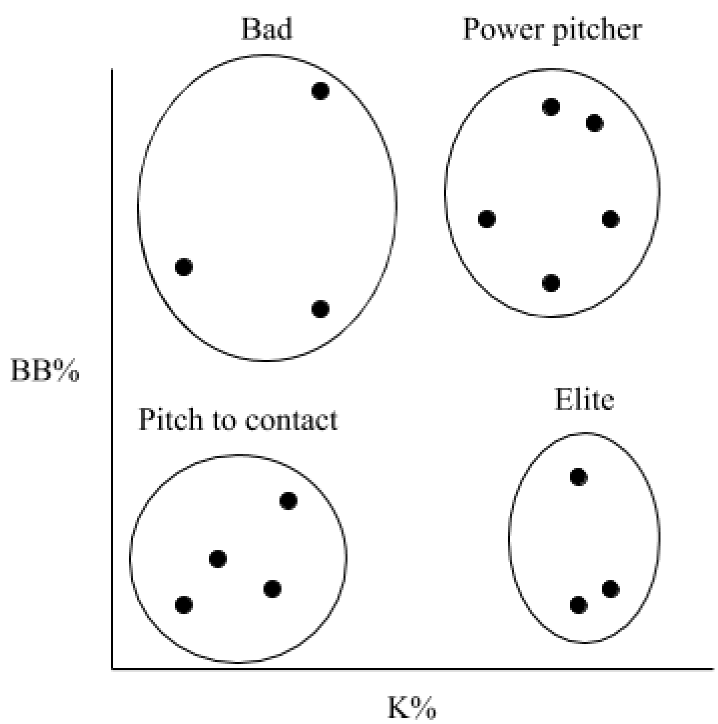
To account for sample size issues (for players that haven't played 500 games), I will use the following formula to adjust:

$$\left(\frac{x}{500} * \text{player avg.}\right) + \left(\left(1 - \frac{x}{500}\right) * \text{MLB avg.}\right)$$

This will assume that a player is MLB average when the sample size is zero, and then as the sample size increases, weight the final statistic more heavily towards the players actual production. Once the sample size hits 500 games, only the players stats will be considered. This is to make sure that small sample sizes don't skew the results of applying this method to an actual team, where a rookie that has gone 3 for 4 wouldn't have an expected AVG of .750.

### 2.1.2 Player-Types

The goal of using player types is to adjust for the type of player that is both batting and pitching. My methodology for this is still a work in progress, but I envision it looking something like the following:




More analysis needs to be completed on the specific statistics to use, but upon creating the player types, I can run simulation to determine the effect of certain player types on other player types to use in my final product.

An example of the result of this would be something along the lines of a table of the following, for a specific type of batter (for example, say a contact, low strikeout and walk batter):

| | **AVG** | **SLG** | **K%** | **BB%** |
|---|---|---|---|---|
| **Bad** | +.030 | +.040 | 0 | 0 |
| **Pitch to Contact** | -.005 | 0 | -0.6% | -0.5% |
| **Power Pitcher** | 0 | 0 | +1.0% | +1.0% |
| **Elite** | -.010 | -.050 | +1.5% | -1.0% |

These values would be compared to what we would expect to come of a pitcher and batter with those specific statistics, using the following formula (in this example, we are comparing a batter with a .450 SLG, pitcher with .500 SLG allowed, compared to the MLB average of .409):

*Odds(Hitter)*: .450/.550 = .818
*Odds(Pitcher):* .500/.500 = 1.00
*Odds(League)*: .409/.591 =.692
*Odds(H) * Odds(P) / Odds(L)* = 1.182
1.182/(1+1.182) = .542 SLG

This gives us an expected SLG for the plate appearance of .542, which makes sense since the batter is an above average slugger and the pitcher is below average in terms of allowing extra base hits (Tango, Tom, Odds-Ratio Method).

By comparing the expected percentage for a matchup and the simulated matchup between different player-types, we can simulate the effect in our lineup generator without having to run a full simulation for every lineup.

### 2.1.3 Simulation

In order to perform the simulations as described above, I am going to use Out of the Park Baseball 17 (OOTP 17), which is an in depth baseball simulation game. As a feature of this game, it allows you to run any number of simulations between two custom teams and allows for control of extraneous influences (park factors, fatigue, etc.).

The most important statistic that will come as a part of this is the amount of runs per game over a large sample size, although additional insights may be discovered as OOTP 17 offers a full statistic summary of simulations. Below is an example from OOTP's website, showing the results of a small size simulation (7 games):


| 5 | Record | 2 |
|---|---|---|
| 35 | Runs Scored | 28 |
| 282 | AVG | .259 |
| 11 | HR | 7 |
| 6 | SB | 0 |
| 28 | Runs Allowed | 35 |
| 3.39 | ERA | 4.66 |
| 7 | HR Allowed | 11 |
| 2.58 | BB/9 | 3.15 |
| 5.43 | K/9 | 5.62 |
| 2.11 | K/BB | 1.78 |
| .716 | Defensive Efficiency | .700 |

ALBUQUERQUE TOREADORS BATTING STATS

| Name | G | AB | R | H | 2B | 3B | HR | RBI | TB | BB | K | SB | CS | AVG | OBP | SLG | OPS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bruno Walker C | 4 | 3 | 2 | 3 | 0 | 0 | 1 | 1 | 6 | 0 | 0 | 0 | 0 | 1.000 | 1.000 | 2.000 | 3.000 |
| David Pineau 3B | 3 | 4 | 1 | 3 | 1 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | .750 | .750 | 1.000 | 1.750 |
| Alejandro Herrera LF | 7 | 28 | 8 | 11 | 1 | 2 | 1 | 4 | 19 | 5 | 3 | 2 | 0 | .393 | .485 | .679 | 1.163 |
| Angel Yánez 1B | 5 | 13 | 3 | 5 | 1 | 0 | 1 | 4 | 9 | 2 | 4 | 0 | 0 | .385 | .500 | .692 | 1.192 |
| Antônio González SP | 1 | 3 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | .333 | .333 | .333 | .667 |
| Andy Skinner 1B | 6 | 18 | 5 | 6 | 0 | 0 | 0 | 2 | 6 | 3 | 4 | 0 | 0 | .333 | .409 | .333 | .742 |
| Justin Barton C | 6 | 22 | 2 | 7 | 1 | 0 | 1 | 1 | 11 | 0 | 2 | 0 | 0 | .318 | .318 | .500 | .818 |
| William Ballard 2B | 7 | 29 | 2 | 9 | 1 | 0 | 0 | 3 | 10 | 1 | 8 | 0 | 0 | .310 | .333 | .345 | .678 |
| Severino Verdades SS | 3 | 7 | 1 | 2 | 1 | 0 | 1 | 2 | 6 | 1 | 1 | 0 | 0 | .286 | .375 | .857 | 1.232 |
| Enrico Alvarez CF | 7 | 32 | 6 | 9 | 0 | 0 | 3 | 4 | 18 | 3 | 0 | 1 | 2 | .281 | .343 | .563 | .905 |
| Aurelio Gil RF | 5 | 15 | 2 | 4 | 1 | 0 | 1 | 2 | 8 | 0 | 2 | 1 | 0 | .267 | .250 | .533 | .783 |
| Paul Lowry 3B | 7 | 23 | 3 | 5 | 1 | 1 | 2 | 4 | 14 | 4 | 6 | 1 | 0 | .217 | .357 | .609 | .966 |

This limits the amount of error that could be introduced by building my own simulator, as well as OOTP 17 being a better simulator than I could build in time for this project.

### 2.1.4 Batting Order Optimization

Using all of this information, I can begin to work on the actual optimizing of the lineup in terms of the maximum number of runs produced per game.

To start, I will simulate a league average lineup (at least 1000 games simulated) against league average pitching to establish a base number of runs scored per game. This league average lineup will be determined by the league averages for each batting order position, as per FanGraphs, and will need to be done with both AL and NL lineups to adjust for pitchers batting (all simulations will need to be done with NL and AL statistics to account for the pitcher batting).

After establishing a base, I will simulate that sam lineup with a Mike Trout/Ted Williams type batter (significantly above league average, wRC+ of at least 200) in each batting order position (again recording the number of runs per game). This will establish the order of importance for scoring runs in the batting order, and will be used later. To adjust this for having a pitcher in the lineup, the nine-hitters statistics will be moved to the eight-hitters spot when the "Mike

Trout/Ted Williams" simulation player is in the ninth position since there has not been an above league average pitcher at hitting since the 1800s (FanGraphs).

Following this, I need to establish which specific traits have the greatest impact on run production at specific positions in the batting order. I will simulate great, good, bad, and terrible versions of each of the following statistics at each spot in the batting order:

AVG, SLG/ISO, BB%, K%, SB%/CS%

I may expand this include more statistics as OOTP 17 allows for, but these categories should at least allow for establishing optimal profiles for each batting order position. The results of this may look something like the following:

| **Pos.** | **AVG** | **SLG** | **BB%** | **K%** | **CS%** |
|---|---|---|---|---|---|
| 1 | + | - | ++ | -- | + |
| 2 | + | + | ++ | - | 0 |
| 3 | 0 | ++ | 0 | + | 0 |
| 4 | 0 | 0 | 0 | + | 0 |
| 5 | + | 0 | + | - | ++ |
| 6 | 0 | 0 | - | 0 | ++ |
| 7 | 0 | 0 | - | 0 | 0 |
| 8 | 0 | 0 | - | - | 0 |
| 9 | 0 | 0 | - | 0 | 0 |

These values will have actual run values attached to them and will all be relative to other positions in the lineup (a better SLG is always good, but it's better in some positions than others).

This table, along with the previous information, will allow us to apply this to a real life roster.

### 2.1.5 Real Life Roster Application

The inputs to this will be the opposing starting pitcher and the available batters for a lineup.

The batter stats then need to be adjusted for the starting pitcher, using the player-type matchup as defined above.

The starting batters will be determined by the maximum wOBA/wRC+ (they're essentially the same stat, just wRC+ is normalized about 100) making sure that all defensive positions are filled. To consider real-life usage, any player which needs a rest day/injured will be omitted from the roster submitted to this algorithm.

Then, using the stat/batting order position table, I will score the player based on how they fit in to each position. If a player has a high SLG compared to the rest of his team, he will receive a high score for any batting order position that benefits from a relatively high slugging, as found in the previous section, repeated for each position and statistic in the lineup. A resulting table might look something like the following:

| **Pos.** | **Player A** | **Player B** | **Player C** |
|---|---|---|---|
| 1 | 26 | 12 | 76 |
| 2 | 58 | 44 | 43 |
| 3 | 12 | 45 | 22 |
| 4 | 47 | 34 | 3 |
| 5 | 56 | 65 | 45 |
| 6 | 12 | 43 | 67 |
| 7 | 86 | 21 | 2 |
| 8 | 34 | 21 | 23 |
| 9 | 23 | 14 | 11 |

This would be done for every player in the lineup. Using the batting order importance found at the beginning of the simulations (Mike Trout/Ted Williams), weight these values based on run scoring importance (ex. if 2 is the most important order position, multiply all 2-hitter values by the highest weight, moving downwards).

With these adjusted values, set the lineup by taking the highest overall value in the table, assigning that player to that batting order position, and then moving downwards until all of the batting order positions are filled. This will present the optimal lineup for run production. Repeat this entire process for the NL, with a pitcher in the batting lineup.

Afterwards, run a simulation with an optimized lineup and compare it to the original baseline that was established to see the change in run production, thus answering the original thesis question. Let's hope that it's significant.

## 2.2 Limitations

This method does present a few limitations.

The opener method, as utilized mainly by the Tampa Bay Rays and is spreading across the MLB, would mean you would have to know the "starter" (pitcher who will eat the most innings) for this method to work. A bullpen game (no true starter, but a bullpen rotation) would also pose a problem, and the average of their bullpen would be used in place of their starter.

Another limitation is the inability to adjust for pinch hitters. As I continue on this project, I may come up with a way to adjust for the quality of the bench, but as of now I have not come up with a reasonable solution.

# 3. Working Bibliography

*FanGraphs Baseball*, www.fangraphs.com/.

"Major League Baseball." *Wikipedia*, Wikimedia Foundation, 2 Oct. 2018, en.wikipedia.org/wiki/Major_League_Baseball#History.

"Moneyball." *Wikipedia*, Wikimedia Foundation, 27 Sept. 2018, en.wikipedia.org/wiki/Moneyball.

Tango, Tom M., et al. *The Book: Playing The Percentages In Baseball*. CreateSpace, 2014.

Tango, Tom M, et al. “THE BOOK--Playing The Percentages In Baseball.” *Inside the Book*, www.insidethebook.com/ee/index.php/site/comments/the_odds_ratio_method/.