# Who Should an MLB GM Take With The First Pick In The Draft?

**ANNA BELL LANSDOWNE, Marquette University**

## 1  INTRODUCTION

Before data science was a prominent force in GM offices everywhere, MLB scouts used to have to rely on their intuition and basic stats to decide who was worth drafting. Intuition included checking whether a player had a weak handshake, small hands or square shoulders [1]. They also avoided players who hit and threw with opposite hands, who had duck feet or who had a less than ideal physique. I hope to instead use data science techniques to decide who is worth drafting. I will be looking at picks from the June Amateur Draft and not the Rule V Draft. I will also be assuming that the ultimate goal of a GM is to win and not to increase ticket sales. I will also be assuming that a player is drafted to play on that team and not to be traded. Finally, I will not be considering intangibles as there are no stats for this and will be assuming that a team could benefit from either a bat or an arm.

## 2  PITCHER VS POSITION PLAYER

### 2.1  Finding a database

The first question I hope to solve is whether it is more beneficial to draft a pitcher or a position player with the first pick. In the past 5 years 3 position players have been taken first (Mickey Moniak, Dansby Swanson, Carlos Correa) and 2 pitchers have been taken first (Brady Aiken and Mark Appel). So is a team more likely to find a Bryce Harper or a Stephen Strasburg with the $1^{st}$ pick? And who is more likely to bring a championship to the organization? My first step was to find a database that contained the data I needed. Baseball-reference.com listed the history of first picks and their stats to 1965 so I created a Python script to put this data into a csv file. This is how the data looked before I did any tidying:

| | Year | Rnd | DT | FrRnd | RdPck | Tm | Signed | Name | Pos | WAR | ... | BA | OPS | G.1 | W | L | ERA | WHIP | SV | Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2016 | 1 | NaN | FrRnd | 1 | Phillies | Y | Mickey Moniak (minors) | OF | NaN | ... | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | HS |
| 1 | 2015 | 1 | NaN | FrRnd | 1 | Diamondbacks | Y | Dansby Swanson (minors)\swansda01 | SS | 0.9 | ... | 0.302 | 0.803 | NaN | NaN | NaN | NaN | NaN | NaN | 4Yr |

Fig. 1. How the data looked before I did any tidying

The first thing I did to clean up this database was to take out any players who had a "N" in the "Signed" column. This is because I want to analyze players who sign with the team who drafts them. For example Brady Aiken was drafted first in 2014 by the Astros but did not sign so I took him out of the data. The next thing I did was to split my data into pitchers and position players. Then with all 3 data sets I took out unnecessary columns including what school they were drafted out of and what round they were picked in because this was the first round for everyone.

### 2.2  Creating a WAR database

From the original data I took out all stats besides WAR because this is the primary stat I want to look at. WAR stands for Wins Above Replacement and can be compared between pitchers and field players. It is harder to compare batting average and ERA for example. The WAR I am using is technically bWAR because WAR does not have a specific calculation and bWAR is the WAR calculated by Baseball Reference. For the pitcher and position player I took

out the data that is not relevant to them. So my 3 (original, pitcher and position in order) sets of data now look like this:

| | Year | Tm | Name | Pos | WAR | Type |
|---|---|---|---|---|---|---|
| 0 | 2016 | Phillies | Mickey Moniak (minors) | OF | NaN | HS |
| 1 | 2015 | Diamondbacks | Dansby Swanson (minors)\swansda01 | SS | 0.9 | 4Yr |
| 3 | 2013 | Astros | Mark Appel (minors) | RHP | NaN | 4Yr |
| 4 | 2012 | Astros | Carlos Correa (minors)\correca01 | SS | 10.1 | HS |
| 5 | 2011 | Pirates | Gerrit Cole (minors)\colege01 | RHP | 9.4 | 4Yr |

**Fig. 2. Original set of data with only the WAR statistic**

| | Year | Tm | Name | Pos | WAR | G.1 | W | L | ERA | WHIP | SV | Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2013 | Astros | Mark Appel (minors) | RHP | NaN | NaN | NaN | NaN | NaN | NaN | NaN | 4Yr |
| 5 | 2011 | Pirates | Gerrit Cole (minors)\colege01 | RHP | 9.4 | 94.0 | 47.0 | 30.0 | 3.23 | 1.20 | 0.0 | 4Yr |
| 7 | 2009 | Nationals | Stephen Strasburg (minors)\strasst01 | RHP | 18.2 | 156.0 | 69.0 | 41.0 | 3.17 | 1.09 | 0.0 | 4Yr |
| 9 | 2007 | Devil Rays | David Price (minors)\priceda01 | LHP | 31.9 | 253.0 | 121.0 | 65.0 | 3.21 | 1.14 | 0.0 | 4Yr |
| 10 | 2006 | Royals | Luke Hochevar (minors)\hochelu01 | RHP | 3.1 | 279.0 | 46.0 | 65.0 | 4.98 | 1.34 | 3.0 | NaN |

**Fig. 3. Pitchers' data with only the WAR statistic**

| | Year | Tm | Name | Pos | WAR | G | AB | HR | BA | OPS | Type |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2016 | Phillies | Mickey Moniak (minors) | OF | NaN | NaN | NaN | NaN | NaN | NaN | HS |
| 1 | 2015 | Diamondbacks | Dansby Swanson (minors)\swansda01 | SS | 0.9 | 38.0 | 129.0 | 3.0 | 0.302 | 0.803 | 4Yr |
| 4 | 2012 | Astros | Carlos Correa (minors)\correca01 | SS | 10.1 | 252.0 | 964.0 | 42.0 | 0.276 | 0.829 | HS |
| 6 | 2010 | Nationals | Bryce Harper (minors)\harpebr03 | OF | 21.5 | 657.0 | 2336.0 | 121.0 | 0.279 | 0.883 | JC |
| 8 | 2008 | Rays | Tim Beckham (minors)\beckhti01 | SS | 1.3 | 151.0 | 408.0 | 14.0 | 0.238 | 0.720 | HS |

**Fig. 4. Position Players' data with only the WAR statistic**

## 2.3  Basic Stats

The next thing I did was get basic stats about the types of players. First I compared the mean of the WAR of the pitchers against the mean of the WAR of the position players. The average WAR of a pitcher taken first overall is 14.29 and the average WAR of a position player taken first overall is 24.36:

```
In [18]: df_Pitchers['WAR'].mean()

Out[18]: 14.292307692307691

In [19]: df_Position['WAR'].mean()

Out[19]: 24.35625
```

Fig. 5. The average WAR of a pitcher taken first overall and the average WAR of a position player taken first overall

## 2.4   High School VS College

I also want to ask the question should a GM pick a player from High School or College? Going back to the past 5 picks, 3 have been from High School (Mickey Moniak, Brady Aiken, Carlos Correa) and 2 have been from College (Dansby Swanson and Mark Appel). So I compared the mean of the WAR of players from a 4-year institution against the mean of the WAR of players from a High School. The average WAR of a player taken first overall out of a 4-year institution was 16.32 and the average WAR of a player taken first overall out of a high school was 27.69:

```
In [26]: df[(df.Type == 'HS')].mean()

Out[26]: Year     1988.500000
         WAR        27.690476
         dtype: float64
```

```
In [54]: df[(df.Type == '4Yr') | (df.Type == 'JC')].mean()

Out[54]: Year     1992.000000
         WAR        16.547826
         dtype: float64
```

**Fig. 6. The average WAR of a player taken first overall out of a high school and the average WAR of a player taken first overall out of a 4-year institution**

To examine these numbers further I took the means of the WARs from pitchers from High School, position players from High School, pitchers from 4-year institutions and position players from 4-year institutions. The results were:

| Type | WAR |
|---|---|
| Pitcher from HS | 0.70 |
| Position Player from HS | 29.04 |
| Pitcher from 4Yr | 16.55 |
| Position Player from 4Yr or JC | 16.55 |

## 3   SIGNIFICANCE

### 3.1   Position Players VS Pitchers

Because we are comparing two different populations (pitchers vs. position players), I wanted to check the significance of the means that I had previously

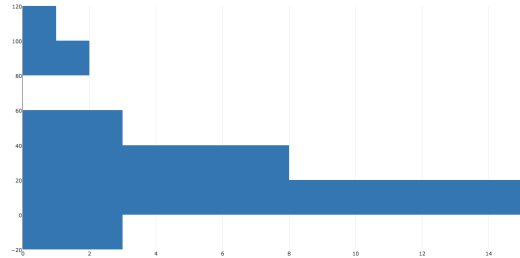**calculated. The first means I compared were just all position players vs. all pitchers.**



**Fig. 7. Distribution of the means of all position players**
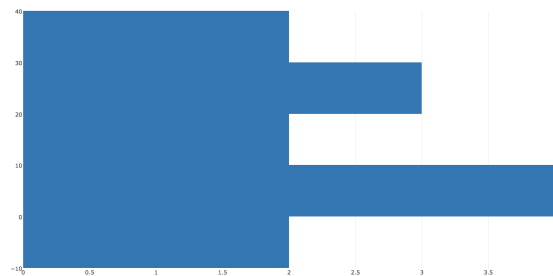


**Fig. 8. Distribution of the means of all pitchers**

```
In [41]:   from scipy.stats import ttest_ind

           position = df_Position.WAR
           pitchers = df_Pitchers.WAR

           ttest_ind(position, pitchers, nan_policy='omit')

Out[41]:   Ttest_indResult(statistic=1.2424590601379675, pvalue=0.22080051149262181)
```

**Fig. 9. Two-sided t test**

**Using a two-sided t test I was able to calculate that the p-value was 0.22, which is less than 0.5 meaning the difference in means is significant. From this we can conclude that it is more likely that drafting a position player with the first pick in the draft will give you a player that brings more Wins Above Replacement than a pitcher.**

```
                          import levene
click to scroll output; double click to hide
         positionNoNan = df_Position[np.isfinite(df_Position['WAR'])]
         pitchersNoNan = df_Pitchers[np.isfinite(df_Pitchers['WAR'])]

         position = positionNoNan.WAR
         pitchers = pitchersNoNan.WAR

         levene(position, pitchers)
```

Out[56]: LeveneResult(statistic=1.7618720391492331, pvalue=0.19139583278358543)

**Fig. 10. Levene test**

**Then, using a Levene test I was able to calculate that the p-value was 0.19, which is less than 0.5 meaning the difference in variance is significant. From this we can conclude that when drafting a position player, the variance in WAR that these players usually bring is significantly bigger than the variance in WAR that a pitcher brings. Position players have the potential to bring massive amounts of WAR, as evidenced by the histogram, while pitchers tend not to have as big a potential.**

**The next means I compared were position players from 4-year institutions vs. position players from high school and pitchers from 4-year institutions vs. pitchers from high school.**
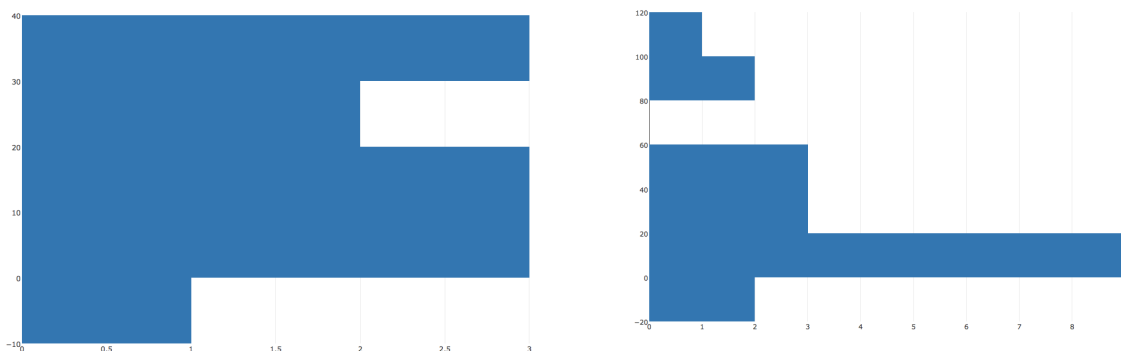


**Fig. 11. Distribution of the means of position players from a 4-year institution (left) vs. distribution of the means of position players from a High School (right)**

```
In [57]: from scipy.stats import ttest_ind

         fouryr = position4yr.WAR
         HS = positionHS.WAR

         ttest_ind(fouryr, HS, nan_policy='omit')

Out[57]: Ttest_indResult(statistic=-1.2362274039635859, pvalue=0.22596595998751204)
```

**Fig. 12. Two-sided t test**

**Using a two-sided t test I was able to calculate that the p-value was 0.23, which is less than 0.5 meaning the difference in means is significant. From this we can conclude that it is more likely that drafting a position player from a high school with the first pick in the draft will give you a player that brings more Wins Above Replacement than one from a four-year institution.**

```
In [57]: from scipy.stats import levene

         position4yrNoNan = position4yr[np.isfinite(position4yr['WAR'])]
         positionHSNoNan = positionHS[np.isfinite(positionHS['WAR'])]

         position1 = position4yrNoNan.WAR
         position2 = positionHSNoNan.WAR

         levene(position1, position2)

Out[57]: LeveneResult(statistic=3.0621728114053282, pvalue=0.090358125274031656)
```

**Fig. 13. Levene test**

**Then, using a Levene test I was able to calculate that the p-value was 0.09, which is less than 0.5 meaning the difference in variance is significant. From this we can conclude that when drafting a position player from a high school, the variance in WAR that these players usually bring is significantly bigger than the variance in WAR that a position player from a 4-year institution brings. Position players from high school have the potential to bring massive amounts of WAR, as evidenced by the histogram, while position players from a 4-year institution tend not to have as big a potential.**
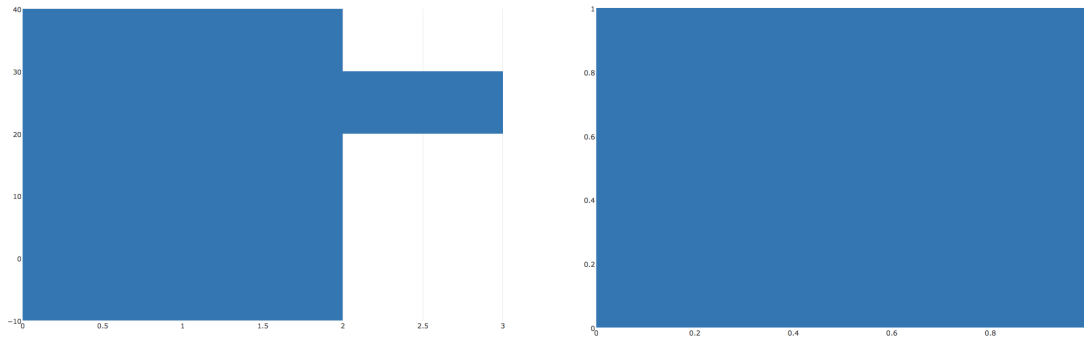
**Fig. 14. Distribution of the means of pitchers from a 4-year institution (left) vs. distribution of the means of pitchers from a High School (right)**

I did not perform a 2-sided t-test or Levene test on these two means because there was not enough data for a pitcher from a high school to get a meaningful result.

## 3.2   Odds Ratio

The next thing I did in order to analyze my data was create an odds ratio for all position players against all pitchers. I wanted to see what the odds were of drafting a franchise player with the first pick. For this odds ratio I decided that a truly impactful position player would have a WAR of over 40. I decided upon this number because according to Baseball Reference in order to have an all-star quality year a player should have a WAR above 8 and I thought an impactful player should have at least 5 years where they are an all-star. For a truly impactful pitcher I lowered the number to 30 because pitchers tend to have more injuries and be more inconsistent so they might not have as many all-star quality seasons. So I wanted to see what the odds were of drafting an impactful player. The odds ratio formula is:

$$\text{Odds ratio} = \frac{PG_1 / (1 - PG_1)}{PG_2 / (1 - PG_2)}$$

**Fig. 15. The odds ratio formula**

**PG1 (The odds of drafting an impactful position player): 6/34 = 3/17**

```
In [24]: df_Position[(df_Position.WAR > 40.0)].count()

Out[24]: Year      6
         Tm        6
         Name      6
         Pos       6
         WAR       6
         G         6
         AB        6
         HR        6
         BA        6
         OPS       6
         Type      6
         dtype: int64

In [26]: df_Position.count()

Out[26]: Year      34
         Tm        34
         Name      34
         Pos       34
         WAR       32
         G         32
         AB        32
         HR        32
         BA        31
         OPS       31
         Type      34
         dtype: int64
```

**Fig. 16. The counts for a position player with a WAR over 40 and all position players**

**PG2 (The odds of drafting an impactful pitcher): 2/15**

```
In [28]: df_Pitchers[(df_Pitchers.WAR > 30.0)].count()

Out[28]: Year     2
         Tm       2
         Name     2
         Pos      2
         WAR      2
         G.1      2
         W        2
         L        2
         ERA      2
         WHIP     2
         SV       2
         Type     2
         dtype: int64

In [29]: df_Pitchers.count()

Out[29]: Year     15
         Tm       15
         Name     15
         Pos      15
         WAR      13
         G.1      13
         W        13
         L        13
         ERA      13
         WHIP     13
         SV       13
         Type     14
         dtype: int64
```

**Fig. 17. The counts for a pitcher with a WAR over 30 and all pitchers**

**So the odds ratio would be (3/17 / 14/17) / (2/15 / 13/15) =**

**1.39**

**It's important to note that if the odds were equal the odds ratio would be 1 and not 0.**

I wanted to use a Fisher's Exact Probability Test to see if this was significant. When I ran this test I got a p value of 1 which is greater than 0.5 and we can conclude this is not significant.

```
In [30]: import scipy.stats as stats

In [31]: oddsratio, pvalue = stats.fisher_exact([[3, 14], [2, 13]])

In [32]: pvalue
Out[32]: 1.0
```

**Fig. 18. Fischer's Exact Probability Test to see if the odds ratio was significant**

From this calculation we can conclude that drafting a very good pitcher or player as the first pick is equally likely, (that is not so likely).

## 4   CONCLUSIONS

In summary, I would advise a Major League Baseball General Manager to pick a position player from a High School. In general, position players provide more wins above replacement than a pitcher does and position players from high school provide the greatest WAR on average with the greatest variance. Position players have the potential to bring massive amounts of WAR as evidenced by the variance of their means. If a general manager wants to take a pitcher, one from a 4-year institution is a safer bet. Lastly, it is equally unlikely to find a superstar position player or a superstar pitcher with the first pick so I recommend trying to find a solid player rather than one that would be considered high risk/high reward like a position player out of high school.

## REFERENCES

[1] Jerry Crasnick. May 29, 2013. MLB draft: The Scouts' eyeball test. *ESPN* http://www.espn.com/mlb/story/_/id/9319990/draft-eligible-players-how-scouts-percieve-their-shortcomings

**Turned in May 2017**