2023

The Value of Playing in College Baseball to Major League Baseball Organizations

Matthew Penn
Syracuse University

The Value of Playing in College Baseball to Major League Baseball Organizations

A Thesis Submitted in Partial Fulfillment of the
Requirements of the Renée Crown University Honors Program at
Syracuse University

Matthew Penn

Candidate for Bachelor of Science in Sport Analytics
and Renée Crown University Honors
May 2023

Honors Thesis in Sport Analytics

Thesis Advisor: ________________________
Dr. Rodney Paul, Director,
Sport Analytics

Honors Director: ________________________
Dr. Danielle Smith, Director

Abstract

The Major League Baseball (MLB) draft allows MLB organizations the opportunity to select from the top amateur players in the country. This thesis investigates the probability of an MLB draftee playing in the Major Leagues based on factors such as where they went to school and where they were drafted. The data were obtained from The Baseball Cube for the first 20 rounds of the Regular June MLB Draft from 1980 to 2010. Players were grouped into hitters and pitchers and categorized based on the level of play, which are High School, National Collegiate Athletics Association (NCAA) Division 1, Junior Colleges, and Other 4 Year. The paper analyzes the distribution of the level of school in each round and applies a logit model with three factors: draft pick, level of school, and handedness. Results show that NCAA Division 1 hitters and pitchers have a higher chance of playing in the Major Leagues than High School hitters and pitchers, and right-handed hitters have a lower chance than left-handed hitters. Therefore, it is concluded that MLB organizations that draft Division 1 players give themselves the best chance for a player to make their Major League roster.

Executive Summary

This thesis analyzes the probability that a drafted baseball player will play in the Major Leagues at some point in their career, taking into account the level of school from where they were drafted. The study uses datasets purchased from The Baseball Cube and filters it to include only players who signed with the team who drafted them when they were drafted. The data are separated into two sets, hitters and pitchers, and the first 20 rounds of the Regular June Major League Baseball (MLB) Draft from 1980 to 2010 are analyzed. The paper focuses on three main factors: draft pick, level of school, and handedness, with only those three variables included in each of the models. The models use interaction terms between each of the three variables as well as a quadratic term for the overall pick. Using these inputs for each player in the dataset, the models output a number between 0 and 1 that represents the probability that the player will make the Major Leagues during their career, with higher numbers indicating an increased likeliness of making the Major Leagues.

The results indicate that players who signed out of an NCAA Division 1 school have a better chance of making the Major Leagues than those who were drafted out of High School. The paper also shows that lefty High School hitters have a better chance to make the Major Leagues than a righty hitter drafted at the same pick. Switch hitting is shown to have the biggest effect at the Junior College level, although the sample size of switch hitters is small. The changing effects of overall draft pick number also indicate NCAA Division 1 hitters clearly have the best shot at making the Major Leagues in the earlier rounds of the draft, while they are more comparable to other school levels in the middle and later rounds.

The study has significant implications for baseball organizations as they prepare for the Major League Baseball draft. Teams need to be able to compare and project players across levels

to evaluate them uniformly. By understanding the probability of a drafted player making it to the Major Leagues, teams can make more informed decisions about their draft choices. Additionally, this study highlights the importance of scouting and player development at the High School and college levels, as these are key factors in determining a player's future success in professional baseball.

Overall, this thesis provides valuable insights into the factors that contribute to a baseball player's chances of making it to the Major Leagues and the effects and differences between levels of school. The findings of this thesis can help teams make more informed decisions during the draft process and provide a foundation for further research into the predictors of success in baseball.

Table of Contents

Table of Figures

Introduction

When preparing for the Major League Baseball draft, organizations have a wide variety of players from which they can pick. Most of those players can be broken down into High School or College players, based on where they played the most recent season, and College players can be grouped even further based on the level of play. As such, teams need to be able to compare and project players across the levels to evaluate them uniformly. The data were collected on players drafted in the first 20 rounds of the first-year player draft and analyzed using logit regressions for hitters and pitchers. Ultimately, when an organization makes a selection, they are picking a player they hope can contribute to winning at the Major League level. To do that, the player must first make the Major Leagues. In this paper, I explore some factors that determine whether a player drafted will make it to the Major Leagues at some point in their career.

Literature Review

At the First-Year Player draft, MLB organizations need to decide whether they want to select a player out of high school or out of college. To do this, they need to understand the player's choice between attending college or signing with a team and how that choice will affect their career long term. In this literature review, I look at articles that discuss topics relating to players' decisions to enter professional drafts and their career performance afterwards. When compiling this review, I looked for papers that dealt with the impact of drafting players with different levels of experience and projecting future performance that could be applied or referenced in relation to MLB players and their careers.

Effects of Entering the Draft Early in Baseball

Winfree and Molitor (2007) look at the decision to enter the draft vs college for high school players, from the player's perspective. The authors argue that it may be beneficial for the player to enter the pros directly after high school. Doing so may increase their chance to make it to the Major Leagues and earn a big signing bonus, while also preventing the risk of injury that comes with playing in college. In the analysis, they use draft round as a proxy for baseball ability. Based on prior research, they determine that each additional year of schooling increases the expected non-baseball earnings by 10%.

They use 2 models, one to predict the current value of future earnings if the player enters professional baseball out of high school and one to predict the current value of future earnings if the player decides to go to college. The data used are from the June MLB drafts from 1965 to 1980, which gives a sample of 7,800 high school players. The reason for using slightly older data is so that the players' entire careers can be analyzed. The variables used include round drafted, primary position, number of years played in the Minor Leagues, year of MLB debut, an indicator variable for whether the player is from the South, number of years played in the Majors, and the team that drafted the player. The number of years played in the Minor Leagues and the year of their MLB debut were obtained from the Old Time Data Company. For financial variables, more recent data were used because the decision being analyzed is for current players (when the article was written). All the financial data are from 2003. The estimate for non-baseball earnings is taken from the US Census Bureau.

A two-stage Heckman procedure is used to estimate the signing bonus of a player who doesn't sign straight from high school, due to the high number of 0 observations that need to be censored. To model the number of years playing in the Minor Leagues, the data are split into two groups, one for players who signed out of high school and one for players who didn't sign out of high school. For the first group, OLS is used, while two-stage Heckman is used for the second. When modeling the number of years playing in MLB, two-stage Heckman is used for both groups.

The results found that it would be beneficial to play in the Minor Leagues instead of college if drafted in the 11th round or earlier. In rounds 12-14, the player has to make a decision between maximizing their earnings and maximizing their chance to play in the Major Leagues.

When looking at maximizing signing bonuses, players should attend college according to MacPhail (2011). This study attempts to determine whether players are better off signing out of high school or going to college. While this is a legitimate study done with modeling from FanGraphs, it provides a possibly different point of view than the academic papers and theses.

The article looks at players taken out of high school in the first 45 rounds of the MLB drafts from 2003-2007 who did not sign, which creates a sample of 1,341 players. The round in which they were drafted after high school was compared to the round in which they were drafted after college. Signing bonuses were estimated based on the round to compare what the players turned down out of high school and what they were offered when they were drafted out of college.

The results of this showed that players received a higher mean signing bonus when drafted out of college, which was surprising. Still, the players were drafted in a later round out of college than when they had been drafted out of high school. Even if taken in the same round, the players would be expected to receive a higher signing bonus out of college than out of high school.

However, because they have no eligibility to return to college like high school and other college players, college seniors have no leverage when negotiating their bonus, leading to a bonus that is around 80% lower than the slot amount (Gordon, 2015). The proposed solution for this problem is a minimum guarantee for college seniors to try to ensure teams can't take advantage of them.

The study looks at the history of the Collective Bargaining Agreement (CBA) and its changes over time and how they have affected college seniors' bonuses. Key variables looked at are the player's signing bonus, the slot value, and the difference between them. The data are collected from Baseball America's draft database and Newsday's baseball salaries database.

For most professions, having a college education results in higher pay throughout your career. This thesis attempts to look at whether the same is true for baseball players (Choi, 2017). Instead, Choi finds that there is no statistically significant effect on salaries for either hitters or pitchers who go to college vs those who don't. But when the sample of players is limited to starting MLB players, a college education lowers the expected salary of hitters by approximately 21%.

The three types of models used in Choi's study were a Heckman Correction Regression, a Probit regression, and a 2SLS regression. The data used in this study are primarily obtained from Sean Lahman's database and is limited to players born in or after 1960. The key variables used for hitters were OBP, SLG, and SB, and each was compared to the average at that player's age. There was also a 1/0 variable for whether the player batted right-handed. For pitchers, the variables were K%, BB%, HR/9, BABIP, FIP, and saves, again compared the mean by age, as well as whether the pitcher was a reliever and righty. Variables that were used for both hitters and pitchers included their year 7 real salary, the year, if they were predicted to go to college, debut age, and 1/0 variables for whether they went to college and whether they were born in the United States.

Hubley (2012) looks at how teams value and project players at the Major League Baseball draft. Signing bonuses have risen dramatically throughout the years and this study attempts to find whether that is due to teams bidding each other up or if it is because teams can accurately assess which players will perform the best and offer those players more money. The results found a positive and statistically significant correlation between signing bonus and future MLB performance. They also found that the effect is larger for pitchers than position players, even though they are generally considered riskier prospects at the draft. When comparing high school and college picks, there is no significant difference in the size of the effect of the signing bonus, indicating that teams don't have a significantly better ability to scout one or the other in terms of future MLB success. However, when looking at the relationship between signing bonus and chance to make MLB, which is significant, there is a statistically significant difference between high school and college picks. The effect of an increase in signing bonus for a college player is associated with about a 4% higher chance than for a high school pick to make the Majors.

The data were collected from Baseball Reference and the final dataset included 3,370 players. Key variables in the data used in the model include the signing bonus, the round in which the player was drafted, number of major league service years, number of MLB games, the position they play, and whether they played in college and at what level. The model used was an OLS model to measure future WAR production at MLB level. Similar models were also run for whether they made MLB and whether they made an All-Star Game.

Spurr (2000) looks at the draft and analyzes how good teams are at evaluating talent and whether the player will make the major leagues. The results found that there was no difference between how different organizations discover talent. They also found that a player was more likely to make the majors if they went to college, especially a college with a good baseball program. At first, there was a market inefficiency surrounding college players, but it was corrected.

Data were collected on players from the June draft from 1965-1983 in various degrees, from where is not specified. Key variables included the draft pick number, if and where the player went to college, their position, and who drafted them. Probit models were run for 1966-1968.

Other Draft Effects in Baseball

Wymore, Chin, and Geary (2016) look at pitchers who had UCL reconstruction surgery prior to being drafted to see whether they have different career paths than those who didn't. The results found that the different groups of pitchers did not reach the majors at statistically significantly different levels and their stats were not different either, although pitchers who had the surgery had an increased rate of being placed on the injured list.

Chi-square tests were used to compare the groups for categorical data and independent-samples t tests were used for the continuous data. Mann-Whitney U tests were also used if the data had nonnormal distributions. Key variables considered were where the player was drafted from, age/height/weight demographics, what level they made it to, and professional stats. The data were collected from the MLB Medical Committee and MLB database.

Sims and Addona (2016) look at how a player's birthdate compared to other players their age affects their chance of making the majors and performance if they make it. Because Little League has a cutoff date of July 31, players born August 1 would be the oldest among players their age. The results found that age and relative age have a negative correlation with making the majors, which indicates that younger players were undervalued. Once in the majors, age/relative age were not significant.

The data were collected from Baseball Reference and The Baseball Cube and key variables included are position, where they were drafted from, when they were drafted, birthday information, and performance statistics. A logistic regression is used to model the probability of making the majors, and a truncated negative binomial regression is used to model the number of games played in the majors, if they make it.

Barden and Choi (2021) look at how teams make decisions on which players to draft in the first year MLB player draft. The results find that having a disadvantage at payroll promotes risk-taking for the chance of hitting big, and on-field success promotes risk-taking early in the draft. This results in a U-shape relationship between pick number and risk taking.

Logistic regressions and standard two-stage least squares are used to model in this project. The key variables are whether the player was drafted out of high school, per capita income of the metro area, team payroll, and round of pick, and the data were gathered from a combination of LexisNexis, MLB, Wikipedia, US Bureau of Labor Economics, and Baseball Reference.

On the front office side of drafting, Roach (2022) looks at the incentives for Front Office personnel at the draft outside of drafting the best player available. Because GMs are judged on the success of their draft classes, they might overvalue players who can reach the majors sooner. The results find that teams who underperform in the previous season tend to draft fewer high school players in the next draft.

The data are collected from Baseball Reference. Key variables looked at include the performance of the team using wins and Pythagorean wins, number of HS and total draft picks, the length of tenure for the GM, and whether the GM was forced out. Several linear regressions were run with that data for the modeling.

Burger and Walters (2009) look at expected return for picks in the baseball draft. The results find that even though many prospects don’t live up to their draft day hype, the ones who do more than make up for the bonuses they were paid. They also found that teams didn't value college players enough compared to high schoolers and overvalued pitchers compared to position players.

The main model used was a log-lin regression looking at the bonus the player was offered. Key variables in the model were the slot value, the year, whether the player went to college, whether the player was a pitcher, and what market the team was in. Data were shared by Jim Callis, who was a talent evaluator and an executive editor of a baseball journal.

Effects of Entering the Draft Early in Other Sports

Garnes (2019) looks at NBA players since the rule changed that required at least one year of high school that have been "One and Done." Their careers are analyzed to see if the rule affected performance or money earned. On average, players who were drafted directly out of high school played 7 more seasons in the NBA than those who went to college for a season (12 seasons vs 5). While the points per game for each group were similar, the high school group was slightly higher and the difference was not statistically significant. Rebounds and assists per game were also similar between the groups. Players drafted out of high school were shown to have a significantly higher career and average yearly salary, but there are a number of factors that could be influencing those numbers such as their careers being finished and inflation.

The data for this paper were collected from Basketball Reference. Key variables collected included salaries, length of career, draft year, and performance stats for players drafted out of high school and after one year in college for players from 1995 to 2017. The models used in this study were pretty simple. They looked at averages for different stats between the 2 groups and then compared their statistical significance with t-tests.

Groothuis, Hill, and Perri (2007) also look at early entry into the NBA draft. Their article looks at players entering the NBA draft and how trends changed with new rules being introduced. They look at whether a player is drafted early because they are talented and don't need as much experience as other players or because they have "signaled" that they are good enough against college competition. The results show that players who enter the draft early play fewer minutes but improve quicker. Because the length of the rookie contract was extended, teams and players have incentive to for the players to enter the draft early.

Data were collected from numerous sites including the official NBA encyclopedia and the NBA website, among others. Variables examined include points, rebounds, assists per game, among other performance statistics, a dummy variable for rookie contract, years of experience, and draft pick. The study uses a log-linear model to look at the effect of the rookie scale salary.

When looking at the league as a whole instead of the players, Grier and Tollison (1994) attempt to find whether drafts increase competitive balance, using the NFL as a case study. They look at teams across years and compare their win percentage to the previous years' draft position. The results find that the draft does its job and tends to promote competitive balance by evenly distributing new talent in the league. Although the location of the data is unspecified, they collected data on winning percentage and draft position, before trades, across several years. The study used a time series model.

Farah and Baker (2021) look at NHL draft picks and their future performance. Players were split up by position and future performance for forwards was accurate in the first two rounds, while not for the rest of the draft or defensemen. This shows inaccuracies in the drafting process and a non-linear relationship between pick number and future performance.

The data were primarily collected from coriscahockey.com. The study used Kruskal-Wallis H Tests and Mann-Whitney U Tests with variables of various performance statistics from offense and defense and the pick number in the draft.

Mulholland and Jensen (2016) look at the projection side of drafting players. Their article attempts to evaluate pass catchers in the NFL, such as WR and TE. The results found that college performance has a lot of predictive power for WRs for their pro careers, along with size and strength factors as well.

The data were gathered from publicly available sources, including nflcombineresults.com, Pro Football Reference and College Football Reference. The key variables that were collected were physical attributes, such as heigh and weight, the college conference they played in, combine stats like the 40-yard dash, and receiving stats in college. Regression models and partitioning decision trees were created for each position, each done multiple times with different outcomes variables.

Projecting Future Performance

Bradbury (2009) attempts to isolate age from other factors to determine an aging curve for hitters and pitchers in Major League Baseball. This could be important when comparing players who might have skipped college because they will be younger when they enter professional baseball than players who go to college. Due to the ability to be able to compete in baseball with less athleticism, it is hypothesized that the peak of baseball players could be later than in other sports.

The data used in this paper are from 1921 to 2006 for players debuting after 1920. The data are obtained from Sean Lahman's database. Limitations set on players in the data set include being between 24 and 35 years old, a minimum of 5000 plate appearances for batters, and a minimum of 4000 batters faced for pitchers. The paper uses a Baltagi and Wu random-effects method for their model to estimate the effects of age.

The results show that a batter’s peak is 28 years old, with 14% of players having their best season in that year and 52% having their best year between 27 and 31. The most common age for a pitcher’s peak was also 28, while the mean was 30.38 years old. Both samples have the mean at a higher age than the mode, which could be due to the chance of injury limiting players from reaching a later peak.

Albert (2002) also looks at how players progress and regress throughout their careers based on age and finds similar results. This article compares hitters at different ages and finds that the median peak age was between 27.1 and 29.8 for each of the 6 decades analyzed, with the above peak of 28 years old right in the middle of that range.

The data were acquired from Lahman’s baseball database and included a total of 473 players. A quadratic regression is used to find the maximum linear weights and the ages at which that value occurs. After, a Bayesian Exchangeable model was run to combine the individual results and attempt to find better results. The variables used included a linear weight formula for singles, doubles, triples, homer, and walks/HBP.

Chandler and Stevens (2012) look at future success of Minor League players. They found that stats such as GIDP/AB, HBP/AB, and BB/AB at the AAA level have little predictive power to what a player will do in MLB. Stats such as AVG, OBP, SLG, and OPS have much more predictive power in AAA than at the lower Minor League levels. They did not find any stats that are consistently high in MLB production predictive power throughout the entire ranks of the Minor Leagues, which would have been useful to Major League organizations.

The data for this paper comes from The Baseball Cube and the My MLB Draft website. Key variables in the study were stats at the AAA level, ML contribution, which was a binary value that represented whether a player established themselves at MLB level, and stats at other levels of the Minor Leagues. Because not every player plays at every Minor League Baseball (MiLB) level, the players were split into different groups based on which levels they played at. A classification tree is used to look at whether or not players had Major League success and whether they were promoted beyond the level or not. Then, using that classification tree, they use the Random Forest technique.

McLeod, Pifer, and Plunkett (2021) look at what the players themselves think of their future as opposed to their stats. Athletes often have unrealistic expectations for how their career will go, which can affect their decision making when signing contracts. This study compares athletes' expectations with a model's expectation about their career and attempts to look at how athletes react to information that estimates how they will do in their career. The results found that athletes were overly optimistic about their career success compared to the model's. After seeing the model, most athletes updated their predictions, with the biggest difference being in the chance that they are completely out of Major League affiliated baseball.

Interviews were conducted with MiLB players where they drew a career tree and compared it to a personalized C5.0 career tree. Key variables that were used for the C5.0 model included biographical information, performance stats, what level of the minors they were in, and their position, which were collected from The Baseball Cube.

Another way to evaluate MiLB players is their ranking, which is what Tymkovich (2012) does. His study looks at Minor League prospects and the correlation between their ranking and future MLB success they might have. The results found a high correlation between the prospect ranking and success, regardless of which ranking site was looked at. Players who went to college were also found to produce more WAR than players drafted out of high school or signed as free agents.

The data came from Baseball America, Baseball Prospectus, Minor League Ball, ESPN, and FanGraphs. Key variables looked at were the prospect ranking, age, whether they went to college, level of Minor League, position, and organization, as well as various performance statistics. Regressions were run for each prospect ranking site to compare their rankings to future success.

Benavidez, Brito, McCreight, and McEvoy (2019) attempt to project a position player's value in the next year given their performance in the previous years. While the model worked occasionally and made progress in accomplishing the goal, it was not as accurate as the authors would have liked due to various factors. Models used in the study included unregularized linear regression, regularized linear regression, support vector regression, and recurrent neural network with long short-term memory. The data collected were from Baseball Reference and included performance stats such as PA, G, Innings, WAA, WAR, and OPS+, and salary.

In addition to ranking and stats, players' intelligence could be looked at to see whether there's an effect on their future performance. Bowman, Boone, Goldman, and Auerbach (2021) do this by looking at the Athletic Intelligence Quotient to see whether it has any effect on baseball players' performance. The results find that AIQ had a statistically significant relationship with both pitching and hitting stats, after controlling for other variables. The players' performance statistics were gathered by a MLB team and made available to the authors of the paper. Key variables looked at in the model include batting stats such as AVG and OPS, pitching stats such as WHIP and FIP, as well as the players AIQ score. Regressions were run separately for each of batters and pitchers.

Chaudhari, McKenzie, Borchers, and Best (2011) try to find whether there is correlation between pitcher balance and their in-game pitching performance for Minor League pitchers. The results find a statistically significant effect from pitchers who have higher lumbopelvic control, so a test might help scouts identify players with a better chance of pitching success.

T-test statistical comparisons were used for each variable to compare the 2 groups. Chi-squared test was performed to determine whether the number of injuries between the 2 groups was different. Key variables were the results to the level belt test and pitching stats from during the season, such as IP, WHIP, K/inning, etc. Level belt tests were performed and collected at spring training, pitching stats were collected from Minor League Baseball's official website

In other sports, Sun, Yu, and Centeno (2017) attempt to project college players' 3PT shooting in the pros when they are drafted. The model ended up working well for average players, but there were many outliers that caused the model to be unreliable. Using more data could help correct some of the shortcomings of the model, but it was still able to work for average players at this level.

4 models were used. First, a linear regression was run as a baseline. Then they ran a weighted linear regression with emphasis on points with less variance. Third was a random forests regression and finally a gradient boosted regression. Key variables in the models were the players

3pt attempts, their 3pt%, their FT%, SOS, and offensive stats from the team that drafted them. The data were gathered from nbadraft.net's big boards.

Other Related Articles

In a study, Miyamoto and Ito (2017) compare younger and older pitchers in their pitching style to determine whether there are significant differences. As pitchers get older and can't throw as hard as they used to be able to, they need to find different ways to prove that they should still have a spot on the roster. In this study, they found that younger pitchers choose to throw straighter pitches more often, while older pitches rely on pitches that have more movement. Although, these are very different types of pitches, they both saw similar success in striking batters out with 2 strikes.

For each of the 13 pitch types defined in the article, a different regression analysis was performed. Pitchers were split into one of two categories: young (20 to 30 years old) or old (31 to 43 years old). Pitch by pitch tracking data was the main data used. The data were gathered from PITCHf/x from the 2015 MLB season.

Krautmann, Gustafson, and Hadley (2000) suggest that one reason for low payment in the Minor Leagues is to make back money from training expenses. It looks at the difference between the value provided by a player and what they are paid. They find that surplus is only obtained from players before arbitration, and more is obtained from star players than mediocre players. They also find that more surplus value is extracted from minorities than white players.

The Minor League stats were gathered from The Sporting News Baseball Register. Free Agency data was gathered from the Chicago Tribune and USA Today. Key variables collected include salary, experience, draft, college, position, and performance statistics, mainly focusing on

RBI. Various regressions and log-lin regressions are run to determine value, cost, and surplus value for players.

Ehrilch and Potter (2021) look at whether winning is more “valuable” with a better offense or with better pitching. This is relevant because it would incentivize teams to strengthen that side of the team if fans preferred one over the other. Results find that fans don’t have a preference for winning one way or the other, just winning in general.

3 sets of regressions are used. OLS Fixed Effects are used with Hausman, Wooldridge and Frees tests. Key variables in the models were offensive WAR, defensives WAR, pitching WAR, lucky wins, whether the team was playing in a new stadium, GDP per Capita and population for the local metro area, and a variable for the tv deal. The data were collected from the Bureau of Labor Statistics, US Census Bureau, StatCan, Wikipedia, Fangraphs, and Baseball Reference.

Conclusion

These papers found that as expected, players who stay in school longer, will on average increase their non-sport earning potential significantly, but the best athletes who enter the drafts early will make more money playing professional sport. Another important conclusion is that most players think they have a better chance at being a successful MLB player than they actually do, which could influence their choice on signing a professional contract or attending college. On the other side, many teams overvalue high school players and pitchers coming into the draft. With the length of the MLB draft changing in recent years, many of these conclusions could be changing as the value on certain types of players may change in MLB front offices.

Discussion of Data

Datasets were purchased from The Baseball Cube and merged together by draft pick for each player. Because players could be drafted multiple times which would create duplicates, I filtered the final dataset to only include players who signed when they were drafted. I separated players into two sets, hitters and pitchers. This was done by looking at whether they pitched or hit more in the Minor Leagues, as opposed to what they were drafted as, due to some players being drafted as two-way players or at multiple positions. I filtered the data to only include the first 20 rounds of the Regular June MLB Draft from 1980 to 2010, inclusive. The reason the data is cut off in 2010 is to allow time to observe the players' careers and whether they have made the Major Leagues. I decided to only look at the first 20 rounds due to MLB permanently shortening the draft to 20 rounds starting in 2020.

The final merged datasets included 5 categories of data, including draft information (year, round, pick number, etc.), biographical (name, height, weight, bat and throw hand, etc.), school-related (name, level, location), and career Minor League and career Major League stats. Level of school was split into four categories, which were the same for hitters and pitchers. The simplest categories were High School ("HS") and NCAA Division 1 ("NCAA D1"), which were each their own level. Two-year junior colleges ("Junior Colleges") were grouped together and consisted of the National Junior College Athletic Association (NJCAA), the California Community College Athletic Association (CCCAA), and the Northwest Athletic Association of Community Colleges (NWAACC). The remaining 4-year schools were NCAA Division 2, NCAA Division 3, and the National Association of Intercollegiate Athletics (NAIA), and they were grouped together for the last category ("Other 4 Year"). In figures 1 and 2 below are summary statistics for each of the sets of players for some of the stats used in the later models.

Summary Statistics for Hitters In Dataset						
	Count	**Minimum**	**Median**	**Mean**	**Maximum**	**Std Dev**
Round	7561	1	9	9.79	20	5.848
HS	2380	1	6	7.68	20	5.629
NCAA D1	3825	1	10	10.20	20	5.755
Other 4 Year	765	1	13	12.43	20	5.322
Junior Colleges	591	1	13	12.25	20	5.221
Overall	7561	1	267	274.57	634	167.99
HS	2380	1	176	214.60	624	162.00
NCAA D1	3825	1	285	286.99	634	164.80
Other 4 Year	765	10	360	340.27	630	152.64
Junior Colleges	591	1	363	350.69	621	150.55
Majors	1803	0	0	0.24	1	0.4262
HS	573	0	0	0.24	1	0.4276
NCAA D1	1022	0	0	0.27	1	0.3882
Other 4 Year	99	0	0	0.13	1	0.4426
Junior Colleges	109	0	0	0.18	1	0.3359
MLB Games	7561	0	0	127.89	3080	372.50
HS	2380	0	0	149.24	2784	412.63
NCAA D1	3825	0	0	108.48	3080	362.03
Other 4 Year	765	0	0	131.63	2986	368.15
Junior Colleges	591	0	0	57.72	2023	237.86

Figure 1: Summary Statistics for Hitters

Summary Statistics for Pitchers In Dataset						
	Count	**Minimum**	**Median**	**Mean**	**Maximum**	**Std Dev**
Round	6904	1	9	9.61	20	5.801
HS	1855	1	7	8.04	20	5.655
NCAA D1	3515	1	9	9.62	20	5.856
Other 4 Year	787	1	12	11.53	20	5.253
Junior Colleges	747	1	12	11.48	20	5.272
Overall	6904	1	261	273.48	633	168.57
HS	1855	1	199	228.71	621	164.08
NCAA D1	3515	1	264	273.56	631	170.20
Other 4 Year	787	4	335	325.51	633	154.19
Junior Colleges	747	4	331	329.47	623	153.48
Majors	1934	0	0	0.28	1	0.4491
HS	568	0	0	0.31	1	0.4610
NCAA D1	1059	0	0	0.30	1	0.4589
Other 4 Year	130	0	0	0.17	1	0.3716
Junior Colleges	177	0	0	0.24	1	0.4255
MLB Games	6904	0	0	45.75	1178	125.33
HS	1855	0	0	51.85	1042	129.73
NCAA D1	3515	0	0	48.14	1119	126.64
Other 4 Year	787	0	0	33.37	1058	125.53
Junior Colleges	747	0	0	32.39	1178	104.50

Figure 2: Summary Statistics for Pitchers

To better understand the data, I looked at the distribution of level of school in each round. Figures 3 and 4 below show how the percentage of players from each level changes as the draft progresses. Both figures show that the percentage of players who sign as High Schoolers starts out at its highest in the earlier rounds, before falling later. The opposite trend is seen for players from smaller 4-year schools and Junior Colleges, while the percentage of D1 players stays relatively consistent throughout.

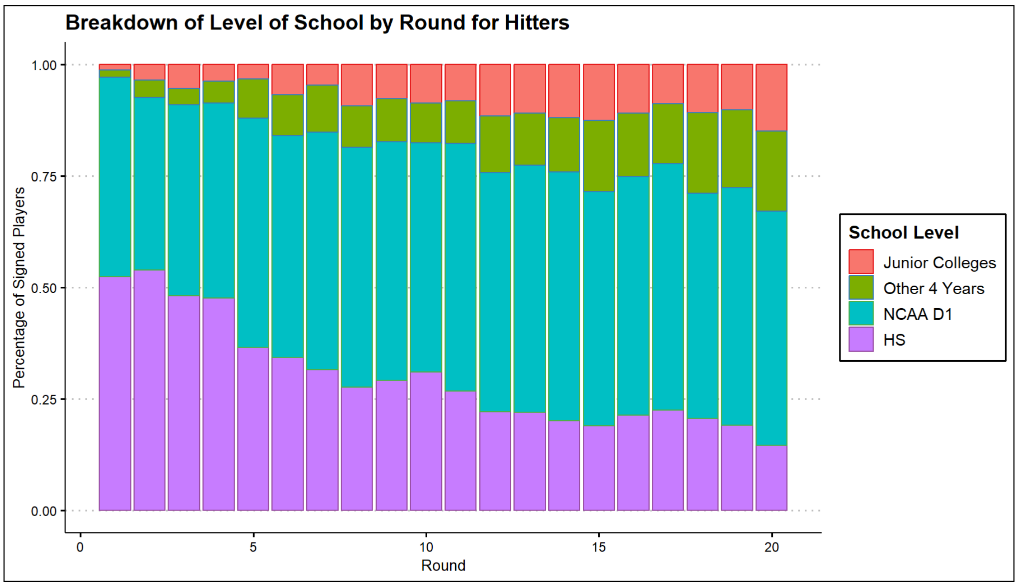

Figure 3: Breakdown of Level of School by Round for Hitters

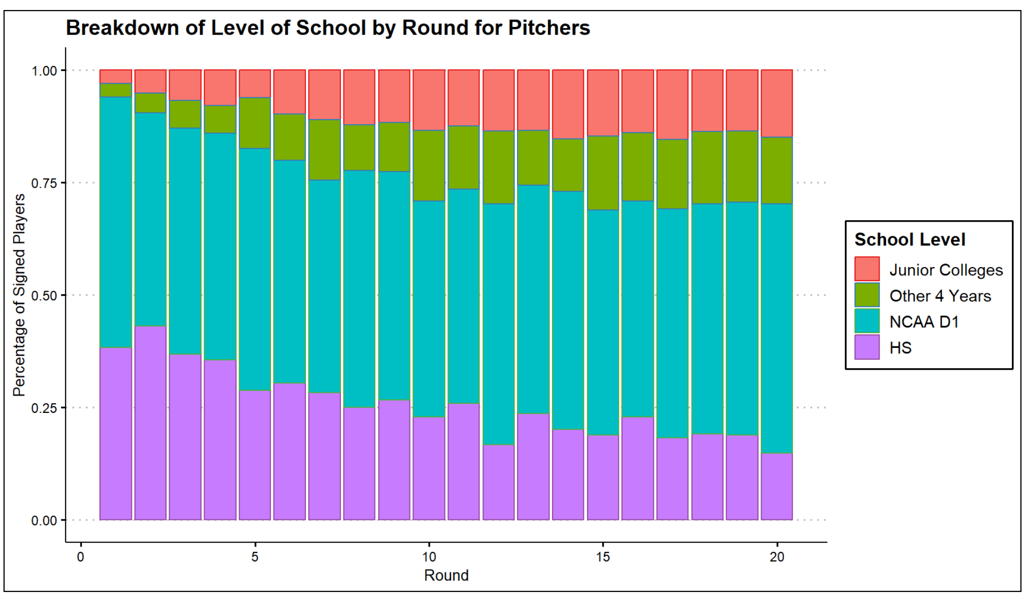

Figure 4: Breakdown of Level of School by Round for Pitchers

Models

Only three factors were included in each of the models: draft pick, level of school, and handedness. For hitters, handedness was determined by where they batted, either right-handed (labelled R), switch (S), or left-handed (L). For pitchers, it was determined by their pitching hand, righty (R) or lefty (L). The only "switch-pitcher" in the sample was Pat Venditte, who I excluded from the dataset. The logit models used interaction terms between each of the three variables as well as a quadratic term for the overall pick. For the results, a positive coefficient is expected for the NCAA D1 dummy variable, which would indicate a better chance of making the Major Leagues having gone to a Division 1 school than being drafted out of High School. A negative coefficient is expected for the overall term, meaning as it increases (i.e., later picks in the draft), the player is less likely to make the Major Leagues in their career.

Hitter Model Results					
Variable	**Interaction**	**Coefficient**	**Odds Ratio**	**P-Value**	
Intercept		0.557	1.745	0.000	***
NCAA D1		0.922	2.515	0.000	***
Other 4 Year		-0.118	0.889	0.784	
Junior Colleges		-0.222	0.801	0.605	
Overall		-0.013	0.987	< 2e-16	***
Overall^2		1.55E-05	1.000	0.000	***
Bats R		-0.371	0.690	0.006	**
Bats S		-0.281	0.755	0.161	
NCAA D1	* Overall	-1.85E-03	0.998	0.221	
Other 4 Year	* Overall	1.02E-03	1.001	0.738	
Junior Colleges	* Overall	3.78E-03	1.004	0.213	
NCAA D1	* Overall^2	-8.03E-07	1.000	0.774	
Other 4 Year	* Overall^2	-3.02E-06	1.000	0.553	
Junior Colleges	* Overall^2	-6.62E-06	1.000	0.176	
NCAA D1	* Bats R	0.204	1.226	0.175	
Other 4 Year	* Bats R	0.098	1.103	0.734	
Junior Colleges	* Bats R	0.245	1.277	0.393	
NCAA D1	* Bats S	0.293	1.341	0.217	
Other 4 Year	* Bats S	-0.277	0.758	0.600	
Junior Colleges	* Bats S	0.738	2.091	0.094	.
Bats R	* Overall	2.43E-04	1.000	0.580	
Bats S	* Overall	1.08E-04	1.000	0.880	

n = 7561

Figure 5: Hitter Model Results Table

For the hitter model, all of the results are compared to a left-handed hitter drafted out of High School and shown in Figure 5 above. While some of the direct interpretations are difficult due to the many interaction terms, there is a clear positive effect for players who signed out of a Division 1 school instead of High School. Another notable result is the negative and statistically significant coefficient for right-handed batters. This indicates that lefty High School hitters have a 31% better chance to make the Major Leagues than a righty hitter drafted at the same spot. This effect decreases slightly later in the draft due to the positive coefficient on the interaction between R and overall pick. From the interaction between handedness and level, switch hitting is shown to have the biggest effect at the Junior College level, with slight significance. Both the overall and its quadratic term are significant, but due to the high number of picks, the effect of a single pick is hard to see. To better understand this, the marginal effects of increasing the pick are shown in Figure 6 below.

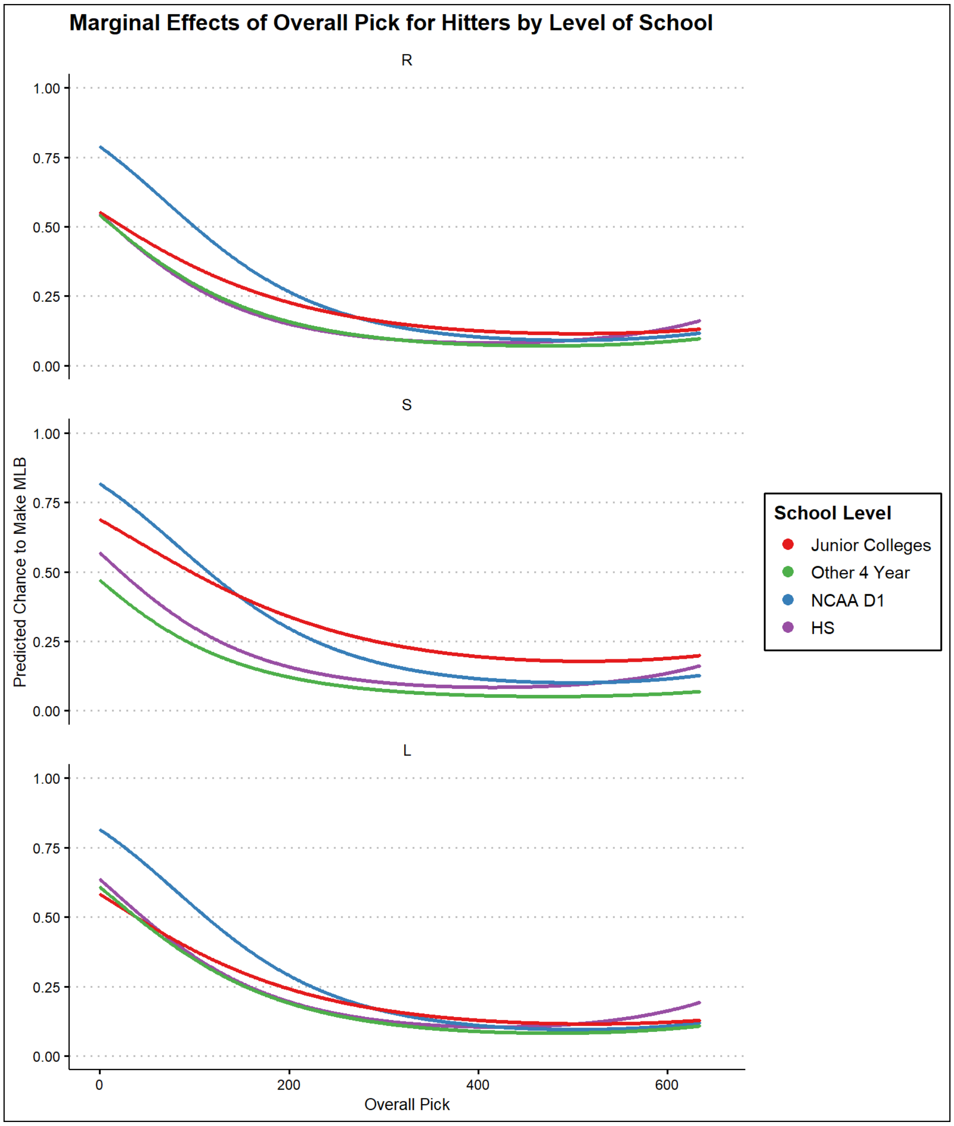

Figure 6: Marginal Effects of Overall Pick for Hitters by Level of School

In this figure, NCAA Division 1 hitters clearly have the best shot at making the Major Leagues for the first ~6 rounds for righty and lefty hitters, but there is more variance among switch hitters, likely due to the sample size in the dataset. In the middle to late rounds of the draft, there isn't a significant difference between any of the levels for any handedness. At the end of the draft, High School players' chances increase over other levels, which could be due to some "lottery picks" becoming MLB players among a smaller sample, as the number of High School players drafted is at its lowest in the final rounds, as shown in an earlier figure.

Pitcher Model Results					
Variable	**Interaction**	**Coefficient**	**Odds Ratio**	**P-Value**	
Intercept		0.709	2.032	0.000	***
NCAA D1		0.559	1.749	0.004	**
Other 4 Year		-0.112	0.894	0.778	
Junior Colleges		0.021	1.021	0.955	
Overall		-0.011	0.989	< 2e-16	***
Overall^2		1.21E-05	1.000	0.000	***
Throws R		-0.277	0.758	0.050	*
NCAA D1	* Overall	-7.49E-04	0.999	0.616	
Other 4 Year	* Overall	1.01E-03	1.001	0.716	
Junior Colleges	* Overall	1.87E-03	1.002	0.466	
NCAA D1	* Overall^2	-1.01E-06	1.000	0.708	
Other 4 Year	* Overall^2	-4.59E-06	1.000	0.332	
Junior Colleges	* Overall^2	-1.03E-06	1.000	0.803	
NCAA D1	* Throws R	-0.195	0.823	0.190	
Other 4 Year	* Throws R	-0.153	0.859	0.563	
Junior Colleges	* Throws R	-0.541	0.582	0.019	*
Throws R	* Overall	1.15E-03	1.001	0.005	**

n = 6904

Figure 7: Pitcher Model Results Table

For the pitcher model, the results are relative to a left-handed High School pitcher. There is still a positive and statistically significant effect for going to a Division 1 school, although the magnitude of the effect is less than with hitters. Both the Right Handed Pitcher (RHP) variable and interaction term between RHP and overall are significant with opposite signs, indicating that earlier in the draft, lefty pitchers are more likely to make the Major Leagues, but later, it is righties with the advantage. For High School pitchers, this switch occurs around the 242nd pick of the draft – just under halfway through the dataset. This is true for all levels other than Junior Colleges, where the interaction term with RHP is the largest and significant, leading to lefties always having a higher predicted chance to make the Major Leagues. The difference in handedness is shown further in Figure 9 below. In Figure 8, the marginal effect of overall pick by level are again shown, similar to Figure 6 for hitters above.

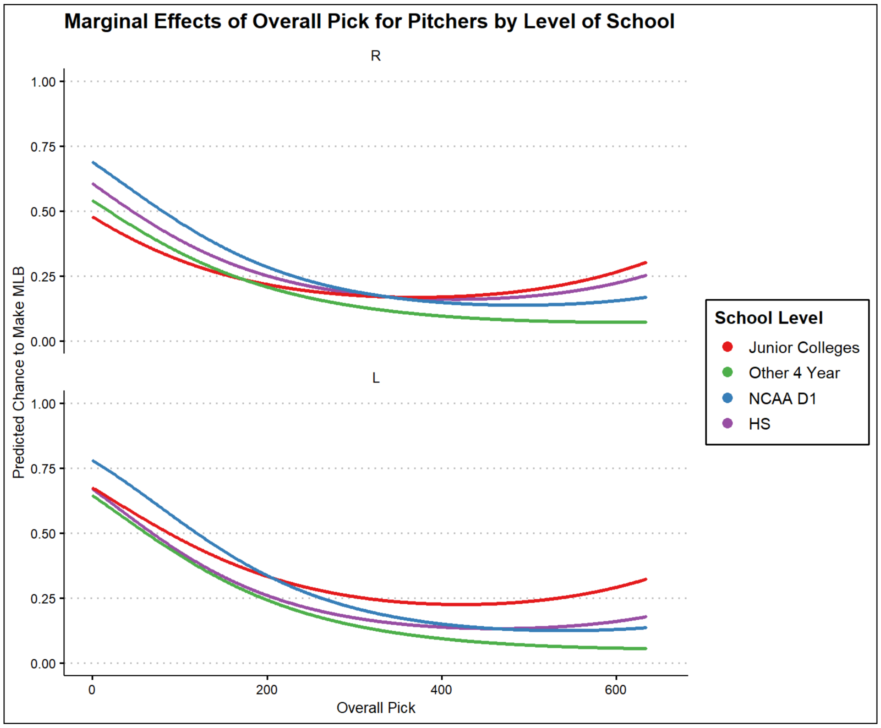

Figure 8: Marginal Effects of Overall Pick for Pitchers by Level of School and Handedness

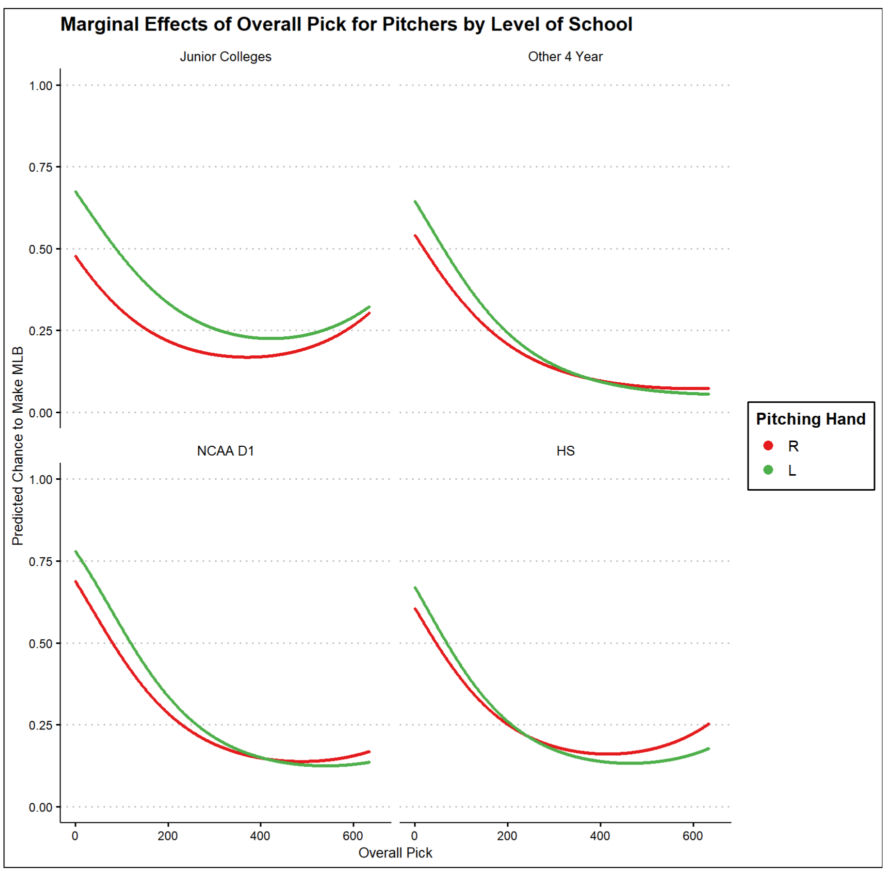

Figure 9: Marginal Effects of Overall Pick for Pitchers by level of School and Handedness

While Division 1 pitchers still have the best chance to make the Major Leagues in their careers, the gap is much smaller than among hitters. Additionally, Junior College pitchers make a peak in the later rounds along with High Schoolers. In the handedness figure, the switch from lefty advantage to righty advantage happens first for High School, as mentioned above, followed by Other 4 Year schools and Division 1. For Junior Colleges, RHP never quite reach the level of Left Handed Pitchers (LHP), although the difference is small at the end of the draft.

Discussion and Conclusion

The percentage of players who are Division 1 remains consistent through the first 20 rounds, but they have a much higher chance of being promoted to the Major Leagues than the other levels in the early rounds of the draft, suggesting an inefficiency in drafting strategies. Another inefficiency is revealed in that left-handed players are more likely to make the Major Leagues than right-handed players drafted at the same spot consistently throughout the draft. For most of the scenarios in the above models, the curve for High School players is similar to that of the smaller 4-year schools and Junior Colleges, with the major difference being the age of the player. For organizations who want the best chance for their draft picks to contribute at the MLB level in the near future, it would make more sense to draft a college player. However, if the organization trusts its player development more than that of colleges, they could opt to draft a High School player and train them for a specific role needed in the organization. Additionally, because the models use overall pick to evaluate the players, organizations should still have a "big board" of prospects they are looking at, using the identified inefficiencies in addition to those criteria. The results of the models are not to say a projected 5th round left-handed hitter should be drafted over a projected 1st round High Schooler, rather that if a team is comparing a Division 1 player to a Junior College player in the middle of the 1st round, the Division 1 player is more likely to eventually make the Major Leagues.

The results of the models for both hitters and pitchers suggest that NCAA Division 1 players have an increased chance of making the Major Leagues compared to other levels of player. For MLB organizations, drafting Division 1 players gives them the best chance at a player who will make their MLB roster. From the player's perspective, if a High School player is able to continue playing at an NCAA Division 1 school or believes they would be able to move up in the draft by attending a different college, they are likely increasing their value to MLB organizations.

References

Albert, J. 2002. “Smoothing career trajectories of baseball hitters”. Bowling Green, OH Unpublished manuscript, Bowling Green State University

Barden, J. Q., & Choi, Y. (2021). Swinging for the Fences? Payroll, Performance, and Risk Behavior in the Major League Baseball Draft, Journal of Sport Management, 35(6), 499-510. from https://journals-humankinetics-com.libezproxy2.syr.edu/view/journals/jsm/35/6/article-p499.xml

Benavidez, S., Brito, S., McCreight, D., & McEvoy, P. (2019, December 13). Prediction of future offensive performance of MLB position players. Retrieved from http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26585755.pdf

Bowman JK, Boone RT, Goldman S and Auerbach A (2021) The Athletic Intelligence Quotient and Performance Outcomes in Professional Baseball. Front. Psychol. 12:629827. doi: 10.3389/fpsyg.2021.629827

Bradbury, John Charles (2009) Peak athletic performance and ageing: Evidence from baseball, Journal of Sports Sciences, 27:6, 599-610, DOI: 10.1080/02640410802691348

Burger, J. D., & Walters, S. J. K. (2009). Uncertain Prospects: Rates of Return in the Baseball Draft. Journal of Sports Economics, 10(5), 485–501. https://doi.org/10.1177/1527002509332350

Chandler, Gabriel and Stevens, Guy. "An Exploratory Study of Minor League Baseball Statistics" Journal of Quantitative Analysis in Sports, vol. 8, no. 4, 2012. https://doi.org/10.1515/1559-0410.1445

Chaudhari, A. M., McKenzie, C. S., Borchers, J. R., & Best, T. M. (2011). Lumbopelvic control and pitching performance of professional baseball pitchers. Journal of strength and conditioning research, 25(8), 2127–2132. https://doi.org/10.1519/JSC.0b013e31820f5075

Choi, Kyung Wan. College Education and Salary in Major League Baseball: Does College Education Lead to Higher Salaries for Major League Baseball Players?. California State University, Fullerton, 2017.

Ehrlich, J. A., & Potter, J. M. (2021). Is offense worth more than defense and pitching? Marginal revenue product and revenue sharing in major league baseball. [Is offense worth more than defense and pitching?] Managerial Finance, 47(6), 760-778. https://doi.org/10.1108/MF-04-2020-0210

Farah, L, Baker, J. Accuracy from the slot: Evaluating draft selections in the National Hockey League. Scand J Med Sci Sport. 2021; 31: 564– 572. https://doi.org/10.1111/sms.13874

Garnes, K. (2019). The Monetary and Social Opportunity Cost of Entering the Nba Draft Directly from High School Versus Completing One Year of College Before Entering the Draft (thesis).

Gordon, Jonathan C., Foul Ball: Major League Baseball's CBA Exploits College Seniors in the

MLB Draft (April 6, 2015). 16 Tex. Rev. Ent. & Sports L. 141 (2015), Available at SSRN: https://ssrn.com/abstract=2590630

Grier, Kevin B., Robert D Tollison, The rookie draft and competitive balance: The case of professional football, Journal of Economic Behavior & Organization, Volume 25, Issue 2, 1994, Pages 293-298, ISSN 0167-2681, https://doi.org/10.1016/0167-2681(94)90016-7.

Groothuis, P. A., Hill, J. R., & Perri, T. J. (2007). Early Entry in the NBA Draft: The Influence of Unraveling, Human Capital, and Option Value. Journal of Sports Economics, 8(3), 223–243. https://doi.org/10.1177/1527002505281228

Hubley, Bobby. (2012). Signing Bonuses & Subsequent Productivity: Predicting Success in the MLB Draft.

Krautmann, AC. & Gustafson, E. & Hadley, L., 2000. "Who pays for minor league training costs?," Contemporary Economic Policy, Western Economic Association International, vol. 18(1), pages 37-47, January.

MacPhail, R. (2011, April 7). Better to sign out of HS or college? part 1. FanGraphs Baseball. from https://blogs.fangraphs.com/better-to-sign-out-of-hs-or-college-part-1/

McLeod, Christopher M., N. David Pifer, Emily P. Plunkett, Career expectations and optimistic updating biases in minor league baseball players, Journal of Vocational Behavior, Volume 129, 2021, 103615, ISSN 0001-8791, https://doi.org/10.1016/j.jvb.2021.103615.

Mulholland, Jason & Jensen, Shane T. (2016) Projecting the Draft and NFL Performance of Wide Receiver and Tight End Prospects, CHANCE, 29:4, 24-31, DOI: 10.1080/09332480.2016.1263095

Miyamoto, Michiko, and Akihiro Ito. "Analysis of Pitching Skills of Major League Baseball Players." International Conference on Applied Human Factors and Ergonomics. Springer, Cham, 2017.

Roach, M.A. (2022) Career concerns and personnel investment in the Major League Baseball player draft. Economic Inquiry, 60(1), 413– 426. Available from: https://doi.org/10.1111/ecin.13012

Sims, J., &; Addona, V. (2016, October). Hurdle models and age effects in the major league baseball draft. Retrieved from https://journals.sagepub.com/doi/full/10.1177/1527002514539516

Spurr, S. J. (2000). The Baseball Draft: A Study of the Ability to Find Talent. Journal of Sports Economics, 1(1), 66–85. https://doi.org/10.1177/152700250000100106

Sun, H., Yu, J., & Centeno, R. (2017, December 16). Projecting three-point percentages for the NBA draft - Stanford University. Retrieved from https://cs229.stanford.edu/proj2017/final-reports/5238212.pdf

Tymkovich, Jay Lyon, "A Study of Minor League Baseball Prospects and Their Expected Future Value" (2012). CMC Senior Theses. Paper 442.

Winfree, Jason & Molitor, Christopher. (2007). The Value of College: Drafted High School Baseball Players. Journal of Sports Economics - J SPORT ECON. 9. 10.1177/1527002506287700. :

Wymore L, Chin P, Geary C, et al. Performance and Injury Characteristics of Pitchers Entering the Major League Baseball Draft After Ulnar Collateral Ligament Reconstruction. The American Journal of Sports Medicine. 2016;44(12):3165-3170. doi:10.1177/0363546516659305http://scholarship.claremont.edu/cmc_theses/442