

WOOD AND ALUMINUM BAT STATISTICAL ANALYSIS FOR THE 2007 MLB AMATEUR DRAFT

PART II OF III

TRANSLATION BACKGROUND INFORMATION, RELEVANT STATISTICS,
LEAGUE DATA, TOP PROSPECT DATA, TRANSLATION FACTORS



*Oklahoma State third baseman Matt Mangini, the 2006 Cape
Cod League batting champion for the Hyannis Mets*

Gabe Gershenfeld

Gregg Reilly, Milton Academy
Sponsor

Chris Long, San Diego Padres
Off-Campus Mentor

Senior Project, May 2007

TRANSLATION BACKGROUND INFORMATION

Three articles published at baseballprospectus.com were very relevant to college/pro translations. The first is entitled “Looking Ahead: Translating College Performance” and written by college baseball expert Boyd Nation on April 20th, 2004. He looks at 288 former NCAA Division 1 hitters who played in Rookie or Short Season A ball and finds the correlation of OPS between college and the minors. After adjusting for park factors and competition, this .37 for both Rookie and A-ball. Keeping the previous adjustments, he then created the following table:¹

Stat	Correlation		Multiplier	
	SS A	R	SS A	R
AVG	0.27	0.26	0.69	0.78
R	0.27	0.28	0.54	0.69
H	0.24	0.33	0.7	0.77
2B	0.13	0.39	0.64	0.77
HR	0.45	0.48	0.31	0.45
RBI	0.38	0.51	0.5	0.64
SLG	0.38	0.4	0.6	0.72
BB	0.45	0.57	0.81	0.98
SO	0.53	0.52	1.33	1.19
OBP	0.36	0.36	0.74	0.84
OPS	0.37	0.37	0.66	0.77

The first things to notice is that the correlations are generally pretty low. Statistically, a strong correlation is .8, and the highest value anywhere on the table is .57. The highest correlations are found in the three true outcomes—walks, strikeouts, and homeruns. These terms are called the three true outcomes because they are almost entirely dependent on the batter/pitcher match up—none of these balls are put in play and fielders are removed as variables. In terms of multipliers, walks remain closest to their aluminum bat level, strikeouts increase, and homeruns are less than half as likely. These results closely follow my statistical hypothesis.

Kevin Goldstein wrote the second and third articles April 13th, 2006 and April 21st, 2006. Goldstein covers minor league and draft prospects for Baseball Prospectus and brings a qualitative scouting aspect to player evaluation (he used to work for Baseball America). His first article, ominously subtitled “When Translating College Statistics is a Bad Idea,” introduces the problem of relying purely on numbers:

“Translating college statistics is one thing, and translating college stats in an attempt to give us a better idea of what a player will become is an exercise in futility. When we look at the college numbers of top prospects and big leaguers, they are, in general, very good. This creates the illusion that translated college statistics will give us a valuable tool in projecting professional performance. The missing piece, however, is all of the outstanding college players who never make it, so we don't look back at their college career... There are definite patterns to the ones who make it and the ones who don't. The patterns, though, cannot be measured in raw statistics. They can, however, be measured in scouting reports.

...At the college level, *how* a player is accomplishing good offensive numbers is far more important than the raw numbers because of the presence of the metal bat. Metal bats can play a significant role in creating 'false power,' as a physically strong player can power a ball out of the park without making solid-centered contact, while that same contact off the handle or end of a wooden bat more often than not leads to an easily played fly ball."²

Indeed, the physical data from Part I along with the multiplier factors from before back up his claim on home runs. Also, the term "solid-centered contact" for wood bats is evidenced by the graph at the end of Part I. Goldstein goes on to give examples of former college players who dominated the NCAA by any statistical measure, but scouts did not like. For instance, in 2003 North Carolina's Jeremy Cleveland led the ACC in batting average, was second in homeruns, and walked more often than he struck out, but scouts saw "an unathletic body and, more importantly, a swing that was far more designed for aluminum than wood."³ Cleveland and other similar players were drafted in the later rounds and did not have professional success, exiting baseball altogether after only one or two seasons.

Goldstein's second article focuses on pitchers. A college player with the pitching equivalent of an aluminum swing possesses the following qualities:

"often [lacks] plus velocity, [has] impeccable commands and the ability to spin a breaking ball or change speeds very well. These types of pitchers can put off the kind of numbers that can fool the stats-only observers into believing that they have a future star on their hands, but in the end, without a true "out pitch" to depend on, these pitchers have little margin for error, and that margin shrinks exponentially with each level they advance."⁴

Given the increased bat speed of aluminum bats, it certainly makes sense that a pitcher who can successfully change speed is going to have amateur success (especially against inexperienced hitters). Goldstein then performs the same test with college pitchers as he did with hitters. North Carolina State's Vern Sterry put up college statistics comparable to Huston Street, Jeremy Sowers, and Joey Devine, but unlike these first round picks and current major leaguers, Sterry had "good command and a plus changeup that he sets up with a fastball that sits only at 85-87 mph."⁵ Sterry was an eighth round pick and is currently out of baseball. However, Goldstein leaves hope at the end of his article for those looking to perform statistical translations:

"By no means am I saying that college statistics are not valuable pieces of information, or that anybody working on translating and/or balancing college numbers should abandon the project. However, I think the true objective here would be to find a way to get a base translation and then adjust by finding a way to quantify the subjective side of things. This would be a tremendous challenge, and one that could only be done by a team, as they're the only ones that have access to professional-level scouting reports for every player. Developing coefficients for each aspect would also require a strict adherence to standardized grading, but the value one could find here borders on groundbreaking. This is the true challenge when it comes to making college statistics useful in evaluating future projection, and it's a monumental task."⁶

So where does my project fit into these three articles? The correlations in Boyd's method are too low to take any adjusted statistics seriously. I also do not have the time or resources (grading on scouting reports is not very standardized) to carry out the standardized grading assessment Goldstein calls for. However, one important thing to take away from Goldstein's last paragraph is that college statistics are still important; a player just needs to be viewed by scouts as having certain qualitative tools in order for his statistics to be relevant. My plan is to quantitatively evaluate a list of top prospects by Baseball America, so that all of the top prospects already have some pro potential and there will be no Cleveland/Sterry washouts. When evaluating the player's college statistics, I will adjust for park factors and strength of schedule as well as find my own league ratios like Boyd Nation. I will also have qualitative reports for all of the players to compare their numbers with. In addition, some of the statistics I use focus on predicting future performance rather than past value.

RELEVANT STATISTICS

I looked at seven statistics for hitters and six statistics for pitchers:

AVG = Batting Average = H/AB

OBP = On-base Percentage = $(H+BB+HBP)/PA$

SLG = Slugging Percentage = TB/AB

ISO = Isolated Power = $(2B+3B+(HR*3))/AB^7$

CON % = Contact Percentage = $(AB-SO)/AB$

SO % = Strikeout Percentage = SO/PA

$I * C$ = Isolated Power times Contact Percentage = $(ISO*F1)*(CON\%*F2)$

For the NCAA in 2006, $F1 = 8.317779086$ and $F2 = 1.22518369$

For the Cape in 2006, $F1 = 11.70901642$ and $F2 = 1.317075425$

For the NCAA in 2007, $F1 = 8.42138665$ and $F2 = 1.228626567$

I chose to look at isolated power multiplied by contact percentage, because the ability to make contact while hitting for power is a useful indicator for success at advanced levels. As seen from the F1 and F2 factors, this ratio is indexed so that the average ratio for each league is set at one; any value above one is above average. This is especially useful in evaluating power hitters, as a high strikeout rate in college indicates offensive struggles are going to come as a pro. While I thought of this stat myself, Chris has looked at some data and saw a correlation between this stat and amateur to professional success.

$K/9$ = Strikeouts per nine innings pitched = $(SO/IP)*9$

$BB/9$ = Walks per nine innings pitched = $(BB/IP)*9$

$HR/9$ = Homeruns let up per nine innings pitched = $(HR/IP)*9$

BABIP = Batting average on balls in play = $(H-HR)/(AB-HR-SO+SFA)$

dERA = Defense Independent Pitching Earned Run Average⁸

ERA = Earned run average per nine innings pitched = $(ER/IP)*9$

BABIP can be a useful tool to measure team fielding but it is viewed as beyond the control of the pitcher and is essentially "lucky," susceptible to random variation. A high

BABIP can often explain an ERA higher than dERA and vice versa. All of the pitching statistics listed are very important in separating the successful college pitchers from the successful college *and* professional pitchers Goldstein wrote about. An intelligent college pitcher without an “out” pitch can put up a low ERA without getting a lot of strikeouts. Such a pitcher would be expected to have a low BABIP and consequently see the BABIP and ERA regress to the league average as they advance professionally. A low BABIP is certainly not a bad thing, but one has to look at the pitcher’s peripherals and dERA along with qualitative scouting reports to get a true sense for his skill. The best way to examine this relationship is by looking at individual players from Part III. For instance, lefthander Nick Schmidt of Arkansas saw his ERA decrease from 2006 to 2007 despite a higher dERA because his BABIP lost 41 points. Despite being ranked the 12th best prospect in the most recent Baseball America prospect list, he is a “high-on-polish low-on-stuff arm”⁹ as his fastball sits at 91, his changeup is above average, and his command is his greatest strength. Any team evaluating him certainly has to take into account his low BABIP because of his physical limitations. I will also sometimes reference BABIP for hitters, as this does regress to a lesser extent and is also an indicator of speed.

Each hitter and pitcher also has his team’s information listed on his evaluation:

APF = Average Park Factor for home and away games (< 100 = pitcher’s environment, >100 = hitter’s environment)

SOS = Strength of Schedule (>100 = above NCAA average)

All of this team data came from boydsworld.com and college baseball expert Boyd Nation.

LEAGUE DATA

The college data collected were incomplete for certain schools. For 2006, hitting data are missing from 7 schools and pitching data are missing from 14 schools. For 2007, hitting and pitching data are missing from the same 19 schools. The data were either completely missing or without certain statistics. These schools have no obvious connection, as some are big conference schools like Missouri and North Carolina State while some are smaller conferences schools like Arkansas-Pine Bluff and Jacksonville State. When I evaluated players from these schools for Part III, I looked up their indicated statistics online.

All of the 2007 College data are up to May 6th, 2007, so it covers about 80% of the season.

For reference, here are the league totals and averages for hitters:

Year	League	PA	AVG	OBP	SLG	ISO	CON %	SO %	I * C	APF	SOS
2006	NCAA	604244	0.292	0.367	0.419	0.120	0.816	0.157	1	100	100
2006	Cape	16774	0.234	0.322	0.323	0.085	0.759	0.205	1	100	100
2007	NCAA	465451	0.290	0.367	0.416	0.119	0.814	0.159	1	100	100
2005-6	R/SSA	421947	0.251	0.330	0.356	0.097	0.772	0.199	1	-	-

League totals and averages for pitchers:

Year	League	IP	K/9	BB/9	HR/9	BABIP	dERA	ERA	APF	SOS
2006	NCAA	129265.3	6.61	3.75	0.70	0.332	4.20	5.07	100	100
2006	Cape	3905.3	7.90	3.75	0.47	0.292	3.66	3.15	100	100
2007	NCAA	100794.3	6.70	3.85	0.68	0.336	4.19	5.13	100	100
2005-6	R/SSA	98871.3	7.79	3.61	0.49	0.310	3.63	3.90	-	-

The majority of the rookie / short-season A-ball data are from 2006, and all of the data are from first year players. The APF and SOS are unavailable for these players. This lack of information is one reason why there is large uncertainty in evaluating statistics from the R/SSA level. Some players in the league come directly from the Dominican Republic having never played on a real baseball field, and there is great variation in different divisions within the leagues. I was originally planning to translate the player's college stats to this level; however, after learning about the extreme variation, I decided to translate the player's college stats to the Cape league. Although the Cape league is still considered a pitcher's league, the stats are well recorded, all of the team's schedules are balanced (they all have a SOS of 100), and the competition is strong (and, just as important, evenly distributed).

The major league BABIP is about .290, so it makes sense for the wooden bat Cape league to have a similar BABIP. Rookie leagues are also, of course, wooden-bat leagues, but as the overall defense of each minor league level improves, BABIP decreases (BABIP is higher in high single-A than double-A which is higher than triple-A). The extremely high college BABIP numbers can probably be attributed to the aluminum bats. It is important to remember dERA was calculated in a major-league hitting environment, so a league average .290 BABIP means a major-league dERA would be very close to the actual ERA. However, a league average .335 BABIP in a college hitting environment means that the dERA will be significantly less than the actual ERA. This explanation can be seen in the actual league average for dERA and ERA, where there is a large gap between dERA and ERA for the NCAA in 2006 and 2007. In the Cape League, where the BABIP is much closer to the major league level, the gap between dERA and ERA is much smaller.

TOP PROSPECT DATA

I individually looked at 63 hitters and 77 pitchers, all top professional prospects. The complete list and source for the group can be found in Part III. Almost all of the players played Division 1 baseball in 2006 and 2007 and a good portion played in the Cape League in 2006. Since these are better players, one can get a better sense for expected performance in the Cape League. Here are the averages these players collectively put up (weighted by plate appearances per hitter and innings pitched per pitcher):

Top 63 Hitters:

Year	League	PA	AVG	OBP	SLG	ISO	CON %	SO %	I * C	APF	SOS
2006	NCAA	15261	0.334	0.416	0.521	0.177	0.839	0.136	1.498	95.72	107.32
2006	Cape	5165	0.254	0.346	0.375	0.117	0.753	0.211	1.360	101.05	100.00
2007	NCAA	12352	0.337	0.433	0.547	0.198	0.852	0.122	1.723	95.91	106.21

Top 77 Pitchers:

Year	League	IP	K/9	BB/9	HR/9	BABIP	dERA	ERA	APF	SOS
2006	NCAA	4703.3	8.12	3.18	0.57	0.306	3.50	3.65	93.71	106.52
2006	Cape	977.3	8.88	3.45	0.41	0.277	3.29	2.55	102.54	100.00
2007	NCAA	4097.3	9.00	2.98	0.45	0.301	3.06	3.16	93.83	105.97

As one might expect, all of these players put up better numbers than their league averages (and played for good teams as the SOS ratings indicate). One possible explanation for the lower than league average BABIP—after all, BABIP is not supposed to be an indicator of skill—is the original selection of players. Some of the players towards the end of the rankings without the physical tools probably put up better ERA than dERA numbers (by NCAA standards) because of a low BABIP. In other words, BABIP helped determine the list and is thus there is a correlation. A low BABIP means that one can expect their BABIP to regress towards the league average in following years, indicating that previous ERAs were uncharacteristically low.

TRANSLATION FACTORS

In Part III, I perform four translations for all of these stats for the 133 players. The main translation I do is to adjust for wooden bats—I do this by essentially making all of the NCAA averages equivalent to the Cape. By finding the ratios of each stat from the NCAA in 2006 to the Cape in 2006, I can create the following table for hitters and pitchers:

Stat	AVG	OBP	SLG	ISO	CON %	SO %
Wood Factor	0.758940295	0.831207397	0.719228377	0.65990787	0.897014862	1.556605906

Stat	K/9	BB/9	HR/9	BABIP	dERA	ERA
Wood Factor	1.093885358	1.084528561	0.708181432	0.904304597	0.93972404	0.698193824

I multiply each NCAA stat with these factors. The Cape stats remain the same, and all of the I*C values are recomputed with the Cape F1 and F2 values. It is interesting to compare this table with Boyd Nation’s original one, because the same stats remain more or less affected by the transition to wood.

My next translation is to adjust for strength of schedule. I divide each of the SOS values by 100 to turn into a multiplier, and then I multiply this value for every stat. For values like ERA and SO % where lower values are better, I subtract the SOS/100 value from 2 and use this number. For example, a SOS of 106 would have an ERA multiplier of 0.94.

I also adjust for experience with a 5 percent factor. The large majority of the draft class are juniors, but there a fair amount of seniors and sophomores. Each sophomore stat is rewarded with a 1.05 multiplier (0.95 for ERA, SO%, etc.) and each senior stat is hit with a 0.95 multiplier (1.05 for ERA, SO %, etc.). While experience isn't explicitly tied to age because of the large number of medical redshirts (especially for pitchers), these multipliers do give a sense for how much projection a player has.

The last translation I do for each player is to adjust for offensive environment. Like SOS, I divide the APF values by 100 to get a multiplier. However, unlike strength of schedule, park factors cannot be expected to affect each stat equally. Park factors are measure of runs, and so each stat's translation depends on it's relationship to runs scored:

- HR/9, dERA, and ERA have a direct relationship to runs scored, and so each stat is divided by the APF multiplier.
- AVG, OBP, SLG, ISO, and BABIP have an indirect relationship to runs scored, and so each stat is divided by the square root of the APF multiplier.
- The ballpark does not directly affect CON %, SO %, K/9, and BB/9, and so these stats remain the same.

All of these translations are fairly subjective, but provide a baseline for predicting how each player will perform with wooden bats in a normalized environment. I also calculate the "wooden" averages for each player, weighted by plate appearances and innings pitched.

¹ Boyd Nation, "Looking Ahead: Translating College Performance," Baseball Prospectus (April 20, 2004), viewed May 3, 2007.
<<http://www.baseballprospectus.com/article.php?articleid=2787>>

² Kevin Goldstein, "Future Shock: When Translating College Statistics Is a Bad Idea," Baseball Prospectus (April 13, 2006), viewed May 3, 2007.
<<http://www.baseballprospectus.com/article.php?articleid=4971>>

³ Ibid.

⁴ Kevin Goldstein, "Future Shock: Translating College Performance, Pitchers," Baseball Prospectus (April 21, 2006), viewed May 3 2007.
<<http://www.baseballprospectus.com/article.php?articleid=4997>>

⁵ Ibid.

⁶ Ibid.

⁷ This is Baseball Prospectus' PECOTA method; traditionally this statistic is just SLG-AVG. This values 2B the same as 3B because the extension of a double into a triple is more often a sign of speed than power.

⁸ The full formula to calculate this statistic can be found at <http://www.baseballthinkfactory.org/mccracken/dipseexpl.html>. This equation is very long, but the important thing to know is that it is calculated from six simple statistics: batters faced, home runs, total walks, intentional walks, strikeouts, and hit batsmen. It is essentially a measure of the pitcher's peripheral statistics.

⁹ Kevin Goldstein, "Future Shock: College Draft All-Stars," Baseball Prospectus (May 18, 2007) viewed May 18, 2007.
<<http://baseballprospectus.com/article.php?articleid=6244>>