# BASEBALL ANALYST

OCTOBER 1982 ISSUE NO. 3

Just a few quick notes from the editor before we get to the business here.

This edition contains four excellent articles, I think:



There is something in this selection of works which truly delights me, which is that most of the articles in this Analyst respond to work that was in the first two editions, and all of the articles here take sabermetric work that has been done before and build on to it, extend it further. They are also, as a group, the most thoughtful, most intelligent, and most productive articles that we have had. I can't tell you how much that satisfies me. The purpose of doing this is to provide a form for sabermetricians, a place for people to have an intelligent discussion about the issues involved, to respond to one another and develop a field of common knowledge. We have here the first sign that this baby's going to walk. She still needs your help--we received about 20 pages of articles in the last two months, and each issue runs 20--but with your help she will make it.

Bill James

# MORE ON THE "TRUE" SLUGGING PERCENTAGE

Jim Reuter

In the August 1982 Baseball Analyst, Jim Morrow sought to replace the conventional 1-2-3-4 total base valuations used in determining a batters slugging percentage with relative weights that would more closely measure the players run production. While that study failed to produce the results desired, it did predict that the conventional total base valuations undervalued homeruns and overvalued doubles.

This problem can be solved by deducing the relative weights used for walks and the various hits in the runs created formula. Runs created was chosen because of its excellent ability to predict run production (this will be shown later). The algebra involved, shown below, is quite simple:

$$RC = \frac{ab}{c}$$ where RC-runs created; a-hits+walks; b-total bases c-at bats+walks.

If the batter hits a single(S):

$$RC_S = \frac{(a+1)(b+1)}{(c+1)} \simeq \frac{ab+b+a}{c} \qquad \text{assuming } a,b,c \gg 1.$$

The effect of the single is then:

$$RC_S - RC = \frac{a+b}{c}.$$

Similarly, the effect of each of the remaining hits and the walk are:

$$\begin{array}{lll}
\text{Double(D)} & : RC_D - RC = \frac{2a+b}{c} \; ; & RC_D - RC_S = \frac{a}{c} \\
\text{Triple(T)} & : RC_T - RC = \frac{3a+b}{c} \; ; & RC_T - RC_S = \frac{2a}{c} \\
\text{Homerun(HR)} & : RC_{HR} - RC = \frac{4a+b}{c} \; ; & RC_{HR} - RC_S = \frac{3a}{c} \\
\text{Walk(W)} & : RC_W - RC = \frac{b}{c} \; ; & RC_S - RC_W = \frac{a}{c}
\end{array}$$

Using the totals from the 1979 and 1980 seasons, the same period the Morrow study was based on, the relative values of a, b, and c are found to be:

$$a/c = .328 \; ; \quad b/c = .359 \; ; \quad (a+b)/c = .687.$$

The weights attached to each successful plate appearance are then:

Walk-.36; Single-.69; Double-1.02; Triple-1.35; Homerun-1.68.

The 1.68 index for the homerun compares quite favorably with John C. Tattersall's conclusion in the 1975 Home Run Handbook that a homerun is worth 1.64 runs and Morrow's conclusion that it was worth 1.74 runs during the 1979 season. Also, a walk is worth about one half as much as a single, just as Morrow found.

Relative weights can now be assigned:

Walk-1; Single-2; Double-3; Triple-4; Homerun-5.

Using the above weights, the resulting slugging percentage was calculated for each major league team during the 1979 and 1980 seasons. These values were correlated with team runs scored. A comparison, shown in Table 1, could then be made between this relation and the other slugging percentages.

Table 1. Correlation Coefficients of Various "Slugging" Percentages.

| Slugging Percentage | Explanation of Term | Correlation Coefficient |
|---|---|---|
| Runs Created | Bill James Slugging % | .954 |
| SP(1,2,3,4,5) | JR Slugging % | .940 |
| SP(0,1,2,3,4) | Conventional Slugging % | .902 |
| SP(1,2,3,6,10) | JM Slugging % | .897 |

The numbers in parentheses refer to the relative weights assigned to walks, singles, doubles, triples, and homeruns respectively. The denominator of the slugging percentages is at bats when the weight assigned to walks is zero and at bats plus walks otherwise.

Table 1 reveals that the 1-2-3-4-5 weighting scheme is a significantly better predicter of a players run production than either the conventional or the Morrow slugging percentages. The Morrow formula overrates the homerun and triple while the conventional slugging percentage totally ignores the walk and overvalues the extra base hit relative to the single. Although runs created is clearly the best predicter, it doesn't allow direct comparisons of the values of the various hits.

With the original goal accomplished, efforts were directed toward comparing the abilities of other batting statistics to predict run production. Table 2 summarizes the results of these studies. A brief description of each of the studies undertaken is given after the table.

Table 2. Correlation Coefficients of Various Batting Statistics.

| | Slugging Percentage | Explanation of Term | Correlation Coefficient |
|---|---|---|---|
| (1) | Runs Created | SP(BJ) | .954 |
| (1) | OB %+SP | On base % + SP | .954 |
| | SP(.36,.69,1.02,1.35,1.68) | Original SP(JR) | .943 |
| (1,2) | SP(1,2,3,4,5) | SP(JR) | .940 |
| (3) | SP(1,1,2,3,4) | Modified SP | .936 |
| (2) | SP(1,2,3,5,7.5) | (SP(JR)+SP(JM))/2 | .917 |
| (2) | SP(.5,1.5,2.5,3.5,4.5) | (SP(JR)+SP)/2 | .913 |
| (3) | SP(.5,1,2,3,4) | Modified SP | .911 |
| 2,3) | SP(0,1,2,3,4) | Conventional SP | .902 |
| (2) | SP(1,2,3,6,10) | SP(JM) | .897 |
| (4) | OB(1,1,1,1,1) | On base % | .890 |
| (4) | BA(0,1,1,1,1) | Batting average | .837 |
| (4) | AIP(0,0,1,2,3) | Adjusted isolated power | .819 |
| (4) | Extra Bases | D + 2T + 3HR | .808 |
| (4) | IP(0,0,1,2,3) | Isolated power | .785 |

The numbers preceding the slugging percentage terms refer to their discussion location below.

(1) The slugging percentage developed in this study is:

$$SP(JR) = \frac{W+2S+3D+4T+5HR}{AB+W} = \frac{\text{Walks} + \text{Hits} + \text{Total Bases}}{\text{At Bats} + \text{Walks}}$$

This is quite close to being:

$$SP'(JR) = \text{On Base Percentage} + \text{Slugging Percentage (conventional)}$$

The new relation is practically the same expression Mitch Albom used to rate ballplayers in the July 1982 issue of Sport magazine. This study shows he was using basically the same formula, within certain approximations, as the runs created relation. It should be noted that while the two expressions perform identically over this short time frame, the runs created formula is superior when a longer time period is considered.

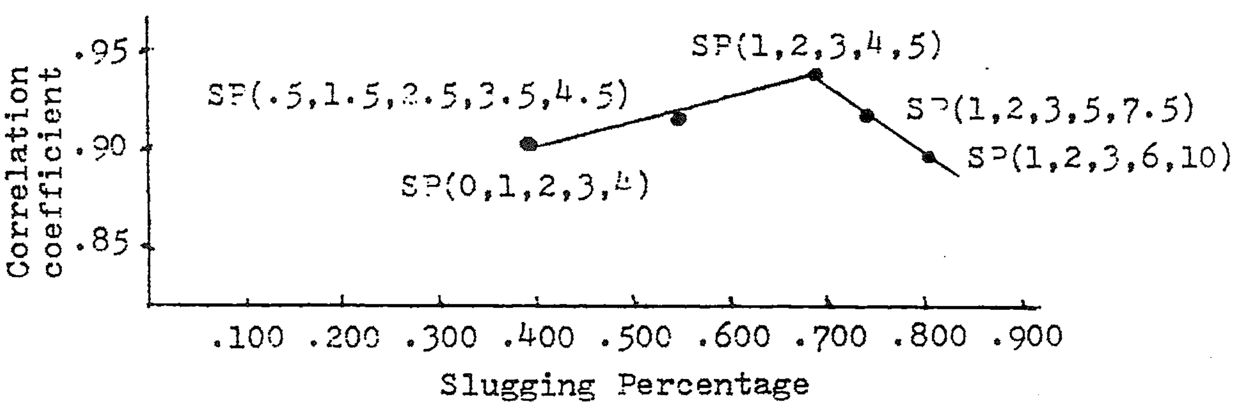
(2) The three slugging percentages originally compared were:

SP(0,1,2,3,4) - Conventional SP
SP(1,2,3,4,5) - SP(JR)
SP(1,2,3,6,10) - SP(JM)

In order to determine whether SP(JR) could be improved, weights midway between the above percentages were considered:

SP(.5,1.5,2.5,3.5,4.5) and SP(1,2,3,5,7.5)

The resulting correlation coefficients were then plotted as a function of the overall slugging percentages for the combined years of 1979 and 1980.




SP(1,2,3,4,5) seems to be very close to the optimum weightage.

(3) The conventional slugging percentage was adjusted to include walks, giving them both one half the value of a single and the same value as a single:

SP(.5,1,2,3,4) and SP(1,1,2,3,4)

The denominator in both cases then becomes at bats + walks. Table 2 reveals that giving walks the same value as a single significantly improves the slugging percentage equation, more so than assigning the more accurate one half value. The reason for this can best be shown algebraically:

$$SP(1,2,3,4,5)-SP(1,1,2,3,4) = \frac{\text{Hits}}{\text{At Bats + Walks}}$$

$$SP(1,2,3,4,5)-SP(.5,1,2,3,4) = \frac{\text{Hits + Walks/2}}{\text{At Bats + Walks}}$$

Giving a walk the same weight as a single is better in this case because it more closely approximates the "true" expression: SP(1,2,3,4,5).

(4) On base percentage was compared to batting average:

OB(1,1,1,1,1) to BA(0,1,1,1,1).

A comparison was also made between the run scoring potential of a high batting average team and a team with a lot of power. Two different power statistics were used for this study: isolated power (SP-BA) and adjusted isolated power ( (SP-BA)/(1-BA) ). Isolated power is the number of conventional extra bases (0,0,1,2,3) per official at bat. Adjusted isolated power is the number of conventional extra bases per out expended. It overcomes the penalty for hitting a single that isolated power contains.

From Table 2, on base percentage is seen to be a significantly better parameter than batting average at predicting run production and is, surprisingly, almost as good as the conventional slugging percentage. It can also be predicted that a high batting average team should score runs with more consistency than a comparable team featuring a lot of power, since the batting average correlation coefficient is higher than the power statistics coefficients. Finally, the correlation order for the power statistics is: adjusted isolated power, extra bases, and isolated power, implying that adjusted isolated power is superior to isolated power as a power statistic.

(5) Recognizing the limitations in drawing conclusions from a study based on such a short time period, a similar correlation study of some of the slugging percentages listed in Table 2 was performed on league averages for the period from 1960 to 1980. In general, the relative order of the statistics did not change. Runs created did emerge as the clear winner, with OB%+SP and SP(1,2,3,4,5) showing virtually identical second best results.

The newly developed slugging percentage (1,2,3,4,5) can be easily expanded to include stolen bases, although this probably makes the use of the term slugging percentage ridiculous. The runs created formula including stolen bases is:

$$RC = \frac{(\text{Hits + Walks-Caught Stealing})(\text{Total Bases + .7xStolen Bases})}{(\text{At Bats + Walks})}$$

Performing the same algebraic manipulations as before, and considering every stolen base attempt to be s stolen bases and (1-s) caught stealing:

$$RC_{SB}-RC = \frac{(.7s)a - (1-s)b}{c}$$ where s-stolen base success rate.

Inserting values for a,b,c: $RC_{SB}-RC = .589s-.359$.

According to this formula, a minimum success rate of 61% is needed to justify an attempted steal.

For a 75% stolen base efficiency, $RC_{SB}-RC = .08$. This corresponds to a value of 1/4 when compared to the other weights in the SP(1,2,3,4,5) relation. The revised slugging percentage then becomes:

$$SP(JR) = \frac{W + 2S + 3D + 4T + 5HR + SB/4}{AB + W}$$

This expression will work well even if the stolen base efficiency is not exactly 75%.

In summary, optimum relative weights have been found to assign to walks and the various hits so that the resulting slugging percentage best describes the batters run production. Although other formulations, runs created in particular, can be found that more accurately reflect run production, they do not allow a direct comparison between the relative values of each type of hit and the walk. As a result of this study, it may be concluded that a homerun is not worth as much as two doubles, but about the same as a double and a single. Further, three homeruns are roughly equivalent to five doubles or seven singles and a walk. A triple has about the same value as two singles while a double is worth a single and a walk. In each case, the increased run scoring potential of one term is offset by the greater run batted in potential of the other.

FINAL NOTE: The same study can be performed using no approximations.

e.g. $$RC_S-RC = \frac{(a+1)\ (b+1)}{(c+1)} - \frac{ab}{c}$$

The new weights are:

SP(.24,.57,.90,1.22,1.55) rather than SP(.36,.69,1.02,1.35,1.68).

The new relation was correlated with team runs scored for the 1979 and 1980 seasons. It did not perform as well as the original formula, either in terms of a slugging percentage(.928 to .943) or total bases(.945 to .949). Regardless, the 1-2-3-4-5 weightage would still be accurate since the new weights correspond to:

SP(.84,2,3.2,4.3,5.4) as opposed to SP(1.0,2,3.0,3.9,4.9)

for the original formula. It can be shown that the new weights are essentially:

$RC_W-RC=\frac{b}{c}(1-\frac{a}{c})$; $RC_S-RC_W=\frac{a}{c}$; $RC_D-RC_S=\frac{a}{c}$; $RC_T-RC_S=\frac{2a}{c}$; $RC_{HR}-RC_S=\frac{3a}{c}$.

## Batting Average Comparisions

I have always been upset with people (usually old TV baseball announcers, in fact upon reflection, ALWAYS old TV baseball announcers) that compare 1930 era batting average accomplishments against current batting average accomplishments. The batting average for the whole National League was .303 in 1930, so batting .300 was hardly note worthy. But sitting in my livingroom, I couldn't get an arguement out of the man on TV, and when I tried to argue the same point with my friends with the same vexation I used with the man on TV, I got pretty much the same response. I decided that the best way to purge my bitterness would be to devise a means to analytically compare a batting average of one season to any batting average in another season. I have been thinking about doing this since 1968, but it wasn't until I was invited to participate in the Baseball Analyst that I got the energy to gather and compile the data necessary. I thank Y'all for inspiring me to do this.

I used the batting averages of all of the players with 150 or more At Bats and determined the arithematic averages (means) and standard deviations on a per season basis. I chose 150 At Bats as a minimum for no specific reason, other than I felt that it a sufficient number to indicate a player's "true" batting ability. I have only completed the American League for years 1901 thru 1980. I was so excited about having this new tool and so dreaded having to compile the same data for the National League that I decided to write up my findings now. Everybody, please keep me in line, for I don't want it to be another 14 years before I finish the National League.

These various means (arithematic averages) and corresponding standard deviations allow for a method of measuring how a given player's batting average for a given season compares to the same player's peers. This measurement is made in standard deviations above or below the mean. What this allows to be done is to compare, for instance, how much better Ty Cobb was than the people he played with and against to how much better Rod Carew was than the people he played with and against.

| | | | | |
|---|---|---|---|---|
| Example -- | Ty Cobb 1911 B.A. | .420 | Rod Carew 1973 B.A. | .350 |
| | League Mean B.A. | .2789 | League Mean B.A. | .2597 |
| | League S. Deviation | .0479 | League S. Deviation | .0307 |
| | (420 - 278.9)/47.9 = 2.94 | | (350 - 259.7)/30.7 = 2.95 | |

Rod Carew's batting .350 in 1973 is approximately equivalent to Ty Cobb's batting .420 in 1911 because the standard deviations above the mean value for each player are approximately equivalent: Ty Cobb's 2.94 standard deviations above the mean to Rod Carew's 2.95 standard deviations above the mean.

This is a pretty amazing claim that a .350 batting average in one season is equivalent to a .420 batting average in another season. This is a difference of 70 percentage points, but I am quite confident that my figures are accurate. I won't be so bold about claiming the figures to be exact.

I think that the chart designating the American League batting average means and standard deviations for the years 1901 thru 1980 is fairly straightforward, and because I have doubts that it really is, I will explain it. Each row contains the following items: the year, the average batting average for the players with 150 or more at bats, the standard deviation in batting average percentage points, the lowest batting average for the season, the number of standard deviations that the lowest batting average is below the mean, the highest batting average for the season, and the number of standard deviations that the highest batting average is above the mean.

I used this chart to calculate on a year to year and career basis the performances of some of the best players that played exclusively in the American League. I weighted each seasonal batting average accomplishment by the player's seasonal number of at bats. I did this because I wanted to discount seasons in which a player did not have very many at bats.

The standard deviations above the mean per at bat nomenclature is inaccurate. I multiplied the at bats in on a seasonal basis and then divided the at bats out from the career total at bats times the standard deviations above the mean figure. So the value is actually career standard deviations above the mean on a per year basis weighted by the seasonal at bat quantity.

The ratings for these players by the aforementioned figure was not startling. In fact, the major point I was trying to prove by this study was that Rod Carew's batting average accomplishments were as good as Ty Cobb's batting average accomplishments.

| Player | Career Weighted S.D. | Career B.A. |
|---|---|---|
| Rod Carew | 2.45 | .332 |
| Ty Cobb | 2.45 | .367 |
| Ted Williams | 2.26 | .344 |
| Joe Jackson | 2.20 | .356 |
| Tris Speaker | 1.82 | .344 |
| Babe Ruth | 1.53 | .342 |
| Lou Gehrig | 1.45 | .340 |
| Joe Dimaggio | 1.42 | .325 |
| Al Kaline | 1.29 | .297 |
| Mickey Mantle | 1.24 | .298 |

Rod Carew's statistics are biased because he is the only player charted that has not completed his career. I am sure that his career weighted standard deviations above the mean value will drop has he continues playing. Without Ty Cobb's last six seasons his career weighted standard deviations above the mean value would have been 2.77.

Since I had all of these yearly batting average figures collected, I decided to use them to determine if it is reasonable to assume that seasonal batting averages for a whole league have a normal distribution. I found that there is no reason to believe that they do not have a normal distribution.

I will give the two extreme seasons as examples: the season that fit normal distribution the best -- 1938, and the season that fit normal distribution the worst -- 1978. Below is a list of the observed number of players that had batting averages within the limits listed, and a list of the expected number of players that should have been within the limits listed if the seasonal batting averages for the whole league had a normal distribution.

| 1938 | Observed | Expected | 1978 | Observed | Expected |
|---|---|---|---|---|---|
| .196 - .215 | 2 | 0.9 | .157 - .176 | 2 | 0.5 |
| .216 - .235 | 3 | 3.4 | .177 - .196 | 3 | 2.8 |
| .236 - .255 | 9 | 9.1 | .197 - .216 | 10 | 11.1 |
| .256 - .275 | 15 | 16.7 | .217 - .236 | 30 | 28.1 |
| .276 - .295 | 21 | 21.1 | .237 - .256 | 30 | 45.1 |
| .296 - .315 | 20 | 18.4 | .257 - .276 | 60 | 46.1 |
| .316 - .335 | 11 | 11.0 | .277 - .296 | 31 | 30.0 |
| .336 - .355 | 5 | 4.6 | .297 - .316 | 12 | 12.4 |
| .356 - .375 | 1 | 1.3 | .317 - .336 | 2 | 3.3 |

I have the same data for all of the American League seasons from 1901 to 1980, and if anyone wants me to submit that, then I will be glad to type that information also.

I used the Chi squared test to evaluate this data and my normal distribution hypothesis. There is about a 2% chance that the 1938 season batting averages do not have a normal distribution, and there is about a 95% chance that the 1978 season batting averages do not have a normal distribution. As I understand the use of the Chi squared test, a 95% chance is generally considered acceptable. Meaning that there is no reason to believe that the collection does not have a normal distribution. Please set me straight if I am wrong.

Lastly, I wondered whether or not league expansion and rule changes concerning the size of the strike zone had any affect upon league average batting averages or league batting average standard deviations.

Expansion Year Assumptions -- The league batting average would increase because of the dilution of the pitching talent. The league batting average standard deviation would increase because more weak players would get to play enough to meet the 150 at bat minimum allowing the better players to look that much better.

Strike Zone Size Change Assumptions -- The league batting average would change inversely to the change in the size of the strike zone. The league batting average standard deviation would change in the same direction as the change in the size of the strike zone; the marginal good players would drop away from the best players and the very bad players would drop away from the marginal bad players creating a wider variance of batting averages.

1950 - Strike Zone Size Decreased.

The league batting average increased by 7 percentage points, but contrary to my assumption the league batting average standard deviation increased by 2 percentage points.

1961 - American League Expands From Eight Teams to Ten Teams.

The league batting average didn't change, but the standard deviation did increase 5 percentage points.

1963 - Strike Zone Size Increased.

The league batting average decreased by 5 percentage points, and again contrary to my assumption the league standard deviation decreased by 3 percentage points.

1969 - American League Expands From Ten Teams to Twelve Teams and the Strike Zone Size is Decreased.

The league batting average increased by 17 percentage points, which supports both my expansion and strike zone size assumptions. The league standard deviation decreased by 3.2 percentage points, but since my expansion and strike zone size assumptions concerning the standard deviation conflict each other I won't draw any conclusions.

1977 - American League Expands From Twelve Teams to Fourteen Teams.

The league batting average increased by 9.7 percentage points and the league standard deviation increased by 2 percentage points.

Conclusions.

My assumption about the affect of strike zone size changes on the league standard deviation doesn't seem to fit. I don't see why making it harder to hit the ball would have a clustering effect on the league batting averages.

Another point I was trying to make with this study was that Carl Yastrzemski's batting .301 in 1968 was not that poor of a batting accomplishment. Well, I was wrong. It was a poor batting accomplishment after all.

Ward Larkin

Note to Dallas Adams re <u>The Effects of Overwork on Rookie Pitchers</u>.

I question the manner in which you grouped the pitchers. Any pitcher that appears in more than 40 games most likely made some relief appearances, and any pitcher that appears in more than 60 games probably pitched them all as a reliever. I think that the groups should be set by a number of innings pitched guideline. I don't feel that a relief pitcher that playes in 65 games and pitches 100 innings is more overworked than a starting pitcher that pitches in 36 games for 240 innings.

American League Batting Average
Averages and Standard Deviations

| Year | Average | Standard Deviation | Low | Deviations Below Mean | High | Deviations Above Mean |
|---|---|---|---|---|---|---|
| 1901 | 284.2 | 37.87 | 206 | 2.06 | 422 | 3.64 |
| 1902 | 279.7 | 39.10 | 185 | 2.42 | 379 | 2.54 |
| 1903 | 257.1 | 40.04 | 185 | 1.80 | 355 | 2.45 |
| 1904 | 250.2 | 36.08 | 173 | 2.14 | 381 | 3.62 |
| 1905 | 247.6 | 36.13 | 164 | 2.31 | 329 | 2.25 |
| 1906 | 253.4 | 42.60 | 144 | 2.57 | 358 | 2.45 |
| 1907 | 251.2 | 37.99 | 166 | 2.24 | 350 | 2.60 |
| 1908 | 245.0 | 32.22 | 168 | 2.39 | 324 | 2.45 |
| 1909 | 252.7 | 36.87 | 162 | 2.46 | 377 | 3.37 |
| 1910 | 247.7 | 42.37 | 124 | 2.92 | 385 | 3.24 |
| 1911 | 278.9 | 47.93 | 151 | 2.67 | 420 | 2.94 |
| 1912 | 273.1 | 42.30 | 194 | 1.87 | 410 | 3.24 |
| 1913 | 262.6 | 40.36 | 180 | 2.05 | 390 | 3.16 |
| 1914 | 256.1 | 38.12 | 163 | 2.44 | 368 | 2.93 |
| 1915 | 254.0 | 35.04 | 190 | 1.83 | 332 | 2.23 |
| 1916 | 257.3 | 35.53 | 187 | 1.98 | 386 | 3.62 |
| 1917 | 253.8 | 39.54 | 179 | 1.89 | 383 | 3.27 |
| 1918 | 257.9 | 39.13 | 140 | 3.01 | 382 | 3.17 |
| 1919 | 277.4 | 39.80 | 141 | 3.43 | 384 | 2.68 |
| 1920 | 287.3 | 43.53 | 198 | 2.05 | 407 | 2.75 |
| 1921 | 296.8 | 42.76 | 167 | 3.04 | 394 | 2.27 |
| 1922 | 293.1 | 37.17 | 223 | 1.89 | 420 | 3.41 |
| 1923 | 287.5 | 40.84 | 177 | 2.71 | 403 | 2.83 |
| 1924 | 295.3 | 36.78 | 175 | 3.27 | 378 | 2.25 |
| 1925 | 296.1 | 43.62 | 195 | 2.32 | 393 | 2.22 |
| 1926 | 289.1 | 36.56 | 192 | 2.66 | 378 | 2.43 |
| 1927 | 294.7 | 37.00 | 221 | 1.99 | 398 | 2.79 |
| 1928 | 285.9 | 36.84 | 204 | 2.22 | 379 | 2.53 |
| 1929 | 291.8 | 37.42 | 200 | 2.45 | 369 | 2.06 |
| 1930 | 291.4 | 38.72 | 185 | 2.75 | 381 | 2.31 |
| 1931 | 284.7 | 34.12 | 209 | 2.22 | 390 | 3.08 |
| 1932 | 281.6 | 34.80 | 201 | 2.32 | 372 | 2.60 |
| 1933 | 278.0 | 33.57 | 195 | 2.47 | 356 | 2.32 |
| 1934 | 282.6 | 37.06 | 159 | 3.34 | 363 | 2.17 |
| 1935 | 288.2 | 29.17 | 192 | 3.30 | 349 | 2.08 |
| 1936 | 299.3 | 35.72 | 197 | 2.86 | 388 | 2.48 |
| 1937 | 287.0 | 37.15 | 191 | 2.58 | 371 | 2.26 |
| 1938 | 289.1 | 33.54 | 201 | 2.63 | 374 | 2.53 |
| 1939 | 284.8 | 31.78 | 211 | 2.32 | 381 | 3.03 |

| Year | Average | Standard Deviation | Low | Deviations Below Mean | High | Deviations Above Mean |
|---|---|---|---|---|---|---|
| 1940 | 278.3 | 35.49 | 186 | 2.60 | 352 | 2.08 |
| 1941 | 274.6 | 37.38 | 190 | 2.26 | 406 | 3.51 |
| 1942 | 263.5 | 32.05 | 194 | 2.17 | 356 | 2.88 |
| 1943 | 255.6 | 30.31 | 188 | 2.23 | 351 | 3.15 |
| 1944 | 269.2 | 31.83 | 201 | 2.14 | 355 | 2.70 |
| 1945 | 261.3 | 29.76 | 194 | 2.26 | 333 | 2.41 |
| 1946 | 262.0 | 33.75 | 196 | 1.96 | 354 | 2.72 |
| 1947 | 260.9 | 34.87 | 157 | 2.98 | 343 | 2.35 |
| 1948 | 270.7 | 31.97 | 203 | 2.12 | 369 | 3.07 |
| 1949 | 270.1 | 29.75 | 196 | 2.49 | 346 | 2.55 |
| 1950 | 277.6 | 31.79 | 215 | 1.97 | 354 | 2.40 |
| 1951 | 266.6 | 33.85 | 159 | 3.18 | 344 | 2.29 |
| 1952 | 260.4 | 27.02 | 197 | 2.35 | 327 | 2.47 |
| 1953 | 270.0 | 30.20 | 196 | 2.45 | 337 | 2.22 |
| 1954 | 264.3 | 31.92 | 195 | 2.17 | 345 | 2.53 |
| 1955 | 265.8 | 36.29 | 194 | 1.98 | 364 | 2.70 |
| 1956 | 266.6 | 34.47 | 173 | 2.72 | 353 | 2.50 |
| 1957 | 261.4 | 34.31 | 181 | 2.34 | 388 | 3.69 |
| 1958 | 259.0 | 33.21 | 170 | 2.68 | 328 | 2.08 |
| 1959 | 261.5 | 30.83 | 181 | 2.61 | 363 | 3.29 |
| 1960 | 264.0 | 26.80 | 194 | 2.61 | 320 | 2.09 |
| 1961 | 264.1 | 31.75 | 185 | 2.49 | 361 | 3.05 |
| 1962 | 261.4 | 29.88 | 159 | 3.43 | 326 | 2.16 |
| 1963 | 256.3 | 26.75 | 172 | 3.15 | 321 | 2.41 |
| 1964 | 256.2 | 27.72 | 196 | 2.17 | 323 | 2.41 |
| 1965 | 249.1 | 28.48 | 163 | 3.02 | 321 | 2.52 |
| 1966 | 247.0 | 27.65 | 155 | 3.33 | 316 | 2.49 |
| 1967 | 244.6 | 30.62 | 167 | 2.53 | 326 | 2.66 |
| 1968 | 235.0 | 34.21 | 135 | 2.92 | 301 | 1.93 |
| 1969 | 252.0 | 30.97 | 164 | 2.84 | 332 | 2.58 |
| 1970 | 259.4 | 32.08 | 190 | 2.16 | 366 | 3.32 |
| 1971 | 255.6 | 31.95 | 177 | 2.46 | 338 | 2.58 |
| 1972 | 246.8 | 32.12 | 140 | 3.32 | 318 | 2.21 |
| 1973 | 259.7 | 30.66 | 179 | 2.63 | 350 | 2.94 |
| 1974 | 257.4 | 30.16 | 151 | 3.53 | 364 | 3.53 |
| 1975 | 257.4 | 33.83 | 147 | 3.26 | 359 | 2.99 |
| 1976 | 254.7 | 31.10 | 180 | 2.40 | 333 | 2.52 |
| 1977 | 264.4 | 33.17 | 162 | 3.09 | 388 | 3.73 |
| 1978 | 259.1 | 29.43 | 171 | 2.99 | 333 | 2.51 |
| 1979 | 268.8 | 30.86 | 167 | 3.30 | 337 | 2.21 |
| 1980 | 267.2 | 32.50 | 178 | 2.74 | 390 | 3.78 |

## EFFECTS OF OVERWORK ON ROOKIE PITCHERS - PART II

by Dick O'Brien

Dallas Adams' article in the August 1982 issue of the *Analyst* really piqued my interest. So much so that I wanted to see if age as a rookie might have any significant affect in reaching the conclusions that were so clearly evident. Having considerably more time on my hands to do so than he, I researched his findings one step further by extending the data base to cover the period 1960-1977, and categorizing the age groups in two year increments.

The results were surprising to me. The younger the rookie pitcher is as a deb, the less rapid is his decline from overwork when he appears in 39 or fewer games. Or so it appears. The data base for the 19 year-olds and younger is probably too small to make any sound conclusions about its career longevity factor.

TABLE I

MAJOR LEAGUE INNINGS PITCHED IN ROOKIE, SECOND, THIRD AND FOURTH YEARS

| GAMES AS ROOKIE | ROOKIE YEAR INNINGS | SECOND YEAR INNINGS | THIRD YEAR INNINGS | FOURTH YEAR INNINGS |
|---|---|---|---|---|
| under 40 | | | | |
| -19 | 1407 | 1721 | 1418 | 1582 |
| 20-21 | 10028 | 10152 | 9793 | 10324 |
| 22-23 | 18304 | 20926 | 18671 | 19025 |
| 24-25 | 14163 | 13201 | 12098 | 11461 |
| 26-27 | 6104 | 4485 | 3619 | 3060 |
| 28-29 | 2608 | 1464 | 1060 | 809 |
| 30- | 81 | 62 | 0 | 0 |
| all ages | 52695 | 52011 | 46659 | 46261 |
| 40 - 59 | | | | |
| -19 | 128 | 318 | 208 | 362 |
| 20-21 | 993 | 651 | 602 | 345 |
| 22-23 | 3107 | 2491 | 2833 | 2460 |
| 24-25 | 5230 | 4315 | 4157 | 3527 |
| 26-27 | 4377 | 3952 | 3566 | 2711 |
| 28-29 | 653 | 481 | 198 | 98 |
| 30- | 350 | 336 | 302 | 229 |
| all ages | 14838 | 12544 | 11866 | 9732 |
| 60 - 79 | | | | |
| -19 | 0 | 0 | 0 | 0 |
| 20-21 | 237 | 99 | 94 | 108 |
| 22-23 | 825 | 828 | 834 | 782 |
| 24-25 | 754 | 664 | 517 | 766 |
| 26-27 | 1440 | 745 | 565 | 392 |
| 28-29 | 0 | 0 | 0 | 0 |
| 30- | 0 | 0 | 0 | 0 |
| all ages | 3256 | 2336 | 2010 | 2048 |

Based on the findings in Table I, it would appear that ages 22-23 are the optimum ones in charting a big league career. When the totals in years three and four are occasionally greater than those in years one and two, it occurs at a young age when one's recuperative powers are considerably better than at a more advanced age.

TABLE II
MAJOR LEAGUE INNINGS PITCHED IN ROOKIE, SECOND, THIRD AND FOURTH YEARS, NORMALIZED TO ROOKIE INNINGS

| GAMES AS ROOKIE | ROOKIE YEAR INNINGS | SECOND YEAR INNINGS | THIRD YEAR INNINGS | FOURTH YEAR INNINGS |
|---|---|---|---|---|
| under 40 | | | | |
| -19 | 1.000 | 1.223 | 1.008 | 1.124 |
| 20-21 | 1.000 | 1.012 | .977 | 1.030 |
| 22-23 | 1.000 | 1.143 | 1.020 | 1.039 |
| 24-25 | 1.000 | .932 | .854 | .809 |
| 26-27 | 1.000 | .735 | .593 | .501 |
| 28-29 | 1.000 | .561 | .406 | .310 |
| 30- | 1.000 | .765 | .000 | .000 |
| all ages | 1.000 | .987 | .885 | .878 |
| 40 - 59 | | | | |
| -19 | 1.000 | 2.480 | 1.630 | 2.830 |
| 20-21 | 1.000 | .656 | .606 | .347 |
| 22-23 | 1.000 | .802 | .912 | .792 |
| 24-25 | 1.000 | .825 | .795 | .674 |
| 26-27 | 1.000 | .903 | .815 | .619 |
| 28-29 | 1.000 | .737 | .303 | .150 |
| 30- | 1.000 | .960 | .862 | .654 |
| all ages | 1.000 | .845 | .800 | .656 |
| 60 - 79 | | | | |
| -19 | xxx | xxx | xxx | xxx |
| 20-21 | 1.000 | .418 | .397 | .456 |
| 22-23 | 1.000 | 1.004 | 1.011 | .948 |
| 24-25 | 1.000 | .881 | .686 | 1.012 |
| 26-27 | 1.000 | .517 | .392 | .272 |
| 28-29 | xxx | xxx | xxx | xxx |
| 30- | xxx | xxx | xxx | xxx |
| all ages | 1.000 | .717 | .617 | .629 |

I question the advisability of carying the data base much before 1960. Relief pitching as a specialty is a rather recent innovation, and when one digs back in the pre-World War II records pitchers usually came on the major league scene at an appreciably older age than they do today.

## Player Development Study

In 1981, Bill James did a study of what each farm system produced in value of performance in the 1980 season. James used his Value Approximation Method (VAM) to measure the 1980 value of each performer, a measure that would credit the production of a Mike Schmidt as considerably more valuable than the production of a Tito Landrum. In his study a player was considered the product of the first organization with which he reached the AAA or major league level, or was included in a major league transaction.

This ranking of the 1980 farm system production is the first column on the accompanying chart. The two most recent expansion teams, Seattle and Toronto, have been deleted from this study as their player development programs have not really had a chance to bear fruit yet.

The basic assumption in this study is that there is a positive correlation between the quality of the player development program and its product's performance in 1980. Assuming that present major league value is a reflection of a farm system's recent past, I went back five years (1976) to count the number of full-time free agent scouts, part-time scouts, A and Rookie League teams, and which organizations were members of the Major League Scouting Bureau.

I broke the twenty-four organizations into thirds - top, middle, and bottom in VAM points produced. Membership in the Major League Scouting Bureau (MLSB) was evenly distributed through the groups, 5 of 8 in each group were members. The teams with MLSB averaged eight full-time, and six part-time free agent scouts. Teams without MLSB averaged 14 full-time and six part-time free agent scouts. By this measure, to bow out of the MLSB which costs $120,000 a year would require hiring six extra full-time free agent scouts. At 1981 figures, that would cost 93 to 132 thousand in salaries alone.

The only advantage suggested by this study in belonging to the MLSB is the dollar savings in being able to get by with fewer full-time free agent scouts. That dollar savings is not that great. If a player development program were to ever decide to drop out of the MLSB, the extra cost to replace the service would not be that great. The information provided by adding extra scouts under more specific direction might actually improve.

With membership in the MLSB evenly distributed between the three groups, we can study the number of scouts employed, without worrying about controlling team membership in the MLSB. There is a very clear trend as a positive correlation between 1980 value produced and the number of free agent scouts.

The four teams employing more than 15 full-time scouts rank 4th, 5th, 6th, and 10th among the 24 teams. The bottom third in production employed fewer full-time and part-time scouts than the other two groups. The middle group employed, on the average, one more full-time scout, but six fewer part-time scouts than the top group. The 1976 averages looked like this:

| Average | Full-time | Part-time |
|---|---|---|
| Top | 11.6 | 10.2 |
| Middle | 12.5 | 3.8 |
| Bottom | 7.6 | 3.6 |
| Total | 10.6 | 5.9 |

The emphasis on part-time scouts in the top group is interesting. Only three teams employed ten or more part-time scouts in 1976, and those three teams ranked 1st, 4th, and 6th in farm production.

There has been a growth since 1976 in full-time scouts employed (14.5, up from 10.6), but the number of part-time scouts has grown phenomenally, nearly doubled (11.3, up from 5.9). Fourteen organizations, better than half the teams, employ ten or more part-time scouts now.

The correlation of the number of farm teams in 1976 with the production of talent is equally dramatic. The difference between the number of farm teams at the A and Rookie League level among the top two groups is significantly higher than the bottom group. In fact, all eight organizations in the bottom grouping were among the 12 teams with two or fewer clubs at the A and Rookie League level.

It is interesting to note that the growth of scouting staffs and farm teams since 1976 has been greatest among the bottom group in production. Call it natural selection. As mentioned, 12 teams had two or less low farm clubs in 1976. Today only four teams have two clubs at that classification.

Since 1976, two teams have cut back from four to three clubs at this level. Eleven have increased to three clubs or more. <u>None</u> have cut back from three to two. Clearly the minimal number should be three. To stick with two lower classification clubs is to ask for trouble.

The major recommendation of this study for a major league ballclub would be this. The decision of whether to continue with or to droup out of the Major League Scouting Bureau is not a financial decision. The costs are roughly similar. The decision should be taken based on the system's judgment as to which path would produce better information. But the size of an organization's scouting staff, both full-time and part-time, does correlate positively with farm production. It would be a dangerous financial decision for any organization to try to scrape by on a small staff. Further, past history shows that a productive farm system needs three clubs at the lower classifications to be productive in 88% of the cases.

This information deleted to protect the confidential source.

| Farm System | 1980 VAM Pts Produced | 1976 Full-time/ Part-time Scouts | 1976 A and Rookie League Teams | 1976 Member MLSB | 1982 A and Rookie League Teams | 1982 Member MLSB |
|---|---|---|---|---|---|---|
| | 242 | 8 - 40 | 4 | yes | 3 | yes |
| | 235 | 2 - 3 | 2 | yes | 4 | yes |
| | 214 | 9 - 2 | 3 | no | 5 | no |
| | 203 | 19 - 16 | 3 | no | 4 | no |
| | 202 | 24 - 0 | 3 | yes | 3 | yes |
| | 186 | 20 - 18 | 3 | no | 3 | no |
| | 184 | 7 - 1 | 2 | yes | 3 | yes |
| | 179 | 4 - 2 | 2 | yes | 2 | yes |
| | 171 | 6 - 4 | 2 | yes | 4 | yes |
| | 168 | 16 - 5 | 2 | yes | 4 | yes |
| | 166 | 29 - 0 | 3 | no | 4 | no |
| | 159 | 8 - 2 | 2 | no | 2 | no |
| | 158 | 12 - 4 | 2½ | yes | 3 | yes |
| | 155 | 8 - 5 | 4 | yes | 3 | yes |
| | 154 | 9 - 6 | 3 | yes | 3 | yes |
| | 151 | 12 - 4 | 3 | no | 3 | no |
| | 149 | 6 - 1 | 2 | yes | 2 | yes |
| | 144 | 4 - 3 | 2 | yes | 4 | yes |
| | 133 | 8 - 7 | 2 | no | 4 | no |
| | 115 | 12 - 4 | 2 | no | 3 | no |
| | 114 | 5 - 1 | 2 | yes | 2 | yes |
| | 106 | 5 - 5 | 2 | yes | 4 | no |
| | 104 | 8 - 4 | 2 | yes | 5 | yes |
| | 103 | 13 - 4 | 1½ | no | 4 | no |
| Average | 162.2 | 10.6/5.9 | 2.5 | 63% | 3.4 | 58% |

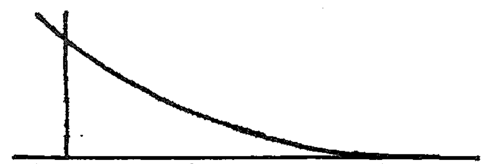
1981 listing of scouts was not done team by team.

The number of full-time scouts rose to 14.5; 11.3 for part-time.

We will close, this time, with a comment from the editor. Ward Larkin observes on page 9 that "there is no reason to believe that (batting averages) do not have a normal distribution". I am on record as saying, and I want to go on record again as saying, that I have never seen anything in baseball which follows a normal distribution curve. Ward doesn't find the reason to believe that batting averages are not normally distributed, I think, for the usual reason--that he expects the distribution to be "normal", and therefore doesn't really look.

My belief is that major league baseball players represent the far right-hand slice of the bell-shaped curve--the best taken from a large bell-shaped population:

BASEBALL TALENT of ENTIRE POPULATION



TALENT of MAJOR LEAGUES

In any one specific performance area, such as batting average, you won't get a sharp "line of non-existence"--a Mendoza line--which causes the distribution to look close enough to bell-shaped to be mistaken for it.

I would note, for example, that in Ward's chart of leaders and lowest in S.D. of B. Average, the highest is further above the mean than the lowest is below it in 47 of the 80 years; the lowest is further away only 32 times, and the distance is equal once. This would be predicted by my theory of the distribution of baseball talent. I would also point out that this occurs despite the fact that Ward uses a very low cut-off point for inclusion in the study, 150 at bats. I would bet that if he re-did the study with a cut-off point of 300 at bats, this would be a more pronounced pattern; 450 at bats, more pronounced yet; and if the study were done for eras and using a cut-off point of 3000 or 4000 at bats, more pronounced yet.

But my main point is that one should not do--and Ward hasn't--a study which simply assumes that talent is normally distributed, because that is exactly the sort of misconception that can lead you to a false conclusion.