AN OFFENSIVE STATISTIC FOR BASEBALL

BY

THOMAS M. COVER
CARROLL W. KEILERS

TECHNICAL REPORT NO. 4
NOVEMBER 20, 1972

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GK-34363

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
AN OFFENSIVE STATISTIC FOR BASEBALL

by

Thomas M. Cover
Carroll W. Keilers

TECHNICAL REPORT NO. 4
NOVEMBER 20, 1972

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT GK-34363

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
An Offensive Statistic for Baseball

Thomas M. Cover†
Carroll W. Keilers‡

1. Abstract

This paper introduces a baseball statistic that plays the role of an offensive earned run average (OERA), which is useful as a way to evaluate hitters. The OERA of an individual is simply the number of earned runs per game that he would score if he batted in all nine positions in the line-up and can be evaluated by following the sequence of at bats of a given batter. This statistic has the obvious natural interpretation and tends to evaluate strictly personal rather than team achievement. Some theoretical properties of the statistic are developed, and we give our answer to the question, "Who is the greatest hitter in baseball history?"

†Stanford University. On sabbatical leave 1971-72, Harvard and M.I.T.
‡Stanford University and Probe Systems.
2. Introduction

There are many factors that make a perfect offensive statistic an unreasonable goal. In particular, there is always a correlation between the team situation and the batter's performance. For example, some of these situations affect the pitching strategy of the opposing team; others influence batting objectives and lead to sacrifice hits or high strike-out percentages. In addition to factors introduced by the team situation, there are certainly subjective differences between players that will be important to managers and coaches. This paper develops the idea of an offensive earned run average under the constraint that the statistic must be simple to understand and easy to calculate. The OERA computation is confined to the application of a simple set of deterministic rules to the sequence of outcomes generated by a particular batter. There are no subjective aspects of the computation.

Existing offensive baseball statistics have certain limitations. Total hits, for example, depend on the number of times that a batter has a chance to perform. Clearly it is a great personal achievement to hit 45 home runs or have 125 RBI's, but in a given situation a manager would prefer to use the player with the highest probability of success. Batting average has the disadvantage that neither power hitting nor the ability to get on base with a walk are considered. Slugging average ignores walks and uses an arbitrary weight for the importance of each kind of hit. The OERA calculation takes into account the probability of each possible batting outcome and evaluates its contribution toward production of earned runs. An interesting result of the computation is the indication that the slugging average exaggerates the importance of extra-base hits, and that the base-on-balls, which is ignored, has almost one half
the value of a single. While the OERA is a normalized average, it has the direct interpretation as an expected number of runs scored. In this respect, it avoids the difficulty of interpreting the importance of the difference between two averages. For example, a .400 batting average is certainly better than a .300 batting average, but it is not obvious how much better it is. An expected score of 5.0 runs per game is 1.0 run better than a 4.0 average, and the implications of this difference are clear. Also, comparison of long-ball and singles hitters is possible.

There have been several very sophisticated attempts at measuring a batter's contribution to run production. The best, in our opinion, is that of Lindsey [1]. He suggests the calculation of a batter's incremental contribution to run production above the league average, where the batter is dropped at random into a random spot in a random game. Empirical statistics gathered from league play are used. Once a simple correction to include the effects of walks and outs is made (which Lindsey apparently omits), the statistic should be an extremely effective measure of a batter's strength. In the context of section 6, Lindsey's modified statistic is given by \( \Delta = P_L(Q_B E_L + R_B - E_L) \), where \( E_L \) is the vector of expected runs associated with the various states of the inning evaluated from empirical league play, \( Q_B \) and \( R_B \) are the state transmission matrix and incremental runs vector derived from the batter's statistics, and \( P_L \) is the probability distribution vector on the 24 states of the inning averaged over league play. The sophistication of the statistic prohibits a simple play-by-play calculation. Also, the result is explicitly league dependent. Lindsey has made additional detailed investigations of statistical data [2] and scoring [3].

Earnshaw Cook [4] presents a "Scoring Index" that is related to run
production in an empirical way. It seems to have few mathematically well defined properties. Lindsey presents a more complete review of Cook [5]. The Mills Brothers [6] have defined a "player win average" that measures the average contribution per at bat to the probability of winning a game. This statistic is highly dependent on the play of one's teammates, perhaps more so than RBI's. It is perhaps a better measure of the level of performance when significant changes in the outcome of the game are possible (i.e. clutch factor) than of offensive power. For additional applications of mathematical approaches to baseball, see also [7], [8], [9], [10], [11], [12], [13], [14], and [15].

3. Definitions and Conventions

The OERA of an individual is simply the number of earned runs per game that he would score if he batted in all nine positions in the line-up and if average baserunners were used. For example, if a player bats on five successive occasions and the outcomes are "out", "walk", "out", "home run", "out", the number of runs scored is two, and the elapsed "game time" is one inning. Over many successive innings of actual play, this player will accumulate several innings of personal "game-time". If, after many at bats, the batter has scored 70 runs in 100 personal innings, his OERA will be \( \frac{70}{100} \times 9 = 6.3 \) runs per game.

It is desirable to compute the numbers of runs per inning according to a specific convention to achieve a simple, deterministic statistic. We have arbitrarily used the following conventions:

1. Sacrifices are not counted at all in OERA.
2. Errors are counted as outs.
3. Runners do not advance on outs.
4. All singles and doubles are assumed to be long. That is, a single advances a baserunner two bases, and a double scores a runner from first base.

5. There are no double plays.

These five rules are not intended to summarize the significance of errors, sacrifices, double plays, or the difference between a long single and a scratch single. They are used simply to make the calculations completely deterministic and independent of scorer judgment, whether the computation is based on observation of the game or on score sheets. However, these five rules can be defended in several ways.

First, the errors should be counted as outs because we are computing an earned run average. Second, when the batter is ordered by his coach to sacrifice in a game situation, there may not be a runner in a position to advance in his personal situation. These two conventions are clearly justified and are consistent with the scoring used for other personal statistics such as batting average and RBI's.

The remaining three conventions are justified on the basis that they do not significantly change the expected number of scored runs. This behavior is caused by several factors. First, the effect of the length of a single on advancing a runner is lowered by the fact that a "short-long" combination will score a man from first as easily as a "long-long" combination. Second, short doubles are relatively rare. Third, although double plays and advances on outs (whether officially scored as sacrifices or not) are very significant in run production, the convention of eliminating both creates offsetting effects. Finally, analysis by computer (discussed in the following paragraph) confirms that OERA is not greatly affected by reasonable changes in the runner-advance conventions.
We therefore have a simple deterministic rule directly computable from score sheets. This measure of batting ability may also be computed from the player's batting statistics. The computation is performed by taking a batter's cumulative statistics (at bats, walks, singles, doubles, triples, home runs, and outs), and computing the probability of each possible hitting outcome. Then, these probabilities are used to generate the expected number of runs for the batter averaged over all possible sequences of hitting performance generated by these statistics. The algorithm is described in Section 6. This average, which requires the specification of only six integer statistics, converges to the play-by-play average associated with the specific record of a batter, under the assumption of independent identically distributed at bats. Thus the computer can generate the OERA from cumulative statistics when play-by-play records are unobtainable. Also, variations in the value of the statistic caused by changes in the runner-advance rules or by slight perturbations of cumulative averages can be evaluated on the computer. The next section supplies an extended example of a play-by-play computation of the statistic.

4. **Example of Play-by-Play Calculation**

The ease of calculation of the OERA is the primary reason this paper was written. The following example illustrates the computation of the statistic.

During the first 11 games of the 1971 season (April 6 through April 19), Reggie Smith of the Boston Red Sox achieved the following sequence of outcomes:
The semicolons denote the breaks between actual games played. For example, Smith had 2 singles and a double in 4 at bats in the opening game, April 6, 1971. The slashes occur every time the batter accumulates 3 outs and denote the breaks between the batter's personal innings. Thus, using the runner advance conventions of Section 3, Smith scores 2 runs in his first personal inning. Now ignoring the sacrifice outs and the incomplete personal inning we have 9 runs scored in 10 personal innings for an average as of April 19, 1971 of .9 runs per inning, resulting in an

\[ (4.2) \quad \text{OERA} = 8.1 \text{ runs/game.} \]

This completes the play-by-play calculation. Smith's empirical statistics through the last slash mark are

\[
\begin{array}{ccccccc}
\text{AB} & \text{H} & 1 & 2 & 3 & 4 & B \\
43 & 13 & 9 & 3 & 0 & 1 & 4,
\end{array}
\]

yielding

\[
p_0 = \frac{30}{47}, \quad p_B = \frac{4}{47}, \quad p_1 = \frac{9}{47},
\]

\[
p_2 = \frac{3}{47}, \quad p_3 = 0, \quad p_4 = \frac{1}{47}.
\]

* Full calculation being planned for final draft.
Computer calculation of the OERA, using the analysis of section 6, yields a computed average of 5.92 runs per game.

The difference between the play-by-play OERA and the computed OERA is almost entirely attributable to the actual order of occurrence of these events. The figures would be identical, except for end effects of partially complete innings, if Smith had generated his outcomes in all possible orders, with replacement, and we had averaged these results. The rather large disagreement of 2.2 runs between the two methods of computation would be typical of small samples only. Longer samples will generally smooth out the effect of big personal innings such as the six run sequence in this example.

5. Possible Modifications of the Conventions

Although the conventions described in Section 3 have the virtue of simplicity, there are many modifications which can be made by a trained scorer. Examples are as follows:

1. Distinguish sacrifice flies and sacrifice bunts.
2. Distinguish short and long singles and doubles.
3. Identify double play balls.

These subjective decisions on the part of the scorer may also be simulated in the computer calculation of the OERA by assigning conditional probabilities to these events.

These performance variations can then be used to explain deterministic data. For example, computer evaluation of the effect of modifications 1, 2, and 3 for baserunner advance on OERA has shown that the difference between using realistic sacrifice, long single, and double play probabilities and using deterministic rules is less than 10% and
is typically 5%. Although these changes are possibly significant in the absolute sense, it is clear that they will not greatly perturb relative rankings of players.

6. Computation of OERA

It is envisaged that the OERA will be calculated in a play-by-play fashion as the batter progresses through the season. However, this section will develop an explicit formula for the OERA in terms of the batting statistics.

Six integer values form the smallest complete set of statistics. The most convenient complete set is

1) Number of Official Times at Bat
2) Number of Singles
3) Number of Doubles
4) Number of Triples
5) Number of Home Runs
6) Number of Bases on Balls.

This data is given, for example, in the Baseball Encyclopedia [16].

From this, we compute a probability spectrum \((p_0, p_B, p_1, p_2, p_3, p_4)\), where \(p_0\) = probability of an out, \(p_B\) = probability of a base on balls, and \(p_1\) = probability of an i-base hit. The state of a baseball inning is described by the number of outs \((0,1,2,3)\) and the positions of men on base \((8\) possibilities). Ignoring the 3-out state, there are 24 such states. Let \(s \in \{1,2,\ldots,24\}\) denote any enumeration of these states with state \(s = 1\) corresponding to no outs and no men on base.

Given a state \(s \in \{1,2,\ldots,24\}\) and a hit \(H \in \{0,B,1,2,3,4\}\), let \(s' = f(H,s)\) denote the new state following the hit, and let \(R(H,s)\) denote the number of runs scored by the hit. We shall need the state transition function
\[(6.1) \quad p(s'|s) = \sum_{H : f(H,s) = s'} p_H \]

and the expected number of runs scored in one at bat, given state \(s\),

\[(6.2) \quad R(s) = \sum_H p_H R(H,s). \]

For example, if \(s = 2\) is the state: 1 man out, man on 2nd; then

\[(6.3) \quad p(2|2) = p_2 \]

and

\[(6.4) \quad R(2) = p_1 + p_2 + p_3 + 2p_4. \]

Also, we have \(p(1|1) = p_4\) and \(R(1) = p_4\), where it will be recalled that \(s = 1\) corresponds to no outs, no men on base.

The following formulae are probably equivalent to those used but not stated in Howard's baseball model [10]. Let \(E(s)\) denote the expected number of runs scored in an inning beginning in state \(s\). Then \(E(s)\) must satisfy

\[(6.5) \quad E(s) = \sum_H p_H \left( E(f(H,s)) + R(H,s) \right) \]

Defining \(Q = (Q_{ij}) = p(j|i)\), we have the equation

\[(6.6) \quad E = QE + R \]

where \(Q\) is a matrix of probabilities but is not stochastic because the three out state is eliminated. \(Q\) is a twenty-four by twenty-four matrix, and \(E\) and \(R\) are vectors of length twenty-four. The solution is
\[
E = \sum_{k=0}^{\infty} Q^k R = (I-Q)^{-1} R
\]

The expected number of runs scored in an inning is given by the first term \( E(1) \) of \( E \), corresponding to beginning the inning with no outs and no men on base. The expected number of runs per game is simply

\[ OERA = 9E(1). \]

The computer easily evaluates (6.7) (20 terms—corresponding to a 20 batter inning—is enough) in the form \( R_k = QR_{k-1}, \quad R_0 = R, \)

\[ E = \sum_{k=0}^{\infty} R_k. \]

This completes the computation. However, we have also used a more ambitious program that yields the probability distribution of runs scored in an inning as well as allowing 9 different batters in the lineup. A description may be obtained from the authors. If a player behaves as a sequence of independent, identically distributed random variables, his personal sequence of outcomes will produce a deterministic OERA that converges to the computer generated value, when the same conventions are used.

7. Pure Performance Bounds

In this section we wish to develop some expressions for the OERA for a batter with a given spectrum \((p_0, p_B, p_1, p_2, p_3, p_4)\), where \( p_0 \) = probability of an out, \( p_B \) = probability of a base on balls, and \( p_i \) = probability of an \( i \)-base hit, \( i = 1, 2, 3, 4 \). In general, the formula for the expected number of runs is an expression involving approximately 100 terms. However, in certain pure cases the formula is simple and informative.
Let us first examine the all homerun hitter. For this batter, 
\[ p_4 = 1 - p_0, \quad \text{and} \quad p_B = p_1 = p_2 = p_3 = 0. \] 
The on-base percentage is 
\[ 1 - p_0 = q_0. \] 
Clearly the number of runs per inning generated by such a batter will be 3 fewer than the number of men up. 
The number of men up in one inning is a random variable \( N \) having a negative binominal distribution given by

\[
\Pr(N=1) = \binom{1-1}{2} p_0^3 q_0^{i-3}, \quad i = 3, 4, \ldots.
\] 

(The number of batters in an inning depends only on \( p_0 \) and not on the quality of the hits.) Note that

\[
\sum_{i=3}^{\infty} \Pr(N=i) = \Pr(N \geq 3) = 1.
\]

The expected number of men coming to bat in the inning is given by

\[
E(N) = \sum_{i=3}^{\infty} i \Pr(N=i) = 3/p_0.
\]

Since 3 of the batters are outs, the expected number of runs scored per inning for an all homerun hitter is simply given by

\[
R_4 = 3/p_0 - 3 = 3q_0/p_0.
\]

The general case for pure hitters is also straightforward. Let the random variable \( N \) denote the number of batters in an inning. Let

\[ R_B, R_1, R_2, R_3, R_4 \]

denote the expected number of runs per inning generated by a batter with on-base-percentage \( q_0 = 1 - p_0 \), all of whose safe at bats are of the type designated by the subscript. For any real number \( t \), define
\[(7.4) \quad (N-t)^+ = \begin{cases} N-t, & N-t \geq 0 \\ 0, & N-t < 0 \end{cases} \]

Now the all-walks hitter scores only \((N-6)^+\) runs in an inning in which \(N\) men come to bat, because in addition to making 3 outs he must always load the bases before he begins to score runs. Similarly, the all-singles hitter scores \((N-5)^+\) runs, and the all-doubles or all-triples hitter scores \((N-4)^+\) runs. Thus we have

\[R_B = E(N-6)^+ \]
\[= \sum_{i=7}^{\infty} (i-6) \Pr[N=i] \]

\[R_1 = E(N-5)^+ = \sum_{i=6}^{\infty} (i-5) \Pr[N=i] \]

\[R_2 = R_3 = E(N-4)^+ = \sum_{i=5}^{\infty} (i-4) \Pr[N=i] \]

\[R_4 = E(N-3)^+ = \sum_{i=4}^{\infty} (i-3) \Pr[N=i], \]

where \(N\) is distributed as in Eq. (7.1). Calculation of these quantities yields the following results, depicted in Fig. 7.1.

All walks: \[R_B = 3/p_0 - 6 + 3p_0^3(1+2q_0+2q_0^2)\]

All singles: \[R_1 = 3/p_0 - 5 + 3p_0^3q_0 + 2p_0^3\]

\[(7.6) \]

All doubles \[R_2 = R_3 = 3/p_0 - 4 + p_0^3\]

All triples: \[R_4 = 3/p_0 - 3.\]
Figure 7-1. Pure Performance Bounds
Finally, the expected number of runs per game is simply $9R_1$, by the linearity of the expectation operation.

Since the number of men left on base equals the number of batters minus the number of outs minus the number of runs, taking expectations yields the expected number of men left on base:

\[
L_i = \frac{3}{p_0} - 3 - R_i
\]

where $R_i$ is given in equation (7.6).

The actual probability distribution of the number of runs scored in an inning is easily obtained by calculation of the distribution of $(N-t)^+$ from inspection of Eq. (7.1). For example, the all-singles hitter has

\[
\Pr[R=0] = p_0^3(1+3q_0+6q_0^2)
\]

\[
\Pr[R=n] = \binom{n+4}{2} p_0^3 q_0^n, \quad n=1,2,\ldots
\]

Incidentally this suggests a lumped negative binomial model rather than the negative binomial model suggested by Lindsey [2, App. A].

It is especially interesting to note the behavior of the OERA for small values of the on-base-percentage $q_0$. Expanding $3/(1-q_0) = 3 + 3q_0 + 3q_0^2 + \ldots$, we have

\[
R_0 = 15q_0^4 + 0(q_0^5) = \binom{6}{2} q_0^4 + \ldots
\]

\[
R_1 = 10q_0^3 + 0(q_0^4) = \binom{5}{2} q_0^3 + \ldots
\]

\[
R_2 = R_3 = 6q_0^2 + 0(q_0^3) = \binom{4}{2} q_0^2 + \ldots
\]

\[
R_4 = 3q_0 + 0(q_0^2) = \binom{3}{2} q_0 + \ldots
\]
It should be noted that the leading term in each case is simply the probability that the inning generates precisely one run. We observe that a weak player (or a weak team) has its weaknesses magnified in the sense that given the same \( q_0 \) the power hitter has a better exponent \( k \) in his run production \( \left( \frac{2+k}{2} \right)^k \).

At the other extreme, for high batting averages \( q_0 \), we see that \( R_i \sim \frac{3}{(1-q_0)} \to \infty \) as \( q_0 \to 1 \), \( i = B,1,2,3,4 \). However, the differences in the expected number of runs converge to constants

\[
\begin{align*}
R_4 - R_3 & \to 1 \\
R_4 - R_2 & \to 1 \\
R_4 - R_1 & \to 2 \\
R_4 - R_B & \to 3
\end{align*}
\]

(7.10)

in the limit as \( q_0 \to 1 \). Thus, for example, an all-singles batter will generate approximately 1 run per inning more than an all-walks batter, in the limit as the modified batting average \( q_0 \) tends to 1.

Finally, the convexity of OERA as a function of \( q_0 \) suggests that it is better to have a team composed of stars and weak batters than a team of mediocre batters. A precise line-up evaluation is the only final answer to this question.

8. The Greatest Hitters

One of the enjoyable aspects of baseball is developing an evaluation system and using it to rank players. The OERA has been applied to a list of candidates to discover the greatest offensive career and the greatest offensive season. The rank is according to the run-producing capability of a line-up composed of identical players with the relevant hitting statistics. Figure 8-1 shows OERA versus slugging average for
the lifetimes (through 1968) of several players, as reported by the

*Baseball Encyclopedia* [16]. Table 8-I lists the precise value of OERA
and slugging average for the top ten players, along with their normal
hitting statistics. The players were selected for this evaluation if
they were mentioned in the top twenty entries in any of the following
lifetime categories: batting average, slugging average, home run per-
centage, runs batted in per game, home run total, or RBI total. Figure
8-2 and Table 8-II show OERA versus slugging average and a table of
statistics for single season performances. Selection of candidates was
from the top twenty entries in the same six performance categories for
single season performances. Three players from the 1971 San Francisco
Giants are included to illustrate the relationship of typical, good
performances to the entries from the record book.

It is clear from the lifetime figure that Ted Williams and Babe
Ruth dominate this category. They also dominate the single season records
when the best year of each player is used. However, some other players
have best years as good as some of Williams' and Ruth's average years.
It should be added that unlisted years of some of the great batters
could appear in the rankings even though they did not qualify on any of
the listings used to generate candidates. It is interesting to note
that, although there is a correlation between slugging average and OERA,
the dispersion of values is as great as six runs per game.

Figure 8-3 presents a comparison between the top seven lifetime
averages, taken from Figure 8-1, and several 1971 season performances.
OERA is plotted versus batting average. The 1971 players were chosen
from two groups: the top ten batting averages from the National League
and fourteen randomly chosen NBL players with three hundred or more at
Figure 8-1. Lifetime OERA versus Slugging Average
## TABLE 8-1

### Lifetime Statistics - Top Ten

<table>
<thead>
<tr>
<th>Players</th>
<th>At Bats</th>
<th>Hits</th>
<th>2B</th>
<th>3B</th>
<th>HR</th>
<th>BB</th>
<th>AV</th>
<th>OBP*</th>
<th>SLGA**</th>
<th>OERA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babe Ruth</td>
<td>8399</td>
<td>2873</td>
<td>506</td>
<td>136</td>
<td>714</td>
<td>2056</td>
<td>.342</td>
<td>.471</td>
<td>.690</td>
<td>12.97</td>
</tr>
<tr>
<td>Ted Williams</td>
<td>7706</td>
<td>2654</td>
<td>525</td>
<td>71</td>
<td>521</td>
<td>2018</td>
<td>.344</td>
<td>.480</td>
<td>.634</td>
<td>12.90</td>
</tr>
<tr>
<td>Lou Gehrig</td>
<td>8001</td>
<td>2721</td>
<td>535</td>
<td>162</td>
<td>493</td>
<td>1508</td>
<td>.340</td>
<td>.445</td>
<td>.632</td>
<td>10.96</td>
</tr>
<tr>
<td>Jimmie Foxx</td>
<td>8134</td>
<td>2646</td>
<td>458</td>
<td>125</td>
<td>534</td>
<td>1452</td>
<td>.325</td>
<td>.427</td>
<td>.609</td>
<td>9.94</td>
</tr>
<tr>
<td>Hank Greenberg</td>
<td>5193</td>
<td>1628</td>
<td>379</td>
<td>71</td>
<td>331</td>
<td>852</td>
<td>.313</td>
<td>.410</td>
<td>.605</td>
<td>9.20</td>
</tr>
<tr>
<td>Mickey Mantle</td>
<td>8102</td>
<td>2415</td>
<td>344</td>
<td>72</td>
<td>536</td>
<td>1734</td>
<td>.298</td>
<td>.422</td>
<td>.557</td>
<td>9.15</td>
</tr>
<tr>
<td>Stan Musial</td>
<td>10972</td>
<td>3630</td>
<td>725</td>
<td>177</td>
<td>475</td>
<td>1599</td>
<td>.331</td>
<td>.416</td>
<td>.559</td>
<td>8.94</td>
</tr>
<tr>
<td>Ty Cobb</td>
<td>11437</td>
<td>4192</td>
<td>725</td>
<td>294</td>
<td>118</td>
<td>1249</td>
<td>.367</td>
<td>.429</td>
<td>.512</td>
<td>8.92</td>
</tr>
<tr>
<td>Billy Hamilton</td>
<td>6268</td>
<td>2158</td>
<td>242</td>
<td>94</td>
<td>40</td>
<td>1187</td>
<td>.344</td>
<td>.449</td>
<td>.432</td>
<td>8.79</td>
</tr>
<tr>
<td>Tris Speaker</td>
<td>10205</td>
<td>3514</td>
<td>793</td>
<td>244</td>
<td>115</td>
<td>1381</td>
<td>.344</td>
<td>.422</td>
<td>.500</td>
<td>8.51</td>
</tr>
</tbody>
</table>

*OBP = On Base Percentage

**SLGA = Slugging Average
Figure 8-2. Single Season OERA versus Slugging Average
<table>
<thead>
<tr>
<th>Players</th>
<th>At Bats</th>
<th>Hits</th>
<th>2B</th>
<th>3B</th>
<th>HR</th>
<th>BB</th>
<th>OBRA</th>
<th>SLGAA*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ted Williams '41</td>
<td>456</td>
<td>185</td>
<td>33</td>
<td>3</td>
<td>13</td>
<td>45</td>
<td>5.49</td>
<td>.735</td>
</tr>
<tr>
<td>Babe Ruth '23</td>
<td>522</td>
<td>205</td>
<td>45</td>
<td>37</td>
<td>41</td>
<td>54</td>
<td>5.42</td>
<td>.764</td>
</tr>
<tr>
<td>Babe Ruth '20</td>
<td>458</td>
<td>172</td>
<td>36</td>
<td>9</td>
<td>16</td>
<td>54</td>
<td>5.28</td>
<td>.847</td>
</tr>
<tr>
<td>Ted Williams '57</td>
<td>430</td>
<td>204</td>
<td>44</td>
<td>16</td>
<td>14</td>
<td>59</td>
<td>5.09</td>
<td>.846</td>
</tr>
<tr>
<td>Babe Ruth '21</td>
<td>420</td>
<td>163</td>
<td>28</td>
<td>1</td>
<td>11</td>
<td>44</td>
<td>5.00</td>
<td>.731</td>
</tr>
<tr>
<td>Babe Ruth '24</td>
<td>455</td>
<td>200</td>
<td>40</td>
<td>5</td>
<td>47</td>
<td>54</td>
<td>5.13</td>
<td>.793</td>
</tr>
<tr>
<td>Rogers Hornsby '24</td>
<td>536</td>
<td>227</td>
<td>43</td>
<td>14</td>
<td>25</td>
<td>47</td>
<td>5.10</td>
<td>.735</td>
</tr>
<tr>
<td>Hugh Duffy '94</td>
<td>539</td>
<td>237</td>
<td>51</td>
<td>15</td>
<td>18</td>
<td>43</td>
<td>5.40</td>
<td>.796</td>
</tr>
<tr>
<td>Rogers Hornsby '25</td>
<td>504</td>
<td>203</td>
<td>41</td>
<td>10</td>
<td>39</td>
<td>41</td>
<td>5.01</td>
<td>.756</td>
</tr>
</tbody>
</table>

*OBP = On Base Percentage  
**SLGAA = Slugging Average
Figure 8-3. Lifetime and 1971 Season OERA's versus Batting Average with Pure Performance Bounds Superimposed.
bats. The performances are superimposed on the pure performance bounds derived in Section 7. The graph is interesting because there is so little correlation among OERA, batting average, and the bounds. The base-on-balls bound is plotted under the convention that walks are counted as hits for the purpose of calculating batting average. In fact, it is the neglect of the base on balls that generates this lack of correlation. Figure 8-4 presents the same earned run averages as Figure 8-3; but they are plotted versus on-base percentage, which does count a walk as a hit. In this case the strong correlation among on-base percentage, OERA, and the pure bounds is obvious. However, one should be aware that the fact that the singles and doubles bounds serve as lower and upper bounds, respectively, to the observed OERA's is not too significant because they are loose. At an on-base percentage of 0.350, they are about four runs per game apart. In the limit of very high on-base percentage, the separation grows to nine runs per game.

Table 8-III gives batting average, on-base percentage, and OERA for the 1971 players used in Figures 8-3 and 8-4. Willie Mays is included in the top ten group because of his unusually high on-base percentage. It is interesting to note that Aaron, Torre, and Mays all had seasons comparing very favorably with the best in baseball history. It is also interesting to note the huge difference between the offensive performance of average players and great players.

9. Discussion

This paper began as an attempt to evaluate the greatest hitters in baseball. The resulting concept of the offensive earned run average is oversimplified because it does not account for situations such as double plays, short singles versus long singles, advancing on outs, etc. These
Figure 8-4. Lifetime and 1971 Season OERA's versus On-Base Percentage with Pure Performance Bounds Superimposed.
### TABLE 8-III

#### A. Random Statistics - 1971: 300 At Bats or More

<table>
<thead>
<tr>
<th>Player</th>
<th>Batting Average</th>
<th>On-Base Percentage</th>
<th>OERA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob Aspromonte</td>
<td>.225</td>
<td>.286</td>
<td>2.73</td>
</tr>
<tr>
<td>Lou Brock</td>
<td>.313</td>
<td>.315</td>
<td>4.52</td>
</tr>
<tr>
<td>Dave Cash</td>
<td>.289</td>
<td>.351</td>
<td>4.45</td>
</tr>
<tr>
<td>Boots Day</td>
<td>.283</td>
<td>.342</td>
<td>4.15</td>
</tr>
<tr>
<td>Alan Gallagher</td>
<td>.277</td>
<td>.339</td>
<td>4.41</td>
</tr>
<tr>
<td>Bud Harrelson</td>
<td>.252</td>
<td>.318</td>
<td>3.15</td>
</tr>
<tr>
<td>Ron Hunt</td>
<td>.279</td>
<td>.351</td>
<td>4.56</td>
</tr>
<tr>
<td>Ed Kranepool</td>
<td>.280</td>
<td>.340</td>
<td>5.28</td>
</tr>
<tr>
<td>Tim McCarver</td>
<td>.278</td>
<td>.338</td>
<td>4.60</td>
</tr>
<tr>
<td>Al Oliver</td>
<td>.282</td>
<td>.317</td>
<td>4.83</td>
</tr>
<tr>
<td>Doug Rader</td>
<td>.244</td>
<td>.302</td>
<td>3.87</td>
</tr>
<tr>
<td>Ron Santo</td>
<td>.267</td>
<td>.358</td>
<td>5.61</td>
</tr>
<tr>
<td>Gary Sutherland</td>
<td>.257</td>
<td>.298</td>
<td>3.18</td>
</tr>
<tr>
<td>Bob Watson</td>
<td>.288</td>
<td>.346</td>
<td>4.79</td>
</tr>
</tbody>
</table>

#### B. Top Ten Batting Average - 1971 (Plus Mays)

<table>
<thead>
<tr>
<th>Player</th>
<th>Batting Average</th>
<th>On-Base Percentage</th>
<th>OERA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe Torre</td>
<td>.363</td>
<td>.420</td>
<td>9.12</td>
</tr>
<tr>
<td>Ralph Garr</td>
<td>.343</td>
<td>.372</td>
<td>6.02</td>
</tr>
<tr>
<td>Glen Beckert</td>
<td>.342</td>
<td>.370</td>
<td>5.49</td>
</tr>
<tr>
<td>Roberto Clemente</td>
<td>.341</td>
<td>.372</td>
<td>6.79</td>
</tr>
<tr>
<td>Hank Aaron</td>
<td>.327</td>
<td>.412</td>
<td>9.97</td>
</tr>
<tr>
<td>Manny Sanguillen</td>
<td>.319</td>
<td>.342</td>
<td>5.10</td>
</tr>
<tr>
<td>Cleon Jones</td>
<td>.319</td>
<td>.384</td>
<td>6.79</td>
</tr>
<tr>
<td>Matty Alou</td>
<td>.315</td>
<td>.351</td>
<td>5.13</td>
</tr>
<tr>
<td>Lou Brock</td>
<td>.313</td>
<td>.315</td>
<td>4.52</td>
</tr>
<tr>
<td>Rusty Staub</td>
<td>.311</td>
<td>.386</td>
<td>7.02</td>
</tr>
<tr>
<td>Willie Mays</td>
<td>.271</td>
<td>.425</td>
<td>8.42</td>
</tr>
</tbody>
</table>
compromises were made so that the statistic would be easily calculated and deterministic. Also, the evaluation considers the offensive capability in the context of a player's own statistics instead of evaluating his incremental aid to his team or to a team composed of league average players. This decision to use only personal statistics is partially motivated by the need for deterministic procedures and partially by the desire to measure intrinsic offensive capability without reference to the other team members. It should be added that the mathematical framework in this paper extends easily to more complicated situations, but the simple play-by-play calculation does not.
FIGURES

Fig. 7.1. Pure Performance Bounds

Fig. 8-1. Lifetime OERA versus Slugging Average

Fig. 8-2. Single Season OERA versus Slugging Average

Fig. 8-3. Lifetime and 1971 Season OERA's versus Batting Average with Pure Performance Bounds Superimposed

Fig. 8-4. Lifetime and 1971 Season OERA's versus On-Base Percentage with Pure Performance Bounds Superimposed
References


