> **If there is one great quest in baseball statistics, it is the search for the best formula for evaluating offensive performance.**
> **—Jim Albert and Jay Bennett, *Curve Ball***

# Who is the Greatest Hitter of Them All?

JOSEPH A. GALLIAN
**University of Minnesota Duluth**

If you want to get into an argument, ask any baseball fan who the greatest hitter is. Along with emotion, bias, and hometown prejudice, you're likely to hear some statistics. Baseball, after all, is the sport that best lends itself to statistical analysis. Using batting average alone to determine the best hitter is too crude since a bases-loaded home run counts the same as a bases-empty single. Other complicating factors in trying to decide who is the greatest hitter include the difficulty of comparing players from different eras, rule changes, changing strategies and even which ballpark a hitter has for his home games. Although there have been many different approaches to answer this question, the results are pretty much in agreement on the top two hitters. In this article I will explain a few of the more interesting methods used to try to answer this question and reveal the top two hitters.

For over 100 years batting averages (number of hits divided by number of at bats) have been used to declare the Batting Champion in each league. But even if we agree to use batting averages as a reasonable way to determine the best hitter it isn't the case that the best hitters are those with the highest batting averages. For one thing, unlike football and basketball, in baseball, the playing fields vary widely. Certain ballparks are "hitters parks" while others are "pitchers parks." (The essential differences are due to the size of the foul territory, not the distances to the outfield fences.) For another, left-handed hitters have an advantage over right-handed hitters. (There are two reasons for this. One is that for both left-handed and right-handed hitters it is easier to hit against a pitcher of the opposite hand, but there are more right-handed pitchers than left-handed ones. The second reason is that left-handed hitters face first base after swinging.) Moreover, rule changes (for example, the size of the strike zone or height of the mound), equipment changes (composition of the ball, batting gloves, batting helmets, etc.), more frequent use of relief pitchers and many other factors result in league averages that vary widely by era. To illustrate, we point out that in 1930 the entire National League average, including the pitchers, was .303 whereas in 1968 Carl Yastrzemski lead the American League with a .301 average. Clearly, it is not the case that the average hitter in 1930 was a better hitter than Yaz in 1968. In an attempt to level the playing field, statisticians have devised ways to adjust for factors such as home ballpark, handedness and different eras. Table 1 shows the top 10 all-time batters from 1900 on using raw averages alone. Note that there is not a single person who has played in the past 40 years! This tips you off that something is wrong. In his book *Baseball's All-Time Best Hitters*, statistician Michael Schell devised a scheme that adjusts for differences in era, ballpark, handedness, and even late-career decline (he ignored all stats after 8000 at bats). Table 2 shows the adjusted list of top 10 hitters according to Schell (through 1997). This time we do have some modern ballplayers but the early players are still disproportionally represented.

| | | | |
|---|---|---|---|
| 1. | Ty Cobb | .366 | 1905–1928 |
| 2. | Rogers Hornsby | .358 | 1915–1937 |
| 3. | Joe Jackson | .356 | 1908–1920 |
| 4. | Ted Williams | .344 | 1939–1960 |
| 5. | Tris Speaker | .344 | 1907–1928 |
| 6. | Babe Ruth | .342 | 1914–1935 |
| 7. | Harry Heilmann | .342 | 1914–1932 |
| 8. | Bill Terry | .341 | 1923–1936 |
| 9. | George Sisler | .340 | 1915–1930 |
| 10. | Lou Gehrig | .340 | 1923–1939 |

**Table 1.** Lifetime batting average—Top Ten.

| | | | |
|---|---|---|---|
| 1. | Tony Gwynn | .342 | 1982–2001 |
| 2. | Ty Cobb | .340 | 1905–1928 |
| 3. | Rod Carew | .332 | 1967–1985 |
| 4. | Joe Jackson | .331 | 1908–1920 |
| 5. | Rogers Hornsby | .330 | 1915–1937 |
| 6. | Ted Williams | .327 | 1939–1960 |
| 7. | Stan Musial | .325 | 1941–1963 |
| 8. | Wade Boggs | .324 | 1982–1999 |
| 9. | Tris Speaker | .322 | 1907–1928 |
| 10. | Willie Mays | .314 | 1951–1973 |

**Table 2.** Lifetime adjusted batting average—Top Ten.

## Slugging it Out

Of course, Schell's adjusted batting average is far superior to the traditional batting average. Nevertheless, it does not take into account power. It is beyond dispute that a home run is better than a single. The traditional way to account for power is the so-called Slugging Average (*SLG*). The slugging average counts a home run the same as four singles, a triple the same as three singles and a double the same as two singles. In short,

$$SLG = \frac{(1\times 1B)+(2\times 2B)+(3\times 3B)+(4\times HR)}{AB}$$

where $1B$ is the number of singles, $2B$ is the number of doubles, $3B$ the number of triples, $HR$ the number of home runs, and $AB$ the number of at bats.

The numerator of this fraction is denoted by $TB$ because it counts the total bases achieved by the hits. As you would expect, this measure of batting performance correlates to team run production much better than batting average does. Table 3 shows the 15 players with the highest career slugging average through the 2001 season (assumming at least 5000 at bats). In contrast to the batting average, notice that about half of the players listed in Table 3 are still active. We believe that the reason for this is that there is now much more emphasis placed on hitting home runs than there was decades ago. This change in emphasis also accounts, at least in part, for the lower batting averages for modern players. (Other reasons for more home runs in recent years include better protection from being hit by a pitch, smaller ballparks, widespread adoption of weight lifting, and the use of nutritional supplements.)

In 2001 Barry Bonds had one of the greatest seasons in the history of baseball. Not only did he break Mark McGwire's 1998 home run record of 70, but Bonds also broke Babe Ruth's slugging record of .847 set in 1920 by slugging .863 and Ruth's 1923 record 170 walks by 7. Bonds will probably hold the slugging record much longer than he will hold the record for home runs. (Other than Bonds and Babe Ruth, who owns four of the top six single-season slugging averages, Lou Gehrig is the only player to break .760 in a single season with his .765 in 1927.)

| | | | |
|---|---|---|---|
| 1. | Babe Ruth | .690 | 1914–1935 |
| 2. | Ted Williams | .634 | 1939–1960 |
| 3. | Lou Gehrig | .632 | 1923–1939 |
| 4. | Jimmie Foxx | .609 | 1925–1945 |
| 5. | Hank Greenberg | .605 | 1930–1947 |
| 6. | Mark McGwire | .588 | 1986–2001 |
| 7. | Barry Bonds | .585 | 1986– |
| 8. | Mike Piazza | .5789 | 1992– |
| 9. | Joe DiMaggio | .5788 | 1936–1951 |
| 10. | Frank Thomas | .5770 | 1990– |
| 11. | Rogers Hornsby | .5765 | 1915–1937 |
| 12. | Larry Walker | .572 | 1989– |
| 13. | Juan Gonzalez | .568 | 1989– |
| 14. | Ken Griffey, Jr. | .566 | 1989– |
| 15. | Albert Belle | .564 | 1989– |

**Table 3.** Lifetime slugging average—Top Fifteen through 2001.

The astute reader will have noticed that there is an important part of offensive production that the slugging average does not take into account—walks. There are a number of schemes that account for extra base hits and walks. One of these called Total Average (*TA*) was devised by the sportswriter Thomas Boswell in 1981 in *Inside Sports*. Total average gives a hitter credit for walks (*BB*), getting on base by being hit by the pitcher (*HBP*) and stolen bases (*SB*) while penalizing him for being caught stealing (*CS*) and grounding into double plays (*GIDP*). Notice that this method rewards ability to get on base, power and speed. Here is the formula for total average

$$TA = \frac{TB + BB + HBP + SB}{AB - H + CS + GIDP}.$$

This formula expresses a ratio of good events to bad events. It favors players like Barry Bonds who hit for average and have power and speed and disfavors players like Mark McGwire who simply hit for power. Total average is the best simple measure of offensive performance. Studies have shown that a team's total average is an excellent predictor of the runs per game scored by the team.

One shortcoming of the total average is that it counts a walk the same as a single but of course a single is more effective at advancing runners on base than a walk. Various refinements of the *TA* formula have been proposed. In one such formula a walk is counted only one-third as much as a single.

Since comparing players from different eras is fraught with complexity, some measures have been devised that compare players to their contemporaries. One of these is called the Offensive Quotient (*OQ*) described by M. VanOverloop in *The Baseball Research Journal* in 1993. To calculate this for a particular player one uses the formula

$$\frac{TB + BB}{Outs}$$

| | | | |
|---|---|---|---|
| 1. | Babe Ruth | 218 | 1914–1935 |
| 2. | Ted Williams | 210 | 1939–1960 |
| 3. | Lou Gehrig | 177 | 1923–1939 |
| 4. | Rogers Hornsby | 177 | 1915–1937 |
| 5. | Frank Thomas | 177 | 1990– |
| 6. | Mickey Mantle | 176 | 1951–1968 |
| 7. | Jimmie Foxx | 173 | 1925–1945 |
| 8. | Ty Cobb | 170 | 1905–1928 |
| 9. | Willie McCovey | 169 | 1959–1980 |
| 10. | Johnny Mize | 168 | 1936–1953 |

**Table 4.** Lifetime Offensive Quotient—Top Ten through 1995.

for the individual and for the entire league. Multiply the ratio of these two by 100 to get the *OQ* for the player. Table 4 shows the all-time leaders in this category from 1900–1995. Since the league average is always 100, Babe Ruth's 218 *OQ* means that Ruth's offensive performance was 118 percent greater than that of his contemporaries.

## Send in the Clones

The measure that I find most interesting is called the Offensive Earned-Run Average (*OERA*) devised by Thomas Cover and Carrol Keilers in a 1977 article in *Operations Research*. The intent of their method is to determine how many runs a team composed of nine identical players would score per game. Say we wanted to calculate the *OERA* for Barry Bonds for his multi-record breaking year 2001. We imagine a line-up for which *every* batter has the same probability of hitting a single as Barry Bonds, the same probability of hitting a double as Barry Bonds, the same probability of hitting a triple as Barry Bonds, the same probability of hitting a home run as Barry Bonds, and the same probability of getting a walk as Barry Bond. In short, the offensive line-up consists of nine Barry Bonds clones!

To illustrate the idea say the first batter gets a walk while the second batter gets a single. This puts runners on first and third (it is assumed that a single advances a runner two bases). If the next two batters make outs the runners are assumed to stay in place. Then a double scores two runs and puts a man on second. Finally, the next batter ends the inning by making an out. So, for that inning, a line-up of 9 Barry Bonds scored two runs. At this rate for an entire season Bonds's *OERA* would be 18.00.

Cover and Keilers achieved this simulation by constructing a $24 \times 24$ matrix $Q$ that accounts for all 24 possible combinations of outs (0, 1, or 2) and men on base ( $\binom{3}{0}+\binom{3}{1}+\binom{3}{2}+\binom{3}{3}$ ). The entries of $Q$ are probabilities, $Q_{ij}$ is the probability of passing from start $i$ to state $j$. For example, if state 7 represents a runner on second base and one out and state 13 represents no men on base and one out, the only way to pass from state 7 to state 13 is for the batter to hit a home run. Thus $Q_{7,13}$ is the probability that the batter being simulated hits a home run. For each state there is a probability that the hitter will produce one or more runs by hitting a single, double, triple or home run or walking. Collect these probabilities into the entries of a $24 \times 1$ vector, $R$, so that $R_j$ represents the expected number of runs scored from state $j$ in one plate appearance. For example, if state 7 is as above, then $R_7 = 1 \times$ (probability of a single) + $1 \times$ (probability of a double) + $1 \times$ (probability of a triple) + $2 \times$ (probability of a home run). After each plate appearance there will be a new state determined by what the batter did in the previous appearance. By iterating this procedure and keeping track of the number of expected runs scored $E$ (a $24 \times 1$ vector) before three outs occur, Cover and Keilers were able to simulate a line-up of nine identical players by using the formula

$$E = \sum_{i=0}^{\infty} Q^i R = (I - Q)^{-1} R.$$

The *OERA* is 9 times the expected number of runs scored in one inning (the first entry of $E$) beginning each time with the state that there are no outs and no men on base.

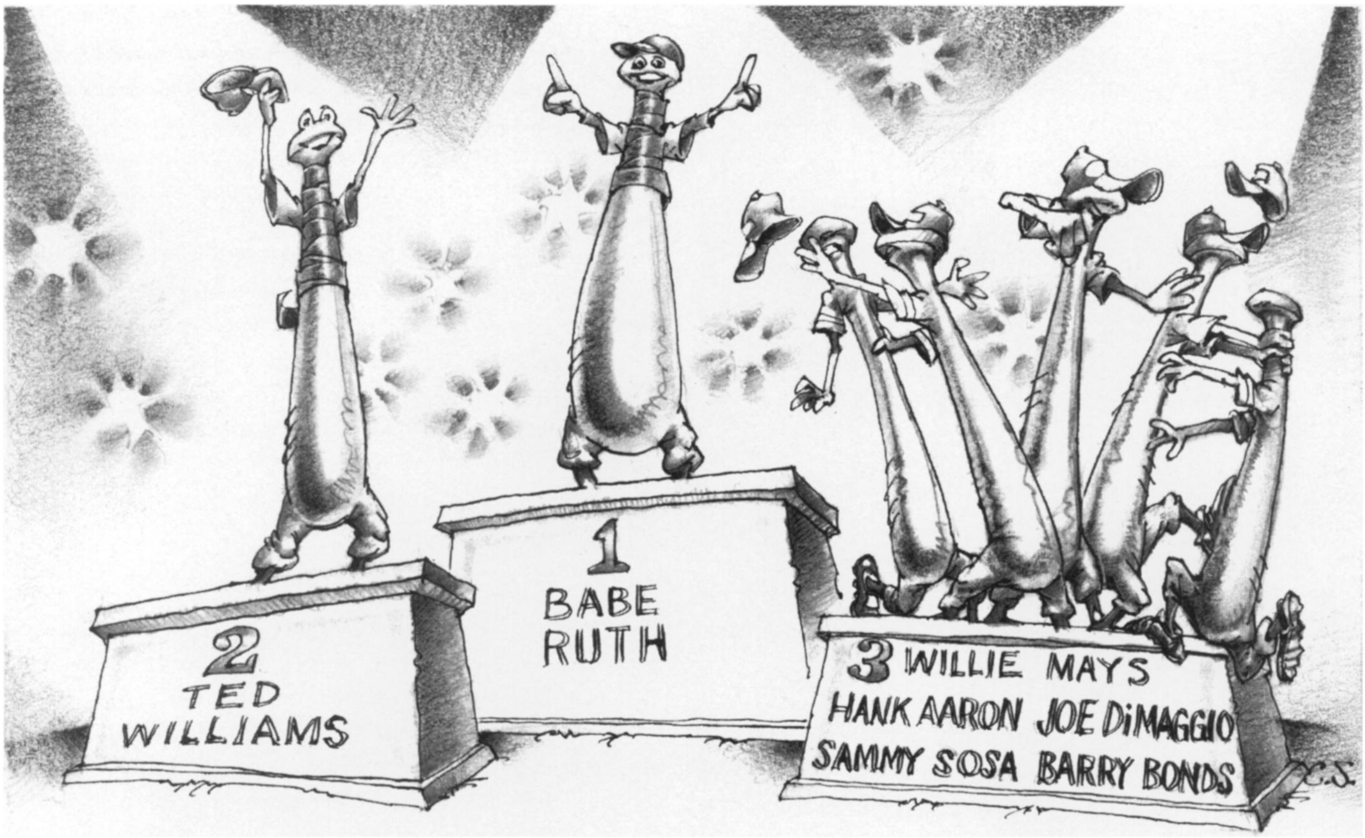
One attractive feature of the *OERA* is that unlike traditional measures of offensive performance such as RBIs and runs scored, it does not depend on the quality of one's teammates.

Harvard statistician Carl Morris and University of Minnesota Duluth graduate student Kai Xu have refined the *OERA* method of Cover and Keilers to more closely model what happens when a batter hits a single or double with a man on base (they do not assume the runner on base always advances two bases on a single or that a man on first always scores when the batter hits a double ).

Tables 5 and 6 show Xu's calculations for those batters with the best single season *OERA* (excluding seasons from the 19th century) and the best career *OERA* up through the year 2001. The astounding thing about Table 5 is the 16 year gap in the two entries for Ted Williams. Williams was 39 years old in 1957. To put these numbers in perspective we mention that typically a team scores about 4.5 runs per game.

| | | | |
|---|---|---|---|
| 1. | Ted Williams | 1941 | 19.54 |
| 2. | Babe Ruth | 1923 | 18.79 |
| 3. | Babe Ruth | 1920 | 18.47 |
| 4. | Ted Williams | 1957 | 17.07 |
| 5. | Barry Bonds | 2001 | 16.98 |
| 6. | Babe Ruth | 1921 | 16.93 |
| 7. | Babe Ruth | 1926 | 16.25 |
| 8. | Babe Ruth | 1924 | 16.00 |
| 9. | Rogers Hornsby | 1924 | 15.73 |
| 10. | Rogers Hornsby | 1925 | 14.90 |

**Table 5.** Season Offensive Earned-Run Average—Top Ten through 2001.




| | | | |
|---|---|---|---|
| 1. | Babe Ruth | 12.91 | 1914–1935 |
| 2. | Ted Williams | 12.83 | 1939–1960 |
| 3. | Lou Gehrig | 10.88 | 1923–1939 |
| 4. | Frank Thomas | 10.09 | 1990– |
| 5. | Jimmie Foxx | 9.86 | 1925–1945 |
| 6. | Rogers Hornsby | 9.76 | 1915–1937 |
| 7. | Barry Bonds | 9.14 | 1986– |
| 8. | Mickey Mantle | 9.06 | 1951–1968 |
| 9. | Hank Greenberg | 9.05 | 1930–1947 |
| 10. | Ty Cobb | 9.00 | 1905–1928 |

**Table 6.** Lifetime Offensive Earned-Run Average—2001.

| | | | |
|---|---|---|---|
| 1. | Barry Bonds | 2001 | 18.86 |
| 2. | Ted Williams | 1941 | 16.59 |
| 3. | Babe Ruth | 1920 | 16.44 |
| 4. | Ted Williams | 1957 | 15.99 |
| 5. | Babe Ruth | 1921 | 15.63 |
| 6. | Babe Ruth | 1923 | 14.86 |
| 7. | Jimmie Foxx | 1932 | 13.88 |
| 8. | Babe Ruth | 1931 | 13.29 |
| 9. | Babe Ruth | 1924 | 13.13 |
| 10. | Babe Ruth | 1926 | 12.97 |

**Table 7.** Season Runs Created per Game—Top Ten through 2001.

The baseball research pioneer Bill James has developed a complex formula called Runs Created per Game (*RC*/*G*) that also attempts to estimate the number of runs scored per game by an entire line-up of the same hitter. Table 7 shows the top 10 single season *RC*/*G* through 2001 using a refined version of James's formula that penalizes a batter for striking out. Barry Bonds's record breaking *RC*/*G* of 18.66 for 2001 is even more impressive than his record setting marks for home runs, walks and slugging average in 2001.

## And the Winners Are...

So, upon looking at the dozens of schemes that have been devised to answer the question "Who is the greatest hitter of them all?" one sees the same two names inevitably come up near the top: Babe Ruth and Ted Williams. Ruth's still-standing 40 batting records dwarfs Ty Cobb's second most 21. There seems to be no clear choice for the third place. Strong cases can be made for Aaron, Bonds, Hornsby, and Gehrig.

Of course, one might ask the question Who is the greatest baseball player of them all? The Official Encyclopedia of Major League Baseball *Total Baseball* uses a rating scheme called "Total Player Rating" that takes into account all aspects of a player's contribution to his team: hitting (adjusted for league average and ballpark), base stealing, fielding, and position (shortstops, second basemen, catchers and third basemen

*Continued on p. 28.*

Using (8) we find that the area of one arch of the diamogon is $A = P + C$ because $C_z = C$. The two arches between the diamogon and the outer polygon have area $2A = 2(P + C)$. In the limiting case when $n \to \infty$ this becomes $2A = 4C$. But $4C$ is the area of the fixed circular disk, which means that the area of the region common to the two diamogons tends to zero. In other words, when $n \to \infty$ the diamogon turns into a diameter of the fixed circle traced twice.

## Tracing Point not at a Vertex

We conclude with an example of a hypotrochogon traced out by a point $z$ not at a vertex of the $n$-gon. We consider $^n/_m = ^1/_2$ and call the hypotrochogon an ellipsogon because the limiting case $n \to \infty$ gives an ellipse. Figure 8 shows an example of a square rolling inside an octagon with the tracing point $z$ inside the square. In this case the ellipsogon traces out two arches, each consisting of four circular arcs.

In the limiting case $n \to \infty$, (9) shows that the area of one arch is given by $A = C + ^1/_2(C_z + C)$, so the two arches fill out a region of area $2A = 3C + C_z$. The limiting configuration of the ellipsogon is an ellipse enclosing an area equal to $4C - 2A = C - C_z$. If the radius of the inner circle is $r$ and if the distance from $z$ to the center of the inner circle is $s$ then $C - C_z = \pi(r^2 - s^2) = \pi(r + s)(r - s)$. The distances $r + s$ and $r - s$ are the lengths of the semiaxes $a = r + s$ and $b = r - s$ of the ellipse, so we get $C - C_z = \pi ab$ the usual formula for the area of an ellipse.

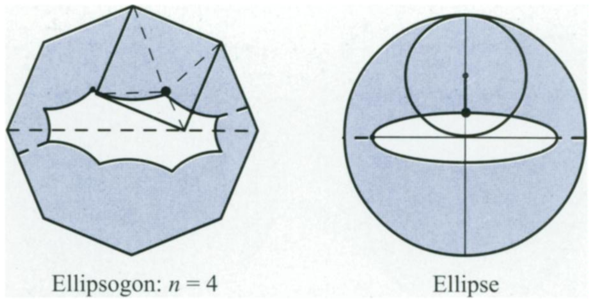



**Figure 8.** An ellipsogon traced by a point inside an $n$-gon rolling inside a $2n$-gon. The ellipsogon becomes an ellipse as $n \to \infty$.

The point $z$ also traces an ellipsogon if it is outside the rolling $n$-gon. If the point $z$ is inside or outside the rolling $n$-gon and then moves toward a vertex, the ellipsogon becomes a diamogon which, in turn, becomes a diameter as $n \to \infty$. ■

## References

1. Tom M. Apostol and Mamikon A. Mnatsakanian, Cycloidal Areas Without Calculus, *Math Horizons*, Sept. 1999, p. 14.
2. Tom M. Apostol and Mamikon A. Mnatsakanian, Sums of Squares of Distances, *Math Horizons*, Nov. 2001, pp. 21–22.

*Continued from p. 16.*

| | | |
|---|---|---|
| 1. | Babe Ruth | 1914–1935 |
| 2. | Barry Bonds | 1986– |
| 3. | Nap Lajoie | 1896–1916 |
| 4. | Rogers Hornsby | 1915–1937 |
| 5. | Ted Williams | 1939–1960 |
| 6. | Mike Schmidt | 1972–1989 |
| 7. | Mickey Mantle | 1951–1968 |
| 8. | Mike Piazza | 1992– |
| 9. | Willie Mays | 1951–1973 |
| 10. | Lou Gehrig | 1923–1939 |

**Table 8.** Total Player Rating—Top Ten through 2000.

count more than outfielders and first basemen; center fielders count more than left and right fielders). If one accepts *Total Baseball*'s rating as reasonable, then the list of the top ten greatest players (excluding pitchers) is given in Table 8. Notice that the Total Player Ratings of Schmidt, Hornsby and Piazza are helped by the positions they play.

During 1999 many polls were taken asking who were the greatest athletes of the 20th century. In the poll of experts taken by ESPN, Ruth was rated the greatest baseball player, Mays was second, Aaron was third, Jackie Robinson 4th and Williams 5th. (Many of the votes Robinson received were because he integrated Major League Baseball.) An Associated Press poll ranked them Ruth, Mays, Williams. Yet another poll had them Ruth, Mays, Aaron, Williams. So there is total agreement with Ruth at # 1 and Mays # 2. Williams and Aaron are three and four or vice versa.

Even though Ruth and Williams were in their primes in the first half of the 20th century, they remain among the most popular players ever. The web site

`www.baseball-reference.com`

(click on "leaders" then on "Most Popular Players" at the bottom of the page) keeps track of the number of hits each player's web page receives. The five most popular players are: Bonds, Ruth, Aaron, Williams, and Mantle. ■

## For Further Reading

Besides Schell's *Baseball's All-Time Best Hitters*, Albert and Bennett's *Curve Ball* is highly readable and describes in great detail how statistical methods are used to measure offensive performance. Baseball records and statistics for players can be found at `www.baseball1.com` and `www.baseball-reference.com`. Access Kai Xu's program for computing *OERA* at `www.d.umn.edu/~jgallian/OERA.html`.