Rating Player Performance -
The Old Argument of Who is Best

Ron Gould
Emory University

October 28-29, 2011
Who is the best quarterback in the NFL?
Who is the best quarterback in the NFL?

- Tom Brady,
Who is the best quarterback in the NFL?

- Tom Brady,
- Aaron Rodgers,
Who is the best quarterback in the NFL?

- Tom Brady,
- Aaron Rodgers,
- Drew Breeze,
Who is the best quarterback in the NFL?

- Tom Brady,
- Aaron Rodgers,
- Drew Breeze,
- Philip Rivers,
Who is the best quarterback in the NFL?

- Tom Brady,
- Aaron Rodgers,
- Drew Breeze,
- Philip Rivers,
- Peyton Manning
Who is the best quarterback in the NFL?

- Tom Brady,
- Aaron Rodgers,
- Drew Breeze,
- Philip Rivers,
- Peyton Manning
- Michael Vick,
How do you attack such a question

Most attempts at player rating (or evaluation) have been done via **Linear Weight Systems**.

Here, you identify various aspects of the game, say $x_1, x_2, \ldots, x_k$ and what you wish to see produced, say $Y$, and you seek constants $c_0, c_1, \ldots, c_k$ so that

$$Y = c_0 + c_1 x_1 + c_2 x_2 + \ldots + c_k x_k.$$
First, we are interested in how quarterbacks are presently "rated".

In particular, we are interested in questions about
First, we are interested in how quarterbacks are presently "rated".

In particular, we are interested in questions about

- How the rating is done.
First, we are interested in how quarterbacks are presently "rated".

In particular, we are interested in questions about

- How the rating is done.
- How it might be modified.
First, we are interested in how quarterbacks are presently "rated".

In particular, we are interested in questions about

- How the rating is done.
- How it might be modified.
- What other things might be important in building any such formula.
Mid 1930’s, NFL begins keeping player stats; Passing Leader is:

- the one with the **most yards**.
History:

Mid 1930’s, NFL begins keeping player stats; Passing Leader is:

- the one with the **most yards**.

- 1938-1940, **highest completion percentage**.
Mid 1930’s, NFL begins keeping player stats; Passing Leader is:

- the one with the **most yards**.
- 1938-1940, **highest completion percentage**.
- 1941, **Special point system**: 1 to 10 points based on your place among the top 10 in each of 6 categories including: TD passes, yards, interception percentage, etc.
Mid 1930’s, NFL begins keeping player stats; Passing Leader is:

- the one with the most yards.
- 1938-1940, highest completion percentage.
- 1941, Special point system: 1 to 10 points based on your place among the top 10 in each of 6 categories including: TD passes, yards, interception percentage, etc.
- For the next 30 years they kept changing from one system to another.
Early 1970s, a special study committee headed by:

- Don Smith of the Pro Football Hall of Fame;
- Seymour Siwoff of Elias Sports Bureau (the official NFL statisticians);
- Don Weiss of the NFL.
Early 1970s, a special study committee headed by:

> Don Smith of the Pro Football Hall of Fame;
Early 1970s, a special study committee headed by:

- **Don Smith** of the Pro Football Hall of Fame;
- **Seymour Siwoff** of Elias Sports Bureau (the official NFL statisticians);
The big change

Early 1970s, a special study committee headed by:

▶ **Don Smith** of the Pro Football Hall of Fame;

▶ **Seymour Siwoff** of Elias Sports Bureau (the official NFL statisticians);

▶ and **Don Weiss** of the NFL

created the passer rating that is now in use in the NFL.
The Formula has 4 parts

\[ A = \left( \frac{COMP}{ATT} - .3 \right) \times 5 \]

\[ B = \left( \frac{YARDS}{ATT} - 3 \right) \times .25 \]

\[ C = \left( \frac{TD}{ATT} \right) \times 20 \]

\[ D = 2.375 - \left( \frac{INT}{ATT} \times 25 \right) \]
The Formula

Let \( mm(x) = \max(0, \min(x, 2.375)) \)

\[
\text{PasserRating} = \left( \frac{mm(A) + mm(B) + mm(C) + mm(D)}{6} \right) \times 100
\]

Linear weighting system: The four parts receive equal weight.

Note: The PasserRating can run from 0 to 158.3.
Note: \( 4 \times 2.375 = 9.5 \)
Rational for the formula

Their Goal
A system that did not depend on how other quarterbacks did.
Devise a formula to give one point for average performance, two\(^+\) points for a record performance, and 0 for really bad performance.
They used the four stats kept at that time (1971-2).

They also used league averages (at that time) for:

Completion percentage (around 50%)

Yards-per-attempt (7),

Percentage of passes resulting in TDs (5%),

Percentage intercepted (5.5%),
Example:

Note: Average yards per attempt = 7 (at that time)

\[ B = \left( \frac{YARDS}{ATT} - 3 \right) \times .25 \]

Contributes 1 at that average as they wanted!
Some Strengths and Weaknesses

- **Strength:** Does not depend on others performances.

- **Weakness:** Scale is ridiculous!

- **Strength:** Applies to a game, season or career!

- **Weakness:** Game has changed greatly from 1970's

- **Strength:** Not dependent upon judgements.

- **Weakness:** This was never meant to be a quarterback rating formula, but rather a passer efficiency formula. Running stats, winning stats, etc. play no role.
Some Strengths and Weaknesses

- **Strength:** Does not depend on others performances.
- **Weakness:** Scale is ridiculous!
Some Strengths and Weaknesses

- **Strength:** Does not depend on others performances.
- **Weakness:** Scale is ridiculous!
- **Strength:** Applies to a game, season or career!
Some Strengths and Weaknesses

- **Strength:** Does not depend on others performances.
- **Weakness:** Scale is ridiculous!
- **Strength:** Applies to a game, season or career!
- **Weakness:** Game has changed greatly from 1970’s
Some Strengths and Weaknesses

- **Strength:** Does not depend on others performances.
- **Weakness:** Scale is ridiculous!
- **Strength:** Applies to a game, season or career!
- **Weakness:** Game has changed greatly from 1970’s
- **Strength:** Not dependent upon judgements.
Some Strengths and Weaknesses

- **Strength**: Does not depend on others performances.
- **Weakness**: Scale is ridiculous!
- **Strength**: Applies to a game, season or career!
- **Weakness**: Game has changed greatly from 1970’s
- **Strength**: Not dependent upon judgements.
- **Weakness**: This was never meant to be a quarterback rating formula, but rather a **passer efficiency** formula. Running stats, winning stats, etc. play no role.
Another Weakness - Or Is It?

2 QB’s - Super Joe and Broadway Joe:

Super Joe takes his team 30 yards for a TD on 3 consecutive 10 yard passes, the final one a TD pass. Rating = 147.9

Broadway Joe throws 2 incomplete passes, then a 30 yard TD pass. Same end result but his Rating = 111.1

Some would say a weakness of the formula - some would say a strength!
The REALLY BIG Problems:

- There is no set of variables that are agreed upon for measuring a QB’s performance.
- There are no generally held beliefs as to how to weigh the various aspects of the game!
- There is no generally held view as to what is even meant by “best”!
- How do you measure the unmeasurable? To build a formula for the “best” quarterback you probably need to consider things like: 3rd down performance, red-zone performance, ability to avoid sacks, play-calling (audibles), overall leadership, and maybe even more!
The REALLY BIG Problems:

- There is no set of variables that are agreed upon for measuring a QB’s performance.
- There are no generally held beliefs as to how to weigh the various aspects of the game!
The REALLY BIG Problems:

▶ There is no set of variables that are agreed upon for measuring a QB’s performance.
▶ There are no generally held beliefs as to how to weigh the various aspects of the game!
▶ There is no generally held view as to what is even meant by ”best”!

How do you measure the unmeasurable? To build a formula for the "best" quarterback you probably need to consider things like: 3rd down performance, red-zone performance, ability to avoid sacks, play-calling (audibles), overall leadership, and maybe even more!
The REALLY BIG Problems:

- There is no set of variables that are agreed upon for measuring a QB’s performance.
- There are no generally held beliefs as to how to weigh the various aspects of the game!
- There is no generally held view as to what is even meant by ”best”!
- How do you measure the unmeasurable? To build a formula for the ”best” quarterback you probably need to consider things like:
  3rd down performance, red-zone performance, ability to avoid sacks, play-calling (audibles), overall leadership, and maybe even more!
Conclusions:

Separating one person's performance from what is clearly a team outcome is difficult.

It is unlikely we can build a formula to rate quarterbacks that everyone would find satisfactory!

The game has changed enough that the old averages used no longer apply. New long term averages could be created, and then reapplied to past performances.

A simpler and more intuitive formula can certainly be devised.
Who is the best hitter in baseball?

We start to run into similar problems as with the QB formula!
What do we mean by best?
Is best:
Is best:

- The top batting average?
Is best:

- The top batting average?
- The top homerun hitter?
Is best:

- The top batting average?
- The top homerun hitter?
- The top RBI producer?
Is best:

- The top batting average?
- The top homerun hitter?
- The top RBI producer?
- Some combination of these? If so, what combination?
Clearly each of the 3 traditional measures clearly produces a "best" player in that category.

Each part is simple and direct to measure.

But overall we again face a number of problems.
Over the years since the end of WWII, there has been a growing movement to use other means to measure player performance.

**Sabermetrics** is the mathematical and statistical study of baseball records.
Many argue that the true offensive value of a player is in the number of runs they create. So any measure of offensive performance should measure run production.

Bill James, (1982) devised the runs created (RC) measure.

\[ RC = (\text{Hits} + \text{BB}) \times \frac{\text{TOTALBASES}}{\text{AB} + \text{BB}}. \]

Rational: Hits + BB counts times reached base and the fraction measures rate at which the runners are advanced.
Example

Johnny Pesky (1942) had 620 ABs, 205 hits, 42 BBs, 258 total bases

Dick Stuart (1960) had 532 ABs, 160 hits, 34 BBs, 309 TBs
Example

Johnny Pesky (1942) had 620 ABs, 205 hits, 42 BBs, 258 total bases

Dick Stuart (1960) had 532 ABs, 160 hits, 34 BBs, 309 TBs

▶ RC(Pesky) = 96.
Johnny Pesky (1942) had 620 ABs, 205 hits, 42 BBs, 258 total bases

Dick Stuart (1960) had 532 ABs, 160 hits, 34 BBs, 309 TBs

- RC(Pesky) = 96.
- RC(Stuart) = 106.
linear weights

First tried by George Linsey (1963) (using recorded data and probability theory)

\[ \text{Runs} = (0.41) \ 1B + (0.82) \ 2B + (1.06) \ 3B + (1.42) \ HR \]
Thorn and Palmer (1984) produced a linear weights formula using all offensive aspects recorded during a game.

\[
\text{Runs} = (0.46)1B + (0.8)2B + (1.02)3B + (1.4)HR + (0.33)(\text{BB + HBP}) + (0.3)\text{SB} - (0.6)\text{CS} - (0.25)(\text{AB - Hits}) - (0.5)(\text{OOB})
\]
**Total Average** (TA)

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

Note: Denominator is a measure of outs made, not at bats.
Scoring Index (DX) (modified in 1971)

\[ DX = \frac{H + BB + HBP + E}{AB + BB + HBP} \times \frac{TB + SB - CS}{AB + BB + HBP} \]

Note:

Compared 10 models of offensive performance including one they created.

Results: Several of the models were very close in performance, including DX (scoring index) and these were far superior to the classic baseball measures.
Using all the data from 1954-1999, they used the variables from Total Average (TA), namely SB, BB, 1b, 2B, 3B, and HR and created a *Least Squares Linear Regression Model*

<table>
<thead>
<tr>
<th>Weight for</th>
<th>SB</th>
<th>BB</th>
<th>1B</th>
<th>2B</th>
<th>3B</th>
<th>HR</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>.159</td>
</tr>
<tr>
<td>LSLR</td>
<td>.16</td>
<td>.36</td>
<td>.52</td>
<td>.67</td>
<td>1.18</td>
<td>1.50</td>
<td>.142</td>
</tr>
</tbody>
</table>
\[
LSLR = \frac{(.52)1B + (.67)2B + (1.18)3B + (1.5)HR + (.16)SB + (.36)BB}{Games}
\]

Note: modified to include HBP as well.
Results of tests

Using the 1954-1999 data:
TA and LSLR produced very similar results.
LSLR had a slightly better RMSE (but it had to!)
When looking at individual years in this period the results were a bit more mixed. LSLR outperformed TA in 35 of the 46 seasons. So there were still seasons when TA was best.
WHY????

LSLR was built on a particular data set and was sure to have a better RMSE on that data set. But on other data sets it is clearly not guaranteed to be best.

This is a problem of any model built from a specific data set, and hence there may be uses for more than one formula.
Albert and Bennett also noted that since RC and DX (and other formulas) can be described as weighted sum of cross product terms, linear regression can be used to better fit the coefficients in these formulas.

Doing this with the 1954-1999 data they improved the RMSE for RC to .140 (better than their LSLR model!).

Independent verification of how well James did in fitting his model.
An attempt at the all-time best

Michael J. Schell wrote

Baseball’s All-Time Best Hitters

Princeton University Press, 1999. After 282 pages of arguments, assumptions, and data analysis, he concluded Tony Gwynn was the best all-time hitter. Clearly no simple argument here! (Ty Cobb 2nd, Rod Carew, Joe Jackson, Rogers Hornsby, Ted Williams, ...)

Ron Gould Emory University Rating Player Performance - The Old Argument of Who is Best