# By the Numbers

Volume 9, Number 2 The Newsletter of the SABR Statistical Analysis Committee May, 1999

## Editor's Comments

Phil Birnbaum, Editor

In his 1982 *Baseball Abstract*, and again in 1983, Bill James defined Sabermetrics as "the mathematical and statistical analysis of baseball records."

In 1985, Craig Wright set him straight.

"Sabermetrics," said Wright, "is the scientific research of the available evidence to identify, study, and measure forces in professional baseball."

It's kind of an unwieldy definition, but it's correct. We don't study baseball statistics. We don't study baseball records.

We study baseball.

In first-year university, I took a layman's statistics course. One of our assignments was to pick up a medical journal, find some understandable statistical study, and report on it. At first, I was dubious – statistical analysis in a *medical* journal? But, of course, the journals were just teeming with statistics. You couldn't turn a page without running into a regression, or a chart, or an equation, or a table. Those medical studies were half a page of medical talk, and two pages of math.

But these doctors, the ones who publish their research – do we really consider them statisticians? Do we say that they work in the field of "medical statistical analysis?" Since they belong to the American Medical Association, do we call them "Amatricians?"

We don't. They are doctors. They are medical researchers. They study medicine, not medical statistics.

You can divide observable truths into two groups: those that occur every time, and those that occur only some of the time. It didn't take long for ancient observers to figure out that if you shove a pointy stick through someone's heart, he dies. But it took a whole bunch of years, and a whole bunch of studies, to prove that smoking can cause lung cancer. Why? Because stick death happens every time. Lung cancer death does not.

Statistics is a tool that helps you analyze things that don't happen every time. Because they don't happen every time, you need some way to measure how often they happen. That means you have to count. How often did smokers get cancer? How often did non-smokers get cancer? You've got to write down all those numbers, then do some statistics to figure out if the difference is significant.

The researcher doesn't care about the math: he cares about the result, the underlying medical truth. The math is only a tool: a tool in the service of medicine.

We don't analyze statistics for the sake of the statistics. We analyze statistics to discover scientific truths about baseball. Like smoking and cancer, the relationship between batting and team wins is not deterministic. What medicine is more likely to cure a losing team: two doubles, or three singles? The Sabermetrician finds the answer the same way as the Amatrician: by counting the patients, and doing some math.

Since the New England Journal of Medicine isn't called "By the Numbers," I don't think we should be either. Let's rename our publication to reflect what we really do. Send or e-mail suggestions for a better name for this publication. We'll run them next issue in this space.

*You can e-mail me at birnbaum@magi.com. Or, you can write me at #608-18 Deerfield Dr., Nepean, Ontario, Canada, K2G 4L1.* ♦

## In this issue

# Committee News

Neal Traven, Committee Co-Chair

## Reviewer requested

Recently, I received a manuscript intended for possible publication in the Baseball Review Journal. Though I've looked it over, I don't feel that I have sufficient background in a particular aspect of statistics to give the article a full review. Therefore, I'm looking for a volunteer to examine this submission and advise me (and, through me, Mark Alvarez) about the potential value of the article.

In particular, I seek someone who's well-rooted in statistical exact methods – whether it is appropriate to use them in the situations proposed by the author, and also whether they are necessary in such situations. Are exact methods applied to the data in a reasonable way? How can the approach in the paper be compared with other models, and are the cases cited by the author as being too complex for exact methods really so difficult to analyze? Finally, is there statistical analysis software in the marketplace that would make it easier to perform the calculations?

I know this request is somewhat vague. However, if you find this request interesting – better yet, if it strikes a responsive chord for you – please contact me to obtain additional information.

## SABR 29 update

As usual, the Statistical Analysis Committee will meet as an entity during the upcoming convention in Scottsdale. When someone from SABR asked me about the committee's needs for our meeting, I requested a 60 minute time slot and a room that would seat about 30 people. Unfortunately, I still haven't heard from either the Arizona or national SABR people about the precise time of our meeting, nor anything about which other committees will be meeting at the same time as us. In any case, I hope I'll see a lot of members at our get-together, and that if you're on another committee whose meeting time conflicts with the SAC you'll see fit to spend at least part of that time with us.

A tentative schedule of research presentations for SABR 29 can be found on the website of the Flame Delhi regional chapter of SABR (the local hosts in Scottsdale), at http://members.aol.com/fdelhichpt/index.html. I saw quite a few SAC members listed among the presenters, including both veteran researchers adding to their portfolios and new names who haven't previously spoken at the national meeting. All in all, there will certainly be lively discussion, both during the research presentation sessions and in lobbies, hallways, and taprooms.

*Neal Traven, 500 Market St. #11L, Portsmouth, NH, 03801; 603-430-8411; baseball@ttlc.net* ♦

## Receive BTN by E-mail

You can help save SABR some money, and me some time, by receiving your copy of *By the Numbers* by e-mail. BTN is sent in Microsoft Word 97 format; if you don't have Word 97, a free viewer is available at the Microsoft web site (www.microsoft.com).

To get on the electronic subscription list, send me (Phil Birnbaum) an e-mail at phil_birnbaum@iname.com. (That's an underscore _ between Phil and Birnbaum.) If you're not sure if you can read Word 97 format, just let me know and I'll send you this issue so you can try

If you don't have e-mail, don't worry–you will always be entitled to receive BTN by mail, as usual. The electronic copy is sent out two business days after the hard copy, to help ensure everyone receives it at about the same time.

# Dogma vs. Open-Mindedness

Mike Hoban

*An article in the last issue of BTN questioned the author's response to those who were critical of his work. Here, he responds to that article, and defends the original research that provoked the comments.*

As many readers may know, there were two lengthy discussions regarding the HEQ rating system (Hoban Effectiveness Quotient) on SABR-L: one took place last September/October and the second this January. I had promised myself that I would say nothing further about the HEQ on SABR-L, at least until after my book is published. Then BTN published some comments from Neal Traven concerning that discussion on SABR-L and I felt that I should try to speak to a few of his points.

First, let me say that I feel that Neal's comments were essentially fair and accurate, for the most part. I want to react to just a few of them because I believe some clarification may be in order. Any quotes that follow will be from Neal's article in BTN.

Neal suggested in his comments that there are some issues regarding "scientific methodology and research collegiality" that I do not seem to understand. Having been an academic for more than forty years, I know a great deal about such matters. If Neal were to examine the SABR-L logs of last September/October more closely, I believe that he would find that one or two of his colleagues (in their rush to judgment to defend *SABR dogma*) were guilty of ignoring virtually any semblance of "research collegiality" or even civility, for that matter. Quite frankly, I was surprised by the unprofessional (and personal) nature of the attacks by two of the members. It was obvious from some of their remarks that the two were outraged by the fact that my book on the HEQ was being published and that an article based on it was printed in the SABR *Baseball Research Journal.* Is this Neal's definition of "research collegiality?"

In general, I have been quite satisfied with the reception that I have received on SABR-L. I have created the HEQ system as a concrete way to question some of the statistical extremes that have been perpetrated on baseball fans over the past two decades. And to challenge the apparent belief that *more complicated must be better*. As a true fan, I welcome the opportunity to discuss and debate all aspects of the game we love. Only the occasional narrow-mindedness bothers me a bit.

### *"He (Hoban) refuses to engage his critics in any meaningful manner."*

This would be a serious criticism – if indeed it were true. In fact, I have answered every question that has been asked on SABR-L. (Although, of course, not to the satisfaction of everyone.) What I refuse to do is to engage in non-productive discussion of points that have no definitive answer. A good example of this would be one that Neal raises in his paper. Indeed, for the sake of fan-friendliness, I gave a walk a weight of 0.5 because it is less valuable than a single which in total bases has a weight of 1. Could I have chosen a weight of 0.6 or 0.4? Of course, but I would have gotten the same response. The fact of the matter is that there is no indisputable "empirical evidence" (despite Pete Palmer's best efforts) to conclusively say how a walk should be weighed in comparison to a single.

But, in terms of engaging critics in a meaningful manner, there is a more significant point here. On at least two occasions, I offered to send information on the HEQ to anyone who wished to have any questions answered – reasoning that not all members on SABR-L were interested in reading all of this discussion. What could be more "collegial" than that? Many SABR members responded and received this information. *Not one of the few members who criticized the HEQ on SABR-L requested this information*. I concluded that their questions were not intended to seek information but to be strictly rhetorical – and why waste the time of SABR-L subscribers by answering rhetorical questions.

### *"Why …has Hoban raised hackles and attracted widespread controversy while Schell's work has drawn little attention …?"*

Yes, Neal, this is the real question. And Neal himself touches on the answer elsewhere in his paper: "… for many of the same reasons that Hoban has been so roundly criticized on SABR-L – failure to adjust for time and place, use of team-dependent quantities like RBIs, and the like." And again, "…drawing conclusions seemingly at odds with SABR conventional wisdom."

This is indeed the heart of the matter and the true reason for the irrational response of a few members on SABR-L. The real issue is that it is very difficult (if not impossible) for some fanatical true believers to accept the fact that there may be alternate paths to the truth other than those pointed out by their own prophets. This has always been the case. Just ask Martin Luther. It is always the tendency of a few to attack the questioner rather than to deal with the questions.

Even though I am well acquainted with the tenets of "SABR conventional wisdom," I have chosen to post my own theses on the cathedral door and suggest that there may be a simpler, just as accurate but more fan-friendly way to compare the careers of position players than those advanced by Bill James, Pete Palmer and others. This is the point that many of the true believers of "SABR conventional wisdom" seem to find unacceptable. Let's not make any mistake about one thing. Player comparison methods such as Runs Created or TPR or HEQ are just attempts to approximate how well a player has performed. There is no right or wrong way to do this. The only things that are real in baseball are the actual numbers that the players have put into the record book. Any attempt to interpret these numbers in a meaningful way is open to all sorts of criticism because there are no clear-cut ("empirical") answers – just intuitive judgments.

Some members have apparently gotten annoyed because in their view I have the effrontery to suggest that those offensive numbers with which the fans are most familiar (and that can be found in any decent daily newspaper) like runs scored and RBIs can be used in a relatively simple manner to approximate how productive a player has been. Is this true heresy? Not really, but it may appear that way to those SABR members who, having adopted the commandments according to "SABR conventional wisdom," feel it to be their duty to their new religion to dismiss those who would have the gall to question these edicts from on high. This is "research collegiality?"

The principal reason why the HEQ has "raised hackles" among a few true believers is because (in their somewhat narrow view of baseball reality) they seem to believe that this approach rejects the statistical advances that have been made over the past twenty to thirty years. (Which it does not.) They choose to believe that no person in his right mind could advocate such an approach unless he was ignorant of those advances. But, if an individual (after his or her indoctrination into "SABR conventional wisdom") is still able to step back and examine these advances, he or she will realize that we are dealing here with ideas and theories and not with facts. To cite one example, it is one thing to demonstrate that different eras produced different levels of offensive production. It is quite something else to suggest that by some form of mathematical manipulation we can *adjust* these numbers in some sort of fair and equitable manner in order to compare players. It seems to be far more accurate and fair to demonstrate what the actual numbers say and then, perhaps, to qualify those results with comments about the era and the park in which the numbers were produced. Mel Ott, for example, emerges as the eleventh most effective offensive player of the century with a HEQ career offensive score of 604 (right behind Barry Bonds at 606). This result is based on his actual numbers. Having said this, advocates of *adjustment* may want to qualify this statement by commenting on the effect of the era and the ballpark in which he played – and this is entirely appropriate. Let's not change the actual numbers to deny what he achieved but rather qualify the actual numbers by explaining those circumstances that we may feel are relevant.

To take another example. For someone to say that numbers such as runs scored and RBIs are unworthy of being used in some way to compare players' achievements because they are "team dependent" is downright silly. Any clear-thinking individual (not blinded by a narrow interpretation of SABR dogma) knows that common sense dictates otherwise. Baseball is a team sport and every accomplishment is team dependent in one way or another. No wonder those who have carried SABR findings to this extreme are dismissed by the majority of serious fans as having surrendered reality in favor of questionable mathematical manipulations.

What the HEQ approach is really suggesting is that some of those who view themselves as *research sabermetricians* have *over-mathematized* the real numbers of baseball to the point where they have lost the attention of 90% of the fans – and that's a shame. In this context, the paranoia exhibited by a few SABR-L fanatics towards the HEQ is somewhat justified. That is, the HEQ, to some extent, does strike at the heart of some of these extreme notions. In attempting to lay some sort of foundation on which to construct a total-season statistic that will be meaningful to the serious fan (and possibly even to the casual fan), the HEQ is saying *enough already*. Let's construct a common-sense based statistic that will approximate how good a season (hitting and fielding) a player really had. (A statistic that will appeal to the serious fan who has not narrowed his vision and seemingly lost touch with reality.) If this is not a valid goal of sabermetrics, then perhaps I have misinterpreted Bill James' definition of the term. Can such a statistic be developed? I am not sure, but I think it can. For more on the subject, please contact me as below.

*Mike Hoban, Dean, Schlaefer School, Monmouth University, West Long Branch, NJ, 07764-1808, mhoban@mondec.monmouth.edu* ♦

Editor's note: As discussion on Mr. Hoban's work attracts a large volume of responses, we won't be able to publish rebuttals to Mr. Hoban here in BTN. Discussion of HEQ, and the preceding article, can be posted to SABR-L, the organization's internet discussion forum. For more information on SABR-L, please visit the SABR website at www.sabr.org.

*Letter to the Editor*

# G. Jay Walker is Still Wrong; and Breakeven Rates

Clem Comly

I join with Phil in his goal of having more dialogue in the newsletter. He wanted me to get this to him a month ago, in case he needed to give G. Jay Walker another inning. Let me just say that I didn't misinterpret Jay's example as his solution to OWAR. My point was that his example of the blemish in his title was no such thing – and that was the flaw I was referring to. I wasn't criticizing the new method he introduced in the last 8 pages of his essay. That method would have stood on its own without Jay discussing this so-called blemish. As to the new method Jay promised for the 1999 BBBA, I have it but haven't had time to read it yet. Fortunately, the Arizona flight will be a long one.

If a policeman's favorite punctuation is a question mark, a commercial spokesman's is an exclamation point, Al Gore's is a period, and Bill Clinton's a comma, mine is a parenthesis (as Jay knows). So, one of the things that came to my mind reading last issue's Tom Ruane (has any one else made the type of mistake I did with Tom by calling him Tom Truane because his e-mail goes to truane?) is the proper use of breakeven points in strategic analysis. I recognize many who hear me are in the choir while many heathen are out of earshot, but I have never seen it addressed in a baseball forum.

The temptation is to compare the breakeven point to the league (or team or player) average. So, to use round numbers, if the steal of second breakeven point is 65% and the 1998 NL is 68%, the NL is doing well. But breakeven is a marginal rate, not an average rate (flashback to micro-economics class). If the runner thinks he can make second 65% or more of the time, he should go. Sometimes the runner calculates he will be safe 75% of the time and he tries for second. Other times, he calculates he will be safe 50% and he holds first.

So, what would an average success rate be for a 65% breakeven? I think a conservative estimate would be that 65% opportunities are twice as common as 75% ones. If we say the success-rate to opportunity is linear, that means our opportunities are 0 at 85% (even though Dave Lopes averaged 90+% in a couple seasons) and 3 out of every four opportunities will fall in the 65-75% success range. That gives us an overall success average of 71.7%. Of course, if we were to says the opportunities are equal across the span of 65-85%, that would give us an average of 75%.

*Clem Comly is co-chair of the Statistical Analysis Committee. 308 Colonial Drive, Wallingford, PA, 19086-6004; ccomly@erols.com* ♦

## Book Reviews Wanted

Every year, a number of books and magazines are published with a Sabermetric slant. Many of our members have never heard of them. Our committee members would like very much to hear when this kind of stuff comes out.

If you own a copy of any baseball book of interest, we'd welcome a summary or a full-length review. Since we've hardly published for the last couple of years, even reviews of older books – say, 1997 or later – would be welcome. The only restriction, please: the book should have, or claim to have, some Sabermetric content.

Send reviews to the usual place (see "Submissions" elsewhere in this issue). Drop me a line if you want to make sure no other member is reviewing the same publication, although multiple reviews of the same book are welcome, particularly for major works. Let me know which book you're doing, so I don't assign the same book twice.

And if you're an author, and you'd like to offer a review copy, let me know – I'll find you a willing reviewer. Sig Mejdal's book review in this issue was made possible by a review copy by the author. You could be next.

# Normalized Winning Percentage Revisited

Bill Deane

*Many ways of adjusting a pitcher's winning percentage for his team's performance suffer a common problem – they do not take into account that it's easier for a pitcher on a .400 team to outperform is teammates than it is for a pitcher on a .600 team. Here, the author re-introduces a statistic that compensates for this bias.*

In the 1996 *Baseball Research Journal* (page 42), I presented something I call "Normalized Winning Percentage," or NWP. After three years' passage of time, and the emergence of a new all-time leader, it seems a good time to revisit this statistic.

To review, NWP projects how a pitcher might perform on a .500-team, thus putting all hurlers, past and present, on an even plane of comparison. The concept starts out by comparing a pitcher's won-lost record to that of his team, neutralizing the impacts of a team's offense and defense on its pitchers' records. This idea is hardly new, but my formula addresses one basic problem others had not: a pitcher on a poor team has more room for improvement than one on a good team. In other words, it's easier for Walter Johnson to exceed his team's win percentage by 100 points than it is for Whitey Ford.

NWP basically measures how much a pitcher has exceeded his team's performance, divided by how much he *could* have done so, and scaling the result as if he had pitched for an average (.500) team. Thus, a hurler who posts a .520 percentage for a .400-team gets credit for the same NWP score (.600) as a .600-pitcher on a .500-team, or a .680-pitcher on a .600-team -- because each has exceeded his team's percentage by 20% of the potential room for improvement.

For a pitcher whose win percentage exceeds his team's, the formula for NWP is as follows: average percentage plus [(pitcher percentage minus team percentage) times (perfect percentage minus average percentage) divided by (perfect percentage minus team percentage)]. Rather cumbersome but, since "average percentage" is always equal to .500 and "perfect percentage" is always equal to 1.000, we can simplify the formula as follows:

$$\text{NWP} = .500 + \frac{\text{Pitcher Pct. - Team Pct.}}{2 \text{ x } (1.000 - \text{Team Pct.})}$$

Incidentally, for a pitcher whose percentage is *lower* than his team's, the converse-NWP formula is applicable: .500 minus [(team percentage minus pitcher percentage) divided by (team percentage doubled)].

### 1998 Leaders In Normalized Winning Percentage, With Wins Above Team
**(Table produced with Assistance from Pete Palmer)**

(Minimum 15 Wins or 20 Decisions)

| *Pitcher, CLUB (LG)* | *W* | *L* | *Pct.* | *Team W* | *L* | *Adj.Pct.* | *NWP* | *WAT* |
|---|---|---|---|---|---|---|---|---|
| John Smoltz, ATL | 17 | 3 | .850 | 106 | 56 | .627 | .799 | 5.98 |
| Roger Clemens, TOR (A) | 20 | 6 | .769 | 88 | 74 | .500 | .769 | 7.00 |
| Rick Helling, TEX (A) | 20 | 7 | .741 | 88 | 74 | .504 | .739 | 6.45 |
| Al Leiter, NY | 17 | 6 | .739 | 88 | 74 | .511 | .733 | 5.37 |
| Kenny Rogers, OAK | 16 | 8 | .667 | 74 | 88 | .420 | .713 | 5.10 |
| David Wells, NY (A) | 18 | 4 | .818 | 114 | 48 | .686 | .711 | 4.64 |
| Pedro Martinez, BOS | 19 | 7 | .731 | 92 | 70 | .537 | .709 | 5.44 |
| Pete Harnisch, CIN (N) | 14 | 7 | .667 | 77 | 85 | .447 | .699 | 4.17 |
| Tom Glavine, ATL (N) | 20 | 6 | .769 | 106 | 56 | .632 | .686 | 4.84 |
| Kevin Brown, SD | 18 | 7 | .720 | 98 | 64 | .584 | .664 | 4.09 |
| Jamie Moyer, SEA | 15 | 9 | .625 | 76 | 85 | .445 | .662 | 3.89 |
| Kevin Tapani, CHI (N) | 19 | 9 | .679 | 90 | 73 | .526 | .661 | 4.51 |
| Dustin Hermanson, MON | 14 | 11 | .560 | 65 | 97 | .372 | .650 | 3.74 |
| Tim Wakefield, BOS (A) | 17 | 8 | .680 | 92 | 70 | .547 | .646 | 3.66 |
| Rolando Arrojo, TB (A) | 14 | 12 | .538 | 63 | 99 | .360 | .639 | 3.62 |

To put the NWP formula into practice, let’s take a look at Pedro Martinez’s 1998 performance for the Red Sox. Martinez compiled a 19-7 (.731) log, while his team was 92-70 overall. Subtracting his decisions, the Sox had a 73-63 record for a .537 percentage. Martinez’s NWP is calculated as follows:

$$\text{NWP} = .500 + \frac{.731 - .537}{2 \times (1.000 - .537)} \quad \text{, or } .500 + .194/.926$$

Martinez’s resultant NWP (.709) was one of the top seven in the majors last year; a list of the 1998 leaders is on the preceding page.

I developed the concept for NWP in the early 1980s. The formula has undergone several minor refinements over the years, and undoubtedly has room for more. NWP’s biggest weakness is that it assumes all pitching staffs to be created equal, so that an average pitcher on a poor staff can appear better than an excellent pitcher on a great staff. While this creates some aberrant single-season results, things tend to even out over a pitcher’s career.

NWP can be, and has been, incorporated into what analyst Pete Palmer calls “wins above team” (WAT), the number of victories a pitcher contributes over what an “average” pitcher might. Palmer revised his formula to include mine in *Total Baseball*. The formula for WAT (for pitchers with higher percentages than their teams) is as follows:

$$\text{WAT} = \text{Pitcher decisions} \times \frac{\text{Pitcher Pct.} - \text{Team Pct.}}{2 \times (1.000 - \text{Team Pct.})}$$

A list of the top 15 twentieth century pitchers in NWP, including WAT, follows. Since joining the 200-win club, Roger Clemens (.657) has supplanted Lefty Grove (.643) for the best career normalized winning percentage, while Cy Young easily retains his record for most wins above team (99.7). Of course, we realize that Clemens (and Greg Maddux, now ranked #9) might drop in the rankings with some sub-par seasons in their waning years. Each of the 15 who is eligible is in the Hall of Fame. As a group, their careers are quite evenly distributed between each decade of the century, as opposed to conventional measures of pitching, which suggest that all of the best hurlers toed the rubber before Warren Harding was president.

### All-Time Leaders (Minimum 200 Wins Since 1900)

| Pitcher | W | L | WAT | NWP |
|---|---|---|---|---|
| Roger Clemens | 233 | 124 | 55.9 | .657 |
| Lefty Grove | 300 | 141 | 62.9 | .643 |
| Grover Alexander | 373 | 208 | 81.6 | .640 |
| Whitey Ford | 236 | 106 | 44.4 | .630 |
| Walter Johnson | 417 | 279 | 90.0 | .629 |
| Cy Young | 511 | 316 | 99.7 | .621 |
| Christy Mathewson | 373 | 188 | 64.9 | .616 |
| Tom Seaver | 311 | 205 | 58.9 | .614 |
| Greg Maddux | 202 | 117 | 32.5 | .602 |
| Juan Marichal | 243 | 142 | 38.7 | .601 |
| Bob Feller | 266 | 162 | 36.8 | .586 |
| Carl Hubbell | 253 | 154 | 34.6 | .585 |
| Joe McGinnity | 246 | 142 | 32.4 | .584 |
| Warren Spahn | 363 | 245 | 45.8 | .575 |
| Ted Lyons | 260 | 230 | 36.2 | .574 |

Although Young and McGinnity started their careers before 1900, they are included because each won at least 200 games after that year; their statistics include pre-1900 records. Statistics for Clemens and Maddux are complete through 1998.

*Bill Deane, P.O. Box 47, Fly Creek, NY, 13337, (607) 547-5786, DizDeane@aol.com.* ♦

# Reaching Base on Errors

Clifford Blau

*Although players with certain characteristics will reach base on error more often than other players, the differences are small, and may be ignored in evaluating a player who is not near one of the extremes.*

In my review of some recent sabermetric posts to SABR-L (*By The Numbers*, February 1999), I commented on one by Tom Ruane. He looked at the relative costs of strikeouts, groundouts, and flyouts. In part because he included times that batters reached base on errors (and unsuccessful fielders choices) in outs, he found that groundouts were the least costly. I implied in my commentary that this information could be useful in creating run scoring models.

Subsequently, Mr. Ruane supplied some data on how many times each hitter had reached base on error (henceforth ROE) from 1980 to 1998. I compared these data to other statistics to see if a model could be developed that would predict a hitter's ROE average.

I initially performed a multivariate regression analysis using the following factors, each on a per-at-bat basis: hits, doubles, triples, home runs, strikeouts, sacrifice flies, stolen bases, and grounded into double plays. I also used the hitter's batting side. Despite finding a strong relationship using a limited data set, once I tried the analysis with the full sample, I found almost no correlation between those factors and ROE. After eliminating all players with fewer than 1000 at bats, the correlation was .264.

I next divided players up into groups with certain characteristics. One such group consisted of right-handed hitters who ground into double plays a lot, while another was left-handed hitters who rarely ground into double plays. I also compared right-handed and left-handed home run hitters who strike out frequently. While the average player in the sample ROE'd 14 times per 1000 at bats, the group expected to reach most often, right handed ground ball hitters, had an average ROE of 15.5 times per 1000 at bats, while the upper cutting left handers ROE'd 10.4 times per 1000 at bats. Using other groups, I found the difference between right handed and left handed hitters to be about 3 or 4 ROE per 1000 at bats. For every extra 100 strikeouts, a batter could be expected to ROE 1.5 fewer times. Speed also had a slight relationship; those stealing bases at three times the average rate had an ROE average .001 higher than normal. A similarly small, opposite relationship exists for slow runners. No relationship was apparent between grounding into double plays and ROE.

Some other authors have looked at this question. In the 1984 edition of the *Bill James Baseball Abstract*, Mr. James studied how often Texas Rangers players reached base on error in the 1983 season. He concluded that right-handed batters ROE almost 30% more often than lefties, and fast runners ROE 12% more often than slow runners. In his 1986 *Abstract*, he reported that on the 1985 Mariners, right-handed hitters reached base just slightly more often than lefties, but fast runners made it 16% more often than slow runners. Mark Pankin, in his article "Subtle Aspects of the Game," used Project Scoresheet data for all major league games from 1984-1992. He found that fast runners and right-handed hitters reach base on errors more often. The advantage for righties overwhelmed the speed factor; slow right-handed batters reach base on errors more than fast lefties do.

In summary, just as evaluating fielders by fielding average is not very meaningful because the differences are so small, the same holds true for hitters. If one is rating a hitter is seems to be at one extreme or the other in ROE, one should keep in mind that the hitter is a little more or less valuable than popular formulas such as Runs Created or Linear Weights would suggest.

*Clifford Blau, 16 Lake St. #5D, White Plains, NY, 10603, proboy@ix.netcom.com* ♦

# "Forecaster" Insightful, but Sabermetrically Light

Sig Mejdal

*This annual book for fantasy players features accurate, valuable predictions and insightful player comments, but sabermetricians may long for a little more of the analysis and research that back those predictions up.*

## For the fantasy player…

Anyone with a calculator can make player projections. In fact, you need look no further than your local bookstore to realize the number of different "calculator-owners" producing these fantasy baseball aids. Instead of asking, "Where can I get player projections?" the fantasy player is now asking "Where can I get the best player projections?"

Projections that are based on systematic, well-researched and validated methods are demanded by the discriminating player, and this is where *Ron Shandler's Baseball Forecaster* comes to the rescue. Ron's background as a forecasting analyst, his knowledge of statistical methods, and the applied research he has collected all combine to produce what seems to be the most accurate and valuable predictions available. In addition to a description of some of the tools used and insights gained form his 13 years of producing the Forecaster, complete projections of every major leaguer and minor leaguer (AA and above), dollar values, forecast risk/certainty, and lists by position are also included.

Ron Shandler's Baseball Forecaster – 1999
Annual Review

By Ron Shandler

Shandler Entreprises LLC, 188 pages, $23.95

Although no comparative analysis with other projection methods is included in this book, his success in the "*Tout Wars – Battle of the Experts*" provides some evidence to the validity of his methods. The answer to "Where can I get the best player projections?" just may be "In *Shandler's Forecaster*."

## For the sabermetician…

I agree with Ron when he writes that the *Forecaster* is one of the few embers that remain from the original fire created by Bill James' *Baseball Abstracts*. In my opinion, the joy of reading the original *Abstracts* was the combination of James' compelling essays and statistical research. Ron clearly has both of these skills. His statistical expertise is clearly evident and his writing is both direct and compelling. I think a great example of his directness is found in his description concerning the problems with small sampling sizes, Ron's succinct comment that even "Mario Mendoza once went 5 for 8" really drives the point home. Unfortunately, only 18 of the nearly 200 pages are filled with his writings. Personally, I would have liked to see much more.

From a sabermetric standpoint, it is easy to see that the research cited and the thought put into the projections are both quite significant. However, I was disappointed by the fact that I was not able to read more about the supporting research behind the projections, and, in particular, the specific methodology used to generate his projections. Ron acknowledges the lack of this information as he writes, "So pardon the lack of support data. Rest assured we're not making it up." Moreover, early on in the book, Ron gives the readers a warning that the realities that he describes "may completely contradict your own values and beliefs. You can buy into them or not – it's your choice." Well, Ron is correct in that much of his "realities" did contradict my beliefs. I *do* want to buy into them – but I need the supporting evidence in order to do that. I *do* believe that they aren't making it up – but as a curious sabermetrician, I still want to read about the research. In all fairness, Ron says that the "data has appeared in our other publications and on our website in the past." But if this book is looked at as a "stand-alone" document, from a sabermetric standpoint and as a first time reader I was left wanting more.

*Sig Mejdal, smejdal@monterreytechnologies.com* ♦

# Methods for Comparing Pitchers Across Eras

Rob Wood

*Choosing a list of all-time greats requires us to have a way to compare performances from different eras. In this article, the author presents a method for normalizing pitching performance, and shows the era-adjusted record of his personal list of all-time greats.*

## Introduction

A favorite hobby of baseball fans and researchers alike is making lists of the all-time greats. By position, by team, by decade, by birthplace, etc. Indeed, SABR has recently conducted a poll to determine members' views on the top 100 players of the 20th century. And every year Hall of Fame voting rekindles old arguments about who was better than whom.

Who is the greatest pitcher ever? Walter Johnson? Lefty Grove? Roger Clemens? Sandy Koufax? In this article I will present a set of methods, a variety of data, and a few opinions on the subject. My goal is to assemble in one place a compendium of data and methods that can be used for comparing pitchers across eras.

For each of several pitching statistics such as ERA and strikeouts, I will present two types of information. First, by calculating the decade by decade league averages, we will see how the statistic has varied over time. From this data each decade's "era effect" will be calculated. These widely varying era effects will demonstrate how important it is to consider the pitcher's era. Second, I will present the all-time career leaders in the statistic, when era effects are fully taken into account. Each pitcher's stats will be adjusted on a season-by-season basis, as described below.

All data used in the article are restricted to 20th century major leagues (NL 1900-1998, AL 1901-1998, FL 1914-1915). This article is based upon research conducted jointly with Bob McCleery, which in turn was based upon research by many SABR members.

While I will frame the discussion in terms of adjusting career stats, the same methods apply to seasonal stats. In fact my career adjustments can be thought of as simply the sum of the individual seasonal adjustments. The focus of the article is on adjusting for a pitcher's era. Just as importantly, data on the relevant pitching statistics will also be adjusted for park effects.

In case you are interested, the article concludes with my personal ranking of the 20 greatest pitchers of the 20th century, based upon the cross-era comparisons. (Of course, your ranking may differ.)

## ERA

Let me begin with the cornerstone of a pitcher's statistical line, earned run average. No other single variable captures true performance quality better than ERA. However, a pitcher's ERA is subject to environmental factors which, if not taken into account, would significantly bias cross-era comparisons.

Table 1 displays the average league ERA by decade in the NL and AL. As you can see, the height of pitching dominance occurred in the 1900's and 1910's. The height of hitting dominance occurred in the 1930's, especially in the AL. Note that the 1990's rival the 1930's as the leading hitter-happy decade of the century.

Table 1: ERA by Decade

| Decade | NL League Ave | NL Era Effect | NL League Ave | NL Era Effect |
|---|---|---|---|---|
| 1900's | 2.88 | 0.78 | 2.84 | 0.76 |
| 1910's | 2.95 | 0.80 | 2.93 | 0.79 |
| 1920's | 3.96 | 1.07 | 4.11 | 1.11 |
| 1930's | 3.98 | 1.07 | 4.58 | 1.24 |
| 1940's | 3.71 | 1.00 | 3.80 | 1.02 |
| 1950's | 3.98 | 1.07 | 3.96 | 1.07 |
| 1960's | 3.57 | 0.96 | 3.59 | 0.97 |
| 1970's | 3.66 | 0.99 | 3.71 | 1.00 |
| 1980's | 3.62 | 0.98 | 4.04 | 1.09 |
| 1990's | 4.01 | 1.08 | 4.44 | 1.20 |
| | | | | |
| 1900-1998 | NL: 3.63 | | AL: 3.80 | ML: 3.71 |

My method of calculating era effects is to divide the league average by the overall 20th century average. For example, the first row in the table shows that the average ERA in the NL 1900's was 2.88. Dividing 2.88 by 3.71, the 20th century major league average ERA, yields 0.78, the decade's ERA "era effect". For variables in which lower numbers denote a more pitcher-friendly environment (such as ERA), an era effect lower than 1.00 denotes a pitcher-friendly era. Similarly, for variables in which higher numbers denote a more pitcher-friendly environment (such as shutouts, complete games, innings, strikeouts), an era effect greater than 1.00 denotes a pitcher-friendly era.

Clearly, if we do not take into account the era in which the pitcher toiled, we would be tempted to claim that all the best pitchers pitched in the early decades of the century, and that all the best hitters played in the 1920's and 1930's.

Era effects are calculated so that we can measure each pitcher's performance relative to his contemporaries (as reflected in league averages). Cross-era comparisons can then be made by comparing how much better, as a percentage, each pitcher was than his own league average. For example, a pitcher with a 3.00 ERA in a 4.00 ERA league (75%) is deemed better than a pitcher with a 2.00 ERA in a 2.50 ERA league (80%).

This method captures the lion's share of what we typically mean by the "era". Other factors, such as wartime or expansion baseball, may not be fully captured by the league average. Essentially, these factors affect both the average and spread (variance) of the distribution. I will not address these issues in this article.

Table 2 presents the all-time leaders in ERA, appropriately measured so that era and park effects are taken into account. A pitcher's career era effect is the weighted average of his individual seasons' era effects, where the weights are the pitcher's innings pitched in each season (the same weighting method is used to derive a pitcher's career park effect). We divide a pitcher's actual career ERA by his career era effect to derive his career "era-neutral" ERA. The final step is to divide the pitcher's era-neutral ERA by his career park effect to arrive at the ultimate "era-neutral park-adjusted" ERA.

Table 2: All-time leaders in era-neutral park-adjusted ERA (min 2000 IP)

| | Actual ERA | League ERA | Era Effect | Era-neutral ERA | Era-neutral Park-adjusted ERA |
|---|---|---|---|---|---|
| Roger Clemens | 2.94 | 4.29 | 1.16 | 2.54 | 2.46 |
| Greg Maddux | 2.75 | 3.91 | 1.05 | 2.61 | 2.46 |
| Lefty Grove | 3.06 | 4.43 | 1.19 | 2.56 | 2.52 |
| Walter Johnson | 2.17 | 3.23 | 0.87 | 2.49 | 2.52 |
| Hoyt Wilhelm | 2.52 | 3.73 | 1.01 | 2.51 | 2.54 |
| Ed Walsh | 1.82* | 2.73 | 0.74 | 2.47 | 2.55 |
| Addie Joss | 1.88 | 2.73 | 0.74 | 2.55 | 2.59 |
| Mordecai Brown | 2.06 | 2.90 | 0.78 | 2.64 | 2.66 |
| Cy Young | 2.19 | 2.95 | 0.80 | 2.75 | 2.73 |
| Christy Mathewson | 2.13 | 2.91 | 0.78 | 2.72 | 2.73 |

The all-time era-neutral park-adjusted ERA leader is Roger Clemens. To review the data on Clemens in the table, entering the 1999 season Rocket had a career ERA of 2.94. During his career, the leagues in which he pitched had an average ERA of 4.29. The next column indicates that this 4.29 is 1.16 times the 20th century average ERA of 3.71 (4.29/3.17=1.16). Dividing Clemens's actual ERA of 2.94 by this 1.16 era effect yields his 2.54 era-neutral ERA. The last column adjusts the era-neutral ERA by taking into account the ballparks Clemens pitched in during his career, using the park effects presented in Total Baseball. Clemens's career park effect is 1.035, a result of runs scored being 7.0% more prevalent in his home parks than league average, and half his games being pitched at home (2.54/1.035=2.46).

My approach of converting a pitcher's ERA (and other stats) to an era-neutral ERA is equivalent to measuring a pitcher's ERA as a percentage relative to the league ERA. For example, Roger Clemens's actual ERA is 68.5% of his league ERA (some have called this "relative ERA"). In essence I take the added step of multiplying relative ERA by the 20th century average ERA to arrive at the era-neutral ERA (68.5% times 3.71 is 2.54). When park effects are taken into account, Clemens's era-neutral ERA of 2.54 is lowered to 2.46. This figure is equivalent to Total Baseball's ERA+, since TB expresses adjusted ERA as a percentage better than league, so Clemens's 2.46 era-neutral park-adjusted ERA is divided into 3.71 to arrive at his ERA+ of 1.51 (TB writes this as 151). My method allows us to compare pitchers using numbers expressed as ERAs, rather than pure percentages, though both approaches are equivalent.

One way to tell whether the approach to "factor out" the era is successful is to see whether a mix of eras shows up on the list of all-time leaders. Such appears to be the case here. To my surprise, Greg Maddux is virtually tied with Roger Clemens at the head of the list. Of course, Clemens and Maddux have not yet experienced any "down side" to their careers that typically accompanies the tail-end of a pitcher's career. Lefty Grove and Walter Johnson, the two main contenders for all-time greatest pitcher, are virtually tied for 3rd and 4th on the list. Ed Walsh, the holder of the lowest actual career ERA, checks in with the 6th lowest adjusted ERA. Hoyt Wilhelm is 5th on the list. However, if a reasonable adjustment is made to reflect a reliever's ERA advantage, Wilhelm falls to 9th on the list.

## Shutouts

Shutouts is another pitching variable subject to era effects. Table 3 displays the average frequency of shutouts by decade. Here an era effect greater than 1.00 denotes a pitcher-friendly environment. 9.4% of all team-games were shutouts in the NL 1900's; equivalent to saying that 18.8% of all games were decided by a shutout, since only one team can have a shutout in any game. The height of shutout mania in recent times was the NL 1960's (Koufax, Drysdale, Marichal, Gibson, Bunning, et al.). Nearly 8% of all team-games were shutouts (nearly 16% of all games) in the decade, with shutouts in zany 1968 accounting for more than 11% of all team-games (or nearly 23% of all games). Is it no wonder that rule changes were implemented after 1968 to get a little more offense into the game?

Table 3: Shutouts by Decade (average frequency per team per game)

| Decade | NL League Ave | NL Era Effect | AL League Ave | AL Era Effect |
|---|---|---|---|---|
| 1900's | 9.4% | 1.35 | 9.7% | 1.39 |
| 1910's | 9.3 | 1.35 | 8.9 | 1.29 |
| 1920's | 5.6 | 0.81 | 5.3 | 0.76 |
| 1930's | 6.3 | 0.90 | 4.4 | 0.63 |
| 1940's | 7.3 | 1.06 | 7.2 | 1.04 |
| 1950's | 6.3 | 0.91 | 6.6 | 0.96 |
| 1960's | 7.9 | 1.15 | 7.4 | 1.06 |
| 1970's | 7.2 | 1.04 | 7.2 | 1.05 |
| 1980's | 5.9 | 0.86 | 5.0 | 0.86 |
| 1990's | 5.7 | 0.83 | 5.1 | 0.73 |
| | | | | |
| 1900-1998 | NL: 7.1% | | AL: 6.7% | ML: 6.9% |

Table 4 presents the all-time leaders in era-neutral park-adjusted shutouts. Walter Johnson, the all-time leader in actual shutouts, heads the list. Walter's 110 actual shutouts are deflated to merely 97 by taking into account the fact that shutouts were 13% more prevalent during his career than in the 20th century major leagues taken as a whole. The Big Train's home parks also aided his cause. Taking the effect of his ballparks into account yields a grand total of 95 era-neutral park-adjusted shutouts.

Since the number of shutouts in any park in any one season presents too small of a sample, I have not estimated a shutout effect for each park for each season. Instead I have used data covering all seasons to estimate a general shutouts park factor based upon the effect the park has on runs scored. As a coarse estimate, I find that the effect a ballpark has on shutouts is about twice as large as its effect on runs scored. For example, if a ballpark depresses runs by 10%, it tends to increase the frequency of a shutout by about 20%. By using this rule of thumb, I have adjusted a pitcher's era-neutral shutout total to derive his era-neutral park-adjusted shutouts.

Table 4: All-time leaders in era-neutral park-adjusted Shutouts

| | Actual Shutouts | Era Effect | Era-neutral Shutouts | Era-neutral Park-adjusted Shutouts |
|---|---|---|---|---|
| Walter Johnson | 110* | 1.13 | 97 | 95 |
| Grover Alexander | 90 | 1.06 | 85 | 86 |
| Roger Clemens | 44 | 0.68 | 65 | 69 |
| Bert Blyleven | 60 | 0.91 | 66 | 67 |
| Red Ruffing | 48 | 0.70 | 68 | 64 |
| Warren Spahn | 63 | 0.97 | 65 | 60 |
| Nolan Ryan | 61 | 0.97 | 63 | 60 |
| Tom Seaver | 61 | 1.03 | 59 | 59 |
| Christy Mathewson | 80 | 1.35 | 59 | 59 |
| Lefty Grove | 35 | 0.67 | 52 | 54 |

Bert Blyleven checks in 4th on the all-time list. Red Ruffing, someone else we do not talk about much as an all-time great, is 5th on the list with 64 era-neutral park-adjusted shutouts. Lefty Grove is 10th all-time with 54. Grove would be higher on this and other counter-stat lists had he "reached" the majors earlier when he was ready.

## Complete Games

Table 5 displays the average frequency of complete games by decade. Over 80% of all team-games were complete games in the 1900's. Complete game frequency has steadily declined until it is currently under 10% in today's game. There are many factors leading to the dearth of the complete game (and why it was so common in the "dead ball" era). The advent of the lively ball required pitchers to bear down on every pitch to every batter. Relatedly, the increased strategic use of relief pitchers has dramatically cut down on complete games. And old-timers will tell you that they just don't make pitchers the way they used to.

Table 5: Complete Games by Decade (average frequency per team per game)

| Decade | NL League Ave | NL Era Effect | AL League Ave | AL Era Effect |
|---|---|---|---|---|
| 1900's | 81.0% | 2.08 | 80.3% | 2.06 |
| 1910's | 57.0 | 1.46 | 58.4 | 1.50 |
| 1920's | 50.1 | 1.28 | 49.7 | 1.27 |
| 1930's | 44.3 | 1.14 | 45.6 | 1.17 |
| 1940's | 41.6 | 1.07 | 44.4 | 1.14 |
| 1950's | 32.8 | 0.84 | 34.5 | 0.88 |
| 1960's | 27.2 | 0.70 | 23.4 | 0.60 |
| 1970's | 22.5 | 0.58 | 28.1 | 0.72 |
| 1980's | 13.1 | 0.34 | 18.0 | 0.46 |
| 1990's | 7.3 | 0.19 | 8.7 | 0.22 |
| | | | | |
| 1900-1998 | NL: 38.0% | | AL: 39.0% | ML: 38.7% |

Table 6 presents the all-time leaders in era-neutral complete games. Note that my research indicates that ball parks do not significantly affect the frequency of complete games (there may be a manager effect, but not a park effect).

Steve Carlton is the all-time leader in adjusted complete games. Carlton had 254 actual complete games, and pitched in an era in which complete games were only 53% as frequent as within the entire 20th century. Thus, Carlton's 254 actual complete games represent 481 era-neutral complete games. Note that this list includes all modern pitchers, no old-timers. This is undoubtedly due to the fact that modern pitchers have more "room", on a percentage basis, to improve on their league's average complete game frequency than did the old-timers.

After all, if the league completes 10% of all games, a pitcher who completes 6 of his 36 starts (16.67%) is deemed to be 66% better than league average. But if the league completes 80% of its games, and you complete all 36 of your starts, you will be deemed by this relative percentage method to be only 25% better than league average.

Table 6: All-time leaders in era-neutral Complete Games

| | Actual Complete Games | Era Effect | Era-neutral Complete Games |
|---|---|---|---|
| Steve Carlton | 254 | 0.53 | 481 |
| Warren Spahn | 382 | 0.82 | 464 |
| Gaylord Perry | 303 | 0.66 | 459 |
| Phil Niekro | 245 | 0.54 | 456 |
| Bert Blyleven | 242 | 0.55 | 437 |
| Nolan Ryan | 222 | 0.51 | 432 |
| Tom Seaver | 231 | 0.54 | 430 |
| Jack Morris | 174 | 0.41 | 425 |
| Greg Maddux | 88 | 0.21 | 409 |
| Fergie Jenkins | 267 | 0.66 | 403 |

A dramatic effect of this "percentage on a percentage" phenomenon can be seen by considering Greg Maddux. Entering the 1999 season, Maddux had all of 88 complete games in his career. Though this total seems rather small, during his career the average NL frequency of complete games has been a paltry 8.3%, or only 21.5% of the historical 20th century average frequency. Thus, Maddux's 88 actual complete games get converted to a staggering 409 era-neutral complete games.

In case you are wondering, Walter Johnson's all-time leading 531 complete games get converted to 364 era-neutral complete games, since his league complete game average frequency was 56% (46% more frequent than the historical average).

I am not sorry to see modern pitchers dominate any list, especially complete games. Indeed it could be argued that a complete game is more valuable in today's baseball for the rest it provides the over-used bullpen. However, I would be interested to hear from anyone who may have a better approach to adjusting complete games for era.

## Innings Pitched

When you look back at the old-timers' pitching stats, you see many 300+ inning seasons, with a few seasons of 400+ innings. As I stated earlier, this is mainly due to the fact that pitchers did not have to throw hard or bear down on every pitch or every batter in the dead ball era. Due to the poor quality of the ball and the controlled swings hitters employed in the era, many hitters in the lineups of the 1900's and 1910's could not really do much damage. So the pitchers would throw the equivalent of today's "batting practice fastball" to these hitters, saving their vigor for the few hitters and occasions in the game requiring it.

Okay, you say, innings pitched stats are surely affected by the era. But how can we adjust innings pitched to account for the era? After all, an inning is an inning, right? Yes and no. While I would agree that an inning is the same from the game perspective, I would argue that a pitcher's seasonal total of innings pitched can be appropriately adjusted for his era.

The method I use is to calculate the average innings pitched for the top 3 pitchers (in innings pitched) in the league. I take this to be a reflection of how easy it is to log a ton of innings. I consider the top 3 rather than just the league leader so as to not overly factor out a phenomenal individual performance.

Table 7: Innings Pitched by Decade (avg of top 3 in league)

| Decade | NL League Ave | NL Era Effect | AL League Ave | AL Era Effect |
|---|---|---|---|---|
| 1900's | 355 | 1.20 | 357 | 1.21 |
| 1910's | 324 | 1.10 | 335 | 1.13 |
| 1920's | 302 | 1.02 | 304 | 1.03 |
| 1930's | 292 | 0.99 | 291 | 0.98 |
| 1940's | 285 | 0.96 | 289 | 0.98 |
| 1950's | 289 | 0.98 | 268 | 0.91 |
| 1960's | 301 | 1.02 | 280 | 0.95 |
| 1970's | 299 | 1.01 | 317 | 1.07 |
| 1980's | 260 | 0.88 | 266 | 0.90 |
| 1990's | 239 | 0.81 | 245 | 0.83 |
| | | | | |
| 1900-1998 | NL: 295.2 | | AL: 295.0 | ML: 295.6 |

Table 7 displays the average number of innings pitched by the league's top 3 pitchers by decade. In the 1900's the top 3 pitchers logged over 350 innings per season (even when fewer games were played per season). The average has declined precipitously until today's top pitchers log less than 250 innings per season.

The table above automatically accounts for different number of games per season throughout the 20th century. Of course, today's season is 162 games. Before the 1961/62 expansion, the season was 154 games. In the first three years of the century, seasons were 140 games long. In addition, strike-shortened seasons (or lock-outs) are also automatically taken into account. Essentially, a pitcher in these strike years is credited with additional innings he would have logged in proportion to the number of innings he was able to accumulate.

In case you are interested, the top 3 in the NL and AL 1910's averaged 333 and 340 innings, respectively, excluding the war-shortened seasons of 1918-1919. Excluding 1972 has no impact on the NL 1970's average, and actually lowers the AL 1970's average. Excluding 1981 raises the average to 268 in the NL 1980's and 273 in the AL. Excluding 1994-1995 raises the average to 251 in the NL 1990's and 256 in the AL. Excluding all of these shortened seasons raises the overall 20th century average to 298.6 innings.

Table 8: All-time leaders in era-neutral Innings Pitched

| | Actual Innings Pitched | Era Effect | Era-neutral Innings Pitched |
|---|---|---|---|
| Nolan Ryan | 5386 | 0.96 | 5593 |
| Phil Niekro | 5403 | 0.97 | 5542 |
| Don Sutton | 5280 | 0.98 | 5399 |
| Walter Johnson | 5924* | 1.11 | 5356 |
| Steve Carlton | 5217 | 0.97 | 5351 |
| Warren Spahn | 5244 | 0.98 | 5346 |
| Gaylord Perry | 5351 | 1.02 | 5241 |
| Bert Blyleven | 4970 | 0.98 | 5051 |
| Early Wynn | 4564 | 0.93 | 4929 |
| Tom Seaver | 4779 | 0.97 | 4902 |

Table 8 presents the all-time leaders in era-neutral innings pitched (I do not calculate an innings pitched park effect). Nolan Ryan leads the pack, barely edging out Phil Niekro. Walter Johnson, the all-time leader in actual innings pitched, comes in fourth since in his era league leaders in innings pitched were able to log about 11% more innings than the 20th century average total.

## Strikeouts

Table 9 presents the average number of strikeouts per team per game by decade. Strikeouts were at their all-time low in the 1920's (fewer than 3 strikeouts per team per game) and have risen steadily until today's NL games average over 6 strikeouts per team. Today's games even surpass the strikeout totals owing to the high mounds and large strike zones of the 1960's.

Table 10 presents the all-time leaders in era-neutral strikeouts (I do not calculate a park effect for strikeouts). Nolan Ryan, the all-time SO king, nips Walter Johnson in era-neutral strikeouts. Johnson compiled over 3500 strikeouts in an era in which batters were truly embarrassed by striking out, and did not take the cuts of today's sluggers (not to mention today's utility infielders). Dazzy Vance is another era's strikeout king who appears high on the list of era-neutral strikeouts. Dazzy did not win his first game in the majors until age 31, and then proceeded to lead the NL in strikeouts for 7 consecutive seasons in the 1920's.

Table 9: Strikeouts by Decade (average SO per team per game)

| Decade | NL League Ave | NL Era Effect | AL League Ave | AL Era Effect |
|---|---|---|---|---|
| 1900's | 3.42 | 0.78 | 3.60 | 0.82 |
| 1910's | 3.60 | 0.82 | 3.79 | 0.87 |
| 1920's | 2.78 | 0.63 | 2.85 | 0.65 |
| 1930's | 3.31 | 0.75 | 3.35 | 0.77 |
| 1940's | 3.49 | 0.80 | 3.63 | 0.83 |
| 1950's | 4.49 | 1.03 | 4.34 | 0.99 |
| 1960's | 5.75 | 1.31 | 5.66 | 1.29 |
| 1970's | 5.31 | 1.21 | 5.22 | 1.19 |
| 1980's | 5.54 | 1.26 | 5.64 | 1.29 |
| 1990's | 6.28 | 1.43 | 5.94 | 1.36 |
| | | | | |
| 1900-1998 | NL: 4.38 | | AL: 4.40 | ML: 4.38 |

Table 10: All-time leaders in era-neutral Strikeouts

| | Actual Strikeouts | League SO Avg | Era Effect | Era-neutral Strikeouts |
|---|---|---|---|---|
| Nolan Ryan | 5714* | 5.44 | 1.24 | 4603 |
| Walter Johnson | 3508 | 3.43 | 0.78 | 4482 |
| Steve Carlton | 4136 | 5.43 | 1.24 | 3338 |
| Dazzy Vance | 2045 | 2.86 | 0.65 | 3133 |
| Lefty Grove | 2266 | 3.23 | 0.74 | 3074 |
| Christy Mathewson | 2502 | 3.63 | 0.83 | 3020 |
| Bob Feller | 2581 | 3.77 | 0.86 | 3000 |
| Bert Blyleven | 3701 | 5.41 | 1.23 | 2998 |
| Tom Seaver | 3640 | 5.46 | 1.25 | 2921 |
| Gaylord Perry | 3534 | 5.44 | 1.24 | 2847 |

Table 11: Walks by Decade (average BB per team per game)

| Decade | NL League Ave | NL Era Effect | AL League Ave | AL Era Effect |
|---|---|---|---|---|
| 1900's | 2.66 | 0.83 | 2.38 | 0.74 |
| 1910's | 2.82 | 0.88 | 3.13 | 0.98 |
| 1920's | 2.78 | 0.87 | 3.28 | 1.02 |
| 1930's | 2.83 | 0.88 | 3.71 | 1.16 |
| 1940's | 3.40 | 1.06 | 3.74 | 1.17 |
| 1950's | 3.40 | 1.06 | 3.77 | 1.18 |
| 1960's | 2.97 | 0.93 | 3.34 | 1.04 |
| 1970's | 3.32 | 1.04 | 3.29 | 1.03 |
| 1980's | 3.20 | 1.00 | 3.24 | 1.01 |
| 1990's | 3.27 | 1.02 | 3.54 | 1.11 |
| | | | | |
| 1900-1998 | NL: 3.06 | | AL: 3.35 | ML: 3.20 |

## Walks

Table 11 presents the average number of walks per team per game by decade.

Table 12 presents the all-time leaders in era-neutral walks (I do not calculate a park effect for walks). Nolan Ryan is far and away the all-time leader in both actual walks and era-neutral walks.

A potentially interesting study could analyze the co-movements between walks, strikeouts, hits, home runs, steals, etc., in order to better understand the game's various eras.

Table 12: All-time leaders in era-neutral Walks

| | Actual Walks | League BB Avg | Era Effect | Era-neutral Walks |
|---|---|---|---|---|
| Nolan Ryan | 2795* | 3.27 | 1.02 | 2737 |
| Steve Carlton | 1833 | 3.23 | 1.01 | 1817 |
| Phil Niekro | 1809 | 3.21 | 1.00 | 1805 |
| Charlie Hough | 1665 | 3.27 | 1.02 | 1631 |
| George Mullin | 1238 | 2.60 | 0.81 | 1525 |
| Burleigh Grimes | 1295 | 2.72 | 0.85 | 1525 |
| Early Wynn | 1775 | 3.75 | 1.17 | 1516 |
| Bobo Newsom | 1732 | 3.70 | 1.16 | 1499 |
| Walter Johnson | 1405 | 3.13 | 0.98 | 1437 |
| Bob Feller | 1764 | 3.93 | 1.23 | 1437 |

## Strikeout to Walk Ratio

Table 13 presents the average strikeout to walk ratio by decade. The strikeout to walk ratio reached its nadir in the 1920's (dipping below 1.0) and has risen since. The apex was reached in the 1960's, but today's baseball experiences just about the same lofty ratio as the 1960's (only slightly below 2.0 in the NL).

Table 13: Strikeout to Walk Ratio by Decade

| Decade | NL League Ave | NL Era Effect | AL League Ave | AL Era Effect |
|---|---|---|---|---|
| 1900's | 1.29 | 0.93 | 1.51 | 1.09 |
| 1910's | 1.28 | 0.92 | 1.21 | 0.88 |
| 1920's | 1.00 | 0.72 | 0.87 | 0.63 |
| 1930's | 1.17 | 0.85 | 0.91 | 0.66 |
| 1940's | 1.03 | 0.74 | 0.97 | 0.70 |
| 1950's | 1.32 | 0.96 | 1.15 | 0.83 |
| 1960's | 1.94 | 1.40 | 1.70 | 1.23 |
| 1970's | 1.60 | 1.16 | 1.59 | 1.15 |
| 1980's | 1.73 | 1.25 | 1.74 | 1.26 |
| 1990's | 1.92 | 1.39 | 1.68 | 1.21 |
|  |  |  |  |  |
| 1900-1998 | NL: 1.42 |  | AL: 1.33 | ML: 1.38 |

Table 14 presents the all-time leaders in era-neutral strikeout to walk ratio (since I do not calculate a park effect for strikeouts or walks, there is none for their ratio). Cy Young heads his first list. Remember that I include only 20th century stats so Cy is at somewhat of a disadvantage when it comes to counter stats. But when it comes to strikeout to walk ratio, Ole Cy tops both actual and era-neutral lists.

Bret Saberhagen, possessor of the 2nd highest strikeout to walk ratio among 20th century pitchers with over 2000 innings, checks in with the 8th best era-neutral ratio.

Although Table 14 does not present the data in that form, a pitcher's era-neutral SO/BB ratio is equal to the ratio of his era-neutral strikeouts divided by his era-neutral walks.

Table 14: All-time leaders in era-neutral Strikeout to Walk Ratio (min 2000 IP)

|  | Actual SO/BB | League SO/BB | Era Effect | Era-neutral SO/BB |
|---|---|---|---|---|
| Cy Young | 3.75* | 1.44 | 1.04 | 3.60 |
| Dazzy Vance | 2.43 | 1.00 | 0.73 | 3.35 |
| Christy Mathewson | 2.96 | 1.28 | 0.93 | 3.19 |
| Walter Johnson | 2.50 | 1.10 | 0.79 | 3.15 |
| Deacon Phillippe | 2.88 | 1.30 | 0.94 | 3.06 |
| Lefty Grove | 1.91 | 0.90 | 0.65 | 2.94 |
| Carl Hubbell | 2.32 | 1.12 | 0.81 | 2.87 |
| Bret Saberhagen | 3.53 | 1.80 | 1.30 | 2.71 |
| Dennis Eckersley | 3.25 | 1.66 | 1.20 | 2.71 |
| Rube Waddell | 2.90 | 1.50 | 1.09 | 2.66 |

## Baserunners per Nine Innings Pitched

Table 15 presents the average number of baserunners (hits plus walks) in every nine innings pitched by decade. Although you can discern the historical trends of the dead-ball era and hitting frenzy of the 1920's and 1930's, the effects are much more muted than other stats we have reviewed. One reason for this is that both hitters and pitchers adjust to the environment. In the dead-ball era, for example, hitters took many pitches, choked up and tried to "hit 'em where they ain't", thereby keeping the number of baserunners up. In today's home run happy era, hitters, not surprisingly, know that one swing of the bat can plate a run. Thus, they swing from the heels and their batting average (and on base percentage) suffers. Baseball is a game of equilibration. When one part of offense is curtailed, another jumps up to partially compensate, and vice versa.

Table 15: Baserunners per Nine Innings Pitched by Decade (hits plus walks/(IP/9))

| Decade | NL League Ave | NL Era Effect | AL League Ave | AL Era Effect |
|---|---|---|---|---|
| 1900's | 11.3 | 0.93 | 10.9 | 0.90 |
| 1910's | 11.3 | 0.93 | 11.5 | 0.94 |
| 1920's | 12.7 | 1.04 | 13.2 | 1.08 |
| 1930's | 12.6 | 1.04 | 13.7 | 1.12 |
| 1940's | 12.4 | 1.02 | 12.7 | 1.04 |
| 1950's | 12.4 | 1.02 | 12.6 | 1.04 |
| 1960's | 11.6 | 0.95 | 11.6 | 0.95 |
| 1970's | 12.0 | 0.99 | 12.0 | 0.99 |
| 1980's | 11.8 | 0.97 | 12.3 | 1.01 |
| 1990's | 12.2 | 1.00 | 12.8 | 1.05 |
| | | | | |
| 1900-1998 | NL: 12.0 | | AL: 12.4 | ML: 12.2 |

Table 16: All-time leaders in era-neutral park-adjusted Baserunners per 9 IP (min 2000 IP)

| | Actual BR/9IP | Era Effect | Era-neutral BR/9IP | Era-neutral Park-adj. BR/9IP |
|---|---|---|---|---|
| Roger Clemens | 10.29 | 1.04 | 9.91 | 9.74 |
| Greg Maddux | 9.97 | 0.99 | 10.07 | 9.78 |
| Walter Johnson | 9.62 | 0.99 | 9.75 | 9.81 |
| Cy Young | 9.13 | 0.91 | 10.02 | 9.97 |
| Addie Joss | 8.73* | 0.88 | 9.92 | 9.99 |
| Bret Saberhagen | 10.23 | 1.02 | 10.05 | 10.04 |
| Babe Adams | 9.83 | 0.97 | 10.18 | 10.18 |
| Ed Walsh | 9.00 | 0.90 | 10.05 | 10.21 |
| Carl Hubbell | 10.50 | 1.04 | 10.14 | 10.21 |
| Grover Alexander | 10.10 | 0.98 | 10.27 | 10.21 |

Table 16 presents the all-time leaders in era-neutral park-adjusted baserunners per nine innings pitched. The park adjustment is similar to that done for shutouts. I have estimated that the effect a park has on the number of baserunners is roughly half its impact on the number of runs scored. For example, a park that depresses runs by 10% depresses baserunners by about 5%.

Roger Clemens and Greg Maddux appear at the top of another list. These two active pitchers get the job done in very different ways, yet both undoubtedly get the job done well. Clemens in particular has had both his era and ballparks working against him. When account is taken of the Rocket's environment, the true splendor of his career shines through.

## Winning Percentage

You might be wondering how a pitcher's era can possibly affect his winning percentage. After all, in every game there is a winning team and a losing team. The whole league plays .500 ball each and every season. You are right; there is no era effect here. The same reasoning implies that there is no park effect either. Both opposing teams, of course, play in the same ballpark.

I want to include a short discussion on winning percentage here since the stat is often used in comparing and evaluating pitchers. I have developed a method, somewhat different from other published methods, to evaluate a pitcher's winning percentage.

A pitcher's winning percentage must be considered in the context of his team. Steve Carlton's 27-10 season for the 1972 Phillies is deemed one of the best seasons ever largely because the Phillies were 59-97 for the season – or only 32-87 in games in which Carlton did not get a decision. But how to properly take into account the team?

I start with two premises. First, as a rough rule of thumb, over the course of the career of a typical starting pitcher, his team's winning percentage is made up of 1/6 his own winning percentage and 5/6 his pitching teammates' winning percentage. In a 162 game season, this split assumes that a typical starting pitcher has about 27 decisions.

Second, I need a "standard" against which to compare the pitcher's winning percentage. Again, as a rough rule of thumb, I use 2/3 of his pitching teammates' winning percentage plus 1/3 of .500. So if a pitcher pitches on a weak ballclub, his standard will be below .500. If he pitches on a club that regularly wins 90 games (a .556 winning pct), his standard will be over .500 (.537 in this case).

Working through the algebra, the difference between a pitcher's own winning percentage and this standard becomes equal to 1.133*Own – 0.80*Team – 166.67. You don't have to understand this formula itself; all you have to do is understand the two premises underlying the formula.

Table 17 presents the all-time leaders in winning percentage relative to team standard, as described above. Entering the 1999 season, Randy Johnson is the all-time leader. Randy currently has a 644 win percentage, his teams over the course of his career have a 492 win percentage, implying that his teammates (using the 1/6 rule) have had a 462 win percentage, making his team standard (using the 2/3 rule) a 474 win percentage. 644 minus 474 is 170, the highest among all 20th century pitchers with 100 or more wins.

Table 17: All-time leaders in Winning Percentage relative to team standard (min 100 wins)

| | Actual Win Pct | Team Win Pct | Teammates Win Pct | Standard Win Pct | Win Pct Differential relative to Std |
|---|---|---|---|---|---|
| Randy Johnson | 644 | 492 | 462 | 474 | 170 |
| Roger Clemens | 653 | 518 | 491 | 494 | 159 |
| Mike Mussina | 667 | 539 | 513 | 509 | 158 |
| Spud Chandler | 717* | 625 | 606 | 571 | 146 |
| Lefty Grove | 680 | 583 | 564 | 542 | 138 |
| Grover Alexander | 642 | 533 | 511 | 507 | 135 |
| Sandy Koufax | 655 | 561 | 542 | 528 | 127 |
| Whitey Ford | 690 | 611 | 596 | 564 | 126 |
| Don Gullet | 686 | 605 | 589 | 559 | 126 |
| Dwight Gooden | 642 | 546 | 527 | 518 | 125 |

Following Randy Johnson are Roger Clemens and Mike Mussina, somewhat of a surprise. Spud Chandler, the all-time leader in actual win percentage, is fourth on the adjusted list.

## Putting it all together

Now that we have generated just about every type of list needed for cross-era comparisons, where does that leave us? I have previously published the formula I use to evaluate pitchers' career values. Essentially, it is a weighting of all the variables described above, plus a few miscellaneous variables (saves, post-season performance, missing seasons due to war, major injury or other factors, and devaluing performance during WWII and the FL).

To be sure the two variables that I consider most important are era-neutral park-adjusted ERA and era-neutral innings pitched. I also give lesser weights to winning percentage relative to team standard, era-neutral strikeouts, era-neutral park-adjusted baserunners per nine innings, era-neutral strikeout to walk ratio, era-neutral park-adjusted shutouts, and era-neutral complete games.

A pitcher's job is to help his team win games. So why don't we just look at his won-loss record? Clearly the offensive support a pitcher receives plays a large role in his W-L. So we look at ERA, abstracting from the team's offense. However, the same reasoning can be applied to ERA. After all, the defensive support a pitcher receives can play a large role in his ERA.

Don't get me wrong. I am an ardent believer in ERA. But I don't believe that it tells the whole story. This is why I and others like to also consider other pitching stats such as strikeouts, complete games, shutouts, baserunners, etc. Only by considering the whole panorama of a pitcher's statistics do I think that we can properly evaluate his true quality.

Table 18 presents my list of top 20 pitchers of all time, according to career value entering the 1999 season. Many of these greats will have appeared on the previous lists.

### Table 18: My All-Time Greatest Pitchers (20th century career value)

| | Era-neutral Park-adjusted ERA | Era-neutral Park-adjusted SHO | Era-neutral CG | Era-neutral IP | Era-neutral K's | Era-neutral SO/BB ratio | Era-neutral Park-adjusted BR/9IP | Win Pct Diff relative to Std |
|---|---|---|---|---|---|---|---|---|
| Johnson | 2.52 | 95 | 364 | 5356 | 4482 | 3.15 | 9.81 | 118 |
| Grove | 2.52 | 54 | 251 | 4009 | 3074 | 2.94 | 10.29 | 138 |
| Alexander | 2.77 | 86 | 320 | 4845 | 2818 | 2.59 | 10.21 | 135 |
| Clemens (a) | 2.46 | 69 | 394 | 3795 | 2362 | 2.53 | 9.74 | 159 |
| Spahn | 3.11 | 60 | 464 | 5346 | 2413 | 1.76 | 10.90 | 76 |
| Feller | 3.01 | 49 | 260 | 3954 | 3000 | 2.11 | 11.10 | 79 |
| Seaver | 2.90 | 59 | 430 | 4902 | 2921 | 2.13 | 10.30 | 115 |
| Mathewson | 2.73 | 59 | 239 | 4098 | 3020 | 3.19 | 10.26 | 120 |
| Maddux (a) | 2.46 | 41 | 409 | 3471 | 1440 | 2.26 | 9.78 | 104 |
| Ford | 2.76 | 43 | 214 | 3495 | 1718 | 1.77 | 11.08 | 126 |
| Hubbell | 2.82 | 39 | 225 | 3636 | 2235 | 2.87 | 10.21 | 96 |
| Blyleven | 3.19 | 67 | 437 | 5051 | 2998 | 2.33 | 10.82 | 33 |
| Ryan | 3.27 | 60 | 432 | 5593 | 4603 | 1.70 | 11.55 | 25 |
| Carlton | 3.22 | 54 | 481 | 5351 | 3338 | 1.85 | 11.43 | 57 |
| G. Perry | 3.16 | 49 | 459 | 5241 | 2847 | 2.03 | 10.93 | 41 |
| Eckersley | 3.26 | 25 | 242 | 3674 | 1909 | 2.71 | 10.26 | 16 |
| Palmer | 2.93 | 50 | 313 | 3882 | 1822 | 1.43 | 10.91 | 79 |
| P. Niekro | 3.27 | 46 | 456 | 5542 | 2672 | 1.50 | 11.50 | 47 |
| Roberts | 3.28 | 47 | 384 | 4879 | 2130 | 2.49 | 10.53 | 58 |
| Koufax | 2.81 | 38 | 190 | 2352 | 1885 | 2.21 | 10.31 | 127 |

(a) denotes active

According to my methods, Walter Johnson and Lefty Grove are very closely grouped and stand apart from the others at the top rung. Currently there is another gap between Clemens and the clustered grouping of Spahn, Feller, Seaver, and Mathewson. Greg Maddux will likely join this group in the near future. Bert Blyleven is a strange case: he is an all-time great according to my methods, yet will likely struggle in his quest to make the Hall of Fame. The fact that Dennis Eckersley appears as the 16th greatest pitcher of all time may imply that I have not properly "sorted" starters and relievers.

## Conclusion

Baseball statistics must be considered in the context in which they were achieved. In this article I have presented data and techniques designed to take into account the era and ballpark in which a pitcher toiled. By using such methods to convert statistics into an era-neutral (park-adjusted) varietal, pitchers from disparate eras can be compared on an even footing.

Every decade of the century is represented on my list of all-time greats. Roger Clemens and Greg Maddux, active stars, compare quite favorably to their historical counterparts. Both currently have cracked the Top 10, and may move up the ladder over the remainder of their careers.

When it comes to cross-era comparisons, some people have argued that it was easier for the old-time stars such as Walter Johnson, Ty Cobb, Honus Wagner, et al., to stand out from their contemporaries than for modern stars. Early baseball was less standardized than today's version, including such factors as scouting, coaching, training, strategies, and management. In addition, it has been argued that the population from which old time stars sprung was "sparser" than today's population, especially taking into account the prohibition against blacks and lesser acceptance of Hispanics and other minorities.

To the extent that there is some truth to these arguments, the relativistic era adjustments I employ will allow the early stars to appear greater than the modern stars. On the other hand, there may be an over-representation of modern stars (in sheer numbers) on the all-time top 20 list. That is, the few old time stars get the relative advantage of being able to stand apart from their era, while the modern stars get the relative advantage of being more prevalent and having longer careers.

Era effects, no matter how sophisticated, can never settle all of these debates, nor would we want them to. Instead my limited hope is that these methods can contribute to baseball fans' and researchers' discussions of the all-time greats by making cross-era statistical comparisons readily accessible.

*Rob Wood is a management consultant in Mountain View, California. He can be reached at 2101 California St. #224, Mountain View, CA, 94040-1686, rob.wood@us.pwcglobal.com.* ♦

J

The above letter is Thomas J. Hanrahan's middle initial. An incorrect initial appeared in the Februrary, 1999 issue of BTN (and then migrated to the SABR Bulletin). I have apologized to Tom, and agreed to avoid this problem in future by just calling him "Tom".

Sorry, Tom. In other errors last issue:

The graph on page 35 did not distinguish between pitchers and hitters. The top line is the pitchers, and the bottom represents the hitters.

In the article by the above-mentioned Tom Hanrahan, the charts on page 30 and 31 are labeled incorrectly. The regression shown as "Chart 1" is actually "Chart 2", and vice-versa.

# Bias in Run Statistics

Phil Birnbaum

*While run statistics like Runs Created and Linear Weights are accurate for teams, they are biased when used for strong and weak player offenses. Here, the author shows where the biases lie, and offers a new corrected statistic for use when accuracy is key.*

Go to the bank for a mortgage and they'll tell you to plan on spending about a third of your income in payments. If you earn $30,000 a year, you should wind up paying about $10,000. Earn $45,000, and you can spend approximately $15,000. And so on.

But that's just a rule of thumb, not an empirically proven formula. The one-third rule breaks down at the ends of the salary scale. Earn $3,000 a year and you won't be buying any real estate, even one with a $1,000 mortgage payment. And, if you're Bill Gates with an annual income of, say, $120 million, you probably won't be paying the bank forty million dollars a year on your half-billion dollar estate. For one thing, you probably won't need a loan, and, for another, there aren't any half-billion dollar houses unless you plan to live in SkyDome or something.

### Table 1: 1984 American League Runs Created

| | Predicted | Actual | Difference |
|---|---|---|---|
| Detroit | 824 | 829 | -5 |
| Toronto | 793 | 750 | 43 |
| New York | 771 | 758 | 13 |
| Boston | 847 | 810 | 37 |
| Baltimore | 697 | 681 | 16 |
| Cleveland | 724 | 761 | -37 |
| Milwaukee | 633 | 641 | -8 |
| Kansas City | 697 | 673 | 24 |
| California | 660 | 696 | -36 |
| Minnesota | 677 | 673 | 4 |
| Oakland | 727 | 738 | -11 |
| Chicago | 680 | 679 | 1 |
| Seattle | 683 | 682 | 1 |
| Texas | 652 | 656 | -4 |
| TOTAL | 10065 | 10027 | -38 |

Which brings us to the topic of run predictor statistics — Runs Created, for instance. Given an offensive line, Runs Created will predict how many runs resulted. And, for middle-class batting lines, it's been well established that it does predict runs quite well. Every year in his *Baseball Abstract*, Bill James would compare teams' actual runs scored to Runs Created.

Take, for instance, the 1984 American League (Table 1). Runs Created correctly picked the worst and best teams, and no team was off by more than about six percent. Five teams, in fact, were within a single percentage point, and the league as a whole was off by only 38 runs out of 10,027.

But here's what I mean by middle class. The best offensive team, Detroit, scored 5.1 runs per game; the worst, Milwaukee, managed "only" 4.0. For teams, those are fairly extreme totals, but for players, they're middle of the road.

Table 2 lists the batting lines of all fourteen teams, normalized to 500 at-bats. If these were fourteen players, they'd be average, really, really average — they're all the typical center fielder having the typical year. Even though they run the gamut of *team* offense, if they were the career lines of a single player, you'd be amazed by his consistency. What you've got here are fourteen of your typical middle-class $35,000 to $37,000 per year mortgage holders. Probably 90% of all players fall outside this range.

### Table 2: Normalized Team Stats, 1984 American League

| | AB | R | H | 2B | 3B | HR | BB | avg |
|---|---|---|---|---|---|---|---|---|
| Baltimore | 500 | 62 | 126 | 21 | 2 | 15 | 57 | .252 |
| Boston | 500 | 72 | 141 | 23 | 4 | 16 | 44 | .283 |
| California | 500 | 64 | 125 | 19 | 3 | 14 | 51 | .249 |
| Chicago | 500 | 62 | 123 | 20 | 3 | 16 | 47 | .247 |
| Cleveland | 500 | 67 | 133 | 20 | 3 | 11 | 53 | .265 |
| Detroit | 500 | 73 | 135 | 23 | 4 | 17 | 53 | .271 |
| Kansas City | 500 | 61 | 134 | 24 | 5 | 11 | 36 | .268 |
| Milwaukee | 500 | 58 | 131 | 21 | 3 | 9 | 39 | .262 |
| Minnesota | 500 | 60 | 132 | 23 | 3 | 10 | 39 | .265 |
| New York | 500 | 67 | 138 | 24 | 3 | 11 | 47 | .276 |
| Oakland | 500 | 68 | 130 | 24 | 3 | 14 | 52 | .259 |
| Seattle | 500 | 61 | 129 | 22 | 3 | 12 | 47 | .258 |
| Texas | 500 | 59 | 130 | 20 | 3 | 11 | 38 | .261 |
| Toronto | 500 | 66 | 137 | 23 | 4 | 17 | 53 | .271 |

And that's the problem. In real life, Runs Created isn't used just to predict runs scored for batting lines like this, for middle-of-the-road players and teams. Where it's used most — in fact, where it's the most interesting — is on the extreme players, the Mark McGwires and the Mario Mendozas. How do we know it works for those guys? We don't. Here, take a look at Barry Bonds' 1993:

| | AB | R | H | 2B | 3B | HR | RBI | BB | K | avg |
|---|---|---|---|---|---|---|---|---|---|---|
| 1993 Bonds | 539 | 129 | 181 | 38 | 4 | 46 | 123 | 126 | 29 | .336 |

The basic Runs Created formula (using 25.5 outs per game) tells us that a team full of Bondses would score 12 runs per game. But would they? Maybe Runs Created is like "Thirty percent of income" — it works for typical cases, but not for extreme cases. Maybe Barry Bonds is like Bill Gates. Maybe Runs Created doesn't work that high up the scale.

How can we find out? There's only one way: clone nine copies of Barry Bonds, make them play baseball, and see how many runs they score. Using my computer simulation (see *By the Numbers, 5.3, Sept. 1993)*, I ran 162 games worth of Team Bonds. Adjusted to 539 at-bats, they performed almost identially to Bonds' actual stats:

| | AB | R | H | 2B | 3B | HR | RBI | BB | K | avg |
|---|---|---|---|---|---|---|---|---|---|---|
| 1993 Bonds | 539 | 129 | 181 | 38 | 4 | 46 | 123 | 126 | 29 | .336 |
| Simulation | 539 | | 183 | 39 | 5 | 45 | | 127 | 29 | .340 |

The ersatz Bonds had 29 steals and was caught 11 times, as opposed to 29/12 in real life.

According to the stolen base version of Runs Created, Team Bonds should have scored 11.6 runs per game; they actually produced only 10.8. Expressed in total runs, in the same chart form as we used for the league:

| | Predicted | Actual | Difference |
|---|---|---|---|
| Team Bonds '93 | 1886 | 1756 | 130 |

That's a huge difference, 130 runs, considering the highest middle-class difference ('84 Blue Jays) was 43 and the mean square difference was only 23. Even after taking into account Bonds's huge totals, it's still significant. 130 runs is 7.4 percent of the total; the 1985 Jays' 43 run difference is only 5.7 percent. So far, Runs Created doesn't look like it works nearly as well for great teams as it does for average teams.[1] (Bonds, on the other hand, still comes out looking OK; against American League opponents scoring 4.5 runs per game, Team Bonds would go 137-25.)

Of course, the one Bonds' season doesn't prove anything about Runs Created. Anything can happen in one season. The Bondses might have just gotten unlucky, and spread their hits out too much, or something. We need more than one season, and preferably more than one batting line, to legitimately conclude anything at all.

To give Runs Created a full test, I took each player in the 1988 American League with at least 10 at-bats, and ran five Bonds tests for each player — that is, five 162-game seasons. Then, I added a random selection of all-time great seasons, to add some bulk to the high-end data. In total, I ran 1,817 seasons of data.[2]

Then, I split the seasons up by "income group" based on on-base percentage plus slugging (OPS), a rough measure of offensive performance. For each group of teams, I compared their "actual" (simulated) runs scored to their Runs Created estimate.

Take a look at Table 3 for the results.

The AL average OPS is about .725, and the middle-class teams clustered around that average hit their estimates almost dead on. But as you go down to the poorer players, those creating only one or two runs per game, the discrepancy increases a bit, as Runs Created is a bit too pessimistic. And, at the very rich end of the scale, Runs Created overcompensates, consistently predicting lots more runs than are actually scored. It seems that the Bonds overprediction was due to a real Runs Created shortcoming, and not just luck. In fact, Runs Created was *more* accurate for Bonds than this data suggests it should have been. Bonds' 1993 OPS was 1.13. At that level, the estimate should have been high by more than a run; it was off by only eight-tenths of a run.

[1] Of course, it's quite possible that Runs Created is perfect, and that the *simulation* is wrong. I don't think that's the case, for a couple of reasons. First, when run on the league's aggregate stats, the simulation predicts actual runs perfectly. And second, the result that Runs Created overpredicts for potent offenses is one that Bill James acknowledges, as we will see in a bit.

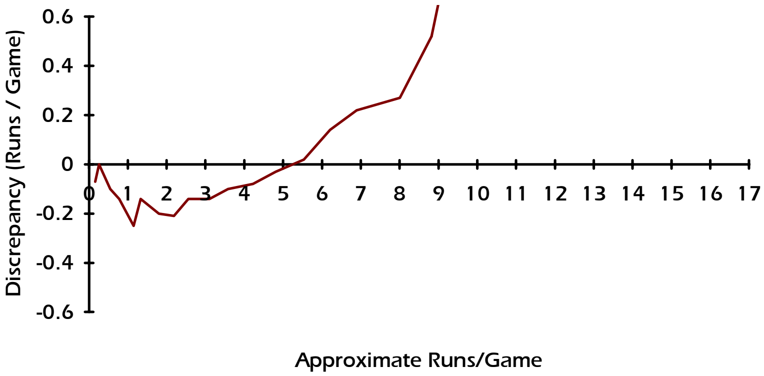
[2] For some reason, I lost three seasons of Ted Williams off the end of the file, which is why there are 1,817 seasons instead of 1,820. That omission shouldn't affect the conclusions of the study.

On an absolute scale, the rich teams are off by a lot more than the poor teams — a run or more, versus a tenth of a run. But on a percentage scale, the figures are pretty close. The 1.050 group, which scored about 10 runs per game, is off by about 9 percent, but the 0.450 group, which scored only 1.7 runs per game, is still off by almost 12 percent. Which error is more serious depends on what kind of research you're doing. For now, I'll deal with absolute errors only, but keep in mind that if you're dealing with low-scoring teams, the smaller discrepancies are no less significant.

Because it's hard to get a feel for the trend from just a list of numbers, I'll put the results up in a graph. What I'll do is this: I'll put OPS — the level of income, or offense, on the horizonal axis, and I'll put the run discrepancy on the vertical. If Runs Created were perfect, the line would be perfectly horizontal down the middle. But it's not, and so we get this:

Table 3: Runs Created accuracy for various levels of offense

| OPS (rounded to nearest .050) | # of teams | Predicted Runs | Actual Runs | Difference |
|---|---|---|---|---|
| 0.100 | 14 | 0.04 | 0.11 | -0.07 |
| 0.150 | 12 | 0.10 | 0.16 | -0.07 |
| 0.200 | 19 | 0.26 | 0.26 | -0.00 |
| 0.250 | 10 | 0.45 | 0.55 | -0.10 |
| 0.300 | 14 | 0.63 | 0.78 | -0.14 |
| 0.350 | 45 | 0.90 | 1.15 | -0.25 |
| 0.400 | 33 | 1.19 | 1.33 | -0.14 |
| 0.450 | 42 | 1.60 | 1.80 | -0.20 |
| 0.500 | 89 | 1.98 | 2.19 | -0.21 |
| 0.550 | 114 | 2.42 | 2.56 | -0.14 |
| 0.600 | 195 | 2.96 | 3.10 | -0.14 |
| 0.650 | 183 | 3.48 | 3.58 | -0.10 |
| 0.700 | 211 | 4.15 | 4.22 | -0.08 |
| 0.750 | 194 | 4.78 | 4.81 | -0.03 |
| 0.800 | 115 | 5.55 | 5.53 | 0.02 |
| 0.850 | 77 | 6.35 | 6.21 | 0.14 |
| 0.900 | 78 | 7.12 | 6.89 | 0.22 |
| 0.950 | 79 | 8.28 | 8.01 | 0.27 |
| 1.000 | 58 | 9.34 | 8.82 | 0.52 |
| 1.050 | 75 | 10.21 | 9.32 | 0.90 |
| 1.100 | 75 | 11.64 | 10.48 | 1.16 |
| 1.150 | 45 | 12.46 | 11.19 | 1.27 |
| 1.200 | 17 | 14.17 | 12.84 | 1.33 |
| 1.250 | 11 | 14.88 | 13.42 | 1.46 |
| 1.300 | 5 | 17.07 | 16.15 | 0.91 |
| 1.350 | 3 | 18.48 | 17.34 | 1.14 |
| 1.400 | 3 | 18.54 | 16.34 | 2.20 |
| 1.450 | 1 | 19.86 | 16.96 | 2.91 |



I've used "predicted runs" as the offensive variable instead of OPS, because it's easier to understand. Seven runs per game means something to you right away; OPS of .900 sends you running to the chart to see if it's any good. (It's about the same thing.)

So what does all this mean? Given that Runs Created isn't perfect, is it close enough? Is 0.2 runs per game, or 0.8 runs per game, so large an error that we should be worrying about it?

I think the answer is yes. A fifth of a run per game is 30 runs over a season, or three wins; even a tenth of a run is almost two wins. Two wins isn't insignificant — replace an average player on your team with a star and you only gain two wins. And no study I've seen has ever proven a manager is worth even as much as two wins per year. If you're running a study to see whether a manager's strategy gained his team runs, and the teams you happen to be using scored four runs per game, well, your conclusions are going to be biased about 16 runs in the manager's favor.

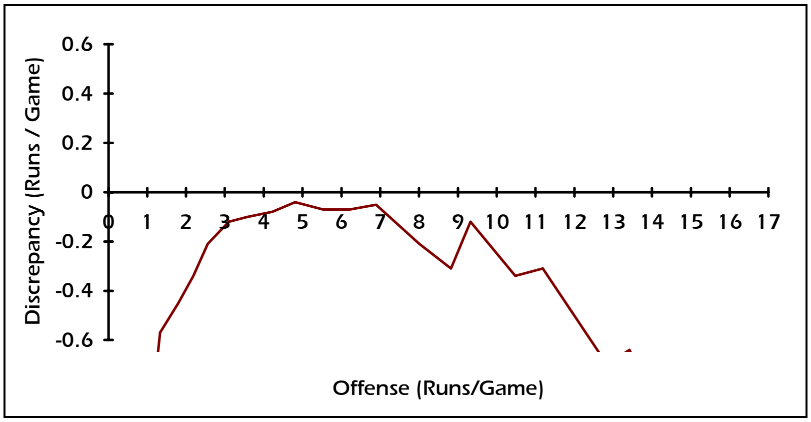
And, of course, at the top and bottom ends of the scale, the errors are much, much worse. Superstars creating 10 runs per game, according to Runs Created, are actually creating only 9, an error of 11 percent. And dweebs below the Mendoza line are being *underestimated* by that same 11 percent or more. For offenses outside the middle class, we need to find another statistic.

## Estimated Runs Produced

Bill James was aware of some of the shortcomings of Runs Created, and he's acknowledged that it's too generous for powerful offenses. In the 1985 *Baseball Abstract*, James introduced Estimated Runs Produced, an alternative statistic invented by Paul Johnson, calling it "more accurate than runs created for [high-offensive] types of players."

ERP, as I'll call it, is still occasionally used as an alternative to Runs Created in some Sabermetric studies. Bill James no longer uses it, though. In a comment he wrote to me in 1988, he reported that the apparent accuracy of ERP didn't hold up when he ran further tests.

But James may have been right the first time — ERP is at least as accurate as Runs Created. Instead of the chart, I'll just run the graph:



It's not perfect, but ERP does seem to keep closer to the center line for more of the offensive range. Taking a bias of 0.1 runs per game, or 16 runs per season, as a reasonable cutoff, you can use ERP between roughly 3 and 7 runs per game. Runs Created, by this standard, is good only betwen 4 and 6.

But ERP drops off drastically for poor offenses. For .083 hitters and the like, stick to Runs Created. For bad hitters, ERP drastically underpredicts; in fact, it often predicts *negative* runs. Runs Created, on the other hand, is never off by more than a quarter of a run for those types of players.

Where ERP shines most over Runs Created is for superstar offenses. As we saw, the Runs Created error takes off beyond 1.0 after about 9 runs per game; ERP is good within half a run all the way to 12 runs per game. It's interesting that ERP and Runs Created are off in *opposite* directions — the former predicts too low, the latter too high. It would seem like averaging the two might give the best result, since the errors partially cancel each other out, but the Runs Created error is so large that you're better off sticking with just ERP.

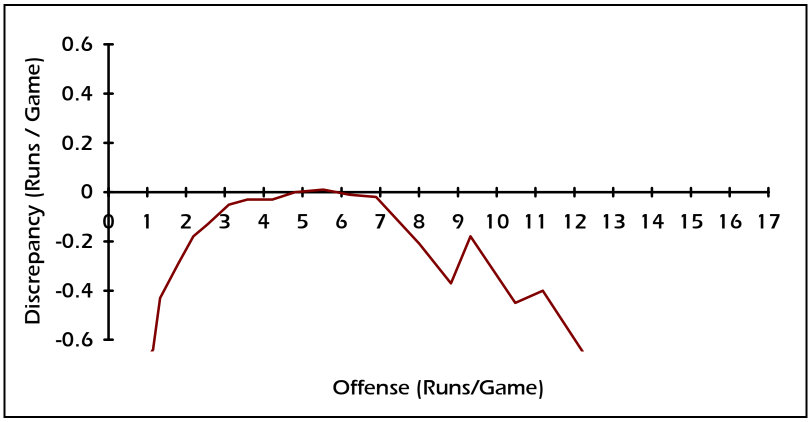
## Linear Weights

We'll see in a bit that Linear Weights, created by Pete Palmer, is the most accurate of the three major statistics, beating out both Runs Created and ERP. But Linear Weights works a little differently from those other statistics, so we'll need to deal with a couple of the technical aspects of the formula before we go on.

First, Linear Weights is denominated in Runs Above Average, not Runs. Runs Created might tell us that a player created 100 runs in a certain season — Linear Weights, on the other hand, would say that the player contributed (for example) 35 runs above an average player, or "+35." To properly compare Linear Weights to the other two statistics, we need Runs. To get Runs from Runs Above Average, we can just add Average Runs. I added "Outs * League average runs per out" to the formula to convert it to runs.

Second, Palmer adjusts the value of the out every year to get the league Linear Weight to zero. This practice has been criticized (with some justification, in my view) on the grounds that any formula can be made accurate if you're allowed to continually adjust it towards the result you're trying to predict. For this study, I fixed the value of the out for the 1992 American League, and used that value throughout. Since the study was based on 1988 players, there shouldn't be a problem.

So, after all that, here's the Linear Weights accuracy graph:



If this graph looks familiar, it's because we've seen it before — it's almost identical to the one for Estimated Runs Produced! And in fact, that's because the (adjusted) Linear Weights formula and ERP are nearly identical themselves. A little basic algebra on Estimated Runs Produced gets it to look almost exactly like Linear Weights:

*Linear Weights Runs = .46(1B) + .80(2B) + 1.02(3B) + 1.40(HR) + .33(BB) + .30(SB) - .50(CS) - .0883(out)*
*Estimated Runs Produced = .48(1B) + .80(2B) + 1.12(3B) + 1.44(HR) + .32(BB) + .16(SB) - .00(CS) - .0984(out)*

The weights are a bit different — ERP gives a little "bonus" for singles, triples, and home runs and pardons caught stealing, but compensates by giving less credit for stolen bases and a higher penalty for outs. But otherwise, they're really just slight variations of the same theme. [3]

But those slight differences add up — for a wide range of teams, in Linear Weights' favor. Palmer's statistic is almost perfect in the 3-7 run range, while ERP was off by about 10 runs per season. For teams below 3 runs, Linear Weights is slightly better; for teams above 10 runs, ERP is a bit more accurate. Overall, Linear Weights wins.

Here, let me show you this way. For each simulated season's line, I compared each statistic to each of the other two. In every head-on comparison, the statistic coming closer to actual runs scored that year scored a "win"; the less accurate was given a "loss". Here are the final standings:

| | W | L | pct | GB |
|---|---|---|---|---|
| Linear Weights | 2049 | 1585 | .564 | -- |
| ERP | 1721 | 1913 | .474 | 328 |
| Runs Created | 1681 | 1953 | .463 | 368 |

It's a bit suprising that Linear Weights turns out to be more accurate than Runs Created, since Pete Palmer never intended it be used to determine the output of teams of offensive clones. "The Linear Weight is supposed to measure how a player would do in an average setting," Palmer says, "not in a specific [team of 9 identical players] setting."[4] Because teams of one outstanding player and eight average players are still close to average teams, Palmer's purposes demand only that Linear Weights be accurate in the middle class — and that it is.

[3] I am not the first to notice the similarity between the two statistics. Clay Davenport (jcd9s@virginia.edu) gave a similar comparison in an Internet posting in the summer of 1994.

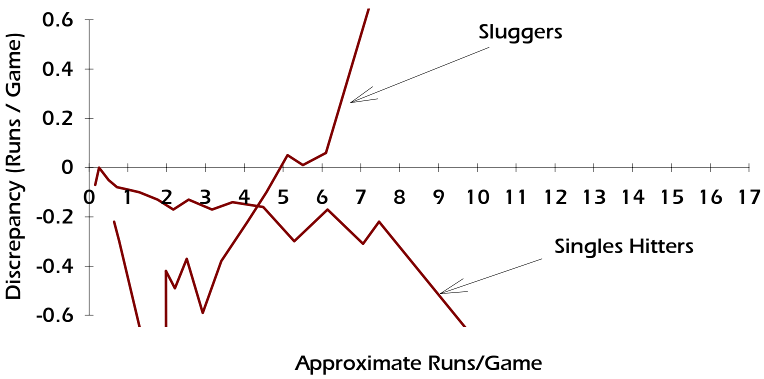
[4] Letter to the author, May, 1994.

## Offensive Types

We've been using OPS, on-base plus slugging percentage, to group the players in the study. But not all players with the same OPS are alike. A player with an on-base percentage of .450 who slugs .450 — a singles hitter with walks — is different from the home-run hitter who slugs .600 but hits only .250 with few walks, even though they both might have an OPS of .900. As a rule, one point of on-base percentage is worth more than a point of slugging percentage — the .450 slugger in the example will produce many more runs than than the .600 slugger.

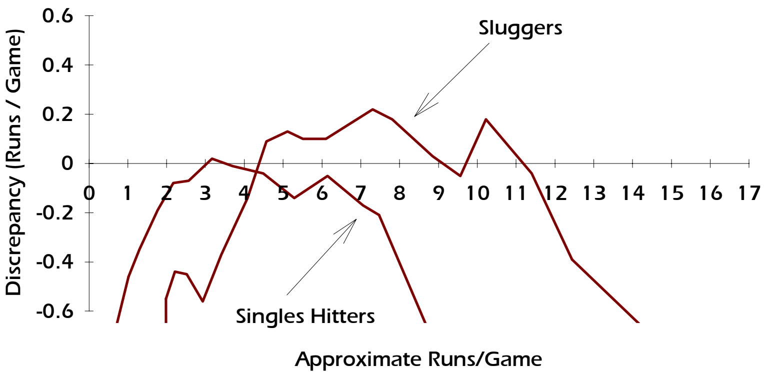
More importantly, the run predictor statistics may not treat different types of players the same way. Runs Created, for instance, predicts too high for players with above average offense. But does it err for Wade Boggs the same way as for Pete Incaviglia?

To find out, I went back to the list of 1,817 seasons and pulled out the most extreme offensive lines on both ends of the scale. For each season, I divided the slugging percentage by the on-base percentage. Players with less than 1.1 were classified as singles hitters; those with more than 1.6 were sluggers. Here's the chart:



There's quite a difference. While Runs Created overestimates good offenses in general, it *underestimates* good singles-hitting offenses. For slugging offenses, it's biased the same way as for offenses in general, but much worse — the graph disappears from sight at seven runs for sluggers; taking all offenses together, it stays in the park until nine. These, of course, are the two extremes; for hitters in the middle, their line will lie somewhere between those two. There's probably some class of hitter between those two types for which the Runs Created line is close to perfect.

Linear Weights isn't nearly as two-faced as Runs Created:



Linear Weights displays roughly the same accuracy curve for each type of hitter, but at different places on the offensive scale. Sluggers are reasonably well predicted between 5 and 11 runs per game, singles hitters between 2 and 7. But both those cases cover most of the players within their types, since sluggers tend to be more productive hitters than average.

## A New Statistic

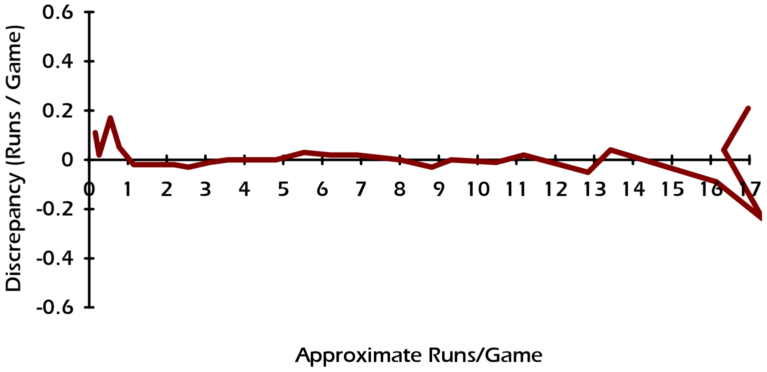
So Linear Weights is the most accurate of the three statistics. But even so, it's unacceptably off for the most extreme offenses, the 1-for-17s and the Ted Williamses. We'd still like to have a straight-line formula, one that's accurate for all offenses, regardless of potency or type. If there were some way to adjust Linear Weights or Runs Created to straighten out its line, we'd have the accurate predictor we're looking for.

Here's what I did. Given an offensive line, I ran a regression to predict the Linear Weights error — that is, I tried to come up with a formula that would predict how much Linear Weights was off. Then, to get actual runs, I just subtracted that formula from Linear Weights.

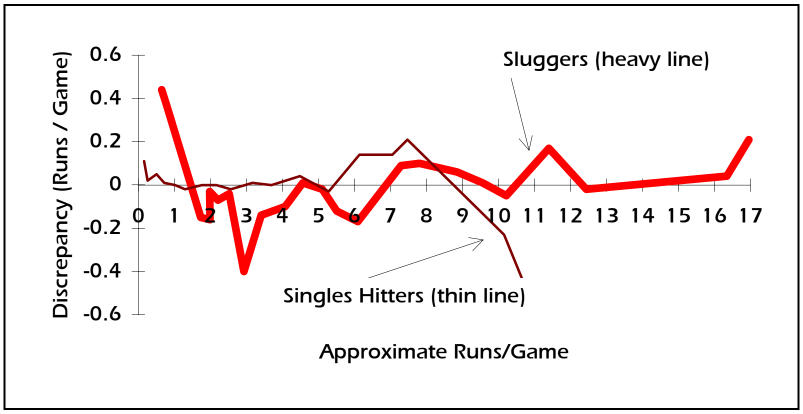
Here's the result, a corrected run predictor formula. I'll call it "Ugly Weights," for obvious reasons:

$$\textit{Runs} = .46(1B) + .80(2B) + 1.02(3B) + 1.4(HR) + .33(BB) + .3(SB) - .5(CS) - [\,.687*ba - 1.188*ba^2 + .152*ip^2 - 1.288*iw*ba - .049*ba*ip + .271*ba*ip*iw + .459*iw - .552*iw^2 - .018]\,(outs)$$

The "ba" is batting average; "ip" is isolated power (extra bases divided by at-bats); and "iw" is what I thought I'd call isolated walks — walks divided by at-bats. It's a mess, but it seems to work. Here's its accuracy curve:



It looks nearly perfect, but it's not quite as good split into sluggers and singles-hitters:



But it's still not bad, with an error of less than 0.1 runs per game most of the way across. The spikes at the ends of the curve aren't much to worry about, since they're based on only a couple of seasons.

So, for players, it looks like Ugly Weights is the way to go. But what about for those forgotten middle-class teams? It's certainly possible that when we tweaked Ugly Weights to work better at the extremes, we made it worse in the middle. Maybe Ugly Weights is good for great offenses, and bad offenses, but the original Linear Weights is better in the middle, for average offenses.

For each team batting line in both leagues, from 1950 to 1992 (not including 1981), I ran their predictions using all four statistics, and compared their actual runs to their predictions. Averaging out each statistic's errors should give zero, if all are accurate for average teams. All four came close:

| Statistic | Average Error | Mean Square Error |
|---|---|---|
| Runs Created | -8 | 26 |
| ERP | -4 | 24 |
| Linear Weights | +2 | 24 |
| Ugly Weights | 0 | 24 |

They're hard to rank on this basis, looking pretty much the same. To draw out the differences, I ran the win/loss test on each pair of stats for each season:

| | W | L | pct | GB |
|---|---|---|---|---|
| Linear Weights | 1449 | 1299 | .527 | -- |
| ERP | 1439 | 1309 | .524 | 10 |
| Ugly Weights | 1438 | 1310 | .523 | 11 |
| Runs Created | 1170 | 1578 | .426 | 279 |

The top three statistics, whose differences are probably just due to luck, seem to have taken turns beating up on Runs Created. For teams, Ugly Weights doesn't seem to be required; Linear Weights will do just fine.

## Conclusion

This simulation study pretty much confirms what we already knew about the existing run predictors as applied to middle-class offenses — they work. If you need to choose one, Linear Weights will give you the best results; Runs Created is about as accurate as Estimated Runs Produced, and both fall a bit below Linear Weights.

But for unusually good or bad offenses, none of the statistics remains accurate, although you can bridge the gap somewhat by choosing the statistic that's the least in error for the type of offense you're analyzing. Or, you can use Ugly Weights, which seems to work pretty well for virtually all types of offense — at least as far as we've tested.

*Thanks to Pete Palmer for reviewing an earlier version of this article and providing many valuable comments.* *Phil Birnbaum, 18 Deerfield Dr. #608, Nepean, Ontario, Canada, K2G 4L1, phil_birnbaum@iname.com.* ♦

## Submissions

Submissions to *By the Numbers* are, of course, encouraged and drooled over. Articles should be concise (though not necessarily short), and pertain to statistical analysis of baseball. Letters to the Editor, original research, opinions, summaries of existing research, criticism, and reviews of other work (but no death threats, please) are all welcome.

Articles should be submitted in electronic form, either by e-mail or on PC-readable floppy disk. I can read most word processor formats. If you send charts, please send them in word processor form rather than in spreadsheet. Unless you specify otherwise, I may send your work to others for comment (ie, informal peer review).

I usually edit for spelling and grammar. (But if you want to make my life a bit easier: please, use two spaces after the period in a sentence. Everything else is pretty easy to fix.)

I will acknowledge all articles within three days of receipt, and will try, within a reasonable time, to let you know if your submission is accepted.

# Informal Peer Review

The following committee members have volunteered to be contacted by other members for informal peer review of articles.

Please contact any of our volunteers on an as-needed basis - that is, if you want someone to look over your manuscript in advance, these people are willing. Of course, I'll be doing a bit of that too, but, as much as I'd like to, I don't have time to contact every contributor with comments on their work. (I will get back to you on more serious issues, like if I don't understand part of your method or results.)

If you'd like to be added to the list, send your name, e-mail address, and areas of expertise (don't worry if you don't have any - I certainly don't), and you'll see your name in print next issue.

Expertise in "Statistics" below means "real" statistics, as opposed to baseball statistics - confidence intervals, testing, sampling, and the like. Some of our statistics experts have quite the credentials - they make my bachelor's degree and half-a-masters look puny.

| Member | E-mail | Expertise |
|---|---|---|
| John Matthew | john.matthew@ca.arthurandersen.com | Apostrophes |
| Jim Box | jim.box@duke.edu | Statistics |
| Rob Fabrizzio | rfabrizzio@bigfoot.com | Statistics |
| Duke Rankin | RankinD@montevallo.edu | Statistics |
| Keith Karcher | kckarcher@compuserve.com | |
| Tom Hanrahan | HanrahanTJ@navair.navy.mil | Statistics |
| Steve Wang | Steve.C.Wang@williams.edu | Statistics |
| Larry Grasso | l.grasso@juno.com | Statistics |
| Keith Carlson | kcarlson@stlnet.com | Economics/Econometrics/Statistics |
| John Stryker | johns@mcfeely.interaccess.com | |