By the Numbers

Volume 9, Number 3 The Newsletter of the SABR Statistical Analysis Committee August, 1999

News

Phil Birnbaum, Editor

BRJ Submissions

Please take a look at Mark Alvarez's note on page 12. Mark is SABR publications director, and one of his jobs is putting together the "Baseball Research Journal." He's been contacted by members of this committee who are concerned about the kinds of statistical articles that BRJ publishes – that is, that they're not up to the standards of the field.

Mark would like to know how this committee could get involved in helping ensure that the statistical pieces he selects are appropriate. One thing he points out is that even if you've already published in BTN, you can still send your article in to Mark. (As BRJ is aimed at a less statistically-sophisticated audience than this newsletter, you might want to change the tone a bit for those who don't have the same background we do.) I'm pleased to hear this is the case. BRJ seems to be a more prestigious place to publish than BTN, and we don't want to create an incentive for researchers to bypass our committee.

Another suggestion might be that the committee review all submissions and make recommendations to Mark. Instead, we might like the idea that members wishing to publish in BRJ first publish in BTN. (That's one way of ensuring that we continue to receive submissions!) Or, you might come up with another suggestion for Mark altogether.

Please think about this, and, if you like, send suggestions here and we'll run them in the next BTN.

"By the Numbers" Name to Remain

It looks like the name "By the Numbers" will stay. The overwhelming consensus from members who replied to last issue's comments is that BTN is a fine name, and there's no need to change it. A selection of those comments appears on page 3.

In this issue

Convention and Thank Yous

It was great to meet so many of you in Scottsdale. As this was my first SABR convention since Cleveland in 1990, I was meeting most of you for the first time. Almost everyone I met mentioned that they appreciated "By the Numbers", and I thank all of you for your kind words. But, again, I point out that there wouldn't be a newsletter without articles from our contributors. The effort involved in doing the research and putting a study together is much more than I do putting it all together. Please, if you like what you see, take a minute to drop a line to our contributors to thank them for their efforts.

Deadline

Next issue's submission deadline is October 24.

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

Convention Report

Neal Traven, Committee Co-Chair

SABR 29 in Scottsdale, Arizona was, for the most part, a rousing success. The topper, of course, was the privilege of being in Bank One Ballpark for Jose Jimenez's no-hitter against the Diamondbacks. For the third time in five years (and second year in a row), I participated on a trivia team which lost its semifinal to the eventual championship foursome. This year, Steve Roney was a fine complement to my usual teammates John Rickert and non-SAC member Jerry Kahn. Maybe next year we'll reach the next plateau.

For the committee meeting at the convention, we had to deal with a small mixup – local organizers had reversed our request for a room to handle 30 people for 60 minutes, originally giving us a large room but only half an hour of meeting time. We ended up keeping the cavernous space, but starting our meeting even earlier Friday morning than originally slated. Somehow, SAC meetings always seem to take place on the early shift.

I count 43 names on the attendance sheets we passed around. Add co-chairs Clem and myself to give a total of 45 members in attendance, a somewhat larger crowd than I'd anticipated. As is customary at such gatherings, we started the meeting by going around the room introducing ourselves and briefly describing our research interests. I must note that when the rotation reached Phil, a round of applause erupted immediately … before Clem and I had the chance to ask for one. A well-deserved acknowledgment of the tremendous improvement in *By The Numbers* (if it's still called that) since he volunteered to edit our publication. Eliciting by far the most discussion of any topic discussed at the meeting was Mike Webber's introduction of the ambitious minor league project he's working on with Lloyd Johnson. A fuller description of their proposed database appears elsewhere in this issue of the newsletter.

For several reasons, among them an early departure time on Sunday morning, I wasn't able to attend nearly as many research presentations as I usually do. Thus, I didn't get to hear very many of the 12 research papers (which might not necessarily have been statistical in focus) presented in Scottsdale by Statistical Analysis Committee members and/or committee meeting attendees:

- John Jarvis – An analysis of the intentional base on balls
- Mike Sluss – Probability of Performance: General relativity for the baseball fan
- Doug Lyons – Out of left field: Adventures of a baseball researcher
- David Smith and Clem Comly – Bulking up baseball: Are expansion teams muscle or fat?
- Steve Krevisky – Vern Stephens: Shortstop and sentinel of the southwest
- David Stephan – Unraveling DNP codes: Why great ballplayers didn't play and missed games
- Bill Gilbert – How does McGwire's 1998 season rank with the best offensive seasons of all time?
- Sky Andrecheck – Ranking the dynasties: A statistical approach
- Tony Blengino – The class of 1981: The historic dominance of the six 35-year-old American League pitchers in 1998
- Jim Vail – The Cooperstown baseline: Hall of Fame statistical standards and errors of selection
- Mike Schell – Improving on Relative Batting Average
- Mark Pankin – Who's right, LaRussa or Gant?

Two additional papers of potential statistical interest – based on my review of their titles and abstracts -- were presented by SABR members who aren't on the SAC mailing list:

- James McMartin – The 25 best hitting seasons: A comparison of Runs Created, Batting Runs, and Runs Generated
- Steve Schulman – Runs Prevented … Revisited

All in all, quite a successful SABR meeting. I hope Boca Raton can do as well as a convention site in 2000. Finally, I'm simply amazed that I made it all the way to the end of this report without once mentioning how HOT it was in the Phoenix area. Supposedly it's the humidity rather than the heat, but a convection oven still feels mighty warm.

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

"By the Numbers" Comments

Committee Members

Last issue, a piece by the editor argued that what we analyze is baseball, not numbers, and so we should consider a new name for this publication. Most of the responses that came in were against a name change. Here are some of them.

Good Point, Bad Example

Nice comments in the editor's notes. If nothing else, changing our name would give you a chance to have them published in the SABR Bulletin by way of explanation, where the wider membership could be gently 'corrected' about what it is we do, and why.

While I agree with everything you said in your editorial notes, you picked an unfortunate example with medicine. When I was in grad school at Cornell, there was an entire department called "Biometrics" -- that is, statistical analysis of clinical trials data. And your colleague Neal Traven used to be my colleague at the University of Pittsburgh, where he was in the Epidemiology department -- that is, statistical analysis of public health data. As you note, medicine is chock-full of mathematical statistics, because of the way effect doesn't invariably follow cause. So full, in fact, that fields have grown up between the general statisticians and the physicians, focusing on those kinds of data that are available in clinical trials or public health studies, and on the best ways to interpret them.

Now, I agree that baseball isn't there yet, but doctors and statisticians may miss the point of your example. An even better example might be meteorology, which has the following similarities with studying Major League Baseball:

1. designed experiments are impossible
2. underlying processes are extremely complex
3. specific prediction is extremely fallible, while general trends can be predicted accurately
4. mathematical models and simulations are essential
5. the people who do this work are interested in weather/baseball, not the math itself

In fact, two of the best sabermetricians I know of, Clay Davenport and Harold Brooks, are both meteorologists. No coincidence, I think.

As for a new name...

Unfortunately, Keith Woolner has already taken the name "Baseball Engineering" for his web site, for precisely the reasons you outline. I lean toward some name that conveys a similar sense of "taking off the back panel to see how the mechanism works". "Mechanics" means something else in a baseball context, though, as does the word "inside". A title like "Behind the Curtain" appeals to me, but would again be misinterpreted as being about the action in the clubhouse, not the action on the field. Hmmm.

"Diamond Crystallography" is probably too obscure, though it appeals to my sense of humor -- it implies both the analytical techniques of X-ray diffraction, and the soothsayer's crystal ball gazing for predicting the future. "Analytical Baseball", or some other title using the word 'analytical' might work -- again, the word 'analysis' has been co-opted by the media to mean 'uninformed blather, generally by Peter Gammons'. Back when the Usenet group rec.sport.baseball spun off a separate group for sabermetric sorts of things, the new group was called rec.sport.baseball.analysis.

"Reverse Engineering" conveys the sense I'm looking for of going from visible behavior to an understanding of how and why the system behaves that way. That's probably not an acceptable title all by itself, though. "Stochastic Baseball" is a beautifully concise description of what we do, but way too obscure for the general public. (You probably know this, but 'stochastic' is from the Greek for "to aim" or "to guess", and refers to randomness or chance in a model. "Aleatoric" is another synonym.)

I don't seem to have been very helpful here. Feel free to use any or none of these ideas, recombine them, throw them out to the membership, or whatever.

Dave Tate

Fine With Me

By The Numbers is fine with me!

S.J. Davis

But We *Do* Do Numbers

I really like "By the Numbers" and don't think we should change. Yes, sabermetrics is a lot besides numbers, but this particular magazine is the product of the statistical analysis committee, and numbers are at the center of what we do (or at least what we do for this particular journal). We don't want to imply that we do the same thing as the SABR Research Journal.

David Shiner

Playing With Numbers is What We Do

With regard to your comments as editor, I feel that what I do is play with numbers, and I don't see a necessity for changing the name of the newsletter.

Clifford Otto

Another BTN Vote

I still like "By the Numbers."

Keith Carlson

Still Mathematical

I vote for the current name. Although you are correct in stating that sabermetric research need not be numerical, most of it has been and will continue to be so.

Charlie Pavitt

We'd Need a Weird Cryptic Symbol

My *second* choice [is "By the Numbers"], but my silly first choice is "The Newsletter formerly known as By the Numbers."

Cappy Gagnon

It Ain't Broke

I would observe that there are plenty of things to do without seeking extra work by trying to fix things that aren't boken. 'By the Numbers' is just fine as a title.

Larry Grasso

Statistics Are The Attraction

Count me as among those favoring the Bill James definition of Sabermetrics. For me, baseball statistics provide a thrill beyond any insight they may provide about "forces in professional baseball."

Long before there was a SABR, I was poring over the numbers, creating lists of seasonal performances that met certain criteria. The criteria started out fairly crude (e.g., power-hitting shortstops who managed, say, one homer every 20 at bats). Over the years, the criteria have gotten more complex as I look for progressively more unusual patterns and create fantasy careers from the seasons that satisfy the criteria. Of course, some players have actual careers that achieve almost uncanny consistency -- Ted Williams is the paramount example.

True, over the years, my contemplation of baseball records has given me some strong opinions on what statistics might be the best relative measures of performance, but that contemplation never depended on the purported scientific or aesthetic value I might derive from it.

Hell, I just like the numbers!

Steve Kotz ♦

Minor League Database

Lloyd Johnson, Mike Webber

Anyone interested in contributing to a database of minor league player statistics for the Museum of Minor league Baseball (backed by the National Association) should contact Lloyd Johnson at 816-822-2516 or e-mail lloydj@msn.com or Mike Webber at 785-242-6638 or e-mail kcbbfan@aol.com. Inputters and original researchers are needed. For additional information, contact Johnson at PO Box 22481, KC, MO 64113.

Improved "Big Bad" Annual a Worthwhile Read

Clifford Blau

While certainly not up to the standards of the Bill James Baseball Abstracts from which it is descended, improved writing this year and several interesting features from guest contributors make this annual worth a look.

The *Big Bad Baseball Annual* (BBBA) is the direct descendant of the *Bill James Baseball Abstract*. However, the strength of Bill James' books was the writing; this has always been the weakness of the BBBA. I am happy to report that this is less of a problem this time around. This is due mainly to the addition of several writers over the years. While Messrs. Malcolm and Hanke still indulge themselves at times in flights of fancy, most of the authors are sticking to business.

The 1999 Big Bad Baseball Annual

By Don Malcolm, Brock J. Hanke, Ken Adams, and G.Jay Walker

Masters Press, 509 pages (paperback), $19.95
ISBN 0-8092-2655-3

The book is organized into five main sections. Two contain general baseball essays, one has several articles concerning the new hitter evaluation method used in the book this year, and the other two are the traditional team essays and player ratings.

Since I don't follow what passes for major league baseball now, I can't comment much on the latter two sections. I will simply note that the team essays generally contain certain features such as park-adjusted hitting statistics, bullpen data, information on minor league prospects, and an article about a "bad" pitcher from the team's past, using the authors' QMAX statistic. (It is this latter feature that contains the book's most outlandish statement, that Forbes Field "was where triples went to die.")

The QMAX statistic gives two scores for each pitcher start, one for "stuff" and one for control. The scores are based on hits minus innings pitched and walks minus innings pitched. The problem I have with it is it tends to ignore the pitcher's durability, since it gives the same score to allowing 1 walk in 5 innings as it does to allowing 5 walks in 9 innings. Two hits in 4 innings is as good a start as 7 hits in 9 innings to QMAX, but not to me.

As for the new hitter evaluation method, known as extrapolated runs, it is simply another linear formula with slightly different weights than the many other such systems. It claims greater accuracy, but admittedly it is more closely correlated with the 17th best such formula than it is with actual runs.

This brings us to what attracted me to the book. Some 22 essays covering various aspects of baseball past and present. Tom Ruane contributes an article how teams were built and destroyed between 1960 and 1975. David Grabiner compares the 1998 Yankees, position-by-position, to several great teams of the past, as Bill James did with the Tigers in the 1985 *Abstract*. Don Malcolm pitches in with a discussion of the high level of walks in 1949, and teams up with David Smith to present a Retrosheet-derived statistical recap of Jackie Robinson's career. This last piece was disappointing to me; I'd have liked to seen something other than the conventional batting statistics. How often did Robinson go from first to third on a single compared to his contemporaries? Did he tend to force errors at an unusually high rate? As a second baseman, he had a terrific ratio of double plays to games; was this due mainly to opportunity, or was he great at converting his chances? Instead of answering questions such as this, all we learn is that Jackie's splits were smaller than normal.

Tom Hull has an interesting piece using percentile rankings of runs created to compare historic ball players and to try to predict the future development of current players. While I feel that the system used should take outs made into account, and the method of combining seasons doesn't give proper credit for longevity, it is a good way to learn new things about players. Ken Adams redoes Stephen Jay Gould's study of decreasing variability in hitting, using on base and slugging averages in place of batting average. He finds that the standard deviation in these statistics continues to decline. Don Malcolm contributes another article looking at competitive balance. Charlie Saeger takes a stab at evaluating fielding statistics from 1941. Another educational essay is Doug Drinen's on relief pitching. As he did in the 1998 edition, he looks at which situations' relievers are used in, and how they perform compared with expected runs tables.

The BBBA gives the reader a lot to chew on, despite some shortcomings. I am glad I purchased it, and can recommend it to others interested in baseball analysis.

Clifford Blau, 16 Lake St., #5D, White Plains, NY, 10603, cliffordblau@geocities.com. ♦

Recent Academic Sabermetric Research

Charlie Pavitt

The author reviews two recent academic articles featuring sabermetric research.

This is the first of what I foresee as an occasional review of sabermetric articles published in academic journals. It is part of a project of mine to collect and catalog sabermetric research, and I would appreciate learning of and receiving copies of any studies of which I am unaware. Please visit the Statistical Baseball Research Bibliography at www.udel.edu/johnc/faculty/pavitt.html, use it for your research, and let me know what I'm missing.

Bruce Bukiet, Elliotte Rusty Harold, and Jose Luis Palacios, "A Markov Chain Approach to Baseball," Operations Research, Volume 45 Number 1, January-February 1997, pages 14-23.

Bukiet, Harold, and Palacios's article is a good example of what happens when people who are skilled statisticians decide to add to the sabermetric literature without performing a serious review of what that literature already has to offer. Bukiet et al.'s reference list makes it clear that they are familiar with the stream of sabermetric research performed by statisticians back in the 1960s and 1970s, but nothing since. As a consequence, they reinvent the wheel, and do so using sub-standard materials.

Bukiet et al.'s goal is to use Markov chains to find the most optimal batting order. They acknowledge the earlier work by Freeze in this regard but not, for example, Earnshaw Cook's two Percentage Baseball books from the 1960s, the first of which should have been familiar to them. They are also apparently unaware of relevant work by Boronico (Baseball Analyst 23), Bennico (Baseball Analyst 31), Fletcher (Baseball Analyst 36), Seifert (Baseball Research Journal 23), and in particular Mark Pankin (By The Numbers Vol. 2 No. 4, Vol. 3 No. 5, and Vol. 4 No. 3 and Baseball Research Journal 21). Now I can understand why academic statisticians may be unaware of what I will call the "hardcore" sabermetric literature. But there is a lesson in this: we who are the producers of that hardcore literature, when in the process of adding to it, need to consider ourselves responsible for being familiar with relevant past work so as to improve on it.

The point is that Bukiet et al. have not improved on that work. For example, to represent batter performance they use D'Esopo and Lefkowitz's Scoring Index. In Thorn and Palmer's *The Hidden Game of Baseball,* this index is shown to be a very poor measure of batter performance; and in my opinion ANY baseball researcher, hardcore or not, is responsible for being familiar with that book. Second, they use incredibly simplistic presumptions about baserunner advancement (e.g., singles always advance runners on first to second), although they are aware of the existence of play-by-play databases with actual baserunning data that are available "for a price" (someone should tell them about Retrosheet). As a consequence, their estimations of runs scored by "real" lineups is 7 percent below actual runs scored data, an unacceptable performance by today's standards.

Leaving that aside, using 1989 National League data, they find the difference between best-possible and worst-possible lineups to be about 4 wins per year. This number is roughly consistent with Pankin's findings, and so the article has some value as a replication of Mark's work. As such, it reinforces the conclusion that managers are worse off staying up all night trying to come up with a slightly better lineup than they would be sleeping so as to be alert for the next night's game.

Ira Horowitz, "The Increasing Competitive Balance in Major League Baseball," Review of Industrial Organization, Volume 12, 1997, pages 373-387.

Unlike Bukiet et al., Horowitz is no neophyte sabermetrician; I am aware of four other articles he has published. In this effort, Horowitz attempts to add to our knowledge about trends in competitive balance between baseball teams over time. Horowitz is aware of economic research demonstrating an increase in competitive balance during the twentieth century, and also cites relevant comments in the Bill James Historical Baseball Abstract. Along with replicating this finding, he attempts to add to it with a precise study of the impact on competitive balance made by significant events that could possibly serve to counterbalance this trend.

Again, we have a case in which an author is unaware of previous relevant work in the "hardcore" literature. Wood and McCleery (Baseball Analyst 39) presented evidence that competitive balance in the American League, although increasing over the long run, has in the short run been cyclical. The Yankee dynasties of the 1920s and 1930s decreased parity, as did the expansions of 1961, 1969, and 1977. In contrast, the Yankee dynasty of the 1950s did not decrease balance. Hadley and Gustafson (By The Numbers Vol. 6 No. 1) showed data suggesting that free agency did not decrease competitive balance, at least during the 1980s.

Horowitz examined both the immediate and long-term impact of six events in both leagues; the end of the use of "dirty balls" in 1920 (although Horowitz's discussion does not mention the importance of Chapman's death in this regard), the introduction of black players (1947 in the N.L. and 1948 in the A.L.), expansions (1961, 1969, and 1977 in the A.L., 1962, 1969, and 1977 in the N.L.), and free agency (1976). Some of Horowitz's findings are interesting. He finds 1920 to reduce balance in the N.L. but increase it in the A.L. 1947 brought a slowing down of the trend toward balance in the N.L.; in the A.L., 1948 led to an immediate decrease in balance that took 10 years to overcome. Both expansions and free agency had little effect in the A.L. but a strong destabilizing impact on the N.L.

Horowitz's findings are not always consistent with earlier work, perhaps because of methodological differences. Wood and McCleery worked with five-year moving averages of team wins so as to negate the impact of the year-to-year variability that Horowitz wanted to measure. Similarly, Hadley and Gustafson compared long-term trends and ignored each year's impact. I do have a concern with one issue: I cannot see how one can tease apart the separate impacts of free agency in 1976 and the 1977 expansion using Horowitz's, and perhaps any other, method.

The inconsistencies between this and earlier studies suggest that the last word on this subject has not be spoken. Horowitz's article is an intelligent, useful contribution that stands alongside the earlier work on this topic.

Charlie Pavitt, 812 Carter Road, Rockville, MD, 20852, chazzq@udel.edu ♦

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 detailed 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 so on.

Member	E-mail	Expertise
John Matthew	jmatthew@totalsports.net	Apostrophes
Jim Box	im.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	

Relativity and Statistical Ratings and Rankings: Three SABR 29 Presentations

Tom Howell

Where does 1998 Mark McGwire rank among all-time seasons? Three presentations at SABR 29 in Scottsdale touched on this question. Here, the author reviews that research and adds some of his own observations.

Introduction

The common theme that caught my attention among the three presentations I review here was this: "Where did Mark McGwire's 1998 season rank all time?" Of course, the answer to this question depends upon on how you want to measure his 1998 season: runs, "theoretical" runs, homerun total, homerun frequency, etc. Each of the three papers reviewed below, besides addressing other issues, provided some insight into the noteworthiness of McGwire's 1998 season. Each presentation is discussed separately and then again in the summary. Of course, I have also included discussion of some of my own favorite tools for player comparisons.

Jim McMartin, "The 25 Best Hitting Seasons: A Comparison of Runs Created, Batting Runs and Runs Generated," presented at SABR 29

Mr. McMartin presents a new relationship to be used as an alternative to the more well known runs created (RC) and batting runs/linear weights (BR/LW) relationships, with the objective being to provide a more useable "fan friendly" (read "simpler") calculation; the new method provides "runs generated" (RG), presumably without sacrificing the accuracy provided by the more complex calculations of RC and LW. McMartin then uses minimum cutoffs of .387 (BA), .690 (SLG) and .480 (OBP) to select 72 hitter-seasons (1900+) from which he culls the final 25 best hitting seasons: each hitter-season is given ranking points for its relative position in each season category (1 for first, down to 72 for 72nd) for 3 offensive measures: RG, LWR (BR +SBR) and RC. The 72 seasons considered were presented in chronological order as an appendix in the handout from the presentation. The top 25 seasons, ordered by the average of the ranking of 3 offensive measures (lowest average being best), were presented separately and used as a forum for comparison of RG with LWR and RC.

Here's the RG formula:

$$RG = 0.35\ (TB + BB) - 0.25\ (AB - H)$$

First, Mr. McMartin made it clear that he was after a more user friendly equation (versus the complexity of RC and LW calculations) to encourage the use of sabermetrics by the average fan, fully recognizing and acknowledging the contributions of James, Thorn and Palmer for RC and LW. Several statistical analyses were provided showing that that RG could be used as a predictor for runs with almost the same reliability as RC and LW, although it fell a little short (accounting for 84% of variation in run production versus 85 and 88%, respectively for LW and RC - based on team data from 1970-1989). In addition, the author quickly eliminated the "new" RC values (1997 STATS All-Time Major League Handbook) from his comparisons for a variety of reasons (see the author's handout for a detailed discussion); he relied upon the RC values published in Total Baseball. Although Mr. McMartin recognized that RC was much more dependent on yearly variations in league run-scoring than RG and LW, and "may be misleading when used in cross-era comparisons," he included it in his 3-factor consensus rankings (this point served as a source for further discussion of RC in the author's handout - not treated here). For analyzing the top 25 seasons, he included HBP, SB and CS (where available).

Mr. McMartin acknowledges several limitations in his RG relationship (shown above), by comparing several individual batter-seasons from his top 25: (a) most of the pre-1920 seasons included in the top 72 candidates were underestimated by RG (based on comparison with LWR) and McMartin interprets this as being due to the RG equation treating a single as roughly equivalent to a walk and that there were many more singles than HR prior to 1920; and (b) he felt that McGwire's RG ranking of #5 all-time (#10 overall by consensus) might be somewhat inflated (*my interpretation*) relative to Stan Musial's 1948 season (#30 by RG, #19 overall consensus); he felt these two seasons were much more comparable to each other than RG indicated, probably because of McGwire's high walk and HR rates relative to his singles (versus Musial). See Table in Summary for top 20 season rankings.

I don't think the weakness in RG resides simply in the "equivalency" of singles and walks in the RG equation; it goes somewhat further than that. RG appears to substantially underestimate run production (by 20-30% compared to LWR) for several top batter-seasons prior to 1920 (as acknowledged by Mr. McMartin), but also for some of the more recent 72 top season candidates, namely Carew 1977 and Boggs 1988. I think this discrepancy is due more to the 0.35 "constant" being applied uniformly to total bases in the RG equation, as opposed to anything

special about pre-1920 single, homerun and walk frequency characteristics. The 0.35 constant seriously penalizes "good" singles/doubles hitters (obviously, many from pre-1920) regardless of their walk or HR tendencies, e.g., 1B = 0.35, 2B = 0.7, 3B = 1.05 and HR =1.4, based on RG. Since a single and double should be valued more closely to 0.5 and 0.8, respectively (at least if you are LW fan), RG undervalues singles by about 1/3 and doubles by about 10%, while matching the triple and HR values (based on LW) very closely. If you examine the Carew 1977, Boggs 1988 and Cobb 1911-15 seasons, all of their "single + double" contributions corresponded to about 60% of the "TB+BB" value with triples and HRs only accounting for 5-13% of total hits. Any superior batter, from any era, who depended primarily on singles and doubles versus HR frequency is shortchanged by RG.

There is no question that RG is simpler to calculate than LW or RC; however, I am not convinced that is enough to make me want to use it on a daily (or weekly) basis to see how my favorite player is really doing. Personally I prefer the use of relative OPS (on base + slugging) as a quick and dirty measure of how a player is doing and how it might project out for a full season (related to PRO (Production) values used in Total Baseball, without park adjustment). Current run production (1999) in the major leagues is about 0.13–0.135 runs/plate appearance which translates to about 85 "runs" for an average fulltime player. If we use 85 "runs" as a baseline for the average player (league OPS of about 0.775 for non-pitchers), then the following relationship (Total Runs) gives an estimate of RC, and if you subtract the baseline 85-run value it gives you an estimate (Net Runs) of Batter Runs (BR, based on Linear Weights); OPS^P = on base % + slugging % for player, readily available from your handy-dandy USA Baseball Weekly.

$$\text{"Total Runs"} = 220 \times (OPS^P) - 85 = \text{(estimate of Runs Created)}$$

$$\text{"Net Runs"} = 220 \times (OPS^P) - (2) \times 85 = \text{(estimate of Batter Runs, LW)}$$

I have not compared these relationships to LW or RC in the more rigorous manner that Mr. McMartin went through for RG, but as far as being simple and quick, I find them more palatable, both theoretically and for ease of calculation; full derivation can be found in the Appendix. To illustrate this "quick and easy" method, players having OPS of 1.000, 1.100, 1.200 and 1.300, respectively, would have full season Total Run (Net Run) estimates of 135 (50), 157 (72), 179 (94) and 201 (116), respectively. Only the latter two values (1.200 and 1.300) generate run values that would merit consideration among the top 10 seasons of all-time, e.g., McGwire's 1998 season OPS of 1.2+.

Mike Sluss, "Probability of Performance: General Relativity for the Baseball Fan," presented at SABR 29

Mr. Sluss suggested a new way to rank performance in a single offensive category: "Probability of Performance" (POP). For player X, the POP statistic represents the probability that an average offensive player would accumulate the same or better record than X. The POP statistic is the negative logarithm of this probability (after subtracting out the player's own contribution to the league average). For instance (the author's example), Rod Carew batted .388 in 616 AB in 1977; the probability that a league average player (.265) would bat .388 in at least 616 at bats is about 0.000000000021. The negative logarithm of this number is 10.67, so Rod Carew's POP for batting average is 10.67, the highest since 1940.

To put this in perspective, an average POP is 0.3, i.e., batter has 50% chance to equal or exceed the performance in question. POP values of 1, 2, 3 and 6 = 10%, 1%, 0.1% and 1 in a million, respectively. Some of the top values for certain offensive performances were mind-boggling.

I'm really intrigued by this method, since it appears to give weight not only to the frequency of what happened (H/AB, HR/PA, etc.) versus the norm, but it also treats the absolute quantity or amount (playing time vs league) accumulated in a given season. We are all used to seeing relative batting averages or HR rates (normalized to the league, with or without park corrections), but these ratios are usually in the range of 1.3 to 1.5 (for batting average) and perhaps up to 5 to 10 for HR rates (depending on the era). POP takes these relative values and really spreads them out in terms of probability. Example: Foxx's 58 HRs in 1932 (6.5 times the league HR%) versus Greenberg's 58 in 1938 (5.4X league HR%). At first glance, relative rates of 6.5 and 5.4 might sound pretty similar. Guess again -- translating these into probabilities, the POP values for these two HR seasons are 28.1 and 24.2, respectively. You might say "Big deal -- 28 and 24 sound similar, also." However, the difference of 4 POP units means that Foxx's achievement was 10,000 times less likely to be achieved by an average player than Greenberg's season -- food for thought.

Sluss defines career POP as the sum of each individual season POP. Since a POP must always be positive, and the worst is zero, this career measure means that a player's career POP can never decrease, even after sub-par seasons near the end of his career.

Selected leaders in a variety of categories are presented below:

```
Batting average, season:      Tip O'Neill 1887, 15.21;  Fred Dunlop 1884, 14.70;   Ty Cobb 1911, 14.31
Batting average, career:      Ty Cobb 166.95;  Nap Lajoie 100.09;  Tris Speaker 97.42
Home run %, season:           Babe Ruth 1920, 44.88;  Babe Ruth 1927, 43.74;  Babe Ruth 1921, 40.70
Home run %, career:           Babe Ruth 428.97;  Jimmie Foxx 190.02;  Lou Gehrig 180.71
Stolen base %, season:        Max Carey 1922, 11.68;  Maury Wills 1962, 9.40;  Vince Coleman 1986, 7.20
```

Mr. Sluss proposes an all-time career hitting team, by position, based the player with highest minimum lifetime POP in all of the three hitting categories of OBP, BA and HR%: Piazza (C), Gehrig (1B), Hornsby (2B), Baker (3B), Wagner (SS), Williams (LF), Mays (CF) and Aaron (RF). The only real surprise might be Baker -- Schmidt presumably loses out because of BA.

Mr. Sluss's rankings for the single-season HR% rankings merit further discussion, given McGwire's 70-HR season (see table below). Perhaps to no one's surpise, Ruth still has the top 6 HR seasons, with McGwire coming in at #9. For comparison purposes I added the absolute HR% rankings next to Mr. Sluss's POP rankings so that you could see the effect of league normalization, frequency and amount inherent in the POP statistic and how the seasonal rankings change as a result (McGwire has 3 seasons in the top 10 absolute value ranking, but only 1 in the POP ranking). In addition, I finally did something I have always wondered about, but never checked. The HR values in the Table after the arrow (e.g. "→ 40*") represent what the seasonal HR values would be after home park adjustment (average of Total Baseball, 4th Edition, and STATS All-Time Baseball Sourcebook, 1998, home run park factors). If the POP rankings were to be based on park-corrected HR%, then Ruth's 1924 season and Foxx's 1932 season may drop in the rankings with Gehrig 1927 and McGwire 1998 having the best chance to move up. Of the seasons not listed, Wilson 1930 (56 → 50*), Maris 1961 (61 → 65*), Mantle 1961 (54 → 58*), Mantle 1956 (52 → 56*) and Kiner 1949 (54 → 48*) might have the best chance to break into the top POP HR% rankings if park adjustments were considered.

Rank	Sluss Ranking (POP - HR%/single season)	Absolute HR% Ranking
1	44.88 Ruth 1920 (54 → 40*)	McGwire 1998
2	43.74 Ruth 1927 (60 → 53*)	McGwire 1996
3	40.70 Ruth 1921 (59 → 47*)	Ruth 1920
4	34.59 Ruth 1928 (54 → 50*)	Ruth 1927
5	31.34 Ruth 1926 (47 → 45*)	Ruth 1921
6	30.53 Ruth 1924 (46 → 35*)	McGwire 1997
7	28.08 Foxx 1932 (58 → 42*)	Mantle 1961
8	27.56 Gehrig 1927 (47 → 42*)	Greenberg 1938
9	26.54 McGwire 1998 (70 → 67*)	Maris 1961
10	24.16 Greenberg 1938 (58 → 51*)	Sosa 1998

* adjusted estimate (± 1) based on HR factor for home park, depending upon whether you use park factors from Total Baseball (4th Edition) or 1998 STATS All-Time Baseball Sourcebook.

Bill Gilbert, "How Does McGwire's 1998 Season Rank with the Best Offensive Seasons of All Time?" presented at SABR 29

The author uses a multi-factor (10) approach to generate a "consensus" ranking of the best offensive seasons of all time. Six of the factors considered are percentage or ratio type parameters [HR%, SLG, OPS, earned base average (EBA), bases per plate appearance (BPA), total average (TA)]; three are single value estimates of "runs above average" [runs generated (RG, see McMartin paper), batting runs (BR) and linear weights (LW = BR + SBR)]; and one factor is a gross value estimate of "runs" [runs created (RC)]. This is a real hodge-podge of parameters (rates of production and absolute values of production), many of which use the same contributing factors over and over; however, the end result may be considered a way of satisfying a variety of different views regarding what is important in being considered a "best" season. As a player, you don't make the cutoff for consideration unless your season ranks in the top 10 in at least one of the above categories - you are then given ranking points for your relative position in each season-parameter category (10 for first, down to 1 for tenth). A total of 25 player-seasons was considered. The usual 3-4 Ruth seasons occupy the top positions, with Ruth's 1921 season rating #1 (total points = 92, average ranking of 1.2 per factor).

McGwire's 1998 season comes in at #7, strongly supported by a #1 ranking in HR% and a #5 ranking in RG. However, if the HR% is put in the context of POP (see Sluss paper), then the #5 ranking would drop to #9. Also, a reasonable case can be made for dropping McGwire's

RG #5 ranking out of the top 10 (to #13), given the discussion of McMartin in his paper). Even with these arguable "adjustments," Mr. Gilbert's consensus ranking would still end up putting McGwire's 1998 season at #9.

Summary

All in all, the two consensus ranking methods (McMartin and Gilbert), both of which do not make any adjustments for park or league norms, do a pretty fair job of identifying the top 10 offensive seasons, at least 9 of them (see Table below). I base this on my own comparison using single season Adjusted LWR "Wins" (Total Baseball) as my benchmark - there is something in me that just wants to put things in relative terms where logical "adjustments" can be made. Total Baseball provides rankings for adjusted batting wins, but I have added in the "stolen base runs" to the calculation to better match the concept that Mr. McMartin was trying to achieve with his RG approach. McGwire's 1998 season placing at #7 and #10 in the consensus rankings compares quite well with his (dare I say "real") ranking of #9 (tied) based on Adjusted LWR Wins.

Consensus Rankings vs Adjusted LWR Wins

Ranking	Gilbert Consensus (3)	McMartin Consensus (10)	Adjusted LWR Wins* (Total Baseball)
1	Ruth 1921	Ruth 1921	Ruth 1921T-1
2	Ruth 1920	Ruth 1923	Ruth 1923T-1
3	Ruth 1923	Ruth 1920	Ruth 1920T-3
4	Ruth 1927	Gehrig 1927	Gehrig 1927T-3
5	Williams 1941	Ruth 1927	Ruth 1927T-3
6	Gehrig 1927	Williams 1941	Mantle 1957
7	McGwire 1998	Ruth 1924	Ruth 1926
8	Ruth 1924T-8	Foxx 1932	Ruth 1931
9	Ruth 1926T-8	Ruth 1926	Williams 1941T-9
10	Foxx 1932	McGwire 1998	McGwire 1998T-9
11	Williams 1957	Ruth 1931	Hornsby 1922
12	Hornsby 1925	Williams 1946	Gehrig 1934
13	[**]	Williams 1949	Ruth 1924
14		Ruth 1930T-14	Hornsby 1924
15		Gehrig 1936T-14	Williams 1942T-15
16		Walker 1997	Williams 1946T-15
17		Gehrig 1934	Mantle 1956T-15
18		Hornsby 1922	Bonds 1993T-15
19		Musial 1948T-19	Cobb 1917
20		Mantle 1957T-19	Mantle 1961

* Adjusted LWR (BR + SBR) converted to Wins

** Gilbert ranked 25 players total, but players ranking from #13 to #25 only "scored points" in 1 or 2 of the 10 possible categories, making these rankings after #12 meaningless.

The only "consensus" top 10 season that may be somewhat "overrated" appears to be Foxx's 1932 season, which does not place in the top 25 for Adjusted LWR Wins; this is not surprising given the lack of any league or park corrections in the McMartin and Gilbert analyses. Similary the relatively high ranking (#16) of Larry Walker's 1997 season by the McMartin consensus is more than a bit unjustified, given that it would not even rank in the top 100 Adjusted LWR Win seasons after adjustment for park and league norms. In contrast, Mantle's 1957 season (NOT his 1956 triple crown season), ranked #6 according to Adjusted LWR Wins, gets short shrift in the consensus rankings of Gilbert (unranked) and McMartin (#19).

Some final comments on Mr. Sluss's POP statistic. I would really like to see his method applied to some version of Adjusted LWR Wins; the latter values are very much "bunched" regarding the top 10 or top 20 individual offensive seasons (note the 4 groups of "ties" in the top 20 in the Table below). POP appears to be a way to spread out and distinguish these great seasons from one another much more than the traditional data would indicate. Since POP deals with frequency (rate) of production as well as the amount, Adjusted LWR Wins would need to be converted to "Wins/Plate Appearance" or some sort of ratio, before applying the POP calculations - this is beyond my limited mathematical prowess, but I offer this as a possible presentation subject for Mr. Sluss at next year's national meeting. Another area of useful

application of POP would be to identify "career years" from a given player's entire playing time. Where a player's career spans high and low scoring eras, simply looking at the traditional data (RC, LWR, BA, etc.) doesn't necessarily make a career year jump out at you. Using POP, this should be relatively easy, if not at least providing ammunition for lively discussion. I wonder how Norm Cash's 1961 season would stand out from the rest of his career using POP? I always thought that John Olerud (based on his 1993 season) was going to be a Norm Cash clone in this respect, at least until he rejuvenated himself in the National League.

Appendix

Derivation of Total Runs and Net Runs relationships (OPS^P and OPS^L = OPS for player and league, respectively; PA = plate appearances):

$(2 \times OPS^P/OPS^L - 1) \times (0.13\text{-}0.135\ R^L/PA^L) \times \approx 650\ PA = 170 \times OPS^P/0.775 - 85 = 220 \times (OPS^P) - 85$ (1999)

For periods where run scoring was scarcer than 1999, e.g., mid-late 1960s, the following relationship is more appropriate:

$(2 \times OPS^P/OPS^L - 1) \times (0.10\text{-}0.11\ R^L/PA^L) \times \approx 650\ PA = 130 \times OPS^P/0.650 - 65 = 200 \times (OPS^P) - 65$ (mid-late 1960s)

Tom Howell, 2614 Sunnyside Ave., Langhorne, PA, 19053-1924, Thomas_J_Dr._Howell@rohmhaas.com. ♦

BRJ Wants Your Contributions

Mark Alvarez, SABR Publications Director

This is an invitation as well as a request for help.

I want to get more and better statistical pieces into the *Baseball Research Journal.* I'd be pleased to receive papers that have previously been published in *By the Numbers* or elsewhere where a majority of SABR members won't have seen them. I'd be equally pleased, of course, to receive new work. The key to me is to upgrade statistical pieces in SABR's primariy publication so that they (1) say something interesting (2) in a manner understandable to all (3) using sound assumptions and calculations.

The background is pretty simple. I'm not especially interested in statistical analysis, and I'm very bad at telling good stat work from bad or repeated or old-fashioned stat work. Members who like stats and understand them and want to see good articles using statistics have not been happy with what I've published (although they have been remarkably polite – to my face at least!). I want to make them, as a significant and important part of our membership, happy. So I'm looking for every possible way to find good articles and weed out bad ones. I'd like the members of the Statistical Analysis Committee to get as involved in this as they would like to get. The most direct possible involvement would be for them to send me their own work, especially those pieces that have already been seen and commented on by their committee colleagues.

Mark Alvarez
malvarez@wtco.net

Catchers – Better as Veterans

Tom Hanrahan

A catcher's main tasks are to call the game, handle the pitchers, and frame the strike zone, skills which are not rated by traditional statistics. Here, the author investigates, and finds that as a catcher moves from rookie to veteran with the same team, the ERA of the pitching staffs he handles improves dramatically.

Introduction

A catcher's main job, everyone knows, is to call a game and handle the pitching staff. Everyone may know this, but in a game that has statistics for virtually everything, there seems to be precious little time and energy devoted to measuring how well catchers do their main job. Rather, we see catchers' defense measured by how many base stealers he throws out, and all of his other defensive skills ("framing" the pitch, setting up the hitter, bringing along the pitcher) are defined anecdotally by the TV announcers. This study is an attempt to determine if catchers' defensive abilities as a whole improve as they mature and adjust to a pitching staff, and to quantify this as much as is possible.

We start by asking the question "what general factors might affect a catcher's ability to handle a pitching staff?" I suggest that his ability might vary with:

- His age and experience
- His familiarity with the pitchers he is catching
- His familiarity with the batters his pitcher is facing

There may be other specific factors like special tutoring under a specific coach, but these kinds of things will not help us answer the question in general. Tests could be set up to measure any one of these. My focus in this study was to see if catchers' defense improved in the whole, as they became familiar with a pitching staff. At the end of this paper, I will look at how the findings here might be dissected into the individual factors.

The Study

How do we best measure a catcher's defensive abilities? I propose the only reasonable answer is the ERA of the team for which he is catching. How could we best isolate the catchers' defensive ability from all of the other factors that cause a team ERA to rise or fall? I attempted to do this by using all of the teams that had the same *primary catcher in consecutive seasons*. I defined "primary" as having caught at least 85 games during a season. I used the years 1946-1987; beginning after the players returned from WWII, and going through the last year in my Baseball Encyclopedia. No adjustments were made for strike years or change in length of the season. I did not use years when a team changed cities (Brooklyn to L.A.).

There were 104 catchers used in the study. The total number of catcher years was 539, representing about 60,000 total games caught. This gave me a large set of matched pairs of teams in consecutive years using the same catcher. I found the team ERA for each year, and compared it to the league average. I also recorded the number of games the team's catcher had caught in his career prior to the start of the season. Obviously there was always some movement of pitchers between years, some hurlers improving or declining, changes in the team defense at other positions, and changes in ballpark dimensions. But if I could get a large enough sample that all of these other factors got washed out in the noise, I would be able to see if the number of games caught by the catcher was an important contributor to the team ERA.

As an example, we will use Bob Boone's career. He caught at least 85 games every year except the strike year of 1981.

In his rookie year, with Boone having caught only 14 games prior to 1973, the team ERA was .33 runs per game higher than the league average. In his second season, this improved slightly to .30 higher than the league average; just .03 runs per game better. Continued improvement was shown in the next two years, after which there was a meandering slow drop off until he retired. We have 7 pairs of years while he was with the Phillies (73-73, 74-75…. 79-80), and 5 pairs of years with the Angels. Even if he had caught full time in 1981, the pair of years 80-81 would not be used in the study because he switched teams (and pitching staffs).

Organizing the Data

I built two groups of data. In the first, I grouped the year-pairs in bins of hundreds of career games caught: 0-99, 100-199, etc. I only used those pairs of consecutive years where the catcher's games crossed from one to the next. Thus, we can use Boone's 73-74, but not 77-78, because he crossed right through the 500's. This grouping was used to focus on changes from one year to the next, so I could build a function over time. Controlling the number of games caught (by 100's) allowed me to use that as the variable that could link one group to the next. There were 306 matched year-pairs using this method. I think this method gives a good deal of organization to the data (it's easy

Table 1 – Bob Boone's Career

Catcher	Year	Team	Previous Games Caught	Team ERA	League ERA	Difference from League	Difference from Previous Year	Difference from 3 Years Ago
B Boone	1973	PHI	14	4.00	3.67	+0.33		
B Boone	1974	PHI	149	3.92	3.62	+0.30	-.03	
B Boone	1975	PHI	295	3.82	3.63	+0.19	-.11	
B Boone	1976	PHI	387	3.10	3.50	-0.40	-.59	-.73
B Boone	1977	PHI	495	3.71	3.91	-0.20	+.20	-.50
B Boone	1978	PHI	626	3.33	3.58	-0.25	Not used	-.44
B Boone	1979	PHI	755	4.16	3.73	+0.43	+.68	+.83
B Boone	1980	PHI	872	3.43	3.60	-0.17	-.60	+.03
B Boone	1982	CAL	1085	3.82	4.07	-0.25	Not used	Not used
B Boone	1983	CAL	1228	4.31	4.06	+0.25	Not used	Not used
B Boone	1984	CAL	1370	3.96	3.99	+0.03	-.22	Not used
B Boone	1985	CAL	1507	3.91	4.15	-0.24	Not used	+.01
B Boone	1986	CAL	1654	3.84	4.18	+0.34	+.58	+.09
B Boone	1987	CAL	1798	4.38	4.46	+0.08	-.26	+.05

to use and see the trends), but I did lose a few of the samples.

Secondly, I tired to compare rookies to veterans directly by comparing years that were somewhat further apart. In the next grouping, I again organized the data into hundreds bins, but this time I compared them not to the previous year, but to their record 3 years ago, having caught for the same team for 4 consecutive years. I did not control how many career games the catcher had 3 seasons prior. Going back to the Boone example, between 1973 and 1976 the Phillies' ERA improved relative to the league by .73 runs per game (from .33 to -.40). So, I recorded one data point for catcher with career games caught in the 300s, a team ERA of -.73 compared to 3 years ago. This second grouping contained fewer points, because not as many catchers started at least 85 games for the same team for this length of time. I chose 3 years as the comparison point because the more years apart, the less data there is, so using a longer time span would be difficult, and the results of the first data grouping suggested that a 3-year span would show noticeable differences.

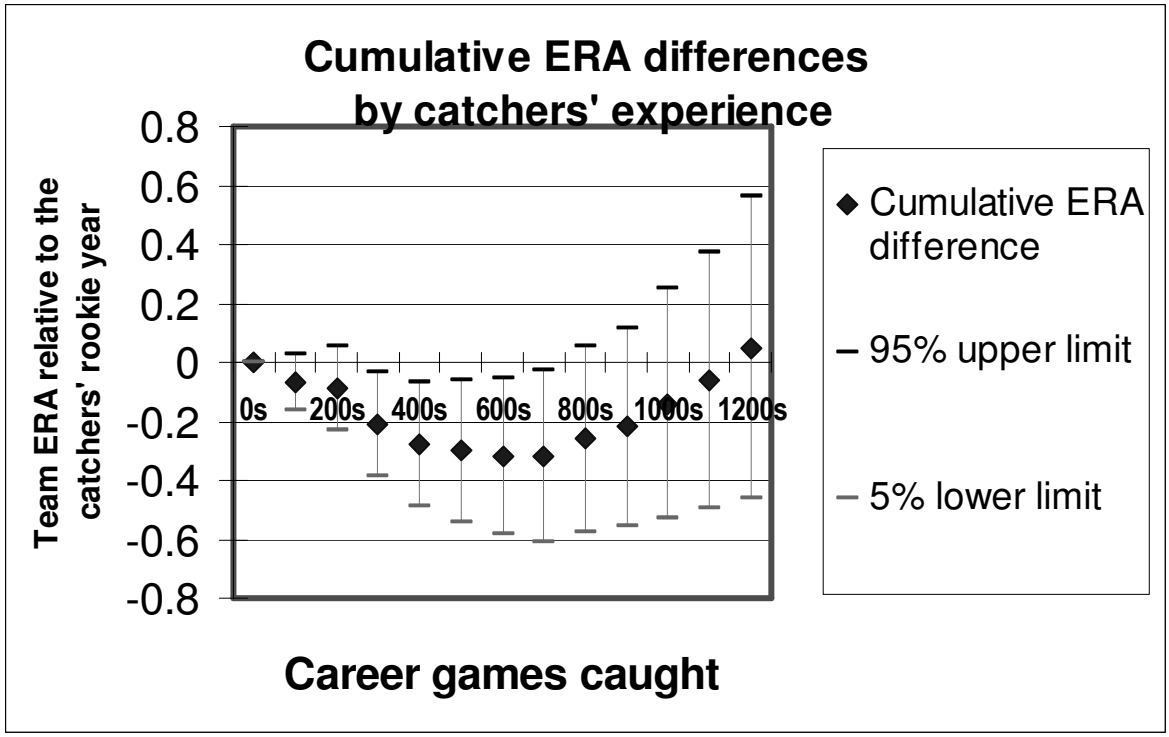
After trying this 3-year comparison, I wound up focusing exclusively on comparing raw rookies to veterans, since this is where the most obvious differences appeared.

Comparing Consecutive Seasons (Grouping One)

I found 49 consecutive year-pairs where the catcher's career games caught went from 99 or less to between 100-199. The average team had an ERA of .07 runs per game lower when the catchers had the extra year (= 100 games) of experience. Table 2 shows the data from every bin. As the amount of data became small for catchers with over 1000 games, I combined the last groups to ensure my sample sizes were at least 15.

Table 2 – Effects of Experience on Catcher ERA

Career Games Caught Between the 2 Years	Number of samples (catcher year-pairs)	Average ERA difference (lower is better)	Standard deviation (SDEV) of ERA differences	Cumulative ERA difference
000s – 100s	49	-.07	.40	-.07
100s – 200s	44	-.02	.43	-.09
200s – 300s	39	-.12	.40	-.21
300s – 400s	29	-.04	.37	-.25
400s – 500s	25	-.02	.36	-.27
500s – 600s	28	-.02	.36	-.29
600s – 700s	22	.00	.34	-.29
700s – 800s	20	+.06	.34	-.23
800s – 900s	15	+.04	.25	-.19
900s – 1000s and 1000s – 1100s	19	+.08 per year	.37	-.11 (1000s) -.03 (1100s)
1100s – 1200s thru 1600s – 1700s	16	+.11 per year (very limited sample)	.27	Sample is too small to make projections for 6 years



The data in Table 2 strongly suggest that the defensive ability of the catchers improves steadily until they have caught somewhere between 400 and 800 games with the same club in the major leagues. The team ERA drops about three tenths of a run per game from the time they have their first full season until they reach this level of maturity. After this there is a slow rise in the team ERA until the catcher retires.*

* Statistician's Corner (ignore this if you're not into the real technical math stuff): The variation (or "noise") in measuring ERAs from year to year is measured by the SDEV (column 4). This typically was about .35 to .40 runs per game. We can use this to measure how certain we are that the average ERA difference is not just a random chance happening, assuming the data is normally distributed, which seems to be a

The right most column of Table 2, cumulative difference, is plotted as chart 1. The bars show the standard deviation of the cumulated difference in ERA from year 1 through the year plotted.

Comparisons Over 3 Years (Grouping Two)

Table 3 shows the catcher year-pairs organized by 100s bins in a different manner. The 280-300s row shows that there were 14 catchers we could use to compare the team ERA between the year when they had between 279-399 games caught under their belts, to the team ERA 3 years prior to that. The average number of games caught in each career 3 years prior is shown. The first row indicates that after 3 years, the team ERA averaged .28 runs per game lower. It also shows that of the 14 teams represented, that 12 of the 14 had a lower team ERA (relative to the league average) when the catcher was a veteran of 280-399 games, as compared to 3 years prior when he had only caught an average of 27 games in his career.

The data in Table 3 is pretty much in agreement with that in Table 2; significant improvement in team ERA the first few years, and a slow decrease in performance toward the catcher's later years. The item that jumped right out at me was the first 2 rows of the right hand column. Out of 38 teams, 32 of them had ERAs that were lower with the catchers who had an extra 3 years of experience. With all of the changes that likely occurred in the team pitching staffs and other defensive changes over the years, this strikes me as remarkable that about 85% of the teams would improve their pitching.

Table 3 – Effect on ERA with games caught

Career Games Caught entering the latter year	Average of Career Games Caught 3 years prior	Number of Samples (Catcher year-pairs)	Average ERA difference (lower is better)	Number of Teams with lower / higher ERA after 3 years
279* - 300s	27	14	-.28	12 / 02
400s	82	24	-.38	20 / 04
500s	172	33	-.15	23 / 10
600s	About 280	26	-.10	16 / 10
700s	About 380	22	+.04	08 / 14
800s	About 480	19	+.16	06 / 13
900s - 1000s	About 630	24	-.02	12 / 12
1100s - 1700s	350 fewer	28	+.07	11 / 17

* 279 was the minimum number of career games caught for any catcher who also was a starting catcher 3 years ago.

As I studied at the 38 catcher seasons involved here, and noticed that the trend was even stronger when using just the catchers who had virtually no previous major league experience. So, I organized the data one last time, using ONLY the catchers who had VERY little experience (less than 50 games) prior to their first full-time year, and making comparison to their "prime" years. Table 2 showed that the catchers' prime seemed to be when he had caught between 400 and 799 games (this is where the cumulative ERA was the lowest). I found all catchers who

a) caught at least 85 games in a season, having had 50 or less career games coming into that year AND
b) caught at least 85 games in other seasons, with the same team, having 400-799 career games caught before these other seasons.

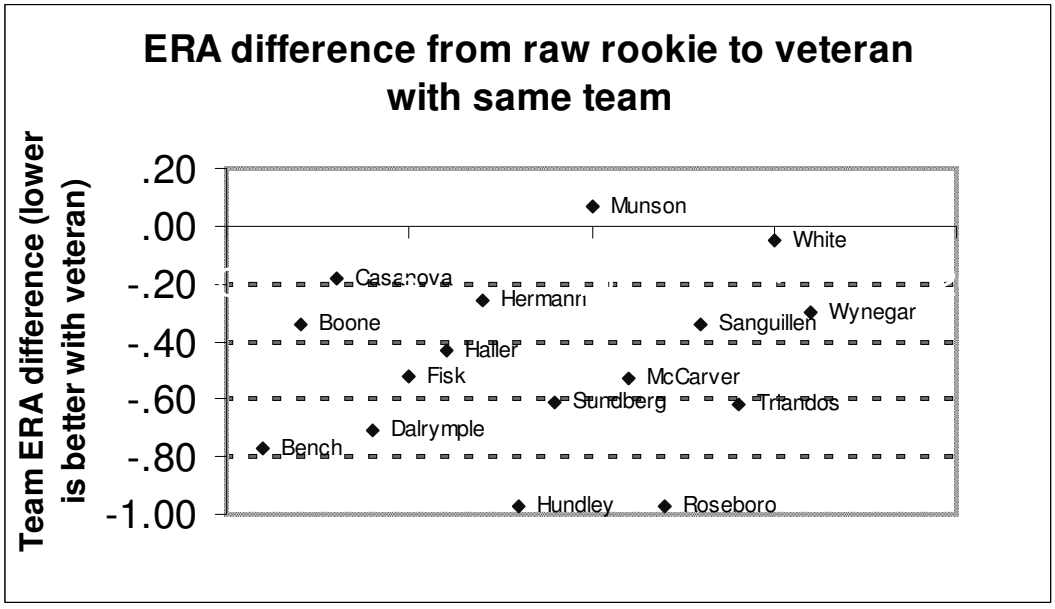
There were 16 comparisons. The teams, "rookie" years and catchers used were:

```
National:     LA  58 Roseboro            Phi 61 Dalrymple    SF 62 Haller
```

reasonable assumption in this instance. In the first row, the average ERA difference is -.07, and is based on 49 samples. The SDEV *of the average* is found by dividing the sample SDEV by the square root of the number of samples; in this case, .40/root(49) = .06. So, the average ERA difference is -.07, plus or minus .06. We can create a "confidence level" that states that the average ERA difference is between -.13 and -.01 with 68% certainty, or between -.17 and .05 (plus or minus 2 standard deviations) with 95% certainty. If we wanted to compare row 1 with row 5 (and we certainly do), by adding the effects we see the cumulative ERA difference is -.30. The SDEV of the cumulative difference is found to be .15 by taking the root sum square of each row's SDEV of the average. Thus, the difference in team ERA between the catchers with less than 100 games caught and those with 400-500 games caught is probably (68% certainly) between –.45 and -.15, according to this set of data. This means that we are not entirely sure that this effect is real, or how large it is, just from this set of data organized in this manner.

	StL 63 McCarver	Chi 66 Hundley	Cin 68 Bench
	Pit 69 Sanguillen	Phi 73 Boone	
American	Bos 52 White	Bal 56 Triandos	Was 66 Casanova
	Chi 69 Hermann	NY 70 Munson	Bos 72 Fisk
	Tex 74 Sundberg	Min 76 Wynegar	

I recorded the team ERA (relative to the league) in the rookie year, and the average team ERA of all years used in the "prime veteran" classification. Of the 16 teams, only ONE had their ERA get worse when the catcher went from rookie to veteran status; fifteen teams had better ERAs with the veteran catchers. The average improvement was .47 runs per game, or 76 runs over a 162 games season! This is very likely a larger difference than importing Ozzie Smith, Willie Mays or Bill Mazeroski in their primes to help your defense. It's even more remarkable when you consider that the ERA comparisons are for the whole season, including the games these catchers did NOT start. Many of these catchers caught three fourths or less of their team's games, so the improvement per game caught might be 30%-40% more! The data for these 16 teams and catchers are graphed in the chart below:



I went back and checked to see what each of these 16 teams' ERA was in the year PRIOR to these catchers being rookies, just to make sure that what I was seeing here wasn't some strange effect, such as a group of all-world defensive catchers (there WERE some mighty fine names in this bunch) helping their teams tremendously while they were in their peak years. Well, these teams had their ERA go UP relative to the league an average of .22 runs per game in the year that they were full time rookies (the years given above in the list). In other words, in their first year, these catchers appeared to hurt their team defensively by a fifth to a quarter of a run per game. Then, over the next 2 to 5 years, their defensive skills improved enough to help their team ERA to go down by almost half a run a game, so their was some net improvement comparing their prime years to the year before they showed up.

I have now noticed that one of the teams in this study was the '58 Dodgers, who moved into a vastly better pitchers' park in 1962, so we shouldn't be surprised that the team ERA improved so much with Roseboro catching as he became a veteran. Still, tossing out one data point won't make that much difference.

Objections

Let's play devil's advocate.

"Maybe this sample is too small and we're seeing some random chance effects."

> Overruled. Already covered this; there's too much data here. When 15 of the 16 teams improve over time…well, if you flip 16 coins, 15 of them will come up heads less than 1 time in 3,800.

"What if the catchers represented an anomalous group of some kind?"

In the second grouping (comparisons over 3 years), obviously catchers who washed out of the majors didn't factor in, since they never reached veteran status. So one could argue that maybe these were the catchers who DID learn how to call a game, and the others did not. But, in the first grouping, we used consecutive year-pairs across every level of games caught, and the same pattern was evident. Overruled again.

"Park factors? Moving over time to a pitching-dominated era? Great hurlers flocking to these teams for a chance to pitch to these guys? A disproportionate amount of good teams and/or catchers in the sample?"

We compared everything relative to the league and within the same teams, to get rid of park and trend effects. These guys were good catchers on good teams…which came first, the chicken or the egg?

General Conclusions

A typical catcher handles a pitching staff better over the course of his first few years in the majors with a club. This is evident by the rather dramatic drop in the team ERA of about a third of a run per game from his rookie season to his prime years with a club.

Specific Conclusions

If you have a veteran catcher who has been with your team for some time, and you're thinking of trading him and calling up the young phenom from AAA, you can expect your pitching results to get worse. Of course, you ought to call him up *sometime*, but don't expect the team to improve right away. How many catchers are offensively 50 runs a year better than their replacement? (Piazza begins and ends the short list).

The differences in catchers' stolen bases allowed are apparently *less* important than his other defensive abilities. The worst throwing catchers in the majors do not allow anywhere near one stolen base per game more than Ivan Rodriguez does.

Why does this happen?

Is it aging, working with certain pitchers, or what?

Well, I don't know; but I think a few of us ought to try to find out.

If I were to try to measure the effectiveness of catchers by age, I would try to control all other factors, and see if catchers how catchers' team ERA varied with the age they broke in as rookies, the ages where they all had an even number of games experience, etc. I do think that this study showed that they was a definite falling off in performance as catchers got older, and so I think we could safely conclude that after a certain age his skills diminish. We ought to be able to measure how much of this degradation in higher team ERA is caused by a lessened throwing ability, or propensity for more passed balls, and infer that any remainder is the ability to handle the staff.

Measuring how much changing teams or leagues could also easily be done. Just check how a team's ERA changes when they swap one veteran catcher who has been with the team for a while with a veteran from a different team. If you limit the study to catchers within the same league, the factor of being familiar with the batters is negated. You could study how importing a receiver from the other league in much the same way.

Even if we were able to isolate the conditions when this happens, we still may not know WHY. Do the veteran catchers "frame" the pitch better? Is it their pitch selection? Something else? I'm afraid the answer to that one may be beyond the reach of mere statistical study, and involve more hands-on research with major league coaches and batteries. But this is only my uneducated opinion.

Other Studies

Craig Wright (and Tom House) wrote a book entitled *The Diamond Appraised*, which is excellent reading. In fact, in hindsight I think reading his research inspired me to do mine. Its first chapter was called Catcher ERA. Craig measured individual catcher's ability to handle a staff by comparing the team ERA in the games when they started as compared with the games when another man was behind the plate. This is probably the best way to quantify a catcher's ability, because you get to hold every other variable (different pitchers or team) constant. He produced many numbers, finding that certain catchers seemed to have definite advantages over others. He also found that "catching is far more a learned skill than a raw talent" and that "rookie catchers show a defensive improvement in their….sophomore seasons." Again, his emphasis was on individual catchers, not on kinds or types, and so there are no hard data from his study in this area, and I hesitate to quote extensively from someone else's work. His general conclusions seem to be right in line with this study.

I prepared an article for the Baseball Analyst in 1988, which dealt with some of these questions. The only data I had available was the annual Bill James Abstract, which listed the current major league catchers and gave the ERA of their team when they caught, along with the overall team ERA. The data I used was only for the years 1982-87. My conclusions from that study were that there was a difference in catcher ERA from rookies to veterans, and that having a veteran catcher swap teams or even leagues didn't seem to have much effect.

Finale

If differences this large show up comparing CLASSES of catchers, does this not suggest very powerfully that the differences of *individual* catchers might be even greater? And if so, wouldn't this be the single most important yet unanalyzed ingredient of determining team success in the game today? Brethren, I issue a call to you all: either prove me wrong, or let us declare that there's gold in them thar hills.

Tom Hanrahan, 21700 Galatea St., Lexington Park, MD, 20653, HanrahanTJ@navair.navy.mil. ♦

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.

Hits and Baserunner Advancement

Dan Levitt

How often do runners score on base hits to various fields? How often are they thrown out by outfielders? Here, the author answers these questions based on Retrosheet play-by-play data.

What happens to the runner on second when the batter hits a single to center; how often does he score? What about the runner on first when the batter doubles to left; does he usually score? The publicly available Retrosheet files for the years 1980 through 1983 allow us to take a look at these questions. Tables 1.1 through 1.3 below summarize where runners end up after a single.

Table 2 – Single With Runner on Second

		<- Runner Destination—Percent ->				
Fielder	# of hits	Out	Second	Third	Home	Bases/hit
NA	1052	1.5%	2.9%	41.5%	54.1%	1.51
LF	5117	4.5%	0.2%	26.9%	68.4%	1.68
CF	5476	2.4%	0.1%	14.8%	82.6%	1.82
RF	4374	4.3%	0.0%	23.9%	71.7%	1.72
INF	2380	4.2%	7.3%	76.6%	12.0%	1.05
Total	18399	3.6%	1.2%	29.9%	65.3%	1.64

Table 3 – Single with Runner on Third

		<- Runner Destination—Percent ->				
Fielder	# of hits	Out	Second	Third	Home	Bases/hit
NA	541	0.2%		2.2%	97.6%	0.98
LF	2840	0.1%		0.0%	99.9%	1.00
CF	3080	0.1%		0.0%	99.9%	1.00
RF	2566	0.1%		0.0%	99.9%	1.00
INF	1106	0.4%		7.2%	92.4%	0.92
Total	10133	0.1%		0.9%	99.0%	0.99

Technical Notes: Totals are for the four seasons 1980 through 1983. The baserunner situations are not mutually exclusive. For example, the runner on first situation makes no reference to runners on other bases--runners may or may not also be on second or third. For the average baserunner advancement calculation (Bases/1B), one base advanced is credited in the case of a runner put out on the basepaths. NA in the Fielder category indicates the data was not available in the Retrosheet database. Fielder is defined as the player who fields the ball. These same notes apply to the doubles tables below.

Table 4 – Doubles with a runner on first

		<- Runner Destination—Percent ->				
Fielder	# of hits	Out	Second	Third	Home	Bases/hit
NA	673	1.5%		50.8%	47.7%	2.46
LF	3118	2.6%		56.9%	40.5%	2.38
CF	1160	4.6%		36.8%	58.6%	2.54
RF	2021	3.6%		58.8%	37.7%	2.34
INF	25	8.0%		80.0%	12.0%	2.04
Total	6997	3.1%		53.6%	43.3%	2.40

Table 5 – Doubles with a runner on second

		<- Runner Destination—Percent ->				
Fielder	# of hits	Out	Second	Third	Home	Bases/hit
NA	399	0.0%		1.5%	98.5%	1.98
LF	1975	0.2%		0.7%	99.2%	1.99
CF	772	0.1%		1.6%	98.3%	1.98
RF	1317	0.2%		2.4%	97.5%	1.97
INF	18	0.0%		11.1%	88.9%	1.89
Total	4481	0.1%		1.4%	98.4%	1.98

The baserunner advance differences between hit locations can be quite dramatic, especially in the case of moving from first to third. A runner from first goes to third nearly half the time on a single to right; this falls to less than 20% on singles to left and 10% on infield singles. A runner on second scores in the 68% to 83% range on any outfield single. Baserunners get thrown out around 2% to 4% of the time, and it appears almost 50% more likely in the case of a runner trying to score from second on a single (3.6%) than trying to move from first to third (2.1%).

Tables 4 and 5 summarize the baserunner results of a double.

In the case of a double, the differences in baserunner advancement are less extreme. Even here, however, a runner from first will score around 40% of the time on a double to left or right but nearly 60% of the time on a double to center. Trying to score from first on a double to center also has the highest incidence of getting thrown out (4.6%) of any of the major categories; nearly as high are runners trying to score from second on a single to left or right.

The next step involves examining how specific players compare with the averages, both as hitters and baserunners. Hopefully, I can get to that analysis in the not too distant future.

Dan Levitt, 4401 Morningside Road, Minneapolis, MN, 55416, danrl@ibm.net. ♦

Evaluating Pitchers' Winning Percentages: A Mathematical Modeling Approach

Rob Wood

If, for instance, a pitcher has a .550 record on a .530 team, what would his record be on an average team? In this study, the author examines and analyzes several published attempts to answer the question, and presents a new model of his own.

Introduction

A pitcher's won-loss record is often considered to be a true indication of how good a season the pitcher had. This article examines the proper evaluation of a pitcher's won-loss record, taking into account the context in which it was achieved. In particular, the quality of his team surely affects a pitcher's won-loss record. Many Cy Young winners won the award largely based upon their offensive support.

Coincidentally, in the previous issue of By The Numbers (May 1999), Bill Deane and I had separate articles dealing with this issue. Deane has developed a formula to project how a pitcher might perform (in terms of won-loss percentage) based upon his own winning percentage and that of his team. About the same time Bob McCleery and I developed a similar formula.

In this article, I want to compare the two formulas and introduce a third formula. The Deane and McCleery-Wood formulas are essentially ad hoc. They seem to give reasonable projections; however, they are not based upon underlying assumptions which could be examined or tested.

My new formula is based upon a mathematical modeling perspective. I first start with a series of reasonable assumptions. Then, based upon these assumptions, I algebraically derive a prediction of a pitcher's true quality, and a prediction of his winning percentage if he pitched on a .500 team. In the final section, I will compare the predictions of the three formulas and draw conclusions.

A Pitcher's Winning Pct

A pitcher's won-loss record, especially over only one season, can be subject to many influences besides the quality of the pitcher's performance. Many pitchers with gaudy won-loss records were fortunate enough to pitch for a great offensive team (and received inordinate offensive support), a great defensive team, or just got real lucky.

In order to give many variables the chance to "even out" over time, this article will focus on the evaluation of a pitcher's career won-loss record. However, there is nothing in the formulas that says they are not equally valid on a seasonal basis.

In this article, I hope to shed light on the information that may be available in a pitcher's won-loss record. I will develop a formula based solely upon standard sabermetric assumptions that estimates the pitcher's true quality given only his winning percentage and his team's winning percentage.

Previous Approaches

Bill Deane developed his Normalized Winning Percentage almost twenty years ago. Recently, Deane reprised his NWP formula and presented the NWP leaders for 1998. (See By The Numbers, May 1999, pp. 6-7).

To understand the formula, we will need to define a few variables. Let WPCT denote the pitcher's own winning percentage (a number between 0 and 1). Let MATE denote his teammates' winning percentage; that is, the winning percentages of his team after removing the decisions of the pitcher in question (again, a number between 0 and 1).

Then, Bill Deane's formula to project what a pitcher would do on a .500 team is:

```
NWP = .500 + [(WPCT-MATE)/(2*(1-MATE))]
```

This formula holds for all pitchers for whom WPCT is not less than MATE.

Bob McCleery and I also introduced a formula dealing with pitchers' winning percentages. (See By The Numbers, May 1999, pp. 18-19). Our formula was also an attempt to take into account the pitcher's teammates. We compared each pitcher's winning percentage to a pitching standard we defined. We thought that a reasonable standard would be a combination of his teammates' winning percentage and .500, with the weights being 2/3 and 1/3, respectively. Thus, our formula to project what the pitcher's winning percentage would be if he pitched on an otherwise .500 team is:

```
MCW = .500 + [WPCT - ((.667*MATE)+(.333*.500))]
```

Below, I will compare the implications of Deane's NWP formula to the McCleery-Wood MCW formula. Suffice it to say here that the McCleery-Wood formula is more "generous" to pitchers who pitch for good teams. As a prototypical example, Lefty Gomez is considered to be just about a league average pitcher according to Deane's formula since Gomez's career winning percentage of .649 is nearly identical to his Yankee teammates' winning percentage of .646 over his career. Deane's NWP formula predicts that Gomez would have had a .505 career winning percentage had he pitched for a .500 team.

On the other hand, the McCleery-Wood MCW formula gives Gomez "credit" for exceeding the .500 mark, even though he could not significantly better his Yankee teammates' winning percentage. Indeed, Gomez is considered to have a winning percentage above his standard by .052 for his career, equivalent to predicting that Lefty would have had a .552 career winning percentage had he pitched for a .500 team.

Both approaches, it should be clear by now, are reasonable formulas that are fun to play with. However, both approaches are essentially numerical in nature having little theoretical foundation.

New Mathematical Modeling Approach: Assumptions

In this article, I would like to develop a formula from first principles, so to speak. Everyone can then see what assumptions lie behind the formula. Sabermetrics is at heart an inferential science. By looking at baseball statistics, we can infer something about baseball, its players, managers, strategies, etc. My new approach will also be inferential. Given some data on a pitcher (and his team), I will set up a set of reasonable assumptions that will allow us to infer something about the true quality of the pitcher.

Please keep the following task in mind. You are told the winning percentage (wins and losses if you want) of two pitchers (for a season or career, say). You are told the winning percentages (wins, losses) of their respective teams over the same period. You do not know their ERAs or any other statistics. Based solely upon this information, you are asked who is likely to be the better pitcher. Ultimately, you are asked to predict what each pitcher would do on an otherwise league-average team.

To answer this question, I will start with the following assumptions, each one fairly reasonable but a simplification nonetheless.

Assumption 1: Winning percentages are generally predicted by Bill James's Pythagorean Formula. The ratio of the square of a team's runs scored to the sum of the square of its runs allowed plus the square of its runs scored is a good predictor of its winning percentage.

Assumption 2: Runs scored by team X in a game with team Y is equally affected by the quality of team X's offense and team Y's defense (pitching and fielding). By symmetry, this implies that runs allowed are also equally affected by one team's defense (pitching and fielding) and the other team's offense.

Assumption 3: Pitching dominates fielding in preventing runs. I will take this to its extreme and assume that all of "defense" is pitching. This is an assumption that can be relaxed subsequently.

Assumption 4: Teams are perfectly balanced. Each team's offense and defense are equally good, relative to the league average. For example, if a team's offense is 10% better than the league average, then its defense (pitching) is 10% better than the league average. Actually, it is the pitcher's teammates that exhibit this balance (i.e., the batters and his pitching teammates). This assumption is likely to be approximated only over the course of a pitcher's career. We will see the implications of relaxing this assumption below.

Assumption 5: Teams' offensive support is perfectly balanced among their pitching staff. Clearly, this is a strong assumption that is often violated when looking at seasonal data, and is more likely to be valid over the course of a pitcher's career.

From these simplifying assumptions, I can derive an estimate of how good each pitcher is, relative to the league average pitcher, and thereby compare any two (or any number of) pitchers. Some of these assumptions will be deemed too severe. They require a degree of "regularity" that is often lacking in real life baseball.

However, remember our task. You are told only the pitcher's winning percentage and his teammates' winning percentage. In the absence of information, assumptions of balancedness and regularity seem appropriate to me, especially since the implications of relaxing the

assumptions (only permitted via this mathematical modeling approach) can be determined in the event that additional information is available.

New Mathematical Modeling Approach: Formulas

Let QUAL denote the pitcher's true quality, relative to the league average runs allowed per nine innings. QUAL of 1.00 denotes a league average pitcher. QUAL less than 1.00 (say, 0.75) denotes a better than league average pitcher (in this case 25% better). Conversely, QUAL greater than 1.00 (say, 1.10) denotes a worse than league average pitcher (10% worse in this case).

Let W500 denote our prediction of the pitcher's winning percentage if he pitched on an otherwise .500 team.

Before I present the key formulas, let me present two sets of antecedent formulas that each use a different extreme assumption on the team's balance.

Assumption 4A: The pitcher's team's offense is league average. So his team's winning percentage (excluding his own decisions) is solely a reflection of the quality of his pitching teammates.

Under this assumption, the formulas for QUAL and W500 do not depend upon TEAM. With assumption 4A of league-average offense, our formula for QUAL becomes:

```
QUAL_A = SQRT[4*(1-WPCT)/WPCT] - 1.
```

Since under assumption 4A the pitcher's offensive support is league average, his winning percentage is an accurate reflection of his true quality, relative to league average.

```
W500_A = WPCT.
```

Let's now look at the opposite extreme.

Assumption 4B: The pitcher's pitching teammates are collectively of league average quality. So his team's winning percentage (excluding his own decisions) is solely a reflection of the quality of his team's offense.

Under this assumption, the formula for QUAL depends upon both WPCT and MATE. In fact, the formula in this case is similar to the previous case, except now it has a new term. With assumption 4B of league-average pitching teammates, our formula for QUAL becomes:

```
QUAL_B = SQRT[4*((1-WPCT)/WPCT)*(MATE/(1-MATE))] - 1.
```

And the formula for his winning percentage on a .500 team becomes:

```
W500_B = [WPCT*(1-MATE)]/[(WPCT*(1-MATE))+((1-WPCT)*MATE)]
```

The higher is WPCT or the lower is MATE, the lower is QUAL and higher is W500, and the better we deem the pitcher.

For the pitchers for whom WPCT>MATE, W500_A is greater than W500_B whenever MATE > .500, and the reverse is true whenever MATE < .500. This can be seen from both the underlying assumptions of the relative strength of the pitcher's teammates as well as the algebraic formulas.

I wanted to show you these two extreme cases for a few reasons. First, they will prepare you for the type of formulas that come next. Second, they will represent the two extremes which can be used for comparison purposes in the next section. Third, it turns out that the formula for QUAL under the assumption of perfect balance (Assumption 4) is exactly halfway between QUAL_A and QUAL_B; however, the formula for W500 is only approximately halfway between W500_A and W500_B due to the non-linearity in the Pythagorean theorem.

The formulas under the assumption of perfect team balance (Assumption 4) are given by:

```
QUAL = {SQRT[(1-WPCT)/WPCT] * [1 + SQRT(MATE/(1-MATE))]} - 1

W500 = 4 / {4 + [((1-WPCT)/WPCT)*(1+(MATE/(1-MATE))+(2*SQRT(MATE/(1-MATE))))]}
```

Of course, the higher is WPCT or the lower is MATE, the lower is QUAL and the higher is W500, and the better we deem the pitcher.

At this point some readers might wonder about the value of such complicated formulas. They have square roots in them and seem to be lacking in intuition. Please bear with me for a little longer. First, the SQRT operator is a direct consequence of the Pythagorean Formula. Second, I will present a more "fan-friendly" version of the W500 formula below.

Comparison of Approaches

At this point in the article we have in our arsenal 6 different formulas dealing with the evaluation of a pitcher's winning percentage. Let me list them here for completeness:

- WPCT, the pitcher's winning percentage
- NWP, Bill Deane's formula for the pitcher's normalized winning percentage, a projection of his winning percentage on a .500 team
- MCW, Bob McCleery's and my previous formula for the pitcher's winning percentage relative to his standard, expressed as a projection of his winning percentage on a .500 team
- W500_A, estimated pitcher's quality assuming his team's offense is league average (as we saw above, W500_A is simply WPCT)
- W500_B, estimated pitcher's winning percentage on a .500 team assuming his pitching teammates are league average
- W500, estimated pitcher's winning percentage on a .500 team assuming his team is perfectly balanced.

I suppose other QUAL formulas could be developed by taking other linear combinations of QUAL_A and QUAL_B (leading to other W500 formulas), depending upon our views of the pitcher's team's actual balance between offense and defense. I choose not to go down that path for two reasons. First, perfect balance is likely to hold in the long run and surely holds for all pitchers taken together (from a league-wide perspective, each run scored by one team is a run allowed by another team). Second, once one jettisons the assumption of perfect balance, one is forced to come up with another assumption. Possibly team runs scored and runs allowed data could be used to derive appropriate weightings, but this would have to be done on a pitcher by pitcher basis for each season. It might be argued that once it becomes that much work, the whole approach expends more effort than generates insight.

Table 1 presents the top 10 pitchers of all-time according to career W500, along with the projections of the pitcher's winning percentage on a .500 team by all of the methods. Table 1 also includes a seventh "mystery" method that I will describe below.

Table 1: Top 10 Projections of Pitchers' Careers Winning Pct on a .500 Team (min 100 wins)

	WPCT	MATE	NWP'	MCW	W500A	W500B	W500*	Mystery
S.Chandler	717	606	641	646	717	622	669	664
R.Johnson	644	462	669	669	644	678	661	663
M.Mussina	667	513	658	658	667	655	661	660
R.Clemens	653	491	659	659	653	661	657	657
L.Grove	680	564	633	638	680	622	651	648
W.Ford	690	596	616	626	690	601	645	642
D.Gullet	686	589	617	626	686	603	644	641
J.Wood	671	569	618	625	671	607	638	636
G.Alexander	642	511	634	635	642	632	637	636
S.Koufax	655	542	623	627	655	616	635	634

In introducing this new method, I am somewhat reluctant to present the numbers related to real pitchers since everyone has opinions on the quality of the pitchers, based upon personal experience, knowledge of their ERAs, number of seasons pitched, etc.

I beseech the reader to try to put aside that "baggage" and try to answer the following question. If you were presented only with the following pitchers' (career) winning percentages and their respective team winning percentages (excluding their own decisions) over that period, who would you deem to be the best of the bunch? What would be your ranking? Why?

Two comments on the data in the table. First, as described in my previous BTN article, the data in the "Mate" column are estimates. The number of decisions that a pitcher has in each season over the course of his career can vary widely. Thus, calculating an appropriate winning percentage of a pitcher's teammates over the course of his career is problematic. My coarse attempt is shown in the table above. Accordingly, the column labelled NWP' may not exactly correspond to the NWP data presented in Bill Deane's BTN article. The differences are irrelevant to the current discussion. Second, the data in the table are as of the beginning of the 1999 season.

Table 1 demonstrates that the various formulas perform similarly when looking at the top of the charts. There is some differences between the formulas for a few pitchers, but generally speaking they paint the same picture.

To gain a greater understanding of the implications of the formulas, let me present another table with selected pitchers. Table 2 includes a few other pitchers that illuminate differences among the methods.

Table 2: Projections of Other Pitchers

	WPCT	MATE	NWP'	MCW	W500A	W500B	W500	Mystery
L. Gomez	649	646	505	552	649	504	573	576
D. Stieb	562	500	562	562	562	562	562	562
N. Garver	451	399	543	518	451	553	499	502

The case of Lefty Gomez is particularly interesting. Gomez pitched on great teams. His Yankee teams had a .646 winning percentage over the course of his career in games in which Lefty did not get a decision (just about a 100-54 season pace). Gomez was able to match this pace, but not exceed it by very much, and checks in with a career winning percentage of .649 (189-102). Of course, he pitched on staffs with other Yankee stars, including Herb Pennock, Red Ruffing, Johnny Allen, and Spud Chandler.

As mentioned earlier in the article, NWP projects Gomez to be a run-of-the-mill .505 pitcher on a .500 team, since he was not able to better his teammates' mark. You can see that NWP' (505) is very close to W500_B (504), the projection assuming league-average pitching teammates. In this case, the only way to get a team winning percentage of .646 with league-average pitching is if the Yankee offense is tremendous. Thus, you can see that Gomez is knocked down quite a bit under that assumption. Take away his tremendous offensive support, so the argument goes, and Gomez would be lucky to be a .500 pitcher. I don't buy this argument since Gomez's pitching teammates included many great or near-great pitchers.

Historical aside: NWP was developed partly to reward pitchers who pitched on good teams to a greater extent than the then-predominant measure of WPCT-MATE. My comments above imply that I feel that NWP does not go far enough in that direction.

MCW credits Gomez with bettering the .500 mark even though he could not better Ruffing, Chandler, et al., thus projecting Lefty to be a .552 pitcher on a .500 team. Note that there is a large difference between a .505 winning percentage and a .552 winning percentage, more than 7 games over a 154-game season.

W500 projects that Gomez is even a better pitcher than MCW projects. Indeed, Gomez is projected to be a .573 pitcher, a pace that would win more than 10 games above what NWP projects over a 154-game season.

Ned Garver, a decent pitcher who toiled for the Browns, Tigers, and Athletics during the 1950's, serves as another extreme. Garver scratched and clawed his way to a career .451 winning percentage (129-157) on a woeful collection of teams whose collective winning percentage (excluding his own decisions) was .399 (a 61-93 pace). According to my estimates, Garver's team winning percentage is the lowest among all 20th century pitchers with 100 wins.

NWP is impressed that Garver could exceed his teammates' ineptitude, and projects him to be a .543 pitcher on a .500 team, far superior to its projection for Lefty Gomez. MCW is harsher to Garver, reasoning that he is not completely at the mercy of his poor teams (remember Steve Carlton in 1972), and projects him to be a .518 pitcher on a .500 team, significantly inferior to Gomez. W500 goes even further. Garver is projected to be a .499 pitcher on a .500 team; just about the definition of a league average pitcher.

The third pitcher listed in Table 3 is merely a confirmation of the obvious. Over the course of his career, Dave Stieb pitched on a .500 team (excluding his own decisions). Of course, in this case all formulas, including NWP, MCW, and W500, will project that his winning percentage on a .500 team will equal his actual winning percentage of .562.

It can easily be shown that if WPCT > MATE > .500, then W500 > MCW > NWP, and that if WPCT > MATE and MATE < .500, then W500 < MCW < NWP. That is, W500 "goes further" than MCW does in comparison to Deane's NWP formula.

The next section uses this insight to simplify the formula for W500 in an attempt to make it more "fan friendly."

Mystery column

Tables 1 and 2 included a "Mystery" column that will now be explained. Before I do so, though, please review the tables again. You should be struck by how close "Mystery" approximates W500, not only in all the top 10 cases listed in Table 1, but even in those hard-to-match extreme cases listed in Table 2.

Based upon the insight that MCW does not go "far enough", I estimated the best standard against which to compare a pitchers' winning percentage in order to approximate W500. As in the MCW formulation, I consider all standards that are linear combinations of the pitcher's teammates' winning percentage (MATE) and .500. I ran a regression on W500 using the data for all 258 pitchers who won 100 games in the 20th century and whose career winning percentages exceeded their teams. It turns out that the best standard weights MATE by .510 and .500 by .490. Recall that MCW weighted these by .667 and .333, respectively.

In the spirit of "fan-friendliness", I round off both .510 and .490 to be .500. I then created W500_M (where the M stands for "mystery") which reflects this best standard:

```
W500_M = .500 + [WPCT – ((.500*MATE)+(.500*.500))]
```

I am confident that the 50/50 standard reflected in W500_M is related to the assumption of perfect balance. MCW implicitly assumes a 67/33 standard weighting of a pitcher's teammates' winning percentage and .500. Based upon the above findings, I conjecture that NWP implicitly assumes a standard weighting more like 90/10. But as shown above, this is equivalent to assuming that the pitcher's offensive teammates are 9 times better than his pitching teammates (relative to league average). I find this to be a very extreme assumption and unwarranted in the vast majority of cases.

The formula for W500_M is very easy to remember and to use. Table 3 shows how remarkably accurate W500_M is in approximating W500 in comparison to the other formulas we have discussed.

Table 3: Average Absolute Errors of Formulas vs. W500
(sample is all 258 pitchers with WPCT>MATE and 100+ wins in the 20th century)

WPCT	MATE	NWP'	MCW	W500A	W500B	W500M (Mystery)
16.6	45.7	13.5	5.8	16.6	16.4	0.4

Table 3 demonstrates the accuracy of W500_M. Over all 258 pitchers who won 100 or more games in the 20th century and whose winning percentages exceeded their teams, the average absolute difference between W500 (the "true" projection) and W500_M (the formula using a 50/50 standard) is less than 1 point (.0004 in decimal representation). This compares quite favorably to NWP' which has an average error of 13.5, and to MCW which has an average error of 5.8. For comparison, I also calculate the average error in the two extreme versions, W500_A and W500_B; both of these average errors are over 16 points.

The complete listing of all the projections discussed in the article for each of these 258 20th century pitchers is available from the author upon request.

Conclusions

In this article, I have used a mathematical modeling approach to develop a new formula to predict what a pitcher's winning percentage would be on a .500 team, enabling us to compare pitchers on different teams and in different eras. My new formula is a direct consequence of a series of reasonable sabermetric assumptions including the accuracy of the Pythagorean formula, pitching being 50% of baseball, team balance, and balanced offensive support within pitching staffs, over the course of a pitcher's career.

Although the new formula in original form had a square root in it and lacked intuition, I was able to convert it into a form that is easy to remember and to use. A pitcher is projected to be better than an otherwise .500 team by an amount equal to the difference between his actual winning percentage and the average of his team's winning percentage (excluding his own decisions) and .500.

I have compared the implications of the new formula (W500) to those of a couple formulas previously developed by Bill Deane (NWP) and Bob McCleery & myself (MCW). We saw that generally MCW does a better job of approximating W500 than does NWP. This is due to the fact that W500 assumes perfect balance between a team's offense and defense, MCW implicitly assumes that a team's offense is about twice as good as the quality of its pitchers, and NWP implicitly assumes that a team's offense is much, much better than the quality of its pitchers.

Rob Wood is a management consultant in Mountain View, California. The article has benefited from discussions with Bill Deane and Bob McCleery. Rob Wood, 2101 California St. #224, Mountain View, CA, 94040-1686, rob.wood@us.pwcglobal.com. ♦

Submissions

Submissions to *By the Numbers* are, of course, encouraged. 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.)

Deadlines: January 24, April 24, July 24, and October 24, for issues of February, May, August, and November, respectively.

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.

Share Your Data

At the convention in Scottsdale, some committee members suggested that we do something to facilitate sharing of data among members. So, if anyone has raw data they can share with other members (your compilation or public domain), let me know, and I'll run a list in this "Announcements" column. Please be specific about what kind of data you have available, and how others can receive it. (Here's my own example: "Full batting stats against for every 1988 American League pitcher. Many categories, including SF, pickoffs. ASCII text, comma delimited. E-mail phil_birnbaum@iname.com.") Also, if you *need* any specific data, you can place your "data wanted" ad here, although you'd probably get a quicker response on SABR-L.

There was another request for a website list; we'll save that for next issue.

Relief ERA: A New Way to Rank Relievers

Sky Andrecheck

Used for relievers, the traditional ERA statistic doesn't consider inherited runners, and whether or not the relief pitcher is successful in preventing them from scoring. Here, the author suggest a new statistic, Relief ERA, which corrects the omission.

As we all know, one of the toughest things to measure in baseball is relief pitching. ERA's are great for starting pitchers, but aren't so good for relievers. A pitcher can enter with two outs and the bases loaded and allow all three runners to score without a scratch on his ERA – hardly fair to the man who comes in with the bags full and gets out of it. Of course, without play-by-play data, it's tough to account for the differences – however, there is one thing we can do to be more fair.

If any of you get *Baseball Weekly*, which I'm sure many of you do, you've probably noticed that there are two unconventional stats listed for relievers. One is the number of inherited runners scored and the other is the number of inherited runners stranded. These are key stats to note. If a man inherits ten runners and allows them all to score, that's a huge difference between the guy who strands all ten of his runners. In fact, assuming they are the same runners, it's a difference of ten runs. That's one whole win.

When the ERA's and number of runners inherited are equal, it's pretty easy to tell which pitcher performed better. However, who has the upper hand between a man with a 4.33 ERA who inherits 42 runners and allows half of them to score, and a reliever with a 4.97 ERA who inherits 22 runners, allowing only 4 to score?

Using a method similar to Pitching Linear Weights, we can find the number of runs saved or blown via inherited runners. On average, relievers allowed 34.59% of inherited runners to come around to score in 1998. (When I say "inherited runners," I mean IR scored + IR stranded. Any runners still on base when a pitcher leaves the inning are not included in the analysis.) Any pitcher allowing less than 35% of runners to score is saving runs; any pitcher allowing more is squandering them. By finding the number of expected runners scored, and subtracting from the actual IR scored, we can find the number of runs saved or blown. We can then add or subtract that number from his earned runs to calculate a new earned run average which is much more accurate for relievers. The formula is listed here where IRSc is the number of inherited runners scored and IRSt is the number of inherited runners stranded:

$$RERA = \frac{9 * (ER + IRSc - .3459 * (IRSc + IRSt))}{IP}$$

To tell you the truth, I felt a little strange about simply adding this new difference onto regular earned runs – after all, a man with 0 earned runs and 1 runner stranded would have a negative RERA – a little bizarre. However, when I thought about it, regular ERA measures the runs you are responsible for, while inherited runners are someone else's responsibility. While ERA measures only how well a pitcher does for himself, Relief ERA measures both how well a pitcher does for himself *and* how well he can cover the last pitcher's butt – and let's face it, covering the previous pitcher's hide is a big part of relieving.

In fact, it is such a big part of relieving that RERA can differ from regular ERA by up to a point or more. Rookie Sean Runyan's ERA looked pretty good at 3.58 over 50 innings, but after factoring in that he allowed 27 inherited runners to score, while only stranding 25, his Relief ERA put him at a more appropriate 5.19. Scott Radinsky also skyrocketed from 2.62 to 3.62 using Relief ERA. On the other hand, Rick Aguilera dropped from 4.24 to 3.44 and Rich DeLucia's RERA was 1.18 lower than his regular ERA of 4.27.

The top 20 pitchers, selected from the four most used relievers of each 1998 staff, are shown in the table for your viewing pleasure. The numbers are league and park adjusted (based on last three seasons) to a national league team in a neutral park to provide a complete ranking of the most effective relievers of the season. The chart shows the IR scored and IR stranded, along with the number of actual runs saved or lost via performance with inherited runners. The final column is the adjusted Relief ERA.

It's hard to argue with RERA's ranking of baseball's top five relievers – Jackson, Hoffman, Urbina, Rivera, and Nen were all incredibly dominant. However, closers already have a stat for them, though not a perfect one, and so we can evaluate closers by their save and blown save totals. RERA is best for evaluating middle relievers because it improves on ERA – the best tool we have so far. RERA ranked John Rocker as the number one middle reliever of 1998. Not surprisingly, he is now a dominant closer. Bobby Howry, the fourth best middle reliever, now closes games for the White Sox. Doug Brocail rated number two and is having another standout season for the Tigers – in my opinion he could easily best Todd Jones for the closer role if given the chance.

As you look down the list, you may find that the pitchers with good regular ERA's tend to get even better with RERA and likewise with poor pitchers. This is because inheriting runners adds extra responsibility to the pitchers and amplifies their performance, making good pitchers even more valuable and poor ones even less valuable.

Table: 20 Top ML Relievers by Adjusted Relief ERA

Name	Team	IP	ER	IR scored	IR stranded	Runs Saved	ERA	RERA	Adj
Jackson	Cle	64	11	4	20	4.3	1.55	0.94	0.84
Hoffman	SD	73	12	4	22	5.0	1.48	0.86	0.93
Urbina	Mtl	69.1	10	2	12	2.8	1.30	0.93	0.94
Rivera	NY A	61.1	13	4	20	4.3	1.91	1.28	1.18
Nen	SF	88.2	15	5	14	1.6	1.52	1.36	1.39
Rocker	Atl	38	9	6	18	2.3	2.13	1.58	1.80
Gordon	Bos	79.1	24	8	30	5.2	2.72	2.14	1.91
Brocail	Det	62.1	19	11	33	4.2	2.74	2.13	1.93
Assenmacher	Cle	47	17	9	33	5.5	3.26	2.19	1.95
Wetteland	Tex	62	14	7	8	-1.8	2.03	2.29	2.00
Howry	Chi A	54.1	19	2	21	6.0	3.15	2.16	2.05
Corsi	Bos	66	19	9	23	2.1	2.59	2.31	2.06
Christiansen	Pit	64.2	18	11	26	2.5	2.51	2.16	2.10
Plesac	Tor	50	21	15	50	7.5	3.78	2.43	2.23
Shaw	LA	85	20	11	23	0.8	2.12	2.04	2.23
Veres	Col	76.1	24	12	27	1.5	2.83	2.65	2.24
Olson	Ari	68.2	23	0	16	5.5	2.02	2.29	2.27
Swindell	Bos	90.1	26	5	39	10.2	3.59	2.57	2.29
Mecir	TB	84	29	11	34	4.6	3.11	2.62	2.30
Seanez	Atl	36	11	1	7	1.8	2.75	2.31	2.33

Top 20 ML relief pitchers as selected from each team's four most used relief pitchers. A complete list is available from the author.

Of course, this system is not perfect, and nothing without using play-by-play data ever will be. However, it consolidates ERA and Inherited Runners statistics very nicely and in my opinion is much more accurate than regular ERA. The RERA is also fairly simple and would be accessable to the average fan because its result is in the same form as the familiar ERA. The main problem with RERA is that all inherited runners are not created equal. A man coming on with a man on 3rd and none out will have a much tougher time in this stat than a man coming on with a man on 1st and two out. Hopefully, these situations tend to cancel each other out for most pitchers, and since most relievers aren't slotted in roles in which they would always be coming into tough jams as the "firemen" of the 60's, 70's and 80's used to do, I would imagine that most pitchers would have similar opportunities to stop opposing runners from scoring. While a play-by-play analysis would be best, the RERA is the best method to evaluate relievers without it.

Sky Andrecheck, 179 Larch Ave., Elmhurst, IL, 60126, andrerhs@aol.com. ♦

Correction

Correction: Sig Mejdal's e-mail address was spelled incorrectly in the previous issue. It should have read *smejdal@montereytechnologies.com*.

Measuring the Accuracy of Runs Formulas for Players

Clifford Blau

While it's easy to verify runs formulas (such as Runs Created) for teams, it's harder to do so for players, since it's difficult to determine the actual number of runs a player is responsible for. Here, the author tries to address that problem by examining individual high- and low-scoring pitching records and team games.

Numerous analysts have developed formulas to predict how many runs will result from any given combination of singles, doubles, walks, outs, etc. These formulas are typically verified using seasonal team data, and most of them are very accurate by that standard. However, they are normally used to estimate how many runs result from the production of an individual batter. The main problem with that is that, since there is no way of determining exactly how many runs an individual is responsible for, the accuracy of the formulas for this purpose is difficult to validate. A related problem, highlighted by Phil Birnbaum in the May 1999 issue of *By The Numbers*, is that the range of offense represented by teams is much smaller than the difference in production by individuals. Therefore, a formula may work very well for a player of average performance, but not for a very good or very bad hitter. He developed a formula that works well over a wider range of player performance.

Some years ago, I did a small study to test the accuracy of runs formulas and, with the introduction of some new methods recently, I decided to revive and expand that study. I applied the formulas to pitchers' statistics. The advantage in this is that we know not only how many singles, walks, homers, etc. a pitcher allows, but also how many runs. Additionally, a pitcher's statistics offer a data set of approximately the same size as a batter's for a season. In theory, the typical error in the estimated runs will be proportionately greater for an individual than for a team, as luck will have fewer opportunities to even out. (While the runs formulas work well for team-seasons, they will not work well for individual games. This is because they assume an average number of uncounted events such as reaching on errors and lost baserunners. They also assume that counted events have a normal distribution, but even for team-seasons there will be some variance from the norm. For smaller samples such as an individual season or team-game, the error will be greater.) A possible drawback with this approach is that the number of runs charged to a pitcher may not reflect the runs deserved, due to the effect of partial innings and relief pitchers. Also, the normal range of performance is narrower for pitchers than for hitters. Few pitchers are good enough to allow under three runs per nine innings for a season, while those bad enough to yield more than seven don't usually pitch long.

I had originally used data from the 1984 and 1985 seasons, and expanded the study to encompass the 1986 and 1987 seasons, since I had data for those years. Fortunately, 1987 was a high offense year, so there were several pitchers who allowed runs at a high rate. The formulas evaluated were Bill James' *Runs Created* (RC)(Technical Version from the 1988 *Abstract*), Paul Johnson's *Estimated Runs Produced* (ERP), *Extrapolated Runs* (XR)(introduced in the *1999 Big Bad Baseball Annual*), and Phil Birnbaum's *Ugly Weights* (UW). These formulas are detailed in the appendix. Since I was lacking double play data, I had to use a simplified version of XR, called *Extrapolated Runs Reduced* (XRR). Results are presented using the standard deviation of the difference between the predicted number of runs and the actual runs, the mean of that difference, and the square of the correlation produced by a linear regression equation (coefficient of determination.)

Table 1 – Runs Formulas Applied to Selected Pitchers' Records

	Standard Deviation	Mean Error	Coefficient of Determination
RC	9.19	5.2	.968
ERP	7.77	0.5	.966
XRR	7.80	-0.8	.967
UW	8.15	-1.9	.968

I selected 119 pitchers, representing a wide range of performance. These pitchers worked an average of 164.4 innings, with a range of 34.3 to 271.7. They allowed a mean of 80.3 runs. The results of this test are shown in Table 1.

Table 2 – Runs Formulas Applied to Selected "Very Good" Pitchers

	Standard Deviation	Mean Error	Coefficient of Determination
RC	8.70	5.1	.949
ERP	7.45	1.9	.949
XRR	7.37	0.5	.949
UW	6.78	1.6	.957

In order to test whether Ugly Weights outperformed the others for very good and very bad pitchers, I divided the sample into four nearly equal-sized groups, based on OPS. The very good group, based on 29 pitchers with a mean of 193.9 innings and 57.6 runs allowed, had results shown in Table 2. The thirty pitchers in the very bad group are in Table 3. They pitched an average of 94.5 innings and allowed 69.6 runs.

Then I hit upon another means of testing these formulas. I went through the major league box scores for July and August of 1993, and compiled the data for high and low offense games into groups of about eighteen games each, to approximate a full-time player. After making nine of these groups for both good and bad offenses, I applied the formulas (using XR instead of XRR this time), and got the results shown in Tables 4 and 5:

Table 3 – Runs Formulas Applied to Selected "Very Bad" Pitchers

	Standard Deviation	Mean Error	Coefficient of Determination
RC	8.05	5.3	.959
ERP	7.77	-1.2	.963
XRR	7.80	-2.2	.966
UW	6.41	-3.2	.967

Table 4 – Groups of good-hitting games

Team	AB	BA	OBA	SA	Runs	ERP	RC	UW	XR
1	711	.353	.430	.592	177	161	184	163	160
2	671	.341	.419	.559	145	143	158	146	143
3	663	.299	.375	.507	139	120	126	117	121
4	652	.314	.393	.472	124	118	124	118	117
5	657	.333	.422	.516	134	133	144	133	133
6	669	.344	.419	.538	136	136	150	137	137
7	673	.330	.397	.612	146	148	162	142	148
8	656	.316	.396	.543	138	132	142	128	130
9	671	.334	.389	.548	138	129	136	127	130
Standard Deviations of differences						9.30	10.32	10.31	9.43
Coefficient of determination						.759	.767	.742	.754

All of the formulas tended to overpredict runs for the weak hitting teams and all but RC underpredicted for the high scorers. A possible problem with this portion of the study is that the groups may be biased due to being selected for the output (runs) rather than the input (hits, walks, etc.) However, I tried to select games based on the input, so hopefully this does not distort the results.

Conclusions

This study has presented two methods of validating runs formulas for individuals. Using both pitchers' opponents batting statistics and groups of team game statistics, I found that Estimated Runs Produced performed best overall, although the differences among the four formulas were fairly small. All of the formulas correlate well with actual runs. However, the typical error in the estimates appears to be roughly 10%, which should be kept in mind when using them to evaluate hitters. It doesn't make sense to produce an estimate of runs created to the nearest .01 of a run when the formula is only accurate to the nearest 10 runs. Based on this study, Ugly Weights may be more useful for very good and very bad hitters. In the second part of the study, Runs Created, the only non-linear formula of the four, was much more accurate in some cases, and much less in others, than the other three. Perhaps an examination of why that is so could lead to a more accurate formula. However, in order to achieve a significant increase in accuracy, data on baserunning, errors, and timely hitting would likely be necessary.

Table 5 – Groups of poor-hitting games

Team	AB	BA	OBA	SA	Runs	ERP	RC	UW	XR
10	574	.202	.260	.282	43	40	41	42	40
11	586	.212	.274	.285	37	44	44	45	45
12	572	.187	.250	.260	36	34	37	34	34
13	578	.220	.282	.282	40	45	46	43	46
14	579	.197	.230	.282	35	33	36	35	31
15	577	.217	.267	.267	31	37	39	34	37
16	587	.233	.303	.334	49	63	61	60	61
17	579	.207	.271	.285	38	43	42	44	44
18	578	.218	.270	.334	43	51	50	52	52
Standard Deviations of differences						6.77	6.36	6.01	6.88
Coefficient of determination						.711	.715	.787	.675
Overall SD's (tables 4, 5 combined)						8.13	8.57	8.44	8.25

Appendix:

Runs Created: (H+W+HBP-CS-GDP)*(TB+(.26(BB-IBB+HBP)+.52(SH+SF+SB))/ AB+BB+HBP+SH+SF)

Estimated Runs Produced: (2*(TB+BB+HP)+H+SB-(.605X(AB+CS+GIDP-H)))*.16

Extrapolated Runs: .50(1B) + .72(2B) + 1.04(3B) + 1.44(HR) + .34(BB+HP-IBB)+.25(IBB) + .18(SB) - .32(CS)-.90(AB-H-K)-.098(K)-.37(GIDP)+.37(SF)+.04(SH)

Ugly Weights: .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) where ip=Isolated Power (Slugging Average minus Batting Average) and iw=walks divided by at-bats

Thanks to Cyril Morong and Phil Birnbaum for their comments on earlier versions of this article. *Clifford Blau, 16 Lake St., #5D, White Plains, NY, 10603, cliffordblau@geocities.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.

See Clifford Blau's review in this issue, or Sig Mejdal's review in the previous issue, for the kind of thing we're looking for.

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.