By the Numbers

Volume 11, Number 1 The Newsletter of the SABR Statistical Analysis Committee February, 2001

Review

"Sabermetric Encyclopedia" – A Good Start

Keith Carlson

Lee Sinins has developed a Sabermetric Encyclopedia (SE). It is on CDROM, and is described, along with sample pages, at http://members.nbci.com/sabermetricencyclope/. Lee set up that website just for that purpose. There is no need to repeat all of the details here. Also, you will find pricing and ordering information there. He offers several packages.

The chief system requirement for the SE is that you have Microsoft Access installed on your computer. You do not need to know anything about running Access; it just has to be there in order for SE to be installed. I found that SE installed easily and quickly, using 45 megabytes of disc space.

The SE has hitting and pitching stats only—no fielding. All of the conventional stats for 1871-2000 are supplemented with a number of sabermetric measures. The major innovation of this program is that individual stats can be figured relative to league average and then sorted. Comparisons can be made in absolute or percentage terms. In fact, every player has a career line of stats that can be compared to a league-average line or a league average at the player's position. This brings up an approximation that might bother some users—each player is listed as having played only one position in his career, and there is no year-by-year position listing. Robin Yount is a shortstop, Stan Musial is an outfielder, Pete Rose is an outfielder, etc.

The SE is useful, fast, flexible and easy to use. There is no apparent maximum limit to the number of players you can sort to a list. It also has a printing capability. One thing Lee does not tell you is that you can copy (Ctrl-C) to a spreadsheet, and, for example, make a chart of a player's stat over time. It is not as quick and colorful as the old Bill James Encyclopedia, but it can be done.

Let me conclude with a summary of what I consider to be the weaknesses of SE:

1. Park effects. Lee gives several stats as park adjusted (runs created above average for hitters and runs saved above average for pitchers, for example). However, he does not give you the park factors.

2. Fielding stats. As previously mentioned, these are not included and the position listing is incomplete. True, these are all available elsewhere, but for the sake of completeness it would be nice to have them included. (Lee did say in one of his emails that he was going to add the fielding stats in the version following the 2001 season.)

3. Federal League stats. No reason is given, but there are no Federal League stats (1914-15).

4. Stat limitations. The user is limited to the stats that Lee includes. This, of course, is a long list, but you cannot create any ratios as you can in Pete Palmer's hand-held encyclopedia.

The pluses for SE far outweigh the minuses. Being named *Sabermetric* Encyclopedia means it is specialized and no attempt should be made to include "everything" like *Total Baseball*. Lee has demonstrated, however, that he is receptive to suggestions for improvement (he has already issued a version fixing some problems that appeared in the initial release). SE is a a great start, delivering capabilities that are not available elsewhere.

Keith Carlson, 3540 Gordon Ave., St. Louis, MO, 63114, kcarlson@stlnet.com ♦

In this issue

Bill James Index

Stephen Roney

Bill James is often cited by researchers in discussions, but typically without the author knowing the exact citation. James' published studies, especially in the Abstracts, are often not easily located.

I am attempting to gather that information (the locations of the various studies) and put it all in one place. The work in progress is located at http://www.members.home.net/sroney3/JamesIndex.

Doing a complete indexing of the Abstracts and other books by Bill James is a lengthy and tedious project which is just getting underway. So I have started with overviews of the books that I have, which are essentially the Tables of Contents, sometimes with a little extra information. I have also listed the articles that James wrote in the Baseball Analyst Newsletter he published, and those that I know about in other books, newspapers and magazines. I know there are many more magazine articles and I would appreciate any citations that anyone can add.

Stephen Roney, 23491 Thornewood Dr., Santa Clarita, CA, 91321, sroney@acm.org ♦

Correction

Last issue's Table of Contents on the front page ("In this issue") misidentified the author of the article "Academic Research: Racial Bias in Hall of Fame Voting." The correct author is Charlie Pavitt.

Neal Traven Address Change

Neal Traven, co-chair of the Statistical Analysis Committee, is moving March 8. His new addresses, regular and e-mail, are:

Neal Traven
4317 Dayton Ave. N, #201
Seattle, WA
USA
98013

beisbol@alumni.pitt.edu

Academic Research: Diversity and Team Performance

Charlie Pavitt

The author summarizes academic research on whether age and race diversity correlate with team performance, both in basketball, where players are very much interdependent on each other, and in baseball, which is more an individual sport.

This is the one of a series of occasional reviews 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.

Thomas A. Timmerman, Racial Diversity, Age Diversity, Interdependence, and Team Performance, Small Group Research, Volume 31, 2000, pages 592-606.

When I am talking with baseball friends unfamiliar with sabermetric research, I often find myself trying to convince them that, although it is played by teams, baseball is essentially an individual sport. At its core is the battle between pitcher and batter, and on a given day the team that wins that battle most often is likely to win the game. This is why our best methods for evaluating individual player performance relate so highly to team performance. Way back in 1974, Jones (Organizational Behavior and Human Performance, Vol. 11, pages 426-451) found individual and team performance in baseball to correlate at over .9 for both leagues between 1947 and 1960, although his use of RBI to represent individual offensive performance can certainly be criticized. Using the sum of points scored, rebounds, and assists weighted by minutes played, Jones found substantially lower correlations of around .6 in professional basketball. One can take this to imply, as did Jones, that the basis for this difference lies in the fact that basketball really is a team sport, such that “the personal relations among the players, teamwork, or the way the team responds in clutch situations are likely to make more difference than they do in baseball” (page 450).

Timmerman’s recent article is rooted in the observation about personal relations. He noted that past management science research on work group competition has led to inconsistent findings concerning the relationship between group diversity and performance, and hypothesized that the critical variable needed to disambiguate these findings is the extent to which the group’s task requires interdependence among group members. Tasks requiring interdependence should be particularly troubling to groups high in diversity, because any interpersonal conflict resulting from diversity would be more damaging than it would be in tasks allowing group members to work independently. As a consequence, he proposed that the relationship between group member diversity and group performance should be stronger in tasks that require group interdependence.

Citing Jones’s conclusions as one piece of evidence in support, Timmerman used professional basketball and baseball performance from 1950 to 1997 to respectively represent tasks high and low in required interdependence. To represent diversity, he calculated indices to measure variation in age and race within a team. For baseball, he found that OBA, SA, ERA, and average team age together accounted for 72 percent of the variance in team winning percentage, and that measures of age and racial diversity had no additional predictive value. For basketball, field goal and free throw percentage, rebounds per game, and average age accounted for only 24 percent of the variance in team winning percentage, and for measures of age and racial diversity to correlate at about -.1 with winning percentage, accounting for 2 percent in additional variance. Based on these findings, we can perhaps conclude that baseball is an individual sport and basketball is not. However, although the results are technically consistent with Timmerman’s hypotheses, given how small their predictive impact was, racial and age diversity do not seem to have much impact on whatever intangibles distinguish between good and bad NBA teams.

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

POP Extended To Slugging Percentage – Who Is Most Definitely Not An Average Slugger?

Peter Ridges

In a previous BTN, Mike Sluss introduced the POP statistic, based on the probability of an average player achieving a specfic batting average. In that article, he wrote that POP would be difficult to extend to slugging percentage. But here, the author takes a stab at doing just that, and presents league-leaders and laggers in the category.

In BTN Volume 10, Number 3 (August 2000), Mike Sluss concluded by pointing out that the exact calculation of POP for slugging percentage is very difficult and requires a supercomputer. Here is what I feel is a sufficiently accurate approximation to the correct answer.

In insurance, statisticians often use compound distributions with 2 variables. N is the number of claims that an individual makes, X is the amount of money for each claim, and the insurance company is interested in S, the total of an individual's claims. The same technique applies to slugging: N is the number of hits for an individual, X is the number of bases on any hit (X=1, 2, 3, or 4) and S is the total number of bases.

If you look in a statistics text for compound distributions, you will find the following formulae for mean and variance (which don't require any assumptions at all):

```
E(S) = E(N) E(X)                     which is fairly obvious
V(S) = E(N) V(X) + V(N)(E(X))^2      which is not at all obvious.
```

Then, if we assume that the actual distribution of a typical hitter's total bases is reasonably close to the normal distribution (I think this is a very robust assumption for a reasonable number of AB), we can use this to estimate probabilities and hence POP.

Let's apply this to 2000 Major League data.

First, find the distribution of N: it seems reasonable to use the binomial distribution. The MLB batting average was

```
45245/167289 = 0.27046.
```

For one at-bat, this gives

`E(N)= 0.27046` and `V(N) = 0.19731.`

Second, find the distribution of X:

```
There were 45245 hits, distributed as follows:
29698 singles (65.64%)
8902 doubles  (19.68%)
952 triples    (2.10%)
5693 homers   (12.58%)
```

Using this suggests

`E(X) = 1.61631` (i.e. the number of bases per hit), `and V(X)=1.03351.`

For a single at bat, we then have

```
E(S)= 0.27046 * 1.61631 = 0.43715, the ML slugging percentage.
V(S)= 0.27046 * 1.03351 + 0.19731 * 1.61631 ^ 2 = 0.79499
```

so the standard deviation of S is `0.89162.`

Next, apply this to Todd Helton's 2000 numbers. In his 580 AB, we would expect 580 * 0.43715 = 253.55 total bases, with a standard deviation of sqrt(580) * 0.89162 = 21.47. He actually had 405 total bases, 151.45 more than a 0.437 slugger would have, which is 7.05 standard deviations. (Pedantic note: we need to apply a "continuity correction", which is always half, giving 150.95). The chances of a 0.437 slugger achieving this are tiny, about 10 to the power of -12, which equates to a POP of 11.98.

In Microsoft Excel, the formula to get from 7.05 (in cell X1, say) to 11.98 is:

```
=-LOG(1-NORMDIST(X1,0,1,TRUE))
```

Table 1 shows, by this method, the 20 players who in 2000 most clearly demonstrated that they are not ordinary (i.e. .437) sluggers.

The table shows that the conventional slugging champion, Todd Helton, is also the hitter who most clearly is not an average slugger. My apologies for the absence of a park-effect adjustment. Note also that Mark McGwire is allowed on the list in spite of his small number of AB. My data excludes players who appeared for more than one team, but I believe this makes no difference.

As it happens, all the players in the chart have at least 200 AB, so the approximation is reasonably accurate. The leading POP for a small-AB player was Esteban Yan, who had one AB and one HR. This gives an approximate POP of 3.528 on the above method. However it is easy to calculate his true POP: 3.4% of AB resulted in homers (12.58% of 0.270) giving a true POP of 1.468.

It is possible to generate exact figures for POP for small AB using an Excel spreadsheet. I did this for up to 50 AB, which required 20 megabytes; by 50 AB I was getting very similar answers to the above figures. The formula is iterative. If P(n,k) is the probability of a player getting exactly n total bases from k AB, then P(n,k)=

```
P(n-4, k-1) * P (HR) +
P(n-3, k-1) * P (triple) +
P(n-2, k-1) * P (double) +
P(n-1, k-1) * P (single) +
P(n, k-1) * P (out)
```

Incidentally, in Mike's analysis he showed that the leading POP (based on BA) in 1930 was only 6.09. This is to be expected -- the range in players' slugging abilities is greater than the range in their batting averages.

Table 1 – Best Slugging POP, 2000

	AB	SLG	POP
Todd Helton, 1B	580	0.698	11.983
Barry Bonds, OF	480	0.688	9.345
Manny Ramirez, OF	439	0.697	9.221
Carlos Delgado, 1B	569	0.664	9.150
Vladimir Guerrero, OF	571	0.664	9.137
Sammy Sosa, OF	604	0.634	7.491
Jason Giambi, 1B	510	0.647	7.217
Mark McGwire, 1B	236	0.746	7.191
Richard Hidalgo, OF	558	0.636	7.118
Gary Sheffield, OF	501	0.643	6.851
Frank Thomas, DH	582	0.625	6.703
Ivan Rodriguez, C	363	0.667	6.264
Jeff Bagwell, 1B	590	0.615	6.163
Alex Rodriguez, SS	554	0.606	5.360
Moises Alou, OF	454	0.623	5.313
Troy Glaus, 3B	563	0.604	5.295
Mike Piazza, C	482	0.614	5.131
Jeff Kent, 2B	587	0.596	5.069
Nomar Garciaparra, SS	529	0.599	4.792
Brian Giles, OF	559	0.594	4.748

Table 2 – Worst Slugging POP, 2000

	AB	SLG	POP
Homer Bush, 2B	297	0.253	0.000088
Pedro Astacio, P	82	0.098	0.000154
Ryan Minor, 3B	84	0.143	0.000660
Mark McLemore, 2B	481	0.316	0.000681
Rey Sanchez, SS	509	0.322	0.000854
Michael Barrett, 3B	271	0.288	0.001407
Rey Ordonez, SS	133	0.226	0.001563
Felix Martinez, SS	299	0.298	0.001637
Vinny Castilla, 3B	331	0.308	0.002022
Alberto Castillo, C	185	0.265	0.002108
Alex Gonzalez, SS	385	0.319	0.002273

The other end of the analysis shows that the inaptly named Homer Bush is the player who is most clearly worse than 0.437. This is well below Pedro Astacio, the worst-slugging pitcher, who can partly blame a small sample size for his low SLG. Table 2 lists Astacio and the 10 worst position players.

Bush's performance is as likely as a POP of only 3.7; as expected, a player slugging 184 below the average will be allowed fewer AB than a player slugging 184 points above 0.437.

Peter Ridges, 25 Cearns Road, Birkenhead, Cheshire, England, pete@ridges.co.uk ♦

Submissions

Phil Birnbaum, Editor

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 (i.e., informal peer review).

If your submission discusses a previous BTN article, the author of that article may be asked to reply briefly in the same issue in which your letter or article appears.

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.)

If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does, and please include your byline at the end with your address (see the end of any article this issue).

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.

Send submissions to:
Phil Birnbaum
18 Deerfield Dr. #608, Nepean, Ontario, Canada, K2G 4L1
birnbaum@sympatico.ca

What Makes a "Clutch" Situation?

Tom Hanrahan

There have been many different situations used as a stand-in for "clutch performance:" batting with runners in scoring position, batting with two outs, batting in the late innings, and so on. In this study, the author suggests that the true clutch situations are those with a high variation of subsequent runs scored – and looks into which situations those are.

Introduction

Any time you watch a ballgame on TV or hear one on the radio, announcers are always quoting statistics that describe clutch hitting: batting average with runners in scoring position, with two outs, close-and-late, etc. The same goes for data in the printed media. Are these stats worthwhile? What WOULD be the best way to describe clutch batting statistics?

This article deals only with what I call "run-dependent" clutch performance. There is also "game-dependent" and "pennant dependent" performance, analysis of which would require different methods.

Approach

I submit that clutch performance is correlated strongly with the dispersion or variation in results for a given situation; in other words, the reason that hitting with the bases loaded is important is because the difference in results between a hit and an out in that situation can mean many runs.

I modeled expected dispersion by situation by using Pete Palmer's Potential Runs For 24 Base/Out Situations (*The Hidden Game of Baseball*, p. 153), along with a model for a generic hitter (see notes for expected probabilities of my "typical" hitter over the last generation). I took the new potential runs for each possible result, weighted them by the batter model probabilities, and arrived with an dispersion of runs scored for every base/out situation.

Example

Suppose our batter is the ultimate Russ Branyan; he either strikes out (80% of the time) or hits a home run (20%). In situation A, with 2 outs and no one on base, the expected runs produced are .095; on average, this situation will yield .095 runs by the time the inning ends. After a home run by Branyan, the expected runs produced is 1.095 (add one for the solo dinger). After an out, it is 0.000 (inning over). The dispersion of this situation, weighted by the fact that Branyan homers 20% of the time, is .44 runs (calculation is the standard deviation of weighted probabilities, using *n* instead of *n-1*, which in this case simplifies to the formula

```
Square root of
{ [.2*1.095^2+.8*.000^2] –
[.2*1.095]^2 } ).
```

If the bases had been loaded with two outs, the grand slam would be worth 4.095 runs more than an out, for a dispersion of 1.64 runs; the sacks full event is almost four times as important as bases empty.

Table 1— Dispersion Of Runs By Base/Out Situation

Runners On Base	Number of outs			Weighted Average
	0 out	1 out	2 out	
Nobody on	.315	.252	.189	.264
1st	.593	.495	.403	.503
2nd	.504	.509	.534	
3rd	.440	.524	.568	
1st & 2nd	.841	.799	.743	.784
1st & 3rd	.689	.800	.772	
2nd & 3rd	.626	.762	.933	
Loaded	.935	1.111	1.138	1.100

Data

Table 1 gives the dispersion of expected runs for each base-out situation. The rightmost column is the average for the number of runners on base, weighted by the approximate number of times that each situation arises with a given number of outs.

Observations

The difference in extremes (bases-full versus empty-with-2-gone) is 1.138/.189 = about 6 times. Of course, you probably figured that these 2 situations were at opposite ends of the spectrum. But sacks-full-with-no-outs really isn't very far behind, according to the chart. In fact, there are *no* situations in which having 2 men on is more volatile (equals important) than bases loaded. Also, 2 on is always more important than 1 on, and 1 on more important than bases empty, *regardless of the number of outs*. Looking at the right hand column, there is an almost perfectly linear (1-2-3-4) relationship between the dispersion of expected runs, and men-on-base-plus-the-batter. Hindsight being 20/20, this makes sense, since with none on the hitter is batting only for himself, while with the bases loaded he represents the fate of 4 men.

Applications

Pinch-hitting

Neither the positioning of the runners (3rd vs. 2nd vs. 1st) nor the number of outs is anywhere near as important as the number of runners. I was surprised to see the importance of many early-inning situations. How many NL managers would let their pitcher come to the plate with no outs and men on first and second (dispersion = .841 runs), yet feel compelled to pinch hit for him with first and third occupied and 2 away (dispersion = .772 runs)?

Clutch measures

The oft-quoted "batting average with men in scoring position" is not, at-bat for at-bat, significantly more important than batting average with runners on. Plus, the sample size is smaller. The same holds even more if you add the "2 outs" qualifier to the former measure. I offer that the best statistic to measure "run-dependent" clutch performance would be to weight on base average and slugging percentage by runners on base.

Lineup selection

This knowledge of dispersion by base-out situations can be applied to look at choices of batting order.

Tom Ruane supplied data that gives the average number of runners on base by lineup position for the 1999 season. Table 2 shows that batters in the cleanup spot came to the plate with an average of .74 runners on, as opposed to leadoff men who only averaged .50 runners on. This means that the typical 4th hitter had 16% more (1.74/1.50) opportunities for producing runs, per plate appearance, than the leadoff hitter did.

Table 2 – Average Number Of Runners On Base By Batting Order Position

Lineup Position	League NL	League AL	MLB average	Effective batting opportunities per game
1	.48	.53	.50	4.41
2	.60	.64	.62	4.65
3	.66	.69	.67	4.69
4	.74	.74	.74	4.77
5	.70	.74	.72	4.60
6	.70	.70	.70	4.43
7	.69	.70	.69	4.30
8	.68	.68	.68	4.16
9	.70	.67	.69	4.07

We can combine this table with the results of Table 1 to produce the expected number of batting opportunities weighted by runner on base for each spot in the lineup. For this exercise, I assume that a typical team will have its leadoff man come to the plate 4.9 times per game, with each successive spot in the lineup coming to the plate 0.11 times per game less. This is what produces the right-most column, expected opportunities per game. For leadoff men, they bat 4.90 times per game, and they hit for an average of 1.5 men (themselves and 0.5 runners). The MLB average is .668 runners on, so I divide by 1.668. This gives an effective number of opportunities to be 4.9 * 1.5 / 1.668 = 4.41.

The table shows that the cleanup position gets the most "effective" opportunities per game. However, it should be noted that there are two factors not reflected in this table which might influence the importance of the leadoff role. First, the leadoff batters (and #2 and #3 men) typically have good hitters following them, so their ability to reach base is magnified in a way not shown here. Likewise, the 5th thru 8th hitters are usually followed by less luminary hitters. Also, the number one hitter often bats with none on and none out (1st inning), which is the more important of the "none on base" situations.

Subjectively, I think this moves the importance of the leadoff role up to above that of the #5 hole. Even so, the differences are not large. Swapping a great hitter for a good hitter in the 4th/5th spots nets the great one another .17 opportunities per game, which might only be 2 runs over the course of a season. Earlier studies concluding that batting order differences are overrated are confirmed using this approach.

Conclusions

- Because the importance of batting situations correlates very closely with the number of runners on base at the time, I offer that the best statistic to measure "run-dependent" clutch performance would be to weight on base average and slugging percentage by runners on base.

- Pinch hit when runners are on. Bases loaded in the 4th inning of a game has got to be more critical than waiting for a man to be on 2nd base in the 8th inning, even if the game is still tied at that point.

- The most important lineup positions are the 2, 3, and 4 spots. But the differences in expected results are not large.

Notes

Typical batter model: walks are weighted as .095 (9.5% of all plate appearances), singles .165, doubles .040, triples .007, home runs .023, outs .670. Of outs,.120 of the .670 are potential DPs, .050 send a runner on third home, an additional .050 also send a runner on 2nd to 3rd, an additional .050 also send a runner on 1st to 2nd, all other leave base runners unchanged. Baserunners advance an average of 1.45 bases on a single, 2.25 on a double, slightly higher with 2 outs than with none. I realize that the typical 2000 MLB hitter is better than this, but since I was using Pete Palmer's data which was extracted from previous years, I used probabilities which reflected lower run environments than we see today.

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

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. The only restriction, please: the book should have, or claim to have, some Sabermetric content.

For a sample of what we're looking for, check out David Shiner's review last issue, or Gabe Costa's review in the issue before that.

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.

Baseball's All-Time Best Hitters: Debate Engaged

Michael J. Schell

The author responds to David Shiner's review of his book last issue. Among the issues re-examined are: SD as a suitable stand-in for league proficiency; pitcher hitting changes over time; and the assumption that there are equal proportions of "great" hitters over baseball history.

David Shiner raised some very interesting questions in his review of my recent book, *Baseball's All-Time Best Hitters*, referred to hereafter as *BABH*. Four adjustments were made to the raw batting average (BA): a mean adjustment, a standard deviation adjustment, a ballpark adjustment and a longevity adjustment (ABs > 8000 were not factored in to a player's career BA). Shiner had problems with two of the adjustments. Concerning the mean adjustment, Shiner wrote: "The main problem with the adjustment for offensive context, which is otherwise well handled, has to do with pitchers' hitting." Shiner then notes that turn-of-the-century pitchers hit much better than their modern day counterparts. Shiner was also concerned about the standard deviation (SD) adjustment and its interpretation as a talent pool measure. He focused especially on the comparison of the AL and NL from 1908 to 1919, when the NL had significantly lower SDs than the AL. The argument presented in *BABH* is that the SD of mean- and ballpark-adjusted batting averages of players with at least 200 ABs in a given year is inversely related to the strength of the league for the year. That implies that the NL was a stronger league than the AL throughout 1908-19, in spite of the fact that the best hitters (Cobb, Speaker, Jackson, Collins and Lajoie) resided in the AL. Moreover, the SD rose as the first four named players entered, leading to the laughable conclusion that they actually *weakened* the AL. Shiner's final major point was stated this way:

> "Early on, Schell states that one of his basic assumptions is that "there is an equal proportion of hitters who are 'great' across baseball history." … I'm not convinced that this assumption can coexist peacefully with [Stephen Jay] Gould's arguments concerning standard deviation, and I'm not sure that Schell is either."

In order to respond to Shiner's questions I need to present some additional data[1]. First, define AB+BB as *net-faced-pitcher*[2], or NFP for short. A *saddlepoint* in the histogram (a local minimum) of AL and NL NFPs from 1901-60 occurs at 2.5 × NFP/G (which is 385 NFPs in a 154-game season) and provides a natural break among groups of players. Consequently, I will separate the players in a given season into 3 groups: regulars (players with ≥ 2.5 NFP/G), substitutes (non-pitchers with < 2.5 NFP/G) and pitchers (players whose primary position is "P" as determined by *Total Baseball, Edition 4*).

The first page of graphs has four panels. The top two show the batting averages of the AL and NL pitchers from 1901-1998[3]. Also shown are piecewise linear fits to the data that will be described in greater detail later. The bottom two panels show the percentage of NFPs that the pitchers obtained year-by-year. The second and third pages show the equivalent plots for "regulars" and "substitutes", respectively.

Batting Average Changes in Hitter Subgroups[4]

For the American League, the batting average of *pitchers* hovered around .180 through 1951, then declined about 2 points per year thereafter to .146 in 1969. The decline in pitchers' batting average in the National League is a bit more complicated. Averages were around .180 from 1901 to 1926, declined by 0.5 points per year to .164 in 1955, then sunk 2.8 points per year to .139 by 1964, rebounded to .152 by 1974 and have been declining by 0.6 since then to .139 in 1998. The r^2 values for the AL and NL pitchers batting average data are .669 and .876, respectively[5].

There was also a decline in the percentage of league NFPs obtained by pitchers during the 20th century. In the AL, the percentage dropped quickly from 10.3% in 1901 to 9.1% in 1914. Then the decline slowed, reaching 8.8% in 1945. From there, the decline picked up speed again and reached 7.2% before inching up to 7.3% in 1972. In the NL, the pitchers held 10.0% of the league NFPs from 1901-1906, before

---

[1] The data for the analyses presented in this response were obtained from the Lahman database (www.baseball1.com).
[2] The term *net-faced-pitcher* was suggested to me by David Stephan, since it is plate appearances minus (HBP + SH + SF). This was used because data are lacking for determining plate appearances across baseball history.
[3] No data are used for AL pitchers past 1972 since the data consist of very small numbers of NFPs.
[4] In this section, the numbers presented are from the regression model fits, not from the raw data. This is done so that I can present overall trends without getting immersed in the inherent variability of individual years.
[5] The r^2 value gives the proportion of variability explained by the model.

skidding to 9.1% in 1908. From there the decline continued dropping to 8.8% by 1929, 7.2% in 1972 and 5.8% in 1998. The r^2 values for the AL and NL are .971 and .986, respectively.

The batting average of AL *regulars* shows a dramatic spike starting at 1908, peaking at 1910 and leveling out in 1921. This is mirrored closely (in inverse fashion) by the dramatic decline in percentage NFP by AL regulars starting in 1907, reaching the nadir in 1909 and peaking in 1921. Notice that no such effect appears in the corresponding NL plots. What are the AL plots telling us? They indicate that the percentage of NFPs garnered by AL regulars took a nosedive starting in 1907 and didn't fully rebound until the end of the dead ball era.

The third group of hitters besides pitchers and regulars are *substitutes*. These are the non-pitchers who had less than 2.5 NFPs/game. Notice that the spike for the percent of NFPs for the AL substitutes is basically a mirror image of the one for the AL regulars. Thus, the sudden drop in NFPs by the AL regulars was picked up by the substitutes, with the pitchers continuing to bat less. There is no inverted spike, however, in the batting averages of the substitutes. Considering the AL graphs together, it appears that the superstars entering in around 1907-11 (Cobb, Speaker, Collins, Jackson) represent the group that destabilized the balance in hitting for average in the AL. This led to an increase in the standard deviation in batting averages noted in *BABH* even though the league was improving not worsening. Over time, other good hitters gradually replaced weaker hitters, leading to a decrease in SD and a decrease in mean-adjusted batting average of the AL regulars.

Other interesting observations can be gleaned from the charts. The percentage of NFPs obtained by AL regulars dropped from 68% in 1921 to 56% in 1971 before leaping back up to 64% by 1974, after the designated hitter replaced the pitcher in the lineup. By contrast, the percent of NFP among regulars in the NL averaged 62% from 1901 until 1954, but has been in decline since that time, to 57% in 1998. The NFPs lost by the NL regulars and pitchers over the second half of the century were picked up by the substitutes, whose percent of NFPs rose from an average of 30% established from 1901-54 to 38% in 1998. Also, note that the batting average for NL regulars did not show a trend from 1901-38, averaging .272. Then, the average climbed to .277 in 1964 before dropping back to .273 in 1980. It has been increasing slowly since that time. Given the gains in batting average of NL regulars, the batting average of NL substitutes declined from .244 in 1921 to .239 in 1974 before jumping up to .248 in 1980. Averages have slowly declined since then.

Table 1. Top 5 Hitters For Average in 1909-14

American League

Player	AB	Years*	MABA	SDABA
Ty Cobb	2996	1909-14	.3946	.3617
Joe Jackson	2124	1911-14	.3712	.3451
Eddie Collins	3251	1909-14	.3458	.3270
Tris Speaker	3253	1909-14	.3457	.3267
Nap Lajoie	2392	1909-10, 12-14	.3440	.3252
Average	2803		.3593	.3370

National League

Player	AB	Years*	MABA	SDABA
H. Zimmerman	2103	1911-14	.3142	.3138
Honus Wagner	3047	1909-14	.3084	.3056
Chief Meyers	1886	1910-14	.3083	.3075
Jake Daubert	2666	1910-14	.3043	.3043
Sherry Magee	2964	1909-14	.3008	.2997
Average	2533		.3067	.3056

* Years when player was a regular

MABA = mean-adjusted batting average
SDABA = mean- and standard deviation-adjusted batting average

Table 2. Performance of Batting Groups from 1909-14

American League

Group	AB	% AB	MABA	Net Hits
Top 5 Hitters	14,016	5.8	.3593	+1,462
Other regulars	126,513	51.9	.2690	+1,761
Substitutes	79,976	32.8	.2373	-1,416
Pitchers	23,091	9.5	.1767	-1,807
Totals	243,596	100.0	.2550	0

National League

Group	AB	% AB	MABA	Net Hits
Top 5 Hitters	12,666	5.2	.3067	+ 655
Other Regulars	139,328	56.7	.2689	+1,932
Substitutes	71,403	29.1	.2432	- 846
Pitchers	22,274	9.1	.1768	-1,741
Totals	245,671	100.1	.2550	0

% AB = percent of ABs of the league
MABA = mean-adjusted batting average
Net hits = difference between the expected number of hits of a league-average .255 hitter and the hits attained by the group

Back to Shiner's Critique

Critique 1: Pitcher's Hitting

Shiner raises a good point in his concern about pitchers' batting average and the effect it may have on the adjusted batting averages in *BABH*. However, while the early pitchers hit for a higher average, they also took a greater share of NFPs than modern hitters. Consequently, no long-term trend in the batting average of non-pitchers is apparent (data not shown). The decline in the pitchers' batting average over time can be interpreted in opposing ways. Some may claim that since modern pitchers bat less they have lost the skill that early pitchers enjoyed. Others may claim that the decline in pitchers' batting average is prima facie evidence that the league has improved. In that regard, the NL might be viewed as being ahead of the AL, since the pitchers' BA declined earlier for them. I think it is reasonable to base the adjustment method only on non-pitcher at-bats.

Critique 2: SD Interpretation

Just as ranking players on the basis of their raw batting averages or mean-adjusted batting averages alone is imperfect (otherwise the best single-season adjusted batting average was obtained by Fred Dunlap in 1884), so too is the SD-adjustment. The question should not be, "is the method *perfect*?", but "is the method *an improvement* ?". My answer to that is an unqualified "Yes". Within each improvement, however, lie clues for its failings. Just as the example of Fred Dunlap shows the limitation of only using a mean-adjustment, so the fact that the entry of players such as Cobb, Speaker, Jackson and Collins yielded a short-term *increase* in the batting average SD poses its challenge to the SD interpretation and Shiner astutely focused on it. Cobb, Speaker, Jackson and Lajoie do earn 4 of the top 11 spots in the list of adjusted top hitters in *BABH*, so it is not clear that AL hitters from 1909-14 were undervalued, as Shiner's critique suggests.

What the AL of Ty Cobb's day suggests is that there are two distinct distributions of hitters (*good* and *poor*), each with their own mean and SD. Then, rather than a model with two parameters, a mean and an SD, we would have a model with five: two means, two SDs and a mixing proportion. A second alternative, which was mentioned in *BABH* is to use three parameters, by adding the skewness of the batting average distribution to the mean and SD. A third alternative, which I am pursuing, is to look at how the league improved over time by noting how players' batting averages changed over time in the same league. If better hitters replaced weaker hitters during some period, players who bridged the transition will fall relative to the average.

Critique 3: The NL was not a better league than the AL from 1909-14

Knowing that the SD interpretation is imperfect, a specific question is whether it is on track enough for us to conclude that the NL had better hitters overall than the AL from 1909-14. Shiner tries to check by comparing players that switched leagues. Such a comparison has to be made very carefully since there are park effects and aging effects that have to be accounted for. Unfortunately, not many players switched leagues during this era (Shiner only identified 10 men with significant playing time in both leagues), severely hampering any comparison. My early investigation into using league-switching data to identify league strength has not yielded any simple, solid answers.

Figure 2 shows that the batting average of AL regulars between 1909-20, and especially between 1909-14 does stand out in an extraordinary way. Let's take a closer look at this 6-year period. Table 1 shows the mean-adjusted batting averages (MABA) of the top 5 hitters for average in the NL and AL in 1909-14, with a minimum of 4 years as a regular[6]. The top five AL hitters had a composite batting average of .3593 after adjusting it to a league-constant average of .255. The top five NL hitters only hit for a composite .3067.

Table 2 places these hitters into the context of their leagues by breaking down the performance by group. It shows that pitchers hit roughly the same in both leagues. So did the remaining regulars after the top five hitters in each league were removed. However, the AL substitutes hit for only a .237 average, compared to a .243 average for NL substitutes. Moreover, substitutes accounted for about 4% more at-bats in the AL compared to the NL. What is the net result of these differences? *Net hits* gives the difference between the expected number of hits of a league-average .255 hitter and the hits attained by the group. Not surprisingly, pitchers in both leagues had roughly 1800 fewer hits than .255 hitters would have had with equal numbers of ABs. However, note that AL substitutes were down 570 more hits than NL substitutes. I claim that it was primarily in the AL substitute group that caused the AL to be inferior in league talent to the NL overall. Since net hits for the entire league are zero by definition, where there are losers, there must be winners. This is where the top 5 AL hitters cleaned up, picking up 1462 more hits than an average hitter in their league, while the top 5 NL hitters were only up 655 hits.

[6] This comparison was suggested to me by Shiner in personal correspondence.

In Table 3, the net hits of substitute players are broken down further by position for player-years with at least 100 NFPs. Player-years with less than 100 NFPs are collapsed into a group called *Bit*, for players who only played a bit. The difference in net hits gained or lost between the AL and NL is given in the final column of Table 3. The most striking difference is that the AL catchers were 227 hits farther below average than the NL catchers. Even though I have referred to these players as "substitutes," catchers in this era were usually below the 2.5 NFP/game cutoff[7]. Additionally, although Bit players in the AL and NL had similar MABAs (.215 and .217, respectively), there were more AL ABs in this group, yielding a net difference in hits lost of -136. In four of the 5 remaining positions, AL players lost around 50 net hits per position.

Table 3. Breakdown of Substitute ABs by Position, 1909-14

	American League				National League				
Pos	N	ABs	MABA	Net Hits	N	ABs	MABA#	Net Hits	Net Hit Difference
1B	38	7,729	.2478	- 56	21	4,353	.2410	- 61	+ 5
2B	36	7,529	.2450	- 75	19	3,783	.2503	- 18	- 57
3B	21	4,988	.2341	- 104	35	6,942	.2475	- 52	- 52
C	100	19,686	.2338	- 417	88	18,819	.2449	-190	-227
OF	95	20,364	.2542	- 16	97	18,413	.2570	+ 37	- 53
SS	25	5,158	.2232	- 164	36	7,232	.2389	-116	- 48
Bit	490	14,522	.2150	- 581	405	11,861	.2175	-445	-136
Total	805	79,976	.2373	-1413	701	71,403	.2432	-845	-568

* Bit – players who had less than 100 NFPs in the season

The SD adjustment in *BABH* accounted for the relative poor performance of AL subs with at least 200 ABs, leading to a downward adjustment of batting average of the top 5 hitters.

The final column of Table 1 gives the batting averages of the players after both an adjustment for the mean (called MABA) and the SD (SDABA) are used. Note that while the averages of the AL players drop 22 points overall, they are still higher than those of the top NL players. When using SDABA rather than MABA, the net hits of the top five hitters drops by 313 hits. This is roughly half of the difference in net hits of AL and NL substitutes plus pitchers. Since Table 2 shows that the top five hitters and the other regulars gained the hits from the other two groups in roughly equal amounts, the adjustment seems to have worked quite well! Thus, this detailed examination still leads me to conclude that the overall hitting ability of NL players was superior to that of AL players from 1909-14.

Statistical Analysis Method

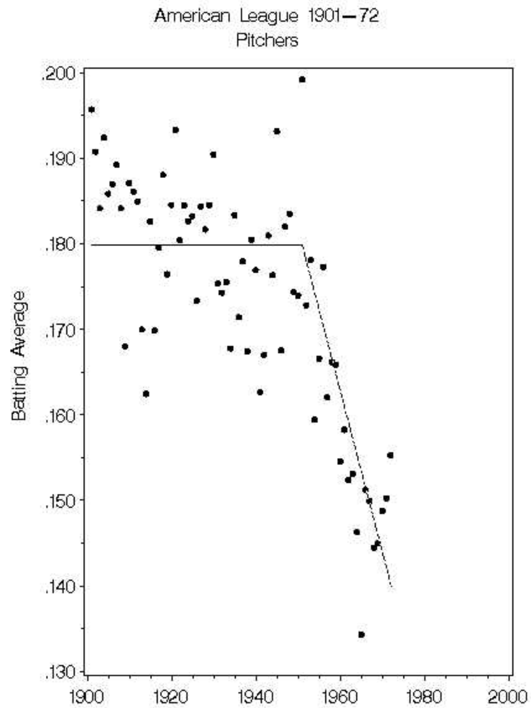
Along with the raw data, the plots in this article provide piecewise linear regression model fits. The analytical method used has been developed by the author and a full description of it is beyond the scope of this article. A brief description, however, is given here. The fit was obtained by starting with a model that links the data from 1901-96 with straight line segments. The only error for the model is generated from the linear regression fitting of the final 3 years (usually 1996-98). Thus, each year is a *knot*, where the slope changes. Then, in a backward elimination procedure, knots are removed sequentially from the model such that, at each step, the r^2 declines as little as possible. The critical statistical issue is to determine when to terminate the knot removal process. Five outlier values were identified: Pitcher % of NFPs – 1902 NL; 1902 AL; 1918 AL; Regulars Batting Average – 1901 NL; Substitutes Batting Average – 1910 NL. The locations of the knots were determined with the outliers excluded.

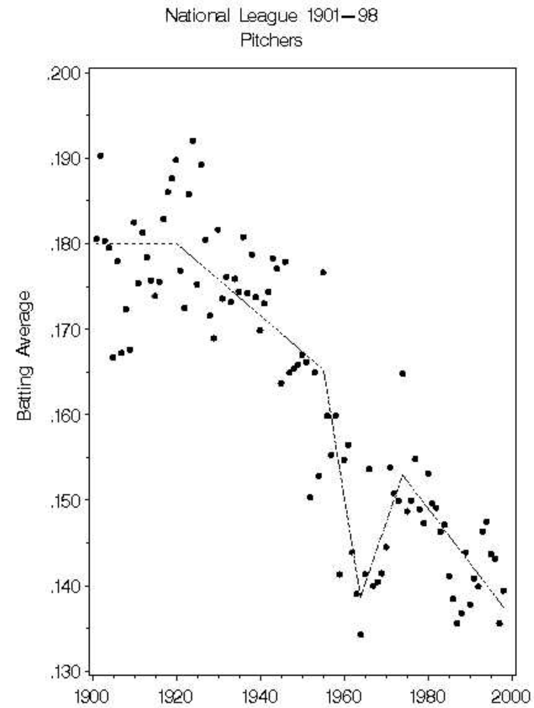
Other piecewise linear modelling algorithms exist. These will often yield a similar degree of fitting but with some knots occurring at different years. Thus, while the model fit shown here does describe some general trends in the data, one shouldn't expect to ascribe each change in slope to a particular baseball event, unless the slope change is striking.

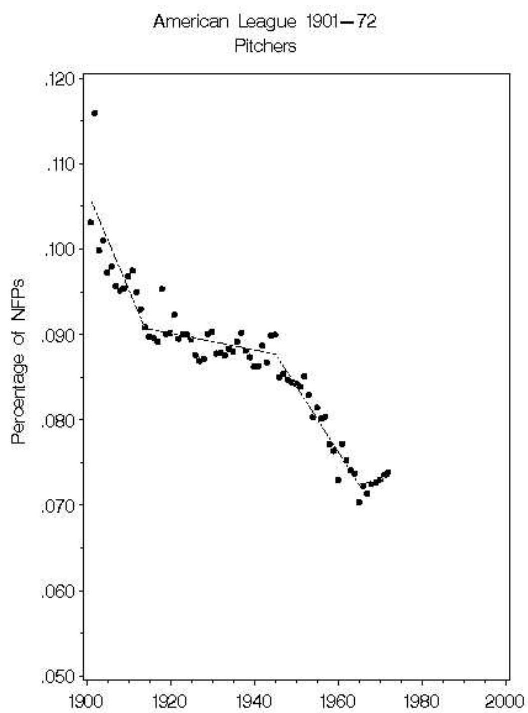
Critique 4: Equal proportion of 'great hitters' across baseball history

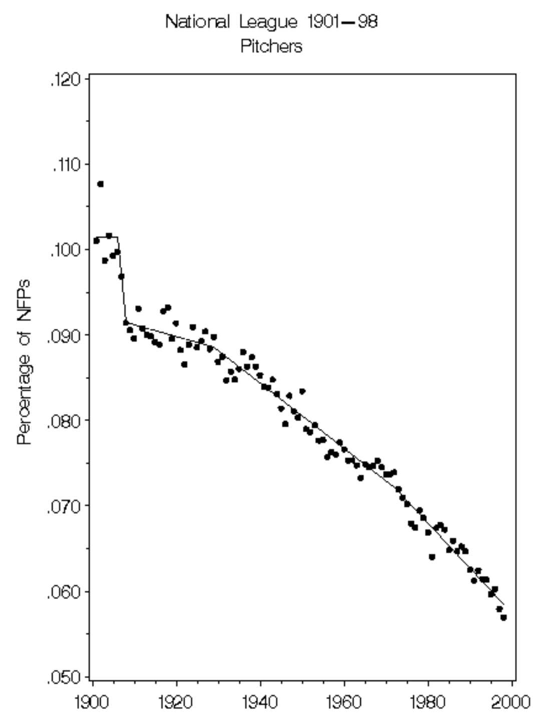
Shiner is correct that the method presented in the book assumes that there is an equal proportion of 'great hitters' across baseball history. He is also correct in surmising that I don't believe the assumption myself. I believe that the proportion of 'great hitters' has been increasing across baseball history. Thus, I believe that players like Cobb, Speaker, Jackson and Collins, as good as they were, are over-valued in my book. Unfortunately, I simply don't yet know how to demonstrate that in a meaningful statistical way. As mentioned above, I am working on ways to identify differences in "average player" ability over time. While I can see some obvious dips, such as occurred during World War II and some gains as well, I haven't yet found an approach that is superior to the SD adjustment. One of the problems in

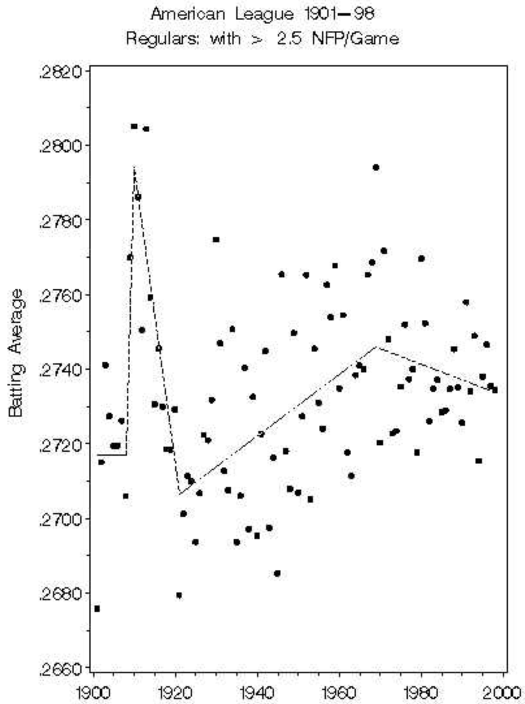
[7] This is the reason why there are 2-3 times more ABs in the catcher position than in the others (counting OF as 3 positions).

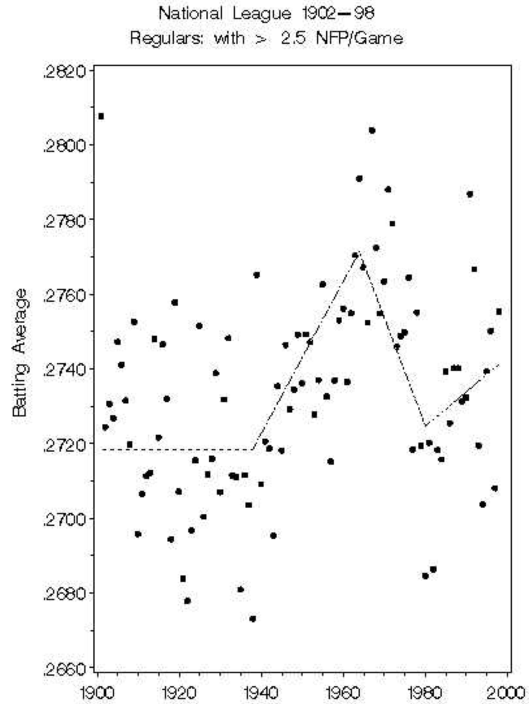

American League 1901-72
Pitchers
Batting Average
.200
.190
.180
.170
.160
.150
.140
.130
1900
1920
1940
1960
1980
2000

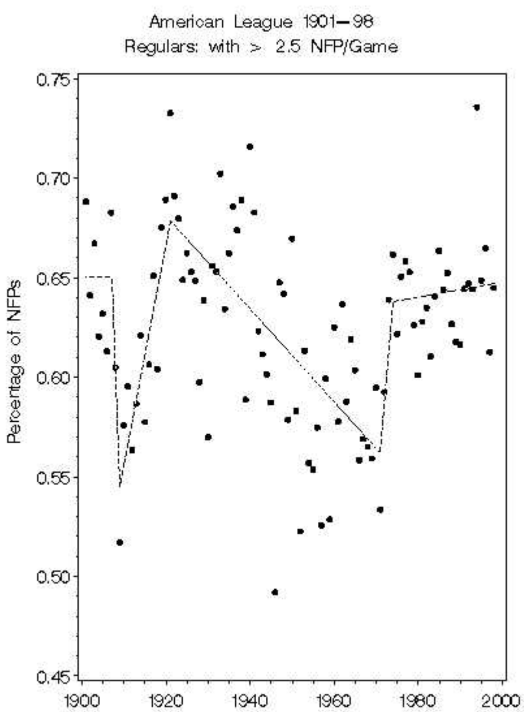

National League 1901-98
Pitchers
Batting Average
.200
.190
.180
.170
.160
.150
.140
.130
1900
1920
1940
1960
1980
2000

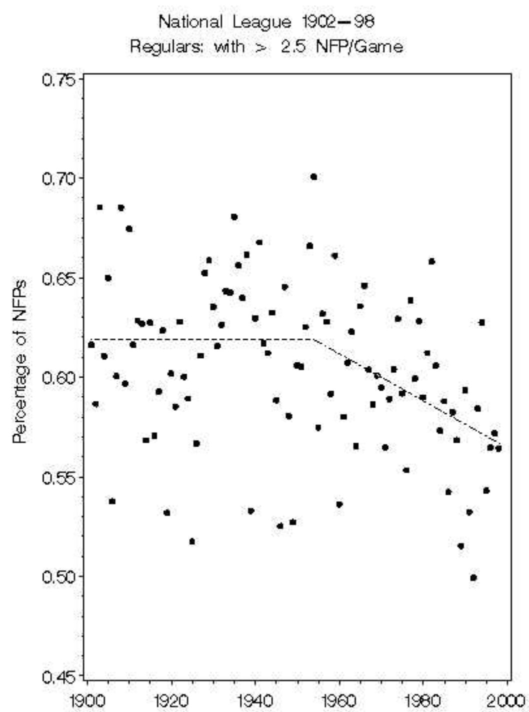

American League 1901-72
Pitchers
Percentage of NFPs
.120
.110
.100
.090
.080
.070
.060
.050
1900
1920
1940
1960
1980
2000

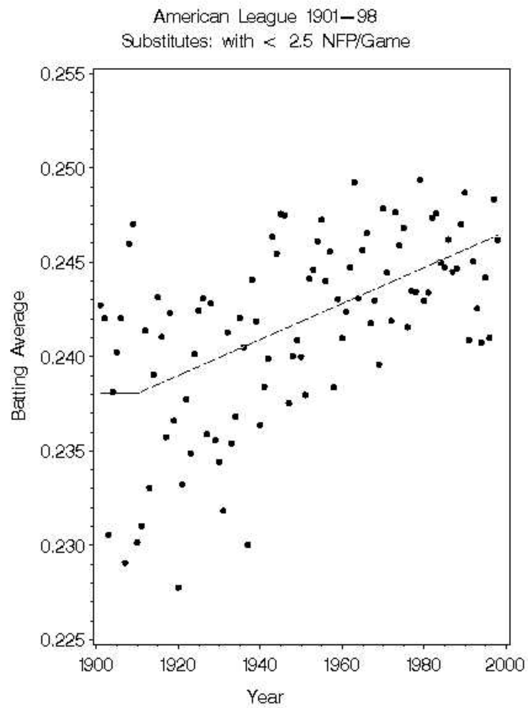

National League 1901-98
Pitchers
Percentage of NFPs
.120
.110
.100
.090
.080
.070
.060
.050
1900
1920
1940
1960
1980
2000

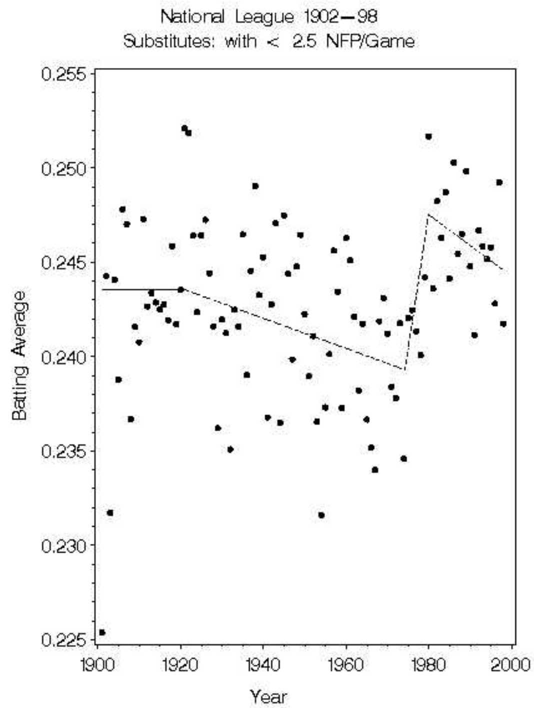

American League 1901–98
Regulars: with > 2.5 NFP/Game
Batting Average
.2660
.2680
.2700
.2720
.2740
.2760
.2780
.2800
.2820
1900 1920 1940 1960 1980 2000

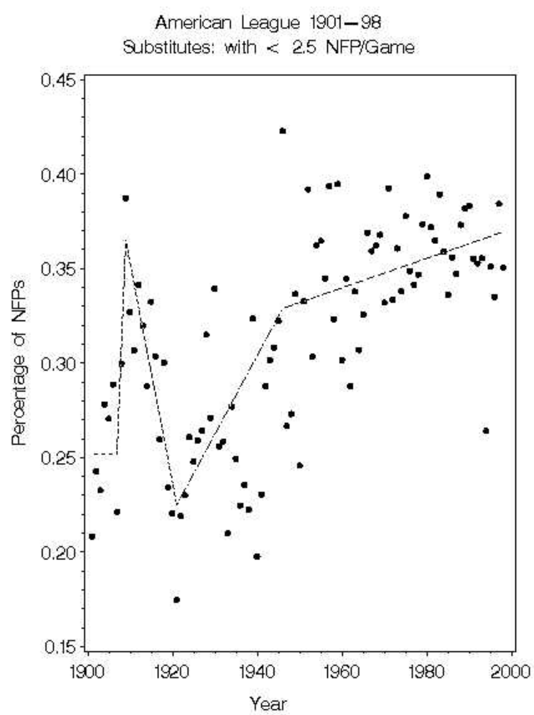

National League 1902–98
Regulars: with > 2.5 NFP/Game
Batting Average
.2660
.2680
.2700
.2720
.2740
.2760
.2780
.2800
.2820
1900 1920 1940 1960 1980 2000

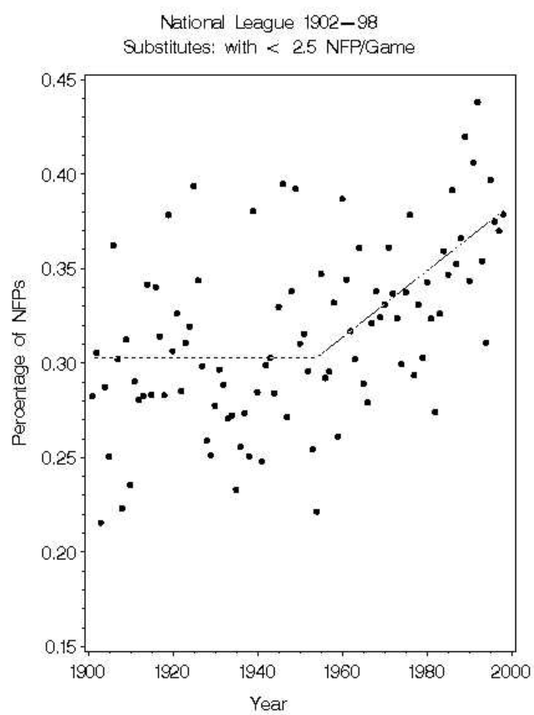

American League 1901–98
Regulars: with > 2.5 NFP/Game
Percentage of NFPs
0.45
0.50
0.55
0.60
0.65
0.70
0.75
1900 1920 1940 1960 1980 2000


National League 1902–98
Regulars: with > 2.5 NFP/Game
Percentage of NFPs
0.45
0.50
0.55
0.60
0.65
0.70
0.75
1900 1920 1940 1960 1980 2000


American League 1901–98
Substitutes: with < 2.5 NFP/Game
Batting Average
Year


National League 1902–98
Substitutes: with < 2.5 NFP/Game
Batting Average
Year


American League 1901–98
Substitutes: with < 2.5 NFP/Game
Percentage of NFPs
Year


National League 1902–98
Substitutes: with < 2.5 NFP/Game
Percentage of NFPs
Year

tracking how a player performs from one year to the next is the conditional nature of the data. Non-established players who hit poorly one year may not return, while other players who hit well one year, perhaps by random luck, return. Then, when their averages drop the following year, they are pushed aside. Methods that compare successive year changes of players are complicated by this "sophomore jinx" effect.

In conclusion, I commend David Shiner for an excellent review. He phrased his critique so carefully that I have a hard time disagreeing with it. However, regarding the SD adjustment and Cobb's AL era, Shiner states: "I don't claim to know how or when to do that, but I do know that Schell's well-intentioned adjustment isn't the answer, at least in this case." While the increase in batting average SD in the AL in the early 20th century troubles me too, I actually think that it works quite well in Cobb's day. The problem is more likely to be in the pre-Cobb AL era. I do believe that the method presented in my book provides our best look at the ability to hit for high average to date. Since Shiner hit at the weak links in my argument, I can't resist making hay with one of his. Shiner writes:

> "If we look at Schell's argument more abstractly, it also fails to convince. Remember Doug Gwodsz? … [If Gwodsz played on a regular basis and increased the batting average SD by his low average], would that indicate that the league in which the Doug Gwodszes of the world played was inferior to the other one? Can you say George McBride?"

My response is that Doug Gwodsz could not have affected any of my SD calculations since he had only 104 lifetime at-bats, and it took him 4 seasons to get them. He is scarcely a match for George McBride, the proud owner of a 5526 at-bat history over 16 years. Can you say "Puh-leeez"?

Shiner understood that *best hitter* in the title of *BABH* referred to *best hitter for average*. Shiner did jump on the subtitle, though, when he said: "In my opinion, we're talking about adjustments that can *help* level the playing field, but we're deceiving ourselves if we believe we'll eventually find that perfect method that will render objective truth." Given that, let me jump on the title of Shiner's review (which was actually written by *BTN* editor Phil Birnbaum). The methodology presented in *BABH* was called "debatable." That could mean that the methodology is not very good. On the other hand, it could mean that it is worthy of debate, which I hope was the intended meaning. The title of Shiner's review was probably chosen for its provocative appeal. Book titles are similarly bold. We could have called it "Baseball's All-Time Best Average Hitters." That would *really* have thrown some people for a loop!

Michael J. Schell is a biostatistics professor in Chapel Hill. Michael Schell, 3607 Jones Ferry Road, Chapel Hill, NC 27516; mjschell@mindspring.com ♦

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
Jim Box	im.box@duke.edu	Statistics
Keith Carlson	kcarlson2@mindspring.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	HanrahanTJ@navair.navy.mil	Statistics
Keith Karcher	kckarcher@compuserve.com	General
John Matthew	john.matthew@home.com	Apostrophes
Duke Rankin	RankinD@montevallo.edu	Statistics
John Stryker	johns@mcfeely.interaccess.com	General
Dick Unruh	runruhjr@dtgnet.com	Proofreading
Steve Wang	Steve.C.Wang@williams.edu	Statistics

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 (http://support.microsoft.com/support/kb/articles/Q165/9/08.ASP).

To get on the electronic subscription list, send me (Phil Birnbaum) an e-mail at birnbaum@sympatico.ca. 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.