By the Numbers

Volume 10, Number 4 The Newsletter of the SABR Statistical Analysis Committee November, 2000

Letters to the Editor

In this issue

Ball-Strike Causation

Phil Birnbaum suggested on SABR-L (2000-10-10) that more runs score after a walk than after a single because the pitchers who give up walks tend to be worse. That insight bears on his article "The Run Value of a Ball and Strike" (BTN, 2/2000). I do not quibble with his run values of each ball-strike count, but with their use to measure what a catcher contributes by successfully "framing" a pitch; that is, by handling a pitch that should be a ball in a way that induces the umpire to call "strike".

Consider the 2-1 count for example: a pitcher who achieves 2-1 with one pitch successfully framed by the catcher tends to be worse than a pitcher who achieves 2-1 with one pitch in the strike zone. Therefore, the difference between average run values of the 3-0 and 2-1 counts overstates what the catcher contributes by framing one pitch (unless successful framing directly affects the pitcher, batter, or umpire in a way that is helpful).

Re: the same article, here is a fourth "way that a catcher can influence a pitcher's effectiveness": (d) by catching foul tips. By the way, a broadcaster noted during this year's playoff that a coach asked Mike Piazza to move up in the box, even though he would not be able to catch as many foul tips. I could not infer whether catching foul tips was Piazza's reason for playing back.

Paul Wendt, Watertown, MA
pgw@world.std.com

“Sinkers Down the Middle” not advised

Re: Tom Hanrahan's comment on hits-per-balls-in-play (BTN, 8/2000):

I support Tom’s call for further research. But his question "Why not just throw sinkers down the middle of the plate?" is not a sound criticism of the plausibility of Clifford Blau's assertion. This is because Blau's assertion is that, given the way pitchers mix (and have mixed) their pitches, there is no difference in ratios of hits to balls in play. If they pitched differently, as Hanrahan suggests, one would expect the results might be different.

This is an example of the elusiveness of "holding all other variables constant" when designing experiments, whether thought-experiments or empirical ones.

Philip Siller, Toronto, ON
Psiller@hexagram.net

Correction

In the August, 2000 issue of BTN, the left side of the second equation on page 28 should read “SA POP”. It reads “Pn”, which is incorrect. Apologies for the editing error.

Letters to the Editor are welcomed, as are all submissions. Authors mentioned in a letter may be asked to comment in the same issue in which the letter appears. See “Submissions,” elsewhere in this issue, for addresses where letters or articles may be sent. ♦

Academic Research: Racial Bias in Hall Of Fame Voting

Charlie Pavitt

The author summarizes academic research on the topic of whether Hall of Fame voting patterns exhibit evidence of racial bias.

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.

Arna Desser, James Monks, and Michael Robinson, Baseball Hall of Fame Voting: A Test of the Customer Discrimination Hypothesis, Social Science Quarterly, Vol. 80, 1999, pages 591-603.

In the February 2000 issue of BTN, I reported on the latest in a series of studies examining racial bias in assignment of fielding positions. This is just one of several topics relevant to discrimination in baseball that has received research attention. There have been demonstrations of racial bias in salary among major leaguers, and even in the value of baseball cards. The present study is the second I have found which attempts to determine whether there has been race-based discrimination in baseball writers' Hall of Fame voting patterns. The first (Findlay & Reid, Economic Inquiry, Vol. 35, 1997, pages 562-578) found underrepresentation among Black and Latin American position players relative to performance measures for voting from 1962 (when Jackie Robinson's eligibility began the relevant era) through 1995. In other words, it took better performance for Blacks and Latins than Whites to receive the equivalent percentage of votes. However, the authors also found evidence that discrimination had disappeared during the latter years of this period.

Desser, Monks, and Robinson also limited their study to position players, in this case between the years 1976 and 1998. They examined the impact of race on three indices: the probability of being chosen for the ballot the first year of eligibility, the percentage of votes in the first year for those chosen, and the aggregate percentage of votes across years of eligibility for those remaining on the ballot. In addition to race, they examined the effect of a large set of other measures, some performance oriented, others not (such as experience as a manager and the opportunity to play in large media centers). They found marginal support, significant at .10 but not at the customary social science value of .05, for discrimination against both Black Americans and Latin Americans for all three indices. However, they concluded that the effect was too small to have had much impact on actual Hall of Fame induction. Some of the other predictors of voting patterns are also of interest. The raw number of HRs, RBI and runs was not as predictive of first-time and overall percentage of votes as the achievement of various milestones (300 HRs, 1350 RBI and 1450 runs). In addition, length of career, All-Star game appearances, and number of MVP awards were strong predictors. Experience as a manager had a significant impact on being chosen for the ballot. Finally, there is evidence that the statistical criteria for being chosen for the ballot is becoming more demanding over time.

As is so often the case in these types of studies, the difference in results between Findlay and Reid (who claimed discrimination had ended by about 1980) and the present authors are quite possibly the result of differences in methodology. Findlay and Reid used more sophisticated statistical procedures, but it could be argued that Desser et al. used a better mix of other measures, which could lead to discrimination standing out more strongly in the analyses. Although both studies imply that the effect is too small to worry about, it is likely that the last word on this issue has not yet been written.

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

"Baseball's All-Time Best Hitters": Debatable Methodology

David Shiner

"Baseball's All-Time Best Hitters," this reviewer says, is interesting and challenging, but flawed in some of its methodology, such as its use of league SD as a proxy for talent. Nonetheless, this well-written book is recommended reading for those with an interest in the subject.

First, let's be clear about what Michael Schell is doing. As he says on page 5 of his recent book, "The search in this book is for the best *hitters* [italics his], that is, the players with the best chance to get a hit in a given at bat." This means that he's looking for the "all-time best hitter" in the old-fashioned sense, where "hitting" is distinguished from "hitting for power." So the main question his book addresses is this: Whose career batting average is the most impressive in major league history?

Baseball's All-Time Best Hitters

By Michael Schell

Princeton University Press, Hardcover, 328 pages, $22.95

ISBN 0691004552

Schell's book is subtitled "How Statistics Can Level the Playing Field." We all know, or at least believe, that Bill Terry's .401 average in 1930 isn't as much superior to Carl Yastrzemski's .301 mark in 1968 as the 100-point differential would seem to indicate (as discussed recently in the pages of *BTN* by Mike Sluss and Rob Wood). That's because the eras in which Terry and Yaz played were very different. Schell tries to account systematically for such differences, adjusting the career batting averages of the game's greatest players so we can tell how they compare.

Since most of this review will consist of criticisms, I should preface my remarks by mentioning that I enjoyed the book. It's well-written and flows better than most works of this type. Schell understands statistics very well; he also understands baseball. He's thoughtful and writes with a light touch. Although I don't always agree with his reasoning or assumptions, that doesn't mean that I consider them outlandish or ridiculous.

Schell makes four major adjustments to each player's career batting average. The first is a longevity adjustment, which essentially means that players aren't penalized for playing for 20 years instead of 15. This helps out players like Henry Aaron and Willie Mays, whose batting averages declined as their abilities did. The second is for offensive context, especially the offensive context of the era, which brings Yastrzemski's feat closer to Terry's. The third is an adjustment for the talent pool; more on that below. The fourth is for each player's home ballpark. While I have problems with some details of the adjustments, I'll restrict my discussion here to a few major points.

The longevity adjustment cuts off a player's batting average after 8000 at bats. While this approach has its drawbacks, Schell demonstrates to my satisfaction that alternative assumptions are at least equally problematic. While it could be argued that those who play long past their prime deserve to have their batting averages discounted as they pile up numbers in other categories, my concern with Schell's approach is at the other end. Since he considers every retired player who amassed at least 4000 at bats, the method appears to favor players who hung 'em up after 5000 at bats rather than 8000 or more. I would think, for example, that the longevity adjustment really helps someone like Joe Jackson, who essentially had no decline phase to his career.

In actual fact, Schell's methods drop Jackson a notch from his unadjusted third-place position. So maybe I'm wrong about that, or maybe Shoeless Joe isn't a good example. Still, when I look at the hitters who rank high on Schell's list without many batting titles, most of them were relatively short-career types: Jackson (4th), Pete Browning (12th), Kirby Puckett (15th), Tip O'Neill (tied for 16th), and so on. Thurman Munson, a major leaguer for ten years, emerges as the best-hitting catcher of all time, ahead of several Hall of Fame receivers whose batting averages were higher, in some cases much higher. Maybe an adjustment that invokes a standard decline, from the end of each player's career ended through a hypothetical 8000th at bat, would help here.

The main problem with the adjustment for offensive context, which is otherwise well handled, has to do with pitchers' hitting. This issue is spotlighted by Schell's discussion of the effect of the designated hitter. As part of his effort to determine each player's batting average relative to his offensive context, Schell invokes a "mean adjustment" to account for the difference between a DH league (the AL since 1973) and a non-DH league (all the rest of them).

I'll explain the rationale for this adjustment by means of an example. In 1990 Lenny Dykstra and Rickey Henderson both posted .325 averages. Without the mean adjustment, Dykstra's average would be more impressive than Henderson's, because the overall NL average was .256 and the AL average was .259. But since the American League is a DH league, the AL average *should* be higher. That's because part of the National League average includes pitchers coming to bat on several thousand occasions, as compared with hardly any in the AL. In fact, since the differential between the leagues is normally more like 8 points than 3, Henderson's feat is actually the more impressive one, all else being equal. And Schell's adjustment accounts for that.

While the mean adjustment works well for the period since the DH was introduced, it raises the issue of whether all other leagues can be considered "constant" simply because there was always a pitcher in the batting order. That's the effect of Schell's mean adjustment; but it's problematic, because the further back into history we travel, the better pitchers used to hit. This can be demonstrated in a number of ways. To take a quick example, if we look at the pitchers from the dead-ball era, every man who won more than 300 games (Young, Johnson, Mathewson, Alexander, and Plank) hit better than .200 over the course of his career. But of the pitchers of the last half-century who won more than 300 (Spahn, Carlton, Sutton, Ryan, Niekro, Perry, and Seaver), the only one with a batting average of .200 or higher was Steve Carlton at .201.

Schell points out that pitchers averaged .145 for the years 1969-72. That's about what Red Ames hit during his dead-ball-era career, and he was considered the worst-hitting pitcher in baseball at the time. Facts of this sort render Schell's use of a uniform average for all pitchers in major league history extremely suspect. While adjusting for the decline in pitchers' hitting would be complicated, I would guess that it's almost as significant an issue as the advent of the DH.

In Chapter 4, "Adjusting for League Batting Talent," Schell makes what I consider the most questionable major assumption in the book. Following the argument made by paleontologist and baseball aficionado Stephen Jay Gould in his book *Full House* and various articles (which, by the way, should be mandatory reading for every sabermetrician), Schell claims that the strength of a given talent pool is inversely correlated with variability in player performance. That leads him to discount the batting averages of players from periods when the standard deviation (SD) in batting averages was relatively large, on the grounds that play during that period must have been correspondingly weak.

Schell's "poster child" for the talent pool adjustment is the American League from roughly 1908 through 1919. He gives the SD of the AL in 1910-14 as .043, while the SD for the NL during that period was a historically more typical .031. Thus, while Ty Cobb was batting around .400 every year, better than any National Leaguer, Washington shortstop George McBride was struggling to clear the Mendoza line, worse than any National Leaguer. Like Mario Mendoza, McBride was a terrific fielder but not much of a hitter. Schell's conclusion (page 91):

> McBride's poor hitting made Cobb's look *relatively* better. The overvaluation of Cobb certainly did not depend *solely* on McBride. Many of Cobb's American League compatriots would likely have ridden the National League bench during the same era and would not make the major leagues today.

The adjustments for this factor are the largest in the book, and they ultimately give Tony Gwynn the nod over Cobb in a photo finish. But that can't be right. Here's why.

As Schell demonstrates, the SD of the National League was actually higher than that of the American during the first few years of the 20th century. That already undermines the rationale for the talent pool adjustment, since even the AL's most diehard advocates concede that the older league was stronger during those years. But while the AL's SD goes up from about 1908 and remains high until the end of the dead-ball era, the NL's goes down during the same period, even though baseball historians have long since assumed that the two leagues had achieved approximate equality by then. Schell argues on behalf of these findings as follows: "When outstanding hitters – Ty Cobb, Eddie Collins, Tris Speaker, and Joe Jackson – joined the American League between 1905-8, the variability increased because they outshone the rest of the league….As the league became stronger, the SD dropped" (page 94).

So is it possible that, contrary to what we normally think, the NL was really stronger than the AL during those years? Despite the ingenuity of Schell's argument, I doubt it. How can adding players the caliber of Cobb, Collins, Speaker, and Jackson make a league *weaker*? Besides, all the other evidence seems to indicate that, if anything, the AL was the stronger league during that period. Their representatives won the World Series most of the time, for one thing. For another, a study of the records of men who played regularly in both leagues during that period belies Schell's claim. There weren't many – players didn't switch leagues much in those days, unless it was to jump to the Federal League or to return after it folded – but there were enough for the purposes of analysis.

I found ten men who played regularly (more than 100 games and/or 400 plate appearances in a season) in both leagues during the 1908-1919 period. Alphabetically, those players were Lena Blackburne, Hal Chase, Bill Hinchman, Dick Hoblitzell, Wade Killefer, Larry Kopf, Ivy Olson, Dave Shean, Harry Wolter, and Rollie Zeider. Taken as a group, they averaged .260 in their best American League season, .275 in their best year in the National. The fact that they hit better in the NL would seem to indicate that the AL was the stronger league at that time.

Furthermore, most of these players' careers fell into a similar pattern. Seven of the ten had their best full-season AL batting average during a ballplayer's normal "peak" age (25-28), then declined, left the league, and resurfaced in the National League to play regularly in their early

30s. Now, if a majority of the players who switched leagues faded out of the American League in their late 20s and came back to play as well or better in the NL in their early 30s, wouldn't that suggest that the AL was the superior league at that time? Or, at least, that Cobb's AL contemporaries would have been just as successful in the National League as in the junior circuit?

Ten players hardly constitute an airtight case, but the rest of the data tends in the same direction. If you take a look at players who didn't quite qualify for the above list, like Gavvy Cravath, Charley Deal, Wally Gerber, and Charley O'Leary, you'll discover that a lot more of them washed out of the AL at their "peak" and later succeeded in the NL than the other way around. Indeed, if you look for evidence that the NL was better than the AL during that period, apart from the Gould/Schell hypothesis, it's awfully hard to find.

If we look at Schell's argument more abstractly, it also fails to convince. Remember Doug Gwodsz? He was the catcher who played a few games for San Diego in the early 1980s. He couldn't hit .200 – heck, he couldn't hit .150 – but whenever Dick Williams sent him out to catch, the Padres won. As Bill James and Craig Wright noted at the time, that was because his defensive contributions more than offset his offensive ineptitude. Now, if some team were to hire a gutsy manager who started playing people like Gwodsz on a regular basis, most of us would probably say something like this: "Finally, a baseball insider who understands sabermetrics." The effect of moves like that, though, would be to increase the standard deviation of batting averages. 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?

I've gone on at some length about this issue because the effect of the talent pool adjustment on the adjusted batting averages is enormous. Every player listed in the book who played before World War II, regardless of era, had his average decline as a result of Schell's methods, while most of those who have played since then went up. The SD adjustment alone costs Ty Cobb a whopping 24 points off his adjusted career batting average, which is practically unaffected by the other three adjustments. For that reason alone, his 20+-point advantage over Tony Gwynn dwindles to less than nothing. And it's the claim of Gwynn's ascendance to the top spot that has brought the book into the public eye.

The SD adjustment also costs Napoleon Lajoie 17 points, but that's less than Cobb or Speaker or Jackson gives up. Essentially, that's because the AL had a lower standard deviation in 1901 than in 1911. But, again, that can't be right. Lajoie was pounding a group of marginal major league pitchers en route to his AL-record .422 mark in '01. The pitchers he and his younger colleagues faced a decade later were far superior.

I'm not saying that the talent pool adjustment is uncalled for. As Schell points out, it's hard to know how else to statistically validate, say, the relative weakness of the Union Association of 1884. That's a case where the method works pretty well, but there are others where we probably need to declare that a particular result is simply a statistical aberration. 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.

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" (page 27). I'm not convinced that this assumption can coexist peacefully with Gould's arguments concerning standard deviation, and I'm not sure that Schell is either. After all, if Gould's argument were true, it would follow that "the reduction in standard deviation demonstrates that there has been an improvement in the overall quality of major league baseball today compared to nineteenth-century and early twentieth-century play," as Schell affirms on page 95. Even after making the adjustments for the talent pool based on standard deviation that lead to the radical demotions of Cobb, Speaker, Jackson, et al, he continues to claim that "my approach still assumes that the best players are comparably good over time" (pp. 98-99). I don't have the space to analyze these comments, but it's evident that Schell is walking a tightrope here. And when he later declares that "[t]he message of this book is that today's hitters are among the best of all time" (p. 251), well, he'd better hope that a safety net has been secured underneath.

Schell's decision to make exactly four adjustments is, of course, open to question. In fact, he raises the issue himself when he notes, several times, that lefty hitters fare better than righties. He's plainly unsure whether to adjust for this. Ultimately he decides not to do so, which seems to me like the right decision. After all, with apologies to Pete Gray and Jim Abbott, we could also adjust for having one hand instead of two. But this does raise the crucial issue of how much it's possible to "level the playing field." Another example is whether there should be a position adjustment, as in Pete Palmer's methods, since outfielders have historically posted higher averages than infielders or catchers (or, for that matter, pitchers). These are not easy questions. The good news is that they don't need to be answered definitively. 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.

I'll conclude by noting some of the really nice features of this book. Schell tosses in some pleasant diversions, like naming the "All-Alabama Team of 1961," which are simply delightful. He makes a number of fresh observations, pointing out for example that the best hitters' park for average in the National League during the 20th century was Forbes Field in the 1960s, at least until Coors came around. I'd never have guessed that. He adjusts for on-base averages, single-season batting averages, and other issues related to his main topic. He explains basic statistical concepts like chi-square and linear regression briefly and effectively. His charts, graphs, tables, and sidebars are to the point and easy to follow.

Despite its flaws, and maybe partly because of them, *Baseball's All-Time Greatest Hitters* is an interesting and challenging book. Pick up a copy. Read it. Mark it up. Argue with it. That, my friends, is how sabermetrics makes progress.

David Shiner, 706 Washington, Knollwood, IL, 60044, cunegonde@prodigy.net. ♦

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

In Defense of Horace Clarke: A Comparison of Two Defensive Measures

W. T. Rankin

From 1969-1973, Horace Clarke of the Yankees and Davey Johnson of the Orioles were rival American League second basemen. Johnson was considered the better defensive player. Was this perception justified? Here, the author examines the relevant statistics, and also comments on the relative merits of two defensive evaluation tools.

Introduction

My first visit to a major league ballpark was a Saturday game between the Twins and Yankees in the late 1960's. Dave Boswell beat the Yankees 1-0 on a three-hitter. The only run scored on an infield grounder to Yankee second baseman Horace Clarke. With runners on first and third and one out, Clarke eschewed a routine double play and tried to tag the runner between first and second. The runner evaded the tag until the batter reached first, eliminating the inning-ending double play and allowing the runner on third to score.

The next day, Ralph Houk, the Yankee manager, defended Clarke in the newspaper report of the game. It was not to be the only time. Gradually, the notion that Horace Clarke fielded poorly sifted into the conventional wisdom of Yankee fans. If only we had a real second baseman like Davey Johnson, the three-time Gold Glove of the Baltimore Orioles.

What I didn't know at the time was that Horace Clarke routinely led AL second basemen in several defensive categories. In fact, Horace Clarke almost always posted better fielding numbers than Davey Johnson. So who was better: Clarke or Johnson?

Although I framed the initial question in terms of Clarke and Johnson, I quickly realized I needed to analyze all AL second basemen of the period to establish a context for the analysis. My interest, therefore, became the best defensive second baseman in the AL during Horace Clarke's tenure with the Yankees -- from 1967 until 1973.[1]

Methods

First, I needed a procedure to compare the defensive abilities of the players. Bill James [1982] proposed two methods: defensive won/lost percentage (DWL%) and defensive wins (DW). Alternately, John Thorn and Pete Palmer [1993] have used fielding runs (FR) to measure defense. Which of these procedures more accurately measures defensive effectiveness?

I confined the study to the regular second basemen (n = 27) listed in the team section of the 8th edition of the Macmillan Encyclopedia [1993].[2] For each player, I recorded seasonal FR from the player register in Thorn and Palmer [1993]. I also used raw fielding data from the Macmillan Encyclopedia to calculate DWL% and DW following James [1982].[3]

In addition to determining the best defensive player in each season, I sought to determine defensive play throughout the period. For each player with three or more seasons of fulltime play in the sample period [n = 13], I calculated DWL% and mean DW by totaling fielding wins

[1] Clarke broke in with the Yankees in 1965, playing 25 games in the field, mostly at third base. In 1966, he appeared in 83 games, primarily at shortstop. Clarke became the regular second baseman in 1967, playing at least 139 games each season through 1973. Clarke's tenure with the Yankee almost exactly overlaps Johnson's tenure with the Orioles.

[2] Some second basemen played other positions, but none to a degree that may have substantially altered the analysis. I also used defensive games played [as opposed to total games played] for all analyses.

[3] The calculations are described on pp. 211 - 213. DWL% is a 100 pt, weighted scale based on the following criteria: double-play percentage adjusted for runners on base, range factor minus double plays, fielding average and team defensive efficiency. DW is DWL% adjusted for games played.

I was able to extract most, but not all, of the necessary information from Thorn and Palmer [1993] to calculate DWL%. The most problematic data were the team defensive efficiency records, which I estimated from the yearly assist totals in the team section. These errors, however, are unlikely to exceed 1% of the players' DWL. I also had difficulty interpolating James' charts. I settled on a fixed schedule; to achieve a given standard, a player had to reach the standard, otherwise he would be assigned the next lower standard. This introduced the possibility of rounding errors approaching 2% of the player's final DWL.

and losses for each season. Average FR was estimated as the arithmetic mean of the seasonal FRs. Although this is almost certainly a miscalculation, Thorn and Palmer [1993] do not offer sufficient information for me to combine FR among seasons in a more meaningful manner.

Results and Discussion

1. *Comparing procedures.* Because FR and DW ostensibly measured the same thing, I expected the two variables to exhibit a strong, positive correlation. The expected correlation did not materialize (Table 1). As a result, I was forced to choose between the procedures. I believe DWL% and DW are more appropriate than FR for assessing defensive effectiveness for the following reasons:

- DW and DWL% make more intuitive sense. For example, Horace Clarke's fielding statistics in 1968 and 1969 were roughly the same: almost exactly 800 putouts plus assists and fielding averages within 2 points [Table 2]. Clarke turned 32 more double plays in 1969 but appeared in 17 more defensive games, so his chances per game were lower in 1969. Still, two pretty good seasons, each characterized by three league-leading totals. According to FR, Clarke went from the best defensive player in the league in 1968 to below average in 1969. I don't doubt the accuracy of the numbers or the importance of the linear weights concept, but I have trouble accepting that the same player, having roughly the same year, can be rated so dramatically different in two adjoining seasons. DWL% and DW, however, rate these as similar seasons -- a DWL% of .640 both years, and DW of 4.39 in 1968 and 4.93 in 1969 (the better season due to the higher double-play rate).

- DW is weighted towards large samples; FR isn't. DW is normalized to a 162 game season, rewarding players who play well for many games. The mean number of defensive games for the ten best DW seasonal totals in the sample period is 149 games. The mean number of defensive games for the ten best FR seasonal totals, however, is 130 games -- a significant difference (p = 0.03, one-tailed t-test).[4] Given a choice, I would prefer the player who plays well for more games.

- Both career DW and DWL% are positively correlated with career longevity. For the 27 players in the sample, both career DW (p = 0.02; R-squared = 0.21) and DWL% (p = 0.01; R-squared = 0.23) regress significantly onto the total number of seasons played during the period. No such relationship exists for FR (p = 0.22; R-squared = 0.06). Players are significantly more likely to lose their jobs as a result of decreases in their DWL% and DW than they are with comparable decreases in FR. The correlation between DW and job loss is quite strong. During the sample period, 14 players lost their jobs. In 1967, Wayne Causey and Pedro Gonzalez were rated last and next to last in DW; both lost their jobs. In 1968, Vern Fuller was rated last in DW. He lost his job in 1969 when he was ranked next to last. The man ranked below Fuller in 1969, John Donaldson, also lost his job. In 1970, Danny Thompson was ranked last and lost his job. In 1971, Eddie Leon was ranked last and lost his job. In 1972, Tim Cullen was ranked last and lost his job. Cullen was only a part-time player (70 defensive games) and probably doesn't count, but Lenny Randle was ranked next to last in 1972 and lost his job. Jorge Orta was ranked last in 1973; he moved to the outfield in 1975.

Table 1: Correlations between FR and DW

Season	p value	R^2
1967	0.18	0.21
1968	0.03	0.46
1969	0.88	0.00
1970	0.16	0.19
1971	0.33	0.09
1972	0.09	0.26
1973	0.25	0.13

Table 2: Horace Clarke's fielding record, 1968 - 1969

Season	PO	A	E	DP	TC/G	FA	FR	DWL%	DW
1968	357*	444*	13	80	5.9*	.984	30	.640	4.39
1969	373*	429*	15	112*	5.2	.982	-4	.640	4.93

PO = putouts, A = assists, E = errors, DP = double plays, TC/G = total chances per game, FA = fielding average, FR = fielding runs, DWL% = defensive won/lost percentage, DW = defensive wins, * = league-leading total.

DW cannot be used uncritically. Clearly, holding a job involves more than defensive ability. In addition, DW may be descriptive, not prescriptive: because DW is normalized to playing time, a poor second baseman may suffer reduced playing time, thereby decreasing his

[4] The mean number of games for the ten best DWL% seasons equals 138 games (no significant differences between either FR or DW). The mean number of games for all player-seasons in the study equals 127 games, but a substantial number of these seasons are clearly part-time players nominally listed as starters by Macmillan [1993].

DW. Compared to FR, however, DW and DWL% offer much stronger correlations between fielding results and career longevity. Compared to DWL%, DW has the added bonus of reflecting playing time. As a result, I will use DW as the primary variable in assessing the defensive ability of the players, and use DWL% and FR as supplementary assessments. I will also use the number of league-leading defensive totals as a supplementary assessment -- in effect, a black ink test of defensive play. Although clearly not as sophisticated as the sabermetric measures, leading the league in a defensive category other than errors seems an unambiguously positive attribute, and certainly suggestive of the best player.

Table 3: Individual fielding records for AL second basemen, 1967-73

1967	DW	DWL%	FR	Leader
Knoop [GG]	4.32	0.550	1	po, dp
Clarke	3.94	0.570	15	*a, tc/g, fa*
Johnson	3.12	0.430	0	
McAuffle	2.80	0.370	-14	*e*
Allen	2.41	0.650	19	
Andrews	2.29	0.320	-6	
Carew	2.18	0.330	-10	
Donaldson	1.61	0.320	-18	
Causey	1.40	0.290	-8	
PGonzalez	0.67	0.180	-5	
1968				
Clarke	4.39	0.640	30	*po, a, tc/g*
Knoop [GG]	4.32	0.580	16	*dp*
Andrews	3.70	0.520	1	
McAuliffe	3.58	0.480	-20	
Johnson	2.62	0.330	4	
Allen	2.54	0.460	5	*fa*
Carew	1.85	0.310	-14	*e*
Donaldson	1.80	0.350	-4	
Alomar	1.33	0.200	-16	*e*
Fuller	1.04	0.210	-12	
1969				
Clarke	4.93	0.640	-4	*po, a, dp*
Johnson [GG]	4.48	0.630	-3	
Green	3.95	0.610	-2	*fa*
Alomar	3.77	0.490	-7	*e*
Knoop	2.91	0.450	24	*tc/g*
Andrews	2.73	0.460	5	
Carew	2.51	0.430	-17	
Allen	2.32	0.420	3	
Adair	1.75	0.300	-22	
McAuliffe	1.67	0.470	0	
Fuller	1.61	0.320	9	
Donaldson	1.04	0.340	0	

DW=defensive wins, DWL%=defensive won/lost percentage, FR = fielding runs.

2. *The best second baseman on a yearly basis*. In 1967, Bobby Knoop led AL second basemen in DW, finished third in DWL%, third in FR, and led two fielding categories. Clarke finished second in DW, second in DWL%, second in FR and led three fielding categories. Bernie Allen led AL second basemen in DWL% and FR, but finished fifth in DW. Clarke and Knoop were clearly the two best second basemen in the AL, but the 0.38 DW differential is substantial. Gold Glove to Knoop.

In 1968, Clarke led AL second basemen in DW by a slight margin over Knoop, but also led in DWL% and three fielding categories. In addition, Clarke's FR of 30 led the American League. Gold Glove to Clarke.

In 1969, Clarke led AL second basemen in DW, DWL% and three fielding categories. Knoop led AL seondbasemen in FR and one fielding category, but finished 2.0 DW behind Clarke. Dave Johnson finished second in DWL%, but, once again, the 0.45 DW differential between Clarke and Johnson is substantial. Gold Glove to Clarke.

In 1970, Sandy Alomar led AL second basemen in DW, DWL% and two fielding categories. Knoop led the American League in FR, but finished well behind Alomar in DW and DWL%. Gold Glove to Alomar.

In 1971, Alomar led AL second basemen in DW and FR, finished second in DWL%, and led one fielding category. Johnson led AL second basemen in DWL% and one fielding category, but finished substantially behind Alomar in DW. Gold Glove to Alomar.

In 1972, Clarke led AL second basemen in DW, finished second in DWL% and FR, and led three fielding categories. Johnson led AL second basemen in DWL% and FR, finished second in DW, and led one fielding category. This is the closest Gold Glove race in the period. Clarke's 0.42 differential in DW is substantial, although it is due to Clarke playing 143 games to Johnson's 116 games. Johnson played better when he was in there, but Clarke played an extra month. I prefer the guy who played. Gold Glove to Clarke.

In 1973, Bobby Grich led AL second basemen in everything. It was easily the best defensive season in the sample; Gold Glove to Grich.

During the seven year sample, the Gold Glove was awarded to the best defender twice and the second best defender four times. The one exception was Doug Griffin in 1972. Griffin ranked fourth in DW and third in DWL%, but was still one of the better defenders in the league. The curious aspect of the Gold Glove award was that, although he led AL second basemen in defense three times, Horace Clarke never won a Gold Glove. Conversely, Davey Johnson won three Gold Gloves although he never led AL second basemen in defense. Sandy Alomar also led Al second basemen in defense twice -- two of the five best seasons in the sample -- without winning a Gold Glove.

3. *The best AL second baseman throughout the sample period*. To answer this question, I calculated mean fielding results for all players with three or more full-time seasons.[5] Grich is the only second baseman excluded by this criterion who could be considered the best defender in the league. Grich had the best single season in the sample, but I would argue his career falls into a later time period: Grich was the best second baseman from 1973 until 1976, when Frank White emerged as the premier AL second baseman.

Of the 13 players remaining in the sample, Clarke ranked first in mean DW per season (Table 4). Griffin led AL second basemen in DWL% by a wide margin. Knoop ranked first in FR/yr. Clarke led the league in fielding categories 16 times.

Clarke, Griffin and Johnson are clearly the best defenders in the sample. Clarke finished first in DW/yr, second in DWL% and first in the number of fielding categories. Johnson was excellent but ranked behind Clarke in all measures. As evidenced by his DWL%, Doug Griffin was better than Clarke when he played, but Clarke played more games -- an average of 148 games/season compared to Griffin's 122 games/season. As always, I prefer the guy who played every day: Horace Clarke is the best second basemen in the sample.

Table 3 continued

1970	DW	DWL%	FR	Leader
Alomar	5.28	0.660	9	*dp, tc/g*
Johnson [GG]	4.70	0.630	1	*po*
Clarke	4.34	0.560	-7	*po, a*
Knoop	3.55	0.570	35	
Kubiak	3.43	0.440	-17	
Leon	3.38	0.450	8	
Rojas	2.84	0.500	1	
Green	2.63	0.400	13	
Cullen	2.45	0.420	27	*fa*
McAuliffe	2.38	0.330	-6	
Andrews	2.34	0.320	-30	*e*
DThompson	1.41	0.230	-12	

1971	DW	DWL%	FR	Leader
Alomar	5.44	0.680	21	*tc/g*
Johnson [GG]	4.84	0.700	-1	*dp*
Green	4.69	0.660	10	
Clarke	4.47	0.580	-3	*po, a*
Griffin	4.35	0.710	7	
McAuliffe	3.48	0.550	7	
Andrews	3.34	0.670	-3	*e*
Rojas	3.27	0.580	-12	*fa*
Cullen	3.04	0.420	20	
Theobald	2.85	0.510	-13	
Carew	2.56	0.360	-25	
Leon	1.88	0.290	0	

1972	DW	DWL%	FR	Leader
Johnson	4.24	0.740	10	*fa*
Carew	4.05	0.590	0	
Griffin [GG]	3.95	0.620	2	
Alomar	3.90	0.510	3	
Brohamer	3.55	0.540	6	
Rojas	3.52	0.520	-2	
Andrews	3.41	0.460	-14	*po, e*
Theobald	2.79	0.500	1	
McAuliffe	2.31	0.390	-7	
Randle	1.19	0.350	1	
Cullen	0.59	0.170	-1	

1973	DW	DWL%	FR	Leader
Grich [GG]		0.820	21	*po, a, dp, tc/g, fa*
Griffin	5.76	0.720	-14	
PGarcia	5.29	0.670	-12	*e*
Clarke	4.65	0.640	8	
Carew	3.99	0.550	-7	
Rojas	3.86	0.570	9	
Green	3.40	0.510	8	
DNelson	3.11	0.450	-12	
Alomar	2.75	0.410	-8	
McAuliffe	2.23	0.430	5	
Brohamer	2.01	0.420	17	
Orta	1.82	0.300	-27	

[5] In calculating period means, I used only full-time, American league seasons. Many players also played part-time for one or more seasons during the period; none of these seasons, however, increased their mean DW or DWL% scores. One pertinent exception was Johnson's 1973 season in Atlanta. Johnson still had good range, but he led NL second basemen with 30 errors. Johnson's 1973 DW of 4.06 would have increased his mean DW to 4.01, but lowered his DWL% to 0.560.

Table 4: Mean fielding records during 1967 - 1973

Player	Seasons	DW	DWL%	FR	LL
Clarke	7	4.48	.612	6.43	16
Griffin	3	4.11	.681	-1.67	0
Johnson	6	4.00	.566	1.83	3
Knoop	4	3.78	.539	19.15	4
Alomar	6	3.75	.503	0.33	1
Green	4	3.67	.547	7.25	1
Rojas	4	3.37	.543	-1.00	1
Andrews	6	2.97	.446	-7.83	-2
McAuliffe	7	2.63	.427	-5.00	-1
Carew	7	2.55	.426	-9.83	-1
Allen	4	2.16	.447	9.00	1
Cullen	3	2.03	.375	15.67	1
Donaldson	3	1.66	.337	-7.33	0

where DW = mean defensive wins per season [James 1982],
DWL% = defensive won/lost percent during the period [James 1982],
FR = mean fielding runs per season [FR from Thorn and Palmer 1993],
LL = total number of league-leading defensive scores [PO+A+DP+TC/G+FA - E].

Table 5: All-time league leaders among ML second basemen

Player	PO	A	E	DP	TC/G	FA	+	-	sum
Mazeroski	5	9	1	8	10	3	35.0	1.0	34.0
Fox	10	6	1	5	7	6	34.0	1.0	33.0
E. Collins	6	4	1	5	2	9	26.0	1.0	25.0
Lajoire	3	3	0	6	6	7	25.0	0.0	25.0
Biggio	5	6	0	1	0	2	24.7	0.0	24.7
Grich	4	3	1	3	4	2	24.5	1.7	22.7
Clarke	4	6	0	2	3	1	22.5	0.0	22.5
H. Reynolds	3	5	4	5	3	0	28.0	7.0	21.0
Doerr	4	3	0	5	4	4	20.0	0.0	20.0
Sandberg	1	6	0	1	1	5	19.0	0.0	19.0
Cutshaw	5	4	0	2	4	3	18.0	0.0	18.0
Gehringer	4	4	2	4	2	6	20.0	2.0	18.0

Totals for seasons after 1960 have been weighted by the number of teams in the league.

PO = putouts, A = assists, E = errors, DP = double plays, TC/G = total chances per game, FA = fielding average,”+” = PO + A + E +DP + TC/G + FA, weighted for number of teams, “-“ = E weighted for number of teams, “sum” = difference between the positive [+] and negative [-] attributes.

Conclusions

A signal characteristic of Horace Clarke’s defensive career was the number of league leading totals. Just for fun, I counted the number of league-leading totals for all regular second baseman between 1901 and 1999. For each player, I summed the positive attributes -- putouts, assists, double players, total chances per game, and fielding percentage -- and subtracted the number of times the players led the league in errors. In this respect, Clarke is among the top ten second basemen of all time (Table 5).[6]

Clarke's tenure with the Yankees lasted only a few games beyond Ralph Houk's. In 1974, Clarke's contract was sold to San Diego -- at the time, the baseball equivalent of Siberia. He played poorly and was released at the end of the season. It was the end of a career that has been defensively underrated.

Literature Cited

James, B. 1982. The Bill James Baseball Abstract 1982. Ballentine, New York.
Macmillan. 1993. The Baseball Encylcopedia. Macmillan, New York.
Thorn, J., and P. Palmer, *eds*. 1993. Total Baseball. HarperCollins, New York.

Duke Rankin teaches ecology at a small, liberal arts college in Alabama. He welcomes your comments. Duke Rankin, 136 Indigo Lane, Calera, AL, 35040-4646; rankind@montevallo.edu. ♦

[6] The number of teams in MLB has increased several times since 1901. To reflect the increased difficulty in leading larger leagues, I multiplied each league-leading performance by a size factor, arbitrarily set at 1.0 for an eight team league [for performances between 1901 and 1960], 1.25 for ten teams, 1.50 for twelve teams, 1.75 for fourteen teams, and 2.00 for sixteen teams.

Does a Pitcher's "Stuff" Vary From Game to Game?

Phil Birnbaum

Does a pitcher's "stuff" differ from game to game, in the sense that it could be predicted in advance whether he will pitch well or not? In this study, the author presents play-by-play evidence as a start to answering this question.

Conventional wisdom says that a pitcher's "stuff" varies from game to game. That is, an average pitcher might have his curveball breaking especially well one start, and have it not working at all five days later. By watching the pitcher warm up (assuming those warm-up pitches are the hurler's best effort), observers, and the pitcher himself, would be able to predict a good outing before the first game, and would be able to predict, before the second game, that the pitcher would struggle that day.

Do pitchers really have variable "stuff", and, if so, how large is the effect? One way to find out would be to check how much a pitcher's game records vary, against how much they would vary if there were no day-to-day variation in stuff. If, for instance, you ran random games of average pitching, you might find that quality starts happen X% of the time, while the pitcher would give up 7+ runs Y% of the time. If real life average pitchers have both great outings and horrible outings *more* than the simulation says they should, this would be evidence of variability of stuff.

In this study, I'm going to approach the question a different way. If the stuff effect does exist, you would expect that a pitcher's performance earlier in the game should allow you to predict how well he would do later in the game. In a simulation, innings are independent, so if a simulated pitcher gives up 3 runs in the first inning, you would still expect him to be average in the rest of the game. But if the pitcher has daily stuff, you would expect the 3-run first to be evidence that the pitcher doesn't have it that day, and you'd expect him to continue his mediocrity as the game goes on.

Using data from Retrosheet, I analyzed every major league game from 1979 to 1990. First, I'll give you the batting records against every starting pitcher in that 12-year span:

	AB	H	2B	3B	HR	BB	K	avg	RC27
All Starters	539	141	25	4	13	48	81	.262	4.30

I've normalized all batting lines to 600 plate appearances, so they'll be easier to compare later. RC27 is Runs Created per Game, using the basic formula and 25.5 batting outs per game.

Here, now, is the batting line for all starting pitchers after a first inning in which they gave up at least three runs:

	AB	H	2B	3B	HR	BB	K	avg	RC27
All Starters	539	141	25	4	13	48	81	.262	4.30
3+ runs in 1st	535	142	25	4	14	50	75	.265	4.51

First, it looks like the effect is small. Intuitively, I would have thought that a pitcher with an ERA of 27 or more in the first inning would be a disaster in the remainder of the game. But he's only .21 runs worse. If we assume this pitcher will, on average, stay in the game four more innings, it costs the team an eighth of a run to leave him in (instead of replacing him with an average pitcher).

But even that .21 runs is overstated. These pitchers give up runs in the rest of the game at the rate of 4.51, versus 4.30 for the average pitcher. But the pitchers in the sample are not average. Starters who give up three runs in the first inning are usually worse than average pitchers – the better than average pitchers do it less often, and the worse than average pitchers do it more often, which weights the sample toward the worse pitchers. To account for this effect, I calculated the RC27 of the average pitcher in the 3+ runs sample, using their season statistics and weighted by the number of plate appearances they contributed to the 3+ line. I'll rerun the above table with that added:

	AB	H	2B	3B	HR	BB	K	avg	RC27	pop
All Starters	539	141	25	4	13	48	81	.262	4.30	4.30
3+ runs in 1st	535	142	25	4	14	50	75	.265	4.51	4.54

The "pop" column is the season RC27 of the population of pitchers who made up that line. Those pitchers who gave up the 3-run inning were, as we expected, worse than average pitchers, by 24 points of RC27. But what is now remarkable is that after giving up three runs in the first, they were *better* pitchers than their season average! Although the difference is very small, this is still the opposite of what we would have expected if there were a stuff effect.

A caveat: even if there were no stuff effect at all, we would still not expect the two columns to be identical, for several reasons:

1. The three-run inning figures into the season average, but not into the 3+ batting line (since the batting line measures only what happened *after* the 3-run inning). And, therefore, we would expect the 3+ line to not match the season line, but the season line with the three run inning removed. Removing a three-run inning from a starting pitcher's line lowers his ERA by 10 to 20 points. This is quite a large difference, and would tend to hide any stuff effect.

2. Just as 3-run innings tend to correspond to pitchers who are worse than average, they also correspond to (a) opposing teams who are better than average, (b) games played in hitter's parks, (c) umpires with a small strike zone, and (d) any other effects (weather, wind, etc.) that would favor the hitter. This would tend to exaggerate any stuff effect.

3. It is possible that pitchers do, indeed, have different stuff day-to-day, but that managers can tell, and are quick to remove a stuffless pitcher. That is, the "3+" column reflects only 3+ pitchers *that the manager chose to leave in the game*. If managers are accurate in knowing when to remove a pitcher for stuff reasons, we would expect to see a much lower stuff effect.

These factors mean that we cannot conclude, from the "3+" line, that there is no effect. However, these factors are likely small, and should not hide a large stuff effect if one exists.

I ran the same study for first innings with 4+ runs, and first innings with 5+ runs. I'll run all these results here, and I'll add one more column for the actual number of at-bats in the sample:

	AB	H	2B	3B	HR	BB	K	avg	RC27	pop	TAB
All Starters	539	141	25	4	13	48	81	.262	4.30	4.30	1168982
3+ runs in 1st	535	142	25	4	14	50	75	.265	4.51	4.54	34819
4+ runs in 1st	537	143	26	4	14	50	74	.266	4.56	4.57	10657
5+ runs in 1st	539	153	29	4	19	52	74	.284	5.58	4.67	1965

The results for the 4+ group are similar to the 3+ results: no obvious stuff effect, with the pitchers performing about as well as expected for their skill level. But, at the 5+ plateau, there does seem to be an effect. After allowing five or more runs in the first inning, starters pitched almost a full run worse than their talent would suggest, giving up 5.58 runs per 9 innings instead of their usual 4.67.

But note the small sample size – the difference came from only 1,965 at-bats, or a bit more than three player-seasons. I did a quick simulation, and found that the standard deviation of RC27 is approximately

$$\frac{17}{\sqrt{AB}}$$

For 1,965 at-bats, the SD works out to about .38, which means our difference is significant at the 95% level. But since this result may be the only significant one out of a series of tests, we should rerun the analysis on other league-years before formally concluding significance. (That is, since we have done three tests already, the chance of getting at least one at the 95% significance level is actually about 85%.)

Walks

Perhaps the non-effect for first-inning runs is partly due to the effects of luck. Even a good pitcher can allow infield hits, or bloop singles, or seeing-eye doubles. Plus, because we didn't control for unearned runs, the runs might have scored on defensive miscues.

But walks are almost completely under the control of the pitcher, and so perhaps we'll see an effect based on walks.

Here are the results for starting pitchers who allowed 2 or more walks in the first inning, as well as batting lines for 3+ walks and 4+ walks:

	AB	H	2B	3B	HR	BB	K	avg	RC27	pop	TAB
2+ walks in 1st	533	139	24	4	13	54	81	.261	4.37	4.41	54073
3+ walks in 1st	529	141	27	4	13	57	78	.267	4.66	4.43	6695
4+ walks in 1st	538	154	33	4	12	51	86	.286	5.24	4.51	439

As expected, all these starters gave up more walks than average after their wild first inning, because pitchers with wild innings tend to have low control to begin with. But the bad 1st doesn't seem to tell us much about what to expect in the remainder of the outing. After 2 walks, we get almost exactly a typical performance. After 3+ walks, we get a bit worse an outing than we expect, but by a very small amount. After a four-walk first, the difference is a bit more significant. We can expect a worse-than-average next few innings from the starter, but not by that much – less than ¾ of a run per game. It's certainly not worth necessarily replacing the pitcher, and overworking the bullpen, just to save half a run or so in a game that you're probably losing already after all those walks.

And the result is probably just random anyway; it's only one standard deviation from expected, which is far from statistical significance. Compare the 4+ batting line to the 3+ batting line. If a player had those results respectively in consecutive years, we'd call him very consistent.

Late Innings

When a starter gets to the late innings of a game, and then starts allowing baserunners, everyone seems to agree that he's getting tired and should be replaced, before his exhausted pitching arm does even more damage. And, almost always, that pitcher is indeed replaced. But what about those times that he's left in the game? Does the expected damage materialize?

Here are batting lines for starters left in the game after allowing three baserunners (hits or walks) in the 7th, 8th, or 9th inning respectively.

	AB	H	2B	3B	HR	BB	K	avg	RC27	pop	TAB
3 runners in 7th	533	140	24	4	14	48	74	.263	4.35	4.07	10802
3 runners in 8th	531	143	22	4	14	48	70	.269	4.50	4.00	3505
3 runners in 9th	532	147	17	3	11	48	72	.276	4.37	3.89	793

We may be on to something here: whether it be in the 7th, 8th, or 9th, when a pitcher is left in after allowing three baserunners, he continues to do worse than his season average if he's left in the game. But, still, none of the differences are statistically significant.

Even in terms of baseball significance, the differences aren't much. Over the season, these pitchers are a bit better than average, which is probably part of the reason their managers let them stay in the game in these situations. Over the remainder of the game, they turn into roughly average pitchers. If the manager isn't worried about their arms, and the game isn't close, there's no reason to turn to the bullpen when you effectively still have an average pitcher out there.

Having said that, this is one result that I would expect to become statistically significant if we had more years to analyze. I suspect that pitchers do get tired in ways managers don't always recognize. But I could be wrong.

No-Hitters

If a pitcher has a no-hitter going, we'd expect that he has good stuff today, and we'd definitely expect that he will continue to pitch well in the coming innings. But it turns out that isn't the case – that he performs at almost exactly his season average after 3, 4, or 5 no-hit innings. Here are the numbers:

	AB	H	2B	3B	HR	BB	K	avg	RC27	pop	TAB
No hits through 3	542	136	24	3	14	47	86	.251	3.98	3.96	21485
No hits through 4	541	132	21	3	13	48	91	.244	3.70	3.89	7355
No hits through 5	543	134	25	3	15	46	90	.247	3.93	3.86	2748

These are good pitchers – note their above-average strikeout rates and good RC27s – and they continue to be good after a few no-hit innings. Not great – just as good as you would have expected for a typical game.

Strikeouts

Bill James has pointed out that strikeout rate is one of the largest factors in determining how a young pitcher will do in the future. Can it be used to figure how the current pitcher will do in the rest of the game?

Here are the batting lines for (a) pitchers who strike out 6 or more in the first three innings; (b) pitchers who strike out 8 or more in the first four innings; and (c) pitchers who strike out 10 or more in the first six innings.

	AB	H	2B	3B	HR	BB	K	avg	RC27	pop	TAB
6 Ks through 3	540	126	20	3	14	51	120	.233	3.53	3.71	6314
8 Ks through 4	534	124	18	4	12	54	136	.232	3.43	3.63	1798
10 Ks through 6	539	124	19	3	12	53	131	.230	3.35	3.47	1420

These are great pitchers – with lots of strikeouts and ERAs in the low 3s. And they are, indeed, even better than that for the next few innings, but the difference is not even close to statistical significance (the top line is about 1 SD better than expected).

And, here, the manager's decisions do not factor in to the situation. For pitchers with a bad first inning, the manager might notice bad stuff and remove the pitcher before he has a chance to show up in our charts. But for these pitchers, the manager's decision-making is out of the picture – the only decision for the manager is to let the guy pitch, which, after 10 strikeouts through six innings, is not a tough call to make.

Conclusions

Going into this study, I expected to find reasonably hefty stuff effects. I thought that average pitchers who give up lots of walks in the first inning are showing they have no control that day, and the results would continue to be disastrous. I was expecting the RC27s to show a big jump, perhaps from 4.5 to 6.5.

But, as we have seen, they did not. There were no big effects either way, and, in general, we can conclude that there's no evidence that the results a pitcher has obtained so far in the game should affect our estimate of what we should expect later in the game.

Go figure.

Phil Birnbaum, 18 Deerfield Dr. #608, Nepean, ON, K2G 4L1; birnbaum@sympatico.ca ♦

Get Your Own Copy

If you're not a member of the Statistical Analysis Committee, you're probably reading a friend's copy of this issue of BTN, or perhaps you paid for a copy through the SABR office.

If that's the case, you might want to consider joining the Committee, which will get you an automatic subscription to BTN. There are no extra charges (besides the regular SABR membership fee) or obligations – just an interest in the statistical analysis of baseball.

To join, or for more information, send an e-mail (preferably with your snail mail address for our records) to Neal Traven, at beisbol@mediaone.net. Or write to him at 500 Market St. #11L, Portsmouth, NH, 03801.

Clutch Hitting And Experience

Cyril Morong

Experience may have a positive influence on performance in certain clutch situations, holding hitting ability in general constant. Here, the author investigates the connection.

Introduction

Does clutch hitting exist? That is, do some hitters hit better in the clutch than others or better in the clutch than they themselves normally hit? If a hitter bats .300 from the seventh inning on we might think he's a great clutch hitter but he might be a .300 hitter all of the time, so his clutch performance is not really so extraordinary. If a hitter bats .400 from the seventh inning on we might think he's a great clutch hitter but wonder why he does not hit .400 all the time if he can do so in the clutch. What would allow him to hit so much better in the clutch? Why would he do better when the pressure is on? Do some hitters do worse when the pressure is on, and if so, why? If a hitter is a good clutch hitter does it mean he simply maintains his normal level of performance in the clutch and does not wilt under pressure?

David Grabiner has done an excellent study on clutch and finds no evidence that clutch hitting exists. The correlations he found of clutch hitting in one year to previous years were quite low (.01).[1] This study takes a different angle.

In general hitters do worse in the clutch (defined below) than they do in non-clutch situations. (More on this in the Data section below.) This study raises the question of whether or not more experienced hitters hit better in the clutch than less experienced hitters.

The Data

Data from the 1995 major league season was used. The clutch situation used was "Close & Late" from the *STATS^{TM} Player Profiles 1996* from STATS, Inc. This occurs when "a) the game is in the seventh inning or later and b) the batting team is either leading by one run, tied, or has the potential tying run on base, at bat or on deck." Only players with 60 or more plate appearances (walks + at-bats) were included. That was the book's qualification for being among the league leaders in this statistic.

How do hitters perform in the clutch in general? In major league baseball in 1995, for example, the slugging percentage (SLG) in clutch situations was .393 while in non-clutch situations it was .421. For on-base percentage (OBP) these numbers were .3346 and .3353, respectively.[2] So hitters generally did worse in the clutch than non-clutch. This was true in 1994 as well. What might explain this is that relief pitchers are often brought in late in the game and a hitter has not faced them yet. Also, a left-handed pitcher might be brought in to face a left-handed hitter and a right-handed pitcher might be brought in to face a right-handed hitter. It is not clear that the pressure of the situation is what makes these numbers lower, because the pitchers are under pressure, too.

With it being more difficult to hit in the clutch, any hitter who simply maintain his normal performance level might be considered to be a good clutch hitter.

Hitters were divided into two groups: the inexperienced, those with less than 2000 previous major league PA's (prior to 1995) and the experienced, those with 2000 or more previous major league PA's. If a player plays every game and gets 4 PA's per game, he would have about 2000 after 3 years.

Performance in both clutch and non-clutch situations was measured using Production (PROD), which is the sum of SLG and OBP.[3]

The Model

An ordinary least squares regression was run using a hitter's PROD in the clutch (CPROD) as the dependent variable and his PROD in non-clutch situations (NCPROD) as the independent variable. Here are the results:

[1] His paper can be found at http://www.baseball1.com/bb-data/grabiner/fullclutch.html This is the website for *The Baseball Archive*.

[2] Only walks, hits and at bats were used to calculate OBP.

[3] PROD, at the team level, is more highly correlated with team runs than batting average or SLG or OBP. So it is a better measure of performance.

The Results

```
N = 175
Adjusted R2 = .187
F-Ratio = 41.11

Variable     Coefficient  T value      P value
Constant          0.248    2.990        0.003
NCPROD            0.657    6.412        <.001
```

The Adjusted R^2 = .187 means that 18.7% of the variation in CPROD is explained by variation in NCPROD. It seems that very little of how well a hitter does in the clutch is explained by is general ability or overall performance . What, then, would explain the other 81.3% of the variation in clutch performance? Some real difference in how players react to pressure? Perhaps. But the number of PA's for each player is low, the minimum being 60. So it does not necessarily mean that clutch performance is all that much different from non-clutch performance.[4]

The second regression was the same as the first accept that a dummy variable was added, with 1 for experienced players and 0 inexperienced. Here are the results:

```
N = 175
Adjusted R2 = .208
F-Ratio = 23.869

Variable     Coefficient  T value      P value
Constant       0.242       2.944        0.004
NCPROD         0.631       6.208        <.001
DUMMY          0.052       2.355        0.020
```

The Adjusted R^2 is a little higher with the addition of the experience variable, but not much. So we can do a better, but only slightly better, job of predicting how well someone will do in the clutch when we include experience. There is little change in the results for NCPROD. The experience variable, however, is statistically significant with a P value of 0.020. Also, the coefficient estimate, in a baseball sense, is significant. At 0.052, it means that a player's CPROD is that much higher if he is experienced. So, instead of maybe a hitter having a CPROD of, say, 0.800, he would have a CPROD of 0.852.[5] That could mean, for example, 26 points more in both SLG and OBP in clutch situations for the experienced hitter.

Experience may teach a player to relax in the clutch and not press. Perhaps he learns how to concentrate better or what pitches to expect. But, whatever the reason, an experienced player, holding ability constant (by using NCPROD to measure this), will do much better in the clutch than an inexperienced player.

Similar results were found using only SLG as a measure of performance and ability instead of PROD.

Conclusions

A variable, experience, was found to be positively and significantly related to clutch hitting. But the results are certainly not definitive or conclusive on the question of whether or not clutch hitting exists or if some hitters have a special ability to do better in the clutch. This study only looked at one season. Player career totals could be studied to raise the number of clutch PA's needed to qualify for the study. This may reduce random noise that could be in the model. Other clutch situations like "men on base" could be studied. Another question is how much difference individual clutch hitting makes in winning games and in the standings. These results indicate that clutch might exist and that it is worth further study.

Cyril Morong is an economics professor in San Antonio. Cyril Morong, 723 W. French Place, San Antonio, TX 78212; CyrilMorong@aol.com ♦

[4] The minimum level of clutch PA's could be raised, but then there are fewer players in the study.

[5] Similar results were obtained when the dummy cutoff was at 1500, 2500 and 3000. At 1000, the dummy was not significant. This means a certain amount of experience is necessary before it helps a player do well in the clutch.

The TPR Chronicle

Rob Wood

The last couple of years have seen vigorous debate on SABR-L on the merits of TPR. Here, the author summarizes parts of that discussion by quoting the posters themselves.

Pete Rose's value to his teams throughout his career is a popular and recurring theme on SABR-L (SABR's e-mail discussion list). In April 1998, this issue was raised in a post to which Bill Deane replied. Deane's post started a long and winding thread on how TPR, Pete Palmer's "Total Player Rating" stat, measures players versus the league average performance. The issues covered by the thread were interesting and far-reaching.

This article serves as a summary of that thread.

All the posts that follow are from 1998 (mostly April-May), with the exception of the final two which are from August 2000. For the sake of brevity, not every post on the subject is re-printed here, and all posts below have been snipped. To facilitate retrieving posts, the date and SABR-L post number is given at the end of each post.

With Bill Deane's last paragraph in a reply to a post comparing Pete Rose to Bill Mazeroski, we were off:

> Although I am a great admirer of Pete Palmer and his Linear Weights system, I disagree with its comparison to "average" rather than "replacement-level" players. The difference in season-to-season comparisons is negligible, but the difference between, say, a 3,562-game career (Rose's) and a 2,163-game one (Mazeroski's), is considerable, and I would venture to say that -- if a corresponding adjustment were made in the Linear Weights system -- Rose would vault ahead of Maz and about 150 others on the all-time list. [April 16, 1998; 8433]

Merritt Clifton replied that Mazeroski was much better relative to his available replacements than was Rose relative to his (what follows is a compilation of two posts):

> Bill Deane seems to think the Pete Rose / Bill Mazeroski comparison should be made on "replacement value," which isn't nearly as quantifiable as linear weights. But fine, let's do it. Replace Bill Mazeroski with Curt Roberts, Gene Baker, and Julian Javier, who would have covered the duration of his career.
>
> Replace Pete Rose with Don Blasingame (1963-1965), Tommy Helms (1966-1967), allowing Tony Perez to play 3b in 1966 and allowing Lee May to play a full season in 1967 instead of platooning for half the year), Mack Jones (1968-1969), and Hal McRae for the rest of his first stint with the Reds; replace Rose with Richie Hebner for his time with the Phillies; replace Rose with Al Oliver as an Expo; replace Rose with more playing time for Perez in his second stint with the Reds. In each case, these are players who were moved -- traded, sold, or allowed to leave via free agency -- essentially to use Rose instead.
>
> The actual TB numbers for the players I postulate as the most likely available replacements for Bill Mazeroski, season by season, multiplied in cases of part-time play up to regular performance: -21.8, over Mazeroski's career. The actual TB numbers for the players I postulate as the most likely available replacements for Pete Rose, calculated the same way: -5.3.
>
> This makes Mazeroski the more valuable player by 16.5 games over their respective careers. That's just about exactly what the Total Baseball rankings indicated even without this exercise, giving Mazeroski a career score of 36.3 and Rose a career score of 20.0. [April 18, 1998; 8461 & April 19, 1998; 8463]

Tom Ruane replied, in a similar vein as had Bill Deane, to a concurrent post comparing the career values of Roy White and Pete Rose (White's TPR is slightly greater than Rose's). After investigating the components of TPR for these two players, Tom wrote:

> Since TPR gives us an estimate of the number of wins ABOVE AVERAGE that a player, in all aspects of his game, contributes to his team, it will always emphasize quality over quantity. And as no player in history had as much quantity as Rose, he suffers more than anyone under such a measurement. Give him some credit for playing regularly for nearly a quarter of a century and he'd leave many players in the dust. A few years ago, I came up with a measurement called PV (player value) which attempted to estimate the number of wins above replacement-value that

> a player contributed over the course of his career. For hitters, it simply took TPR and added a win for every 400 plate appearances. There's an explanation of how I came up with this as well as the formula for pitchers in my 1994 BRJ article. Using this formula, Rose had a career PV of 59.0 with White far behind at 40.0. Now, I was using PV to evaluate trades, expansion drafts, and the productivity of farm systems, and while I'm convinced it is well suited to those areas, I'm not as sure it's what we want to use to evaluate Hall of Fame candidates. [May 2, 1998; 8747]

Bill Deane then gave details of how he would suggest adjusting TPR to measure relative to a replacement level player:

> One valid criticism of Total Player Rating is that it compares players with average players rather than replacement-level players. A team of players with zero TPR's would theoretically finish the season 81-81. But, if the team plane went down and the club had to replenish the roster with Triple-A players, would that team also finish 81-81? Of course not, but TPR assumes that it would. TPR implies that an average player has no value, regardless of how long he plays.
>
> So, how do we adjust TPR to compare with replacement-level players? First we have to assign a value to that term: how would a team of replacement-level players fare at the major league level? Some analysts I've discussed this with feel the value is in the .333 range -- meaning that a team of replacement-level players would go about 54-108. That sounds about right to me; the worst teams of the past 35 years have been right around that level.
>
> If it is right, that means that the difference between a team of average players, and a team of replacement-level players, is about 27 wins per 162-game season. Of those 27 wins, let us assume that 13½ are attributable to the team's offense, and 13½ are attributable to its defense, including pitching. Let's further assume that pitching accounts for two-thirds (or nine wins here) of a team's defensive performance. That leaves 18 wins which we can attribute to the non-pitchers, or a little more than two wins per position. Still with me?
>
> So, I'm suggesting that a man who plays 162 games at a field position, with a TPR of zero, has actually contributed at least two wins to his team, in comparison with a replacement-level player. For a few reasons, I believe the actual number is upwards of 2.5 wins, but I would rather err on the sides of conservatism and simplicity in making my point here.
>
> So, how do we "correct" players TPR's under this set of assumptions? Simply credit each non-pitcher with two extra wins per 162 games played, or .0123 wins per game. Call it a longevity bonus. Here's how my revised list of all-time TPR leaders (among non-pitchers, through 1996) looks:

138.6	Babe Ruth
130.5	Hank Aaron
129.1	Willie Mays
128.5	Ty Cobb
124.8	Nap Lajoie
120.9	Tris Speaker
115.7	Honus Wagner
114.0	Ted Williams
109.0	Rogers Hornsby
108.1	Mike Schmidt
107.9	Stan Musial
106.4	Rickey Henderson
105.7	Mickey Mantle
104.7	Eddie Collins
103.7	Frank Robinson

> Pete Rose, incidentally, moves up from 178th among non-pitchers by the "pure" TPR rating, to 53rd using this one. If the longevity bonus value were increased from two wins to 2.5 per 162 games, Rose would rank 41st all-time. Sad to say, neither Roy White, nor most of the others Merritt Clifton would have traded Rose for, makes the cut under my system. Again, I make no claims as to the scientific accuracy of my system, but I do submit that it is a step toward improving TPR. [May 4, 1998; 8799]

Tom Ruane then replied that Bill Deane's system to adjust TPR is similar to the one that Tom developed:

> It turns out that Bill Deane's method of adjusting TPR from wins-above-average to wins-above-replacement is very similar to mine. He figured that a replacement team would be 27 games under .500; I guessed that they would be 26

> games under. He split his games 18-9 between pitchers and hitters; I went with the old adage that pitching is 38.46 percent of the game and split them 16-10. He used games played to distribute these wins; I used plate appearances. He didn't give us the breakdown on the pitching wins, but I assigned a win for every 150 innings and another for every 150 relief games. Of course, the fact that our two, rather unscientific, systems are so close doesn't make them any more accurate. Perhaps it just means that it's not only the great minds that think alike.
>
> One more thing: when evaluating a player's career, I converted any negative wins-above-replacement to zero. I didn't want to treat a terrible player who was forced against all reason to play regularly as worse than an inferior player cut during spring training. Blame the manager or the organization, I thought, not the player. I might make an exception to this rule for Rose, however, since for some of the seasons where he dipped below a replacement-level player, he was the manager penciling his name into the lineup card. [May 5, 1998; 8817]

Ed Coen also commented on the issue of negative TPR values:

> In very simple terms, here's why I have a problem with normalizing to the average player instead of "replacement level." Let's say player A and player B have had similar careers up to age 30 and both have a lifetime TPR of, say, +10. At this point player A loses a few steps, reducing his range in the field and cutting down on his doubles and triples. He becomes a "below average" player. Good enough to make the starting lineup for his team, but there are better players around. After age 30, he scores a -5 TPR. Player B, on the other hand, really hits the skids. He only plays another half season before getting released and leaves baseball altogether. During the half-season, he scores a -1 TPR. In this example, player B ends up higher than player A, 9 to 5.
>
> My point is, playing in the majors even if you're below average, is better than not playing at all. Yes, a "replacement" player is harder to define than an average one. But just because something is difficult, it doesn't mean it shouldn't be done. An approximation of the right answer is better than an exact wrong answer. This is not about rewarding mediocrity. It's about removing the penalty for longevity. [May 7, 1998; 8868]

Clifford Blau presented a good case for measuring relative to replacement level:

> In the minors, there is a large pool of players who are more or less as good as the worst players in the majors (since there are a lot more people with that amount of ability) who are available for little or nothing. Those are the replacement level players. As long as a player is one of the 700 (or whatever the number is now) best baseball players, he has positive value. If Buckner had died during his career, his team would have had to replace him with one of those replacement level players, and would have won fewer games as a consequence. That is why it is proper to adjust TPR to reflect value above replacement level. That way, it shows the player's real value to his team. [May 7, 1998; 8870]

John Pastier opined that the replacement level is not easily defined and is therefore not a workable statistical tool:

> This looks like a step toward a definition of a replacement player, or replacement level. But the 750 people on ML rosters today are not exactly the 750 best baseball players around. Some better ones are still in the minors, stuck behind a big-club star at that position. Others are better but not yet recognized as better. Still others are too expensive, so they've been let go. Within the pool of roster players, some pretty good ones are in their manager's doghouse, and not getting their deserved playing time, while others are getting more than they deserve. Look at Heathcliff Slocumb this year, an alleged closer with a walk per inning and a 14+ ERA. Is he the replacement-level standard for relievers? The "market" does not function with total efficiency. To say that every ML player at a given moment has positive value seems excessively generous.
>
> And if Buckner had died the day before opening day in 1988, surely the Angels would have come up with someone who could slug better than .209. If he had retired at the end of 1986, surely the Bosox would have found a 1B who could slug better than .322.
>
> All this is to point out how tricky it is to identify a replacement level player. Every player, from superstar to superslob, was supposedly a replacement level player when he came up. Some RPs were clearly better than the players they replaced, while others were worse.
>
> What does that say about this concept as a statistical measuring rod? Some people on the list think that an RP is literally the actual guy who takes over a specific failing player's job, while others think it's someone who will allow his team to play .333 ball. Still others may have yet different ideas about the meaning of the term. I think that RP is a nice metaphor, but not a workable statistical tool. [May 8, 1998; 8896]

Clifford Blau then posted a rhetorical question to the critics of replacement value:

> Merritt Clifton and John Pastier, and perhaps others, have recently commented that players with negative TPR are costing their teams wins. Thus, a player with a longer career and lower TPR is always less valuable than one with a shorter career and higher TPR. If this is so, can someone explain to me why major league teams don't immediately release all of their below average players and replace them with average or better players? [May 10, 1998; 8938]

John Pastier replied to Clifford Blau with several good points of his own:

> Well, there are several parts to the answer. One, the part that Clifford might have in mind, is that it isn't always possible in the real world to find and insert those players immediately. But that's only one of many reasons. Others include: (1) Poor player evaluation methods might mask a non-producer in the short run. Not every team has an Eddie Epstein running daily spreadsheets on all the players in the majors. (2) The player may be a big fan favorite in the declining phase of his career, so benching or dumping him might be inadvisable or impossible until he declines so much that even the fans are reconciled to his departure. (3) It costs money to improve a team. Some teams don't have the bucks or don't want to spend them. Even teams with big bucks can't strengthen themselves at every position every year. (4) A team may be so good that it doesn't need to replace a regular who is costing it games. Or there may be an even worse regular whose replacement is a higher priority. (5) He's his own manager, and wants to keep playing. (6) Managerial politics other than #4. (7) Multi-year contractual obligations might mean having to play someone until a satisfactory trade can be arranged. (8) That phenom in AAA is not considered quite ready. (9) Injuries force the team to play a stop-gap.
>
> By definition, about half the active players at a given moment have to be below average. Few teams can assemble rosters made up predominantly or entirely of above average players. That fact does not give a below-average player positive value, however. It just makes him less detrimental than an even worse player.
>
> A player who is below average at a given position is not helping his team compete, although he may be helping them avoid being terrible. A 40th- or 35th-percentile player would probably bolster the '62 Mets, but so what? If a team has a below-average player who is helping them win games (I concede that possibility), that means the team is even worse than the below-average player. It doesn't mean that the player has "positive value". Given a team bad enough, a poor player could help them win games over the .050, or .100, or .250, or even .375 level. That's not a reasonable yardstick, however -- .500 is a better one, and has the added advantage of relatively easy computability. We all can grasp, and calculate, what an average player is. [May 11, 1998; 8959]

Mike Emeigh pointed out that the average vs. replacement level issue is related to the overall distribution of baseball talent and performance:

> I have a very different problem with normalizing to the average player from that which Ed Coen has - the distribution of talent in baseball isn't a "normal" distribution (in the statistical sense), but is truncated at the low end of the curve.
>
> If talent in baseball were truly normally distributed, there would be about as many players who are below average as there are who are above average. In fact, there are generally "more" below average players than there are above average players - and up to a certain point (AKA "replacement level"), there are generally more players who are below average by a specified amount (say, 5%) than there are players who are above average by that same amount. Beyond the "replacement level" there are "no" players who are below average by the amount that corresponding players can be above average - 40% below average players aren't good enough to play in the majors, whereas 40% above average players are Frank Thomas (the current incarnation, that is).
>
> Bill James once noted that talent in baseball represents the far right-hand end of the normal distribution curve - with the truncated part of the curve representing the baseball talent of the broader populace at large, the great majority of whom are incapable of playing at the major league level. I believe that's the appropriate model of the actual performance of baseball players, and that normalization techniques used in calculating TPR distort the relative worth of players by fitting non-normally-distributed data to a standard normal distribution. [May 11, 1998; 8954]

Peter Morris also cited Bill James on the issue:

> Bill James agreed with Bill Deane's assessment of negative TPRs. In the Historical Baseball Abstract, page 302, he writes: "Pete [Palmer]'s methods always look toward the center, with the result that he considers an 'average' -- a

> player at the center -- to have no value. That is quite incorrect, because an average player has very real value to a team; pennants are lost almost every year because of the inability of organizations to fill slots with average players. A typical pennant-winning team will have one truly outstanding player, three or four other good players, and the rest of the positions covered by players who are more or less ordinary or 'average'; most have below-average players on them at one spot or another. The problem with 'average' is that it's an imaginary line; it has no real significance for a team. The line against which a player's value should be measured is the replacement level."
>
> I concur, and would add that, while it doesn't much matter if you choose to represent a below-average season as a negative, it becomes very inaccurate to apply that to a player's lifetime value. [May 12, 1998; 8967]

Tom Ruane then looked at data to investigate who really are these players with negative TPR's (one hypothesis was that they were aging veterans that risk-averse teams play rather than unproven youngsters):

> What is the age of the players hurting their teams with negative TPRs? To look at this, I took all players with TPRs less than -1 and determined their ages. I broke these ages into 5 groups: 24 or younger, 25-28, 29-32, 33-36, and over 36.
>
> Here [is a percentage breakdown of] the ages of bad players from 1981 to 1993:
>
> | Year | <=24 | 25-28 | 29-32 | 33-36 | >=37 |
> |---|---|---|---|---|---|
> | 1981 | 14.7 | 39.2 | 31.4 | 13.7 | 1.0 |
> | 1982 | 14.5 | 44.8 | 30.3 | 8.3 | 2.1 |
> | 1983 | 17.4 | 37.6 | 28.2 | 12.1 | 4.7 |
> | 1984 | 21.6 | 33.3 | 23.5 | 17.3 | 4.3 |
> | 1985 | 19.3 | 37.3 | 26.7 | 12.0 | 4.7 |
> | 1986 | 21.5 | 37.5 | 24.3 | 9.0 | 7.6 |
> | 1987 | 20.3 | 37.0 | 29.0 | 6.5 | 7.2 |
> | 1988 | 18.5 | 42.0 | 19.7 | 14.6 | 5.1 |
> | 1989 | 20.5 | 46.2 | 23.7 | 5.8 | 3.8 |
> | 1990 | 14.6 | 41.7 | 30.5 | 11.3 | 2.0 |
> | 1991 | 15.2 | 40.2 | 28.0 | 15.9 | 0.6 |
> | 1992 | 19.9 | 38.1 | 26.7 | 13.6 | 1.7 |
> | 1993 | 17.8 | 38.3 | 29.4 | 10.0 | 4.4 |
>
> There are (on average) more bad players under 25 than over 32 and the majority of these players are under 29. In other words, these players are not tolerated because GMs love paying fat contracts; they're tolerated because a) they are cheap, and b) they might get better. Of course, some of these players never get better, but in those cases I suspect that the memo from the scouting department alerting the front office to this fact simply got misplaced. [May 12, 1998; 8974]

Larry Grasso advocated the replacement level adjustment to TPR:

> When you use average as the baseline, players start at zero and their performance will move them up or down. Marginal players typically won't be too far from zero because they have not had the opportunity to cost their team many games. By judicious selection of playing time (use only as a platoon player, or as a defensive replacement under certain conditions) a marginal player may end up with positive value that would disappear if the team were ever forced to play the player full time. A marginal player hitting a hot streak may end up with a positive rating because the streak represents so much of their record. The rating is due more to chance than to ability.
>
> Another problem is that we are talking about human beings playing baseball, not robots welding automobile chassis. There is a lot of variability inherent in performance. Teams stick with players who have had below average seasons because they hope they will return to a higher level of performance. Jim Abbott went 2-18 with a -4.2 TPI in 1996. What were they thinking! Why did they let him pitch? Well, they hoped he might sort himself out and pitch the way he did in 1995, when he had a 1.5 TPI, or better still, the way he'd pitched for them in 1991 and 1992. Even hall of famers or near hall of famers can have average or below average years, and not always just at the beginning and end of their careers. [May 15, 1998; 9036]

Merritt Clifton, in a series of posts, reviewed more than 40 actual replacement incidents and concluded that they were, on average, about average:

> Every year there are many surprises, offsetting the emergency replacements that don't work so well. The key points are that the typical "replacement" regular is not a bum but a rookie, or a minor league veteran who didn't previously get a good chance, or wasn't ready for it, and that about a third of these guys then hold a regular job, having proved to be much better than expected.
>
> Willie Mays was a replacement too. Sure, he was off to a hot start in Minneapolis, but he was a kid, and a lot of kids start hot in a new league. Ahead of him was an All Star, Bobby Thomson. But then Hank Thompson wasn't hitting and went on a binge too many. Reserve Artie Wilson was sent down. [Hank] Thompson was benched, [Bobby] Thomson was switched to third, and the rest is history.
>
> "Replacement value," in short, is a purported bottom line that just plain doesn't exist. Replacements are, on average, a grab-bag with the same cumulative TPR value (0.0) as everyone else who wins a major league job. [May 28, 1998; 9273]

Jesse Thorn used a "off the face of the planet" argument for replacement levels:

> If your first baseman disappears off the face of the planet, his role won't be filled by a league average first baseman. If, say, Henry Rodriguez of the Cubs broke his leg skydiving tomorrow, he wouldn't be replaced in left field by an average player -- he would be replaced by a fourth outfielder, Brant Brown or something. So Rodriguez' value is not in his relationship to the average, but his relationship to his *replacement*.
>
> Obviously, an adjusted TPR and a standard TPR are both valuable tools -- but they measure different things. With a replacement value, you are measuring an accumulated value to the team. Without, you are measuring a relationship to the league average -- we might call it a level of greatness. Neither is better than the other, and they are not mutually exclusive, they simply measure different things. [May 29, 1998; 9292]

Merritt Clifton replied to Jesse with a summary of his views:

> This is where "replacement value" advocates err: by failing to look at what actually happens. Regulars are rarely replaced by career reserves for more than a couple of games; almost always they're replaced by prospects, and as I keep pointing out, about a third of them make good, contributing positive TPR or at least TPR no worse than that of the players they replace, to raise the actual value of the replacements to average: 0.0.
>
> An example of as close a case as ever happened of a player dropping off the face of the planet occurred in 1967 when Jack Hamilton beaned Tony Conigliaro. Conigliaro was replaced on the roster with free agent Hawk Harrelson, the 1968 AL RBI champion. Harrelson was replaced after his release from the Athletics a week or so earlier by Reggie Jackson, called up from Mobile.
>
> Regulars are rarely replaced by career reserves for more than a couple of games; almost always they're replaced by prospects, and as I keep pointing out, about a third of them make good, contributing positive TPR or at least TPR no worse than that of the players they replace, to raise the actual value of the replacements to average: 0.0. [May 29, 1998; 9303]

Clifford Blau suggested that Merritt Clifton's 40 replacement incidents were not representative of the concept of a "replacement level":

> Honestly, I despair of ever getting Merritt to see the light on this issue. The main stumbling block seems to be that he apparently believes that replacement players are, on the whole, as good as the average major leaguer. One flaw in his thinking, IMHO, is that he looks mostly at voluntary replacement. Usually, when a major leaguer is replaced, it is because the team thinks the replacement is better. Therefore, often, a replacement player (not a replacement level player) helps his team improve. [May 30, 1998; 9317]

Tom Ruane presented the working definitions of "replacement level" used by several leading analysts:

> Many people have been looking recently into the value of replacement-level players. When I tried to determine this a few years ago, I figured that the first-year performance of expansion teams was one way to approximate this. The first ten such teams had a winning percentage of .365. Now, these teams were better than my hypothetical group of replacement-level players for two reasons: 1) they contained several players (Richie Ashburn, Dick Donovan, Maury Wills, Rusty Staub, ...) who would've played regularly if there had been no expansion, as well as many others who, while not regulars, would've at least been on major league rosters. And 2) because of the first reason, the teams these expansion squads played were weaker than they would otherwise have been.

> So I picked a .340 percentage; Bill Deane settled on .333. Pete Palmer favored .350. My feeling is that .365 is a reasonable upper-bound and anything somewhat less than that would be in the right ballpark. Remember: when a team needs to replace a single player, you are usually looking at the best (and often only good) replacement player available. This is not an accurate indication of what ALL the starters on a team are contributing by playing regularly. Had the Marlins only lost Moises Alou, they might have been able to find someone close to an average player to replace him. The wholesale exodus has caused them to approach (surprise, surprise) a .333 winning percentage. To use one of Merritt Clifton's anecdotes, when the Red Sox lost Tony Conigliaro in 1967, they were lucky enough to replace him with Ken Harrelson -- a player who had been playing regularly for both the Senators and A's. Had they lost Yastrzemski and Reggie Smith as well, Boston fans would probably have spent 1968 watching the likes of Jose Tartabull and Joe Lahoud (or George Thomas and Al Yates) in their place. [May 31, 1998; 9333]

Keith Woolner, the developer of "Value Over Replacement Player" (VORP), checked in with a lengthy post advocating and explaining his use of replacement levels (I recommend those interested read the entire post):

> Replacement level is the *expected* level of performance the average team can obtain if it needs to replace a starting player at minimal cost. Individual replacements can perform above or below the expected level, but that does not change what the expectation was at the time of the decision.
>
> Specific teams may have better-than replacement level players available in their own systems. This does not change the concept of replacement level -- it shows that team context is important when evaluating particular decisions. If *all* teams had better-than replacement-level players easily available, then that would indicate that your level is set too low.
>
> We define a replacement level player as one who hits as far below the league positional average as the league backups do relative to league average, who plays average defense for the position, and is a breakeven base-stealer and baserunner. VORP is the number of runs contributed beyond what a replacement level player would contribute if given the same percentage of team plate appearances.
>
> So what level of performance does a VORP=0 team represent? Well, using the 1998 season MLB statistics to date, a all-replacement-level team would hit about .235/.300/.356 and have a RA of 5.85. The Pythagorean projection over 162 games would be 44-118, for a .271 winning percentage. This is comparable to the performance of the worst teams in history (e.g. '62 Mets who went 40-120 for a .250 Win%). [May 31, 1998; 9344]

John Rickert extensively reviewed the case of Willie Mays being brought up by the Giants in 1951 who Merritt Clifton had used as an example of a "replacement" player. In the summary of his reply, John writes:

> Mays was called up because he was hitting .450+ [in Minneapolis], not because the Giants wanted to bench [Hank] Thompson. Mays was more of a "displacement player" than a "replacement player" (the terminology "replacement player" is somewhat unfortunate, because those "displacement players" do replace someone in the lineup, but I didn't create the terminology). I'm willing to believe that "displacement players" are about average or above average, but still suspect that "replacement players" are below average. [June 1, 1998; 9359]

Tom Ruane also replied to Merritt Clifton's historical review of replacement incidents:

> Merritt recently listed a series of players who (I suppose) were called into emergency service and responded with (for the most part) average seasons. Actually, all this adequately establishes is that my hypothetical team of replacements would have at least one average player. Granted. But what would the rest of the team look like? Remember, the 26 wins above replacement were doled out among ALL the regulars (position players and starters/closers) based upon my belief that a TEAM of replacement players would have a winning percentage of .350. So let's take one of Merritt's examples, the 1968 Red Sox. We'll leave Ken Harrelson (and his 3.5 TPR) on the team. Who would fill out the rest of the starting lineup after we remove the others starters? Here's a guess:

```
1B  Dalton Jones (-2.5 TPR in 354 at-bats)
2B  Syd O'Brien  (in AAA... -2.0 career TPR per 162 ML games)
SS  Jerry Adair (-1.6 in 208 at-bats)
3B  Carmen Fanzone (in AA... -1.9 career TPR with 588 at-bats)
RF  Ken Harrelson (3.5 TPR)
CF  Jose Tartabull (-1.0 TPR in 139 at-bats)
```

```
LF  Joe Lahoud (-0.9 TPR in 78 at-bats)
C   Russ Nixon (-1.3 TPR in 85 at-bats)
```

> I don't even want to think what their bench would've looked like. Of course the Red Sox probably would've gone outside of their organization to fill a few of these spots. Gary Geiger, Dick Schofield, Dick Nen and Ken Boyer were all available. Of course, none of them were particularly good.
>
> I'm not going to talk too much about the pitchers (how does a starting rotation anchored by Gary Waslewski, Jerry Stephenson and Dave Morehead sound?), because I'm sure that almost all of you get the idea. [June 1, 1998; 9364]

Michael Roca suggested that part of the confusion is due to different definitions of "replacement" players:

> I think the big discontinuity is the difference between "replacement level," the mathematical abstraction used for calculations, and "the player who replaces somebody in my lineup." An organization may have several players at a position who are above the calculated replacement level or none. By the same token "average" is a mathematical abstraction too. But nobody insists on finding a particular player who must serve as the average player and an organization may have several average players at a position or none. [June 3, 1998; 9395]

Merritt Clifton argued that the shape of the baseball "talent funnel" actually makes the replacement level very near the average of major leaguers:

> "Replacement value" proponents hold that players of average major league ability are scarce relative to jobs available, and that therefore a TPR of 0.0 should be credited as a positive. Their assessment of player ability as "positive" seems to begin at approximately -2.2 TPR. However, in real-life sudden emergency replacement situations, the average value of replacements given the chance to play regularly keeps turning up as average: 0.0. Why?
>
> What "replacement value" theorists are really misunderstanding is that the shape of the baseball talent funnel does not really match the shape of the baseball talent curve. The talent funnel looks something like this:
>
> 20,000 high school teams have 160,000 starting eight players. 2,000 college teams have 16,000 starting eight players. 150 entry-level pro teams have 1,200 starting eight players, and will use about 3,600 position players in all. At all of the above levels, annual turnover among the starting eight runs around 90% per year. But at the AA level, the caliber of talent suddenly gets an extra squeeze. 30 AA pro teams have 240 starting eight players, and will use about 600 position players in all. The rate of annual turnover drops to about 50% per year.
>
> 30 AAA pro teams have 240 starting eight players, and will use about 600 position players in all. The rate of annual turnover stays at about 50% per year. About 80% of the starting eight players will play in the major leagues at some point in their careers. About half will play regularly, and about 25% will play regularly for more than one year.
>
> 30 major league teams have 240 starting eight players, and will use about 600 position players in all. The rate of annual turnover in the regular lineup is about 25% per year. The rate of annual turnover among all position players is about 40% per year.
>
> Of the 240 major league starting eight players, about 80 are strongly positive TPR (1.0 or better), 80 are average TPR (within 1 either way of 0.0), and 80 are negative TPR. These are the first candidates to be replaced--but the turnover rate also supplies enough talent to replace some of the "average" players too.
>
> In addition, by this point the talent funnel has already gone through the tightest squeezes: about half of all the regular players in AAA at any given time could at least fill major league reserve jobs at the so-called replacement level, and might do better with opportunity, and about 25% are future "average" or better major league regulars.
>
> In terms of ability, this talent pool is nearly equal to the reserve pool on major league benches at any given time--with the primary difference that the best prospects for future regular duty tend to be in AAA, playing, instead of in the majors, sitting.
>
> It is obviously very hard to replace a player of established positive TPR at his established level, but the total pool of potentially average replacements is huge, certainly 2-to-1 and maybe as high as 3-to-1 relative to the opportunities to be average at the major league level, which accounts for the already high turnover and successful turnover rates, and allows for additional successful turnover (TPR 0.0 plus) in emergency situations. [June 3, 1998; 9396]

Steve Wang proposed a new definition of “replacement level” player that avoids many of the previous confusions:

> I think part of the confusion in this debate stems from a literal interpretation of the term "replacement level". Instead of thinking of replacement level as "the level of play of a typical replacement player", perhaps we should look at the definition from the opposite side: define replacement level as "the level of play typical of a player who *clearly needs to be replaced*”.
>
> For example, take the Yankees bullpen. Willie Banks is a replacement level pitcher. Although he pitched well last year, his career ERA is 4.94 in 503 IP, and this year he had a 10.05 ERA in 14 IP when the Yankees cut him. His career TPI (a Total Baseball equivalent of TPR for pitchers) is -5.
>
> Willie Banks is a zero. He has no value to the Yankees because they can easily obtain someone as good as he is *at virtually no cost* -- by claiming someone off the waiver wire, by trading a marginal AAA player, or whatever. That's the line that defines value in the major leagues. If Graeme Lloyd were to pitch the way he has for another decade, TPR would still say that he'd had no value to the Yankees, and that doesn't make sense. If the Yankees didn't have Lloyd, they'd have to trade a commodity to get someone else to do his job, and that gives him value. If the Yankees didn't have Willie Banks, they could replace him at no cost; hence Banks has no value. That's why value to a major league team is determined by the replacement level, not by the average. An average player has value. [June 7, 1998; 9458]

John Rickert then presented the results of a comprehensive study he did on 204 emergency replacement incidents. He found the average replacement to have negative TPR and that there were about twice as many negative TPR replacements as positive, and he did not find an age skew for replacements:

> As promised, I've looked at several "emergency replacement" scenarios. Tom Ruane kindly supplied me with a list of all players this century whose AB dropped by at least 200 one year and rose by at least 200 the next. (e.g. Tony Conigliaro 1967-9, Ozzie Guillen 1991-1993) I then went through "The Sports Encyclopedia: Baseball" and found those who suffered injuries or were hold-outs (I excluded military service from my final sample.) I compiled the list of replacements and eliminated a few cases where I could find no replacements (some "utility infielders" and players going downhill who had been traded). This left 204 cases, in which there were 285 replacement players. For these 285 replacement players, the mean TPR was -.497, with a standard deviation of 1.322, giving a 95% confidence interval for the mean TPR of (-.65,.-34). The median was -.5.
>
> The number of times each range of values was attained:

+3.0 - +3.9	3
+2.0 - +2.9	11
+1.0 - +1.9	22
0.0 - +0.9	62
-1.0 - -0.1	90
-2.0 - -1.1	69
-3.0 - -2.1	21
-4.0 - -3.1	6
-5.0 - -4.1	2

> The top replacements were

```
1968 Bos AL RF Ken Harrelson +3.5 for Tony Conigliaro
1981 Chi AL SS Bill Almon +3.1 for Todd Cruz (who played 1982 in Sea.)
1991 Pit NL 3b Steve Buechele +3.0 for Jeff King (one of several replacements)
1971 Atl NL LF Ralph Garr +2.9 for Rico Carty
1994 Mil AL SS Jose Valentin +2.9 for Pat Listach
1984 SF  NL RF Dan Gladden +2.9 for Jack Clark
```

> Of these, Garr, Valentin and Gladden were "raw", the rest had "played" 100-999G. For each of these players, the TPR was the highest of their career (though I'm hoping that Valentin will surpass his +2.9 several times in the future).
>
> The bottom replacements were

```
1985 Oak AL SS Donnie Hill -4.1 for Tony Phillips
1938 Chi AL SS Bose Berger -4.1 for Luke Appling
1978 Bal AL LF Carlos Lopez -3.8 for Al Bumbry
1970 Min AL 2b Danny Thompson -3.5 for Rod Carew (one of several replacements)
1931 Det AL 2b Mark Koenig -3.4 for Charlie Gehringer
```

> Thompson was "new", the rest had "played" 100-999G. For each of these players, the TPR was the lowest of their career, though Hill matched his -4.1 in 1987. [June 8, 1998; 9479]

John Pastier voiced his concern about “replacement” level:

> One problem with Replacement Player Level zero is that, unlike absolute zero [a player who never reaches base as a hitter and never fields a ball successfully] and TPR zero, it's unmeasurable. Advocates have offered many numbers as definitions over the last month or two, including a player who will induce his team to play at a .333 level, or at a .340 level, or at a .361 level, or at some level over .400, among those that I can remember -- and there were even more.
>
> My opinion is that if you can't measure it, it's not a useful stat, and maybe not even a real stat. The good thing about TPR zero is that it's less conceptually murky (it's an average player performance weighted by actual playing time, or it's the league average, whichever you prefer), and it's eminently measurable. RPL zero is neither convincingly definable, nor measurable, except in an arbitrary way that even its advocates don't seem to be able to agree on. [June 9, 1998; 9503]

Merritt Clifton summarized his views on the matter:

> What I have pointed out is that the average value of real-life replacement players is average. This is not a hypothetical matter; this is a matter of one or more teams per year actually having to permanently replace a regular who is suddenly and unexpectedly lost for the season, or forever, and generally coming up with a prospect who plays at an average first-year level of -- according to the case studies John Rickert and I have done -- roughly TPR -0.5, with a positive average career TPR value.
>
> Even in the very rare cases where teams have had to replace multiple players, these norms have held up. And even if a whole starting lineup were to disappear, the average career TPR value of the next best starting lineup available from the farm system still tends to be average, as does the average first-season or equivalent TPR value of those players.
>
> That's what the 1968 Red Sox model showed. Of course we don't actually know what would have happened if Carlton Fisk, Carmen Fanzone, et al., had been handed regular jobs in 1968, instead of later or never. But it is reasonable to believe that players like Fisk, who performed at an All Star level as soon as they got the opportunity, would be capable of performing at least an average level of performance a year or two sooner had the chance come sooner. [June 11, 1998; 9565]

Mark Armour disagreed with Merritt Clifton’s reasoning and conclusions:

> The entire point is not whether Fisk *became* a great prospect, it is whether Fisk would have been considered (at the time) a viable replacement if the Red Sox needed to field of team of replacements for the 1968 season. The answer, obviously, is no. He would not have been considered, since he had yet to play a single professional game, and I doubt that he would have been able to hit.
>
> It is easy to read the Baseball Encyclopedia and see who turned out to be a good player, or who had a good year. However, that it not the same thing as going back to a point in time and determining that certain decisions were obvious or that certain events are inevitable. [June 20, 1998; 9725]

Kevin Myers responded to a Tom Ruane post citing the 1998 Florida Marlins as a good example of “replacement value” players (recall the Marlins dumped all their stars after winning the 1997 World Series to cut their payroll):

> I would disagree with Tom’s assessment. The Marlins are not a good test case for this theory. When replacing Charles Johnson or Gary Sheffield, they have not looked for the *best* replacement, but the cheapest. There is a big difference. If they were looking to replace Charles Johnson with an equal or better player (or at least one perceived

> to be), they would have kept Mike Piazza. I feel that discussions of replacement value only have true value, especially when defining the concept, if we look at the norm rather than the exception. [June 12, 1998; 9573]

Larry Grasso disagreed with Kevin Myers and gave his reasons:

> I think Kevin is missing the point of the replacement level discussion. Replacement level as it has been discussed on this listserv and as the concept is used in determining VORP or making replacement level adjustments to TPR refers to a player readily available off the waiver wire or available at little or no cost. Remember, the idea is to come up with a zero value baseline for major league level play. If the Marlins have to give up a valuable player like Charles Johnson or expend a great deal of cash to obtain a replacement, that replacement is not very likely to be a "replacement level" player. The fact that the Marlins have chosen not to replace their World Series champion roster with players of equal value but instead have traded for prospects while filling their roster with marginal or undeveloped talent (some of whom may EVENTUALLY turn out to be stars at some point in the future) is precisely what makes them an appropriate example of replacement level.
>
> I'm thinking that it may be clearer to conceptualize replacement level by looking at players who leave. Players who are demoted to AAA or are cut or designated for assignment. Since this would likely be a low value, an average from empirical data that included players who voluntarily retire after a good year or who play well as an injury replacement but are nevertheless sent down for further seasoning would probably still yield a reasonable estimate of replacement level. [June 28, 1998; 9838]

Bill Deane revisited the Marlins at the end of the 1998 season:

> Back in the spring, there was a long and loud discussion here about "replacement level." I am embarrassed to recall that I started it off, estimating that the value of replacement-level is .333 -- meaning that a team of such players would play about .333 ball. Others chimed in with their own estimations of that value, ranging from .000 [upwards]. Tongue somewhat in cheek, I finally suggested that we define replacement-level as "whatever percentage the 1998 Marlins finish up with." The Marlins finished the season 54-108 (.333). [September 29, 1998; 11337]

The thread basically died out at that point and people moved on to other issues (like the merits of HEQ). There were however a few recent posts touching upon the issue, again, including the ones that follow. In late summer of 2000, the issue of Pete Rose's true value to his teams re-surfaced on SABR-L. Not having been on the list back in 1998 and therefore unfamiliar with the extensive discussion held then, I (Rob Wood) pointed out that TPR under-values Rose since it measured relative to league average rather than replacement level. I gave a crude estimate that the difference is probably about 2-3 wins per season for a full-time player. These posts on recurring themes prompted Tom Ruane to suggest that people check the SABR-L log before submitting a post on a recurring topic to determine if they have anything new to add to what has already been said. Indeed, Tom suggested that a compilation of the most recurring threads be made available – which is what I am doing here for the TPR: Average vs. Replacement Level issue. Anyway, to get back to our story, …

John Pastier, someone who was on the list back in 1998 and participated in that discussion, replied to my post with an interesting hypothesis and analytical approach (note that John seems to now embrace, or at least tolerate, a replacement level adjustment to TPR):

> According to the Bill Deane method suggested a year or two ago, which assumes that a team of replacement-level players will play .333 ball, 3 games per year per player would be the right average number over a 162-game season. (And about 2.85 for a 154-game season.)
>
> My question is this -- could it be that at the skill positions, replacement players are farther below the league average than normal, and therefore there should be a higher credit at those positions? And that the replacement-level corner outfielders, first basemen, and DH's are closer to the league average, and therefore those positions should show a lower credit? My intuition says this may be the case. (Let's call this the PSRL theory, for position-sensitive replacement level.)
>
> Another question is: how do you do this for pitchers? Which leads me to think that a full-time position player does not make up 1/9th of his team's value -- maybe it's more like 8%, with the pitching staff dividing up the other 36% of the total. (And make the necessary adjustments for the DH in the AL.)
>
> OK, now I'm trying to zero in on a better preliminary method –
>
> a) Offense = defense.
> b) Position players and the DH = 100% of AL offense, or 50% of the total game credit, which means about 5.5% of total game credit for a full-time hitter.

> c) Pitchers = 72% of defense, or 36% of the total game credit (a convenient and not totally implausible number, and I'm willing to listen to other opinions), so a pitcher who logs 1/6th of his team's innings (a real workhorse with about 240 i.p.) will earn 6% of the total game credit.
>
> d) Fielders = 28% of defense (some small amount of which accrues to pitchers as fielders), meaning that on average a position player earns a little over 3% of the total game credit. But this is not divided evenly, so that a shortstop will get perhaps 4% to 5% of game credit, while a first baseman gets perhaps 2% or even a bit less.
>
> e) Based on the above, a full time shortstop will get, say, 10% of game credit, and a full time 1B will get, say, 7%, combining offense and defense.
>
> f) Since we're allocating 27 games as the total difference between a team of average players and one of replacement players, a full-time SS might earn about 3.3 games above his TPR, while a full-time 1B might earn about an extra 1.9 games, and a full time DH about 1.5 games. This does not reflect my hunch about a possible PSRL theory, which would make the SS bonus even higher and the 1B bonus even lower.
>
> How does this affect Rose's TPR bonus? It would almost surely reduce it to below the intuitively plausible but simplistic 3 games per year, since pitchers take a bite out of that figure (IOW, the pitching staff as a whole earns more than 1/9th of the bonus), and since Rose played only a small fraction of his games at a high skill position). As a guess, I'd give him 2.2 games per 162, or roughly 48.4 more over his career, which would more than triple his 20.0 TPR in TB5. Impressive? Yes. Best all-time bonus? Maybe not. Ozzie Smith would earn about 52.5 points, Luke Appling about 49.3, and Cal Ripken probably more than 50. Granted, there's a lot of approximation and supposition here, but I think the basic logic is defensible. [August 8, 2000; 24507]

(Note that there are two separate issues John is discussing above. First, skilled position players are more valuable (defensively) to a team so that when apportioning a certain number of wins due to defense among a team's players, the skilled position players get more. Second and separate, John wondered whether the skilled position players essentially deserve an *additional* reward due to a greater "distance" between the average and the replacement level at skilled defensive positions.)

Tom Ruane posted TPR distribution data by position to test John Pastier's position-sensitive replacement level hypothesis (this remains an interesting and open issue):

> To test John Pastier's hypothesis, I computed the TPR by position (using where a player appeared most frequently as his position) by team from 1987 to 1998. (Prior to 1987, I don't have the outfielders broken down by position.) I next took the average of the three worst teams' TPR. My feeling was that if replacement players were worse at skill positions, this would be reflected in the TPR scores of the bottom-rung major league players at the same position. Here's what I found:

Year	C	1B	2B	3B	SS	LF	CF	RF
1987	-3.5	-2.4	-3.4	-2.8	-3.5	-3.1	-4.1	-3.8
1988	-1.9	-3.7	-2.6	-3.6	-3.4	-4.4	-3.5	-3.4
1989	-2.6	-3.3	-4.9	-3.4	-4.2	-2.8	-4.1	-3.9
1990	-2.5	-3.1	-4.3	-5.6	-2.4	-4.5	-3.4	-4.9
1991	-2.9	-3.7	-4.2	-4.7	-4.5	-5.1	-3.9	-3.0
1992	-2.5	-3.8	-3.9	-3.4	-3.8	-5.1	-4.0	-5.4
1993	-2.7	-3.5	-4.5	-3.3	-4.4	-6.0	-3.6	-6.1
1994	-2.9	-2.7	-2.2	-3.1	-2.5	-4.2	-3.0	-2.9
1995	-2.7	-3.2	-3.5	-3.6	-2.4	-4.8	-3.5	-2.5
1996	-2.9	-4.0	-3.9	-4.0	-3.9	-3.7	-5.2	-3.6
1997	-2.4	-3.2	-4.0	-3.7	-4.3	-3.6	-3.9	-3.4
1998	-2.7	-3.8	-4.1	-4.6	-3.4	-5.9	-3.9	-4.5
Total	-2.7	-3.4	-3.8	-3.8	-3.6	-4.4	-3.8	-4.0

> So a "skill" position (catcher) has the BEST "worst" players and the corner outfielders have the WORST "worst" players. This is pretty much the opposite of what I would have expected to find had John's theory been true. Not that this necessarily proves anything, but I thought it was interesting anyway.
>
> Note that this is the sum of ALL the players on a team at each position. As a result, the closer to average scores on the part of the catchers is not caused by their playing fewer games than those players at other positions. [August 9, 2000; 24527]

Concluding Remarks

So you can see that we came full circle. We started with a discussion of Pete Rose and returned again to Charley Hustle at the end. Along our journey, we learned a lot about TPR, “value", roster positions, “replacement level” players, etc. Several good studies were conducted to investigate issues that arose.

I hesitate to try to summarize this thread. But I do feel comfortable drawing a few conclusions and I’ll even try to “reconcile” the two sides of the argument with a general model below. Let me frame the issue this way. Suppose we are trying to compare two hypothetical players’ careers. Joe played for 10 years at a very high level and has a career TPR of 30. Pete, on the other hand, played at a high level for 15 years and has a career TPR of 25.

Based upon pure TPR, Joe seems to have had the greater career – he contributed more to his teams over and above the league average performance over the course of their respective careers. But the issue asks: is this the right way to think about it? After all, in the 5 years after Joe retired and Pete was still playing, Joe’s team had to replace him with someone else. How good do we expect this other guy to have been?

Replacement level proponents argue that Joe’s team would not have been able to easily find a league average player at a moment’s notice and without significant time, cost, and effort (e.g., scouting, development, trades, trying different players, etc.). Thus, the proponents argue that, essentially, Pete should be credited for those extra 5 years for playing above the replacement level, and they have estimated this replacement level to be around 2-3 wins per season for each full-time player (in the numerical example, I call it 2 wins for simplicity).

So, according to the proponents’ view, Joe actually contributed 30+2*10=50 wins above replacement level in his career, and Pete contributed 25+2*15=55 wins above replacement level in his career. That is, Pete picks up an extra 10 wins for the 5 years he played after Joe retired, and these 10 wins catapult Pete ahead of Joe in career value.

Opponents of this line of argument, led by Merritt Clifton, argued that a review of actual roster changes indicates that the pool of available average major leaguers is much larger and more easily accessible than the proponents claim. The players who replace injured or slumping players turn out to have, on average, average (TPR of 0.0) major league careers. The corollary is that Joe’s team would have likely been able to replace him after Joe retired for those extra 5 years with an average major leaguer. Thus, this argument goes, Pete deserves no additional credit above whatever he was able to contribute over the league average performance during those 5 years. Joe’s career TPR of 30 being greater than Pete’s TPR of 25 is the proper way to evaluate their respective career values.

I think that both camps have valid points which are not necessarily contradictory. The replacement level proponents define the replacement level to be that level of play that is available at a moment’s notice for negligible cost (e.g., waiver wire). By various methods, proponents concluded that a team of replacement level players would likely play around .333 ball (54-108 in a 162-game season). Opponents argued that teams can field higher quality (in fact league average quality) replacements, and do so all the time. But, I would contend that they do so only at a significant expenditure of time, money, effort, etc., that could have been used in other ways to make the team better. And being able to come up with league average replacements is likely to be possible for only one or two simultaneous replacements (i.e., every team has some “slack” built into their line-up and a new player can be accommodated by moving existing regulars to other positions).

The way I will try to reconcile the two camps is to introduce the element of time into the discussion. For the moment, suppose that the replacement level proponents are right that a team playing with replacement level players would play .333 ball for the first season. Now grant the opponents claim that this team will approach .500 over time “naturally” as these players develop, normal trades are made, prospects are found and promoted to the major league team, etc. After all, the 1998 Marlins did play .333, but the 1999 Marlins played .395 and the 2000 Marlins played .491. The woeful 1962 Mets played .250 but even they won the World Series in 1969 in their eighth season.

Here is a simple mathematical model that can help us estimate value above replacement level (over and above the value from playing better than league average). Let R be the number of wins that a league-average player would contribute to a team of replacement level players in the first year. From above recall that this was estimated to be 2-3 wins. Let N be the number of years that it would take a team of replacement level players to reach .500 through “natural” channels. Suppose that the team gets better linearly over time.

Under these simple assumptions, working through the algebra yields the following formula for the cumulative number of additional wins a player should be credited for playing a number of years (over and above the wins represented in TPR for his play relative to league average):

$$A_y = \left[\frac{yR}{2(N-1)}\right]\left[2N-(y+1)\right]$$

(for 1 <= y <= N)

where y is the number of years played by the player in question, and A is the cumulative additional value ascribed to him over those y years.

Note, of course, that A is capped at the formula value when y=N; that is, the player cannot accumulate any additional value, by construction, beyond N years.

Let me show how the formula works in three cases. First, according to the replacement level proponents, both R and N are large. It is easy to see that the limit of Ay as N gets large is yR, meaning, of course, that the player should be given R wins each year he plays.

Second, according to the replacement level opponents, both R and N are small. Again, it can easily be shown that Ay=0 as either R or N goes to 0. That is, no additional value should be given in this case.

Third, let me present a "middle" case. Let R be 3 and let N be 20. From the formula above, Table 1 presents the cumulative additional values in this case.

Table 1

Number of Years	Cumulative Additional Value Above Replacement
1	3.00
2	5.84
3	8.53
4	11.05
5	13.42
6	15.63
7	17.68
8	19.58
9	21.32
10	22.89
11	24.32
12	25.58
13	26.68
14	27.63
15	28.42
16	29.05
17	29.53
18	29.84
19	30.00
20 or more	30.00

To use this table, we would add the amount in the right column corresponding to the number of full-time seasons each player played in his career to his "pure" TPR figure. Returning to our hypothetical players, Joe, who played 10 years, would get an additional 22.89 to add to his TPR of 30, arriving at 52.89 career wins above replacement. Pete, who played for 15 years, would get an additional 28.42 to add to his TPR of 25, arriving at 53.42 career wins above replacement.

Of course, the values for R and N were chosen arbitrarily, and other values may be more appropriate, assuming this type of model itself is appropriate. And, as many people have pointed out in the discussion above, the value for R is likely to be sensitive to the position the player plays.

Hopefully, the discussion was informative and this compilation useful as we move forward.

Rob Wood, 2101 California St. #224, Mountain View, CA, 94040-1686, rob.wood@us.pwcglobal.com. ♦

Baker Bowl's Impact on Batting Average

Ron Selter

Park factors for batting average are currently available only for seasons 1983 to date. Here, the author presents BA park factors for the 1929-1937 Baker Bowl, information which is here available for the first time.

Park factor data for HRs and runs scored are currently available for the entire history of the major leagues. However, park factors for batting averages (BA) are available only for recent years as team and individual player home and road data have only been compiled since 1983. For seasons prior to 1983 the effect of a home ballpark on team BA has only been inferred.[1] This study seeks to measure the impact of a particular park (Baker Bowl 1929-37) on team BA. Baker Bowl, the home park of the Philadelphia Phillies (PHL), was selected as the subject ballpark for analysis because it was a most asymmetrical ballpark with a reputation as a great hitter's park.

Baker Bowl was an early Classic Major League ballpark (built 1895), and used by the Phillies until mid-season 1938. In the 1925-37 era, it was a ballpark with a sharply asymmetrical playing field -- a short/medium LF fence and a very short RF with a high wall and screen. (Baker Bowl's original dimensions, at the foul lines, were RF 272 ft. and LF about 390 ft.) In the time period under study, RF (at the foul poles) was 280.5 ft. and LF 341. The average distance (average over the entirety of each field) was 296 for RF and 354 for LF in Baker Bowl in the 1929-37 time period. The corresponding average distances for the NL at this time were 349 for RF and 370 for LF. Unlike Ebbets Field, which it roughly resembled, Baker Bowl was not well regarded while in use, nor fondly remembered after its passing. In the 1930s it was lightly attended, well-rusted, and noticeably rundown.

During the time period under study, Baker Bowl was known as a bandbox of a ballpark with lots of high scoring games. For example, in the 1929 season Baker Bowl contests had per-game averages of 33% more runs scored and 72% more HRs than the average NL ballpark. Baker Bowl was also the home park for the 1930 PHL pitching staff, which managed to post the highest ERA of any 20th Century major league team. Thus, there was evidence to support the widespread view that Baker Bowl itself, along with the contributions of the PHL pitching staff, was the cause of much of the offensive exploits occurring in Baker Bowl.

Methodology

Data on team BA for games vs. Philadelphia at Baker Bowl and at the opponent's home parks were obtained by aggregating the box scores for all games, as taken from The Sporting News and the Los Angeles Times for the 1929-37 seasons. The aggregated team at-bats and hits for each season's games vs. PHL at Baker Bowl were used to derive (from the total PHL Opponents AB/H found in *Total Baseball*) the opponents' AB/H at other NL parks. These box scores are not the official NL game records and may differ in slight details from a compilation made using the official NL game records. A comparison of box score compiled data, for individual batters in games at Baker Bowl, vs. data compilations made using the official NL records (from Bill Deane) showed an error rate of less than 0.5%. The total data set consists of nine seasons encompassing nearly 700 games at Baker Bowl and slightly more at other NL parks.

PHL team batting data were also complied using the same method for the 1929-37 seasons (home and road). The PHL data included the compilation of extra-base hits and included home and road batting data for selected LH batters and seasons.

Results

Not surprisingly, the opponents' composite team BA was higher at Baker Bowl than at home for every season in the nine seasons studied. The data are shown below in Table 1:

Table 1: Opponents' BA – Baker Bowl vs. Home

Year	Baker	Home	Differential (Points)
1929	.3389	.2980	40.9
1930	.3586	.3332	25.4
1931	.2997	.2895	10.2
1932	.2969	.2763	20.6
1933	.3212	.2637	57.5
1934	.2932	.2828	10.4
1935	.3128	.2758	37.0
1936	.3046	.2782	26.4
1937	.3219	.2716	50.3
1929-37	.3155	.2857	29.8

The average opponents' BA differential of 29.8 points (9.9%) for Baker vs. Home can largely be explained by the 10.1% smaller park size at Baker Bowl. The year-to-year variations in the Baker/Home BA differentials are believed to be largely random. An exception may be the 1931 season. In 1931 the NL for began using a baseball with noticeably raised seams and, it is suspected, less tight winding. The overall effect was to sharply reduce NL BA and HRs. It would appear that with a less-lively ball, fewer hitters were able to reach the closer Baker Bowl fences. In subsequent years changes in the composition of NL batters and pitchers, as well as changes in the other NL parks, and after 1934 the ball again, makes it very difficult to draw conclusions about the reasons for the year-to-year variations in the Baker/Home BA differentials.

The results for the Phillies' team BA were somewhat more extreme. The most obvious difference is the much larger impact of Baker Bowl on PHL BA. Table 2A shows the nine seasons of BA data and the Home/Road record of the Phillies is shown in Table 2B:

Table 2A: Phillies' BA – Home vs. Road

Year	Home	Road	Differential (Points)
1929	.3499	.2794	70.5
1930	.3410	.2892	51.8
1931	.3040	.2556	48.4
1932	.3291	.2532	75.9
1933	.2952	.2536	41.6
1934	.3082	.2610	47.2
1935	.2912	.2475	43.7
1936	.2913	.2475	43.8
1937	.2879	.2603	27.6
1929-37	.3102	.2635	46.7

Table 2B: Phillies' Home/Road Record (W-L-T)

Year	Home	Road
1929	37 - 37	32 - 45 - 1
1930	35 - 42	17 - 60 - 2
1931	40 - 36	26 - 52 - 1
1932	45 - 32	33 - 44
1933	32 - 40	28 - 52
1934	35 - 36	21 - 57
1935	35 - 43 - 1	29 - 46 - 2
1936	30 - 48	24 - 52
1937	29 - 45	32 - 47 – 2
1929-37	320 –359 - 1	242 –455 - 8

The absolute size of the BA differentials were large (exceeding 70 points in 1929 and 1932), and I believe are largely responsible for the Phillies' better record at home than on the road (where they were nearly always awful).

What does explain the much larger PHL Home/Road BA differential? I believe this is in part evidence of adaptive behavior on the part of the PHL batters. Players with 77 games a season at Baker Bowl (or other wildly odd and/or asymmetrical parks) had a much larger incentive to adjust their batting styles and techniques to make use of the cozy RF wall than visiting players who had only 11 games per season in Baker Bowl. An alternative explanation is the large BA differential was due to the inherent home park advantage plus the park size effect. The total home/road BA differential for the 1929-37 time period, 46.7 points, is far larger than the estimated average home park advantage of 6-8 BA points.[1] Including the park size effect (29.8 points for the visitors), and allowing a 6 point home park effect produces an estimate of 35.8 points vs. the actual differential of 46.7 points. Data for part of this time period (1929-31) suggest that only a small part (less than 10%) of the larger PHL Baker/Road BA differential was due to the mix effect (PHL having a larger proportion of LH batters than other NL teams (36% vs. 28-33%).

The problem is that the inherent home park advantage, the home teams asymmetrical roster mix, and the batters' adaptive behavior are all intermixed and have not yet been estimated separately.

The possible differential impact of Baker Bowl on LH vs. RH batters was also examined. The data on opponents' BA for LH batters is much more limited than for Team BA. Data was laboriously collected for all visitors' LH batting regulars for the 1930 and 1931 seasons. Regulars were defined as those players listed as regulars in the Macmillan Baseball Encyclopedia for the respective seasons. All other non-LH Regulars, which comprises all reserves, pitchers, and RH batting regulars, are as a group called OTHER. The OTHER group is nearly entirely but not completely made up of RH batters. The comparative BA data for LH Regulars and the OTHER group, at Baker Bowl and for home games vs. PHL, are shown in Table 3.

Table 3: Phillies' Opponents' BA – Baker vs. Home (LH/RH)

	Baker			Home		
	AB	H	BA	AB	H	BA
LH Reg.						
1930	753	310	.412	707	264	.373
1931	685	236	.340	630	177	.320
1930-31	1438	546	.380	1337	441	.330
OTHER						
1930	2189	745	.344	2093	669	.281
1931	2099	598	.285	2033	594	.292
1930-31	4288	1343	.313	4126	1263	.306

As one would expect, given the cozy RF wall at Baker Bowl, LH batters had a much greater Baker/Home BA differential (50 points vs. 7 points) than the OTHER group (which is essentially RH batters). The 1930 BA for LH Regulars (.412) at Baker Bowl could well be the highest full season opponents' BA ever recorded.

The possible differential impact of Baker Bowl on PHL LH vs. RH batters was also examined. BA data for all PHL LH and RH batters (including reserves and pitchers) were collected for the 1929, 1930, and 1931 seasons. The comparative PHL home BA data (at Baker Bowl) and on the road are shown in Table 4:

[1] Private research by the author for team BA in the 1930s produced estimates of the inherent home park advantage, adjusted for differences in park size and asymmetrical configuration, of 3-6 BA points.

Table 4: Phillies BA – Home vs. Road (LH/RH)

	Home			Road		
	AB	H	BA	AB	H	BA
LH						
1929	1024	396	.387	1004	311	.310
1930	986	368	.373	1127	362	.321
1931	990	329	.332	1059	262	.247
1929-30	3300	1093	.364	3190	935	.293
RH						
1929	1633	507	.310	1823	479	.263
1930	1833	594	.324	1721	459	.267
1931	1658	476	.287	1668	435	.261
1929-31	5124	1577	.308	5212	1374	.264

The dropoff in BA from 1930 to 1931 for the Phillies LH batters on the road was remarkable. Their batting average went from .321 to .247. It would appear that the less-lively ball in use in 1931 made batting techniques that still worked well for LH batters at Baker Bowl were now on the road producing below league-average results. For the 1929-31 time period, the Phillies LH batters had roughly double the BA Home/Road differential of their RH batters.

Appendix – Complete Home/Road Batting Lines

PHL-NL, home

HOME	G	AB	H	2B	3B	HR	BA	Slug
1929	76	2657	903	169	11	86	.3399	.509
1930	77	2819	962	187	9	72	.3413	.491
1931	76	2648	805	175	18	51	.3040	.441
1932	77	2805	923	198	20	86	.3291	.506
1933	72	2520	744	137	11	45	.2952	.412
1934	71	2505	772	166	16	28	.3082	.421
1935	79	2727	794	146	9	52	.2912	.409
1936	78	2753	802	148	16	69	.2913	.432
1937	74	2539	731	126	10	65	.2879	.422
Totals	680	23946	7436	1452	120	554	.3105	.451

PHL-NL , road

ROAD	G	AB	H	2B	3B	HR	BA	Slug
1929	78	2827	790	136	40	67	.2794	.427
1930	79	2848	821	158	35	54	.2883	.425
1931	79	2727	697	124	34	30	.2556	.359
1932	77	2705	685	132	47	36	.2532	.377
1933	80	2741	695	103	30	15	.2536	.329
1934	78	2713	708	120	19	28	.2610	.350
1935	77	2715	672	103	23	40	.2475	.347
1936	76	2712	736	102	30	34	.2714	.369
1937	81	2885	751	132	27	38	.2603	.364
Total	705	24900	6555	1110	285	342	.2633	.372

PHL-NL Opponents, Baker

Baker	G	AB	H	BA
1929	76	2800	949	.3389
1930	77	2942	1055	.3586
1931	76	2784	834	.2996
1932	77	2870	852	.2969
1933	72	2721	874	.3212
1934	71	2609	765	.2932
1935	79	2906	909	.3128
1936	78	2915	888	.3046
1937	74	2771	862	.3111
Total	680	25318	7988	.3155

PHL-NL Opponents, Opponents' Home Parks

Home	G	AB	H	BA
1929	78	2664	794	.2980
1930	79	2800	933	.3332
1931	79	2663	771	.2895
1932	77	2667	737	.2763
1933	80	2613	689	.2637
1934	78	2603	736	.2828
1935	77	2694	743	.2758
1936	76	2667	742	.2782
1937	81	2714	737	.2716
Total	705	24085	6882	.2857

Ron Selter, 1430 E. Walnut Ave., El Segundo, CA, 90245; henneseltr@aol.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. 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 this issue, or Gabe Costa's review last issue.

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.

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	kcarlson@stlnet.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.