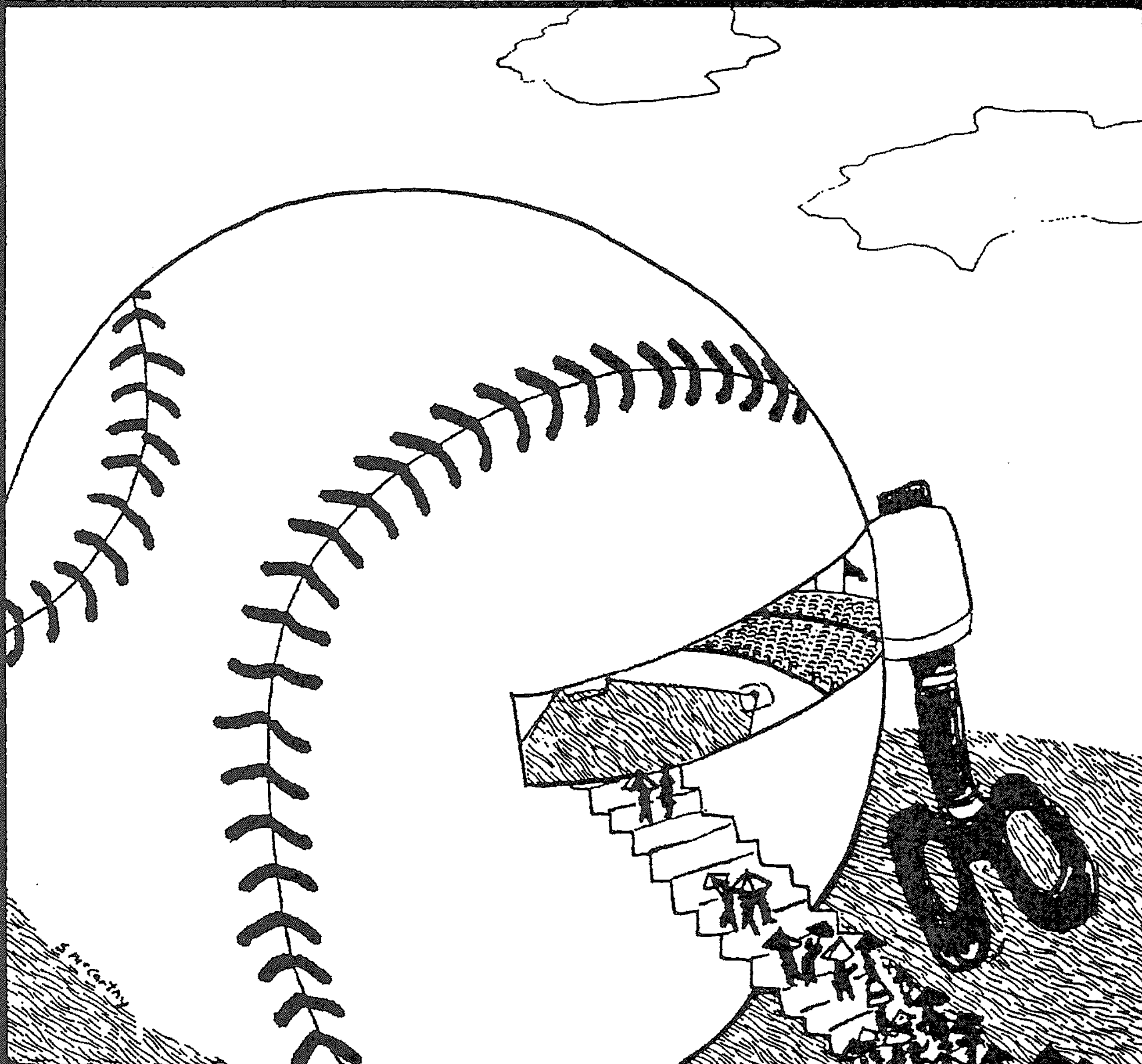
baseball ANALYST

April 1984 The Journal of Sabermetrics Issue # 11



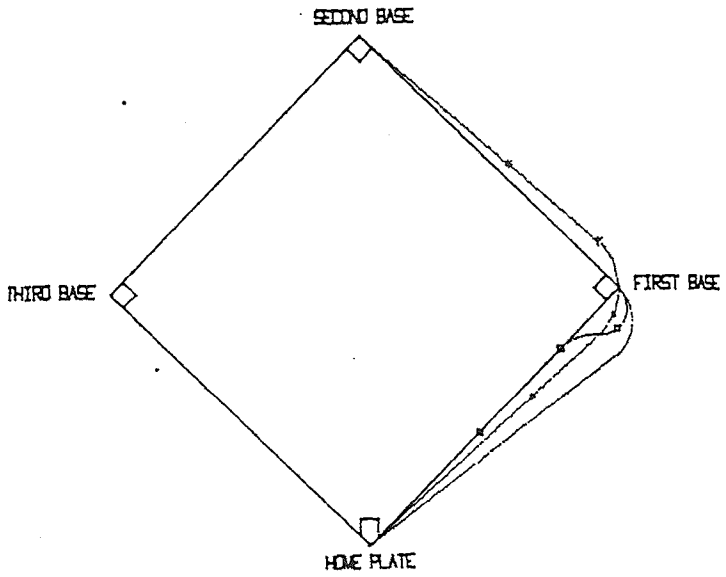
OPENING DAY

To the Editor:

This is just a brief note in response to a brief note written by John Schwartz in the October 1983 Analyst. John wondered about the best way to circle the bases. His question rang a bell. I recalled seeing some data on it once upon a time, but I couldn't remember where. Finally, I remembered. The data I remembered were used as an example in a statistics textbook on nonparametric statistics. It took me a while to rediscover the book, but here is the information.

The study was performed by Woody Woodward in 1970 as part of a Masters Thesis from Florida State University. Incidentally, that same year Woodward had 8 doubles and 3 triples for the Reds; all opportunities to put his research into practice. He compared three methods of rounding first base illustrated in the figure below. The fastest method appeared to be the "wide angle" method which is represented with the solid, smooth curve. The next best was the "narrow angle" method represented with the asterisks. The worst method was the "round out" method pictured with the diamonds in the figure.

Charles Hofacker



A NEW MEASURE OF GREATNESS
by Daniel Greenia

Throw away computers and three-page-long formulas. There's a much easier way to measure a player's greatness: column inches! Using only the 1984 Baseball Register and a ruler I've discovered which active players are most likely to make the Hall of Fame.

The list:

R. Jackson	13-3/4	S. Carlton	9-3/4	S. Garvey	9-1/4
P. Rose	13-1/4	J. Palmer	9-3/4	J. Kaat	9-1/4
J. Morgan	12-1/4	M. Schmidt	9-3/4	G. Brett	9-1/4
J. Bench	11	G. Nettles	9-1/2	R. Carew	9-1/4
Campaneris	10-3/4	T. Perez	9-1/2	G. Perry	9
Yastrzemski	10-1/4	R. Fingers	9-1/2	T. John	9
N. Ryan	10-1/4	D. Lopes	9-1/2		

Simplistic? Beautifully so. Baseball analysis? Well, more freak show. Accurate Hall of Fame predictor? I'll bet at least 18 of these 20 are in by the time I'm Rick Ferrell's age. (Note: One page = 8 inches.)

World Headquarters
945 Kentucky Street
Lawrence, KS 66044

$12.00 per annum

All Submissions ready for reproduction, please

-baseball- ANALYST

April 1984

Founder/Publisher: Bill James
Editor: Jim Baker
Business Manager: Susie McCarthy

Issue no. 11

IN THIS ISSUE:

ABOUT THE COVER: The cover this issue was done by Susie McCarthy

A SPECIAL PLEA: The Baseball Abstract Library needs old Sporting News, Reach and Spalding Guides from the turn of the century to date. We are paying the going rate for these items and would appreciate any help you can give us. If you have any guides, or know of anybody that does, please write us (or call 913-749-2998 9-5cst, or 913-843-3119 after 5pm) and give a description that includes year and condition. We promise prompt payment!

MAYS vs. AARON: A NEW LOOK

By Bill Deane

Who was the greater ballplayer: Willie Mays or Henry Aaron? It is a question we have heard discussed dozens of times.

Since I didn't see either one of them in their primes, I usually just politely nod when the subject comes up, and others recount some of Mays' sensational catches and Aaron's dramatic home runs. Growing up in New York state, I have found that the majority of fans I have listened to believe that Mays was unquestionably the better of the two, defending their choice with almost religious zeal. Aaron's supporters seem less emotional about the subject, a contrast that mirrors the difference of styles between the flamboyant Mays and the low-key Aaron.

Bob Uecker, in a rare serious passage of his book, The Catcher in the Wry, called Aaron "the most underrated player of my time, and his." Uecker cites three reasons why Aaron seldom got the recognition he deserved: "(1) He didn't play in New York or Los Angeles; (2) he was too predictable, meaning that most years he would get his forty homers, drive in a hundred runs and hit .320... no surprises here; (3) he lacked showmanship. His cap didn't fly off when he caught a fly ball, and after a game you never saw him eat twenty hot dogs and wash them down with a six-pack of beer."

The cold statistics, which are all I have to rely on, seem to support Uecker's endorsement. Aaron had roughly 500 more hits, 100 more home runs, and 400 more runs batted in (to be exact, 488, 95 and 394, respectively). My interpretation of this was that, while both were outstanding players, Aaron must have maintained a higher level of achievement for a longer period of time. To test this theory, I did a study in which I assigned arbitrary guidelines to define a "productive" season (20 HR, 80 RBI, .280 Avg.), a "star" season (30-100-.300), and a "superstar" season (40-120-.320). The results:

CATEGORY	AARON	MAYS
PRODUCTIVE		
20 HR SEASONS	20	17
80 RBI SEASONS	18	14
.280 SEASONS	18	16
ONE OF THREE	21	18
TWO OF THREE	18	15
ALL THREE	17	14
STAR		
30 HR SEASONS	15	11
100 RBI SEASONS	11	10
.300 SEASONS	14	10
ONE OF THREE	18	13
TWO OF THREE	15	11
ALL THREE	7	7
SUPERSTAR		
40 HR SEASONS	8	6
120 RBI SEASONS	7	3
.320 SEASONS	8	3
ONE OF THREE	13	8
TWO OF THREE	8	4
ALL THREE	2	0

Some observers maintain that Aaron also held the edge in the intangible asset of "killer instinct", a contention substantiated by comparing the post-season records of the two players. Mays, in 25 post-season games, had a .247 batting average and a meager .337 slugging percentage, with just one home run and 10 RBI. Aaron, in 17 fall contests, slugged 6 home runs, knocked in 16 runs, batted .362, and had an awesome .710 slugging mark.

These facts, it would seem, shift the burden of proof to the Mays supporters.

One category in which Mays outshines Aaron is stolen bases, but not by as much as one might think: Mays stole 338 bases, Aaron 240. Mays had a high success ratio (77%), but so did Aaron. Hank led the N.L. in SB percentage in 1963 (31 thefts, 5 caught stealing, 86.1 %), 1966 (21-3, 87.5%), and 1968 (28-5, 84.8%).

But the two arguments I hear most in Mays' behalf are (1) Mays was the far-superior outfielder, perhaps the best in history, and (2) Aaron had the advantage of playing in the home run haven of Atlanta, while Mays battled the nightmarish winds of Candlestick Park. I set out to examine the validity of these two arguments.

ARGUMENT ONE: DEFENSE. First, I took the Baseball Encyclopedia and went through the year-by-year outfield statistics of each player. The first thing I noticed was that the stats do not support anyone's claim of Mays as the best ever. In 22 years, he led the league only once each in putouts and assists, and never led in fielding average.

Aaron played in right field for the bulk of his career, so he obviously didn't have as much opportunity for putouts as center fielder Mays; but Aaron did top all league right fielders in putouts five times. He had nine or more assists for 15 consecutive seasons, but never led the league in that category; nor did he ever lead in fielding average.

Mays led center fielders three times each in assists and errors. Aaron led right fielders in errors twice.

The bottom lines: Mays, in 2843 outfield games, had 7095 putouts, 195 assists, 141 errors, and a .981 fielding average. Aaron, in 2760 games, had 5539 putouts, 201 assists, 117 errors, and a .980 average. Aaron compares well in every category except putouts, where Mays' average of 24% more putouts per game is chiefly attributable to the differences of positions (although, in fairness, it must be said that center field is certainly the more important position).

Interestingly, when Aaron was shifted to center field for one full season (1961), the stats of the two players were very similar. Mays, in 153 games, had 385 putouts, 7 assists, 8 errors and a .980 FA. Aaron, in 154 games, had 377 putouts, 13 assists, 7 errors and a .982 FA.

ARGUMENT TWO: HOME FIELD ADVANTAGE. For this one, I went through year-by-year league home and away homer records. The results were somewhat surprising (see next page).

It is certainly true that Aaron's nine years in Atlanta helped his home run totals. But, how come nobody mentions how much his sixteen seasons in Milwaukee hurt his homer totals-- so much so that it virtually nullifies the Atlanta effect?

And, if Candlestick drastically reduced Mays' home run stats, it certainly isn't evident from this chart. In only four of his 14 seasons there did he hit more home runs on the road than at home.

And here's the kicker. I took the twenty seasons (1954-73) that both Aaron and Mays were active in the major leagues; I subtracted all home runs that either

player hit in his own park, AND those he hit in the other player's park, making it as even a comparison as possible. The result: Aaron, 315 homers; Mays, 275 homers.

None of this will change the minds of any staunch Mays supporters, but it might at least send them looking for new arguments.

HANK AARON vs WILLIE MAYS: Career Home/Road Home Run Breakdowns

	MAYS		AARON	
YEAR	HOME	ROAD	HOME	ROAD
1951	13	7		
1952	2	2		
1953	(Military)			
1954	20	21	1	12
1955	22	29	14	13
1956	20	16	15	11
1957	17	18	18	26
1958	16	13	10	20
1959	16	18	20	19
1960	12	17	21	19
1961	21	19	19	15
1962	28	21	18	27
1963	20	18	19	25
1964	25	22	11	13
1965	24	28	19	13
1966	16	21	21	23
1967	13	9	23	16
1968	12	11	17	12
1969	7	6	21	23
1970	15	13	23	15
1971	9	9	31	16
1972	3	5	19	15
1973	4	2	24	16
1974			11	9
1975			4	8
1976			6	4

TOTALS	HOME	ROAD
Mays-Polo Grounds (1951-57)	94	93
-Candlestick (1958-71)	234	225
-Shea Stadium (1972-73)	7	7
-TOTAL	335	325
Aaron-Milwaukee (1954-65)	185	213
-Atlanta (1966-74)	190	145
-Milwaukee (1975-76)	10	12
-TOTAL	385	370

* *

NOTE: In the December 1983 issue of Baseball Analyst, I did a piece analyzing pitchers' winning percentages, and introducing "Pitcher Performance Percentage", whose formula was:

$$PPP = .500 + \left[\frac{\text{Pitcher's Pct.} - \text{Team Pct.}}{2(1.000 - \text{Team Pct.})}\right]$$

Since my chart rated the best pitchers in this category, and only three of the 64 200-game winners in this century had lower win percentages than those of their teams, I omitted the converse of my formula, which is ONLY FOR PITCHERS WITH LOWER WIN PERCENTAGES THAN THEIR TEAMS. Since at least one reader spotted this omission, here is the "Converse-PPP" formula:

$$\text{Converse (PPP)} = .500 - \left[\frac{\text{Team Pct.} - \text{Pitcher's Pct.}}{2(\text{Team Pct.})}\right]$$

* *

THE BEST FIELDING THIRD BASEMEN SINCE 1925

Dan Finkle

In a previous article I showed some measures for the best fielding second basemen since 1925. This time we'll ta', a look at third basemen. The results are very different. One second baseman, Bill Mazeroski, stood above the rest like a photo of Gene Conley next to Fred Patek; no one third baseman excells in the same way. On all the various measures of best fielding second basemen only one currently active player, Manny Trillo, appeared; but the third basemen's listings are dominated by players in the 1983 box scores.

The Fielding Index (FI) uses all the fielding information available in the Baseball Guide. Chances per game is the fundamental concept but corrections are made for differences in the pitching talents of each team. The FI for third basemen has four components: Fly Ball Index, Range Index, Double Play Index, and Misplay Index.

The Fly Ball Index is not used in the FI for either second basemen or shortstops because standard statistics do not enable us to separate putouts made at the base from fly ball putouts. Fly ball putouts are a legitimate measure of fielding skill; putouts at the base, with the exception of the pivot on the double play, are usually routine performances requiring no skill at the position. Skill at the double play pivot is measured in the Double Play Index. Putouts at the base are rare occurences at third base as compared to second base because the routine force play does not occur at third base. So the Fly Ball Index for third basemen is mostly line drives and pop flies, good tests of fielding skill.

The FI is calculated for performers with at least 100 games at the fielding position. Players are compared to the average for the league and year. In that way, comparisons among years may legitimately be made. One tacit assumption is made: That overall performance from year to year is the same. Although that assumption is not exact, it appears to be sufficiently valid and in consonance with the accuracy of other data and assumptions used in calculating the FI. When comparisons are made over a player's entire career, as in this report, the possible inaccuracies become negligible.

To identify the best fielding third basemen we will look at performance over four, eight, and twelve consecutive years -- where consecutive means years performing as a regular, not necessarily consecutive calendar years.

Four years does not a career make. But a look at performance over four years can give us the picture of great performers at the peak of their careers.

BEST FIELDING THIRD BASEMEN FOUR CONSECUTIVE YEARS

Bell	1978 - 81	1.195
C. Boyer	1961 - 64	1.165
Robinson	1966 - 69	1.151
Schmidt	1974 - 77	1.146
Nettles	1970 - 73	1.139

Gus Bell's baby boy, Buddy Bell, leads the group by a considerable margin. Notably, three of the top five bests were in the seventies, and the players are still active today.

Before looking further into the dominance of today's third basemen as fielding leaders, let's look at the best fielding third basemen for an eight year span. Since 1925 forty six men, excluding those currently active, have been regulars at third base for at least five years. The average career length for these is 8.2 years. Eight years may, therefore, be an ideal measuring period to identify the best.

BEST FIELDING THIRD BASEMEN EIGHT CONSECUTIVE YEARS

Bell	1975 - 82	1.130
Schmidt	1974 - 81	1.128
C. Boyer	1961 - 70	1.123
Robinson	1962 - 69	1.118
Nettles	1970 - 77	1.112

The same names appear, order somewhat changed. The differences in their FI ratings are small.

The fielders rated just below these come from earlier eras: Ron Santo of the sixties; Ken Keltner of the forties; Art Whitney, Harland Clift, and Merril May of the thirties; and Pie Traynor and Willie Kamm of the twenties.

In considering the fielding skill dominance of current-day third basemen, I speculated that the strategies of the game had changed in recent years to lead to more opportunities for chances at the hot corner. Alternatively, I wondered if the spectacular diving stops we associate

with Brooks Robinson and others, plays that have become more visible because of the marvels of TV replay, could represent a new style and skill among third basemen.

To test these hypotheses I counted the percentage of putouts, assists, double plays, and errors made by third basemen, shortstops, and second basemen in the two major leagues for the seven complete or partial decades from 1925 to 1982. The results do not seem to support the hypotheses.

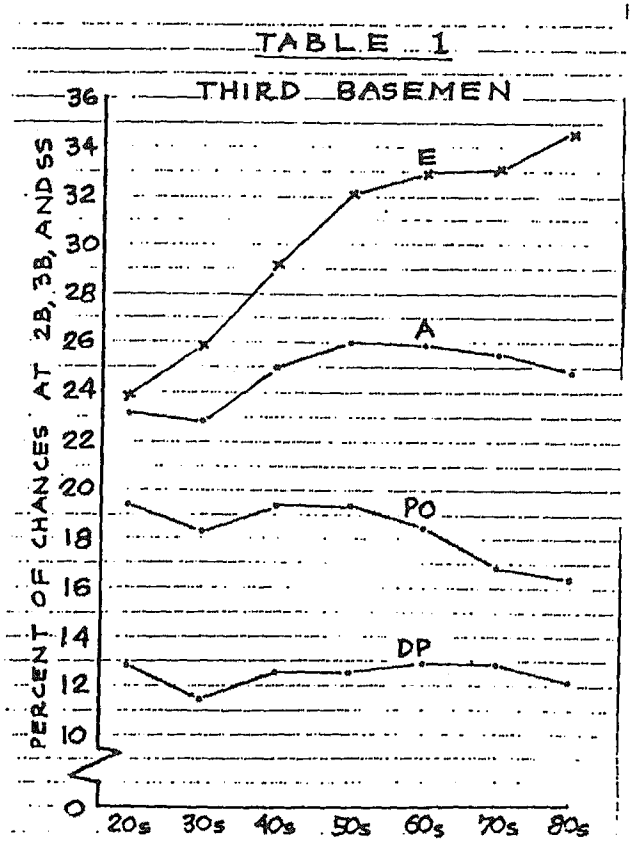


Table 1 is a summary of the results. Assists ranged from a low of 23% in the 20's and 30's to a high of 26% in the 50's and 60's. These gains were made at the expense of second basemen, whose portion dropped as third base assists increased. The shortstop portion remained relatively stable.

The percentage of putouts at third base has been falling since the 50's. At second base the percentage of putouts increased. The proportion of double plays has remained stable.

The proportion of third basemen's errors has been sharply increasing. Rather than confirm, these facts seem to controvert the hypotheses.

I am left with the conclusion that the predominance of great fielders in the current group of third basemen is that we are in a golden era of third basemen. It seems more a mystical than a statistical conclusion, but one analogy occurs to me that may be supportive. Look at the great first basemen of the thirties. On anybody's list Gehrig, Foxx, and Greenberg have to be ranked among the top five first basemen of all time -- and they were contemporary. Add to them Terry and Bottomley and you have five of the eight first basemen in the Hall of Fame, all active in the same years.

I'm willing to leave it at that and enjoy it. Folks, we are living in the era of great third basemen and we may never see the like again.

Now let's look at fielding performance over twelve years to give greater emphasis to longevity of talent.

BEST FIELDING THIRD BASEMEN TWELVE CONSECUTIVE YEARS

Robinson	1960 - 71	1.105
Nettles	1970 - 82	1.085
Santo	1961 - 72	1.075
Mathews	1954 - 65	1.007
Bando	1968 - 79	.958

This is a very different list than that for four and eight years. Neither Buddy Bell nor Mike Schmidt had played for twelve years by 1982. When and if they do, they are likely to enter this select group. Clete Boyer had only eight years as a regular at third base. Eddie Mathews and Sal Bando appear on this list not because of exceptional fielding skill but because of their longevity.

It turns out that a number of good fielding third basemen had eleven but not twelve years as regulars. If we had chosen eleven years as the measure, the top three would remain in the same order. Replacing Mathews and Bando would be Stan Hack and Harland Clift. Mathews falls to

This group is dominated by players from the twenties and thirties. Another look at Table 1 will show why the Fly Ball Index is higher for players from an earlier era. I would be interested if someone could explain what changes in the format of the game have produced this result.

THIRD BASEMEN WITH BEST RANGE INDEX

Four Consecutive Years		Eight Consecutive Years	
Bell	.642	Schmidt	.609
Schmidt	.616	Bell	.600
Nettles	.608	Nettles	.582
Santo	.594	Santo	.577
C. Boyer	.589	Rodriguez	.560

Third basemen who are still active players lead in the Range Index. Combined, the Fly Ball Index and Range Index are a general measure of fielding mobility. Three names appear on both lists: Buddy Bell on the four and eight year list of both measures, Ron Santo, and Clete Boyer. A new name, Aurelio Rodriguez, has not appeared previously but he is ninth on the list of best fielding third basemen.

THIRD BASEMEN WITH BEST DOUBLE PLAY INDEX

Four Consecutive Years		Eight Consecutive Years	
C. Boyer	.141	Schmidt	.129
Schmidt	.139	Robinson	.126
Robinson	.132	C. Boyer	.123
Keltner	.128	Keltner	.116
Nettles	.124	Nettles	.115

Ken Keltner was the standout fielder of the forties. He not only ranks high in double play skills but also eighth, ninth, or tenth in fly ball and range indices. He is the seventh best fielding third baseman.

THIRD BASEMEN WITH BEST MISPLAY INDEX

Four Consecutive Years		Eight Consecutive Years	
Robinson	.258	Robinson	.242
Keltner	.251	Keltner	.234
Kamm	.244	Kamm	.231
Whitney	.239	Whitney	.223
Bell	.234	C. Boyer	.215
		Rodriguez	.215

eighth on the eleven year list, Bando to tenth.

Players with the longest careers as third base regulars are Brooks Robinson 17, Eddie Mathews 15, and Ed Yost, not an outstanding fielder, 14.

Putting these data together I have calculated, as I did with second basemen, an index of best fielding third basemen considering both longevity and high FI.

BEST FIELDING THIRD BASEMEN

Robinson	1.123
Nettles	1.112
Santo	1.105
Bell	1.098
Schmidt	1.084

This result confirms the conventional wisdom. Brooks Robinson is the greatest fielding third baseman to this time. The three players on this list who are still active may yet prove even greater as fielders than Mr. Robinson.

The appearance of Ron Santo on the list may surprise some who thought of him primarily as a fine hitter. In my earlier discussion of second basemen I gave a cry of anguish at ignoring Bill Mazeroski in the Hall of Fame. The case for Ron Santo can be made on the basis of not only his outstanding fielding but also his exceptional offensive abilities I think there are three players whose absence from the Hall of Fame is startling, based not on the need for recognition of fielding that I argued before, but simply based on the present (unwritten) rules for choosing hall of famers. Two are familiar to most connoscenti: Arky Vaughan and Ernie Lombardi. I suggest that the same kind of case can be made for Ron Santo.

The next step in considering the fielding performance of third basemen is to look at each factor index.

THIRD BASEMEN WITH BEST FLY BALL INDEX

Four Consecutive Years		Eight Consecutive Years	
Traynor	.276	Traynor	.263
Kamm	.260	Kamm	.249
Whitney	.252	Whitney	.234
Bell	.235	Bell	.227
Clift	.234	C. Boyer	.227
Santo	.234		

continued on page 20

Statistical Procedures for Baseball Research I: Correlation and Simple Regression

Charles Pavitt and Elaine M. Gilby

In the most recent issue of the Baseball Analyst, Dick O'Brien presents some suggestive evidence that number of double plays has no relationship to the amount of runs a team gives up. However, as he presents no direct evidence that the two measures are associated, his evidence can at best be regarded as circumstantial. At the end of the article, he asks rhetorically how else one would "put to rest for all time the myth of the importance of the double play." The goal of this article is to show just how one would do this. We will discuss two basic methods for relating two quantities ("variables"), correlation and linear regression.

Correlation is a measure of the association between two variables; in other words, the degree to which they "go together". Consider the following two columns of imaginary fielding statistics, with fielding percentage defined normally:

Player	Assists	Pct.
A	400	.980
B	375	.960
C	350	.940
D	325	.920
E	300	.900

Note, for a start, that as one column increases, the other does likewise; we call this a positive correlation. Further, each increase of 25 assists corresponds to a .02 increase in fielding percentage. Thus, the correlation or association between the two variables is perfect, and we could conclude, for example, that assists measures fielding percentage unerringly. We shall represent this association with the number +1.0. Next, consider these columns with percentage redefined as errors divided by chances:

Player	Assists	Pct.
A	400	.020
B	375	.040
C	350	.060
D	325	.080
E	300	.100

In this case, as assists increase by 25 points, error percentage decreases by .02. This is still a perfect correlation, but a negative one, represented as -1.0. Of course, fielding percentage and error percentage are opposites, so this result is no surprise. The point is that the sign of a correlation in this type of situation is the result of the way in which the columns are defined. Here, it is the degree of association, whether positive or negative, which is of interest. Finally, consider this last pair of columns:

Player	Assists	RBI
A	400	75
B	375	95
C	350	85
D	325	90
E	300	80

There is no association between assists and RBI, the latter does not measure the former, and the relationship can be represented by the number 0.

Thus, the correlation coefficient is a measure of the association between two variables, with a range of +1.0 to -1.0. In real life, correlations will never be perfect, unless two alternative measures of the same thing (such as fielding and error percentages) are mistakenly related. Any correlation more extreme than +/-.7 must be considered very strong, enough so that the two variables are indistinguishable for most practical purposes. Any correlation greater than +/-.3 can be quite useful.

The degree of usefulness can be determined by squaring the coefficient. This figure represents the extent to which the two variables overlap. Consider Diagrams 1 through 4 at the end of this paper. Imagine that each circle represents the "space" taken up by variables A and B. We call this space the variable's variance (amount of dispersion of scores around the variable's mean value), and the square of the correlation measures the amount of variance in each variable "accounted for" by the other [1]. Diagram 1 represents a correlation of 0; there is no overlap between the spaces of A and B. In Diagram 2, the variances totally overlap; the correlation is either + or -1.0, and its square is +1.0 in either case. Diagram 3 represents a correlation of about +/-.7; its square is about +.5, so each variable accounts for half of each other's variance. Diagram 4 represents a correlation of about +/-.3, which means an overlap of about .1. This may not seem like much, but with all the many factors which may affect any one variable (think of everything that can affect the outcome of a baseball game), a consistent relationship of this size can be quite meaningful.

Three more points before moving on. First, as in any statistical procedure, the correlation coefficient based on any sample is only an estimate of the degree to which the variables are actually associated. This estimate will always be in error. Luckily, the degree to which it is in error decreases as the size of the sample it is based on increases. Second, always remember that a correlation can only represent a linear relationship between variables. If, hypothetically, we expect number of errors to be greater for fielders with the least number of chances (showing lack of skill) and for fielders with the most number of chances (showing great aggressiveness) than for fielders with an intermediate number of chances, the correlation coefficient is unable to represent this curvilinear relationship. Third, high correlations between variables can be spurious. In other words, their relationship is artifically inflated due to the effect of a third variable. For example, the number of successful stolen bases and the number of times caught stealing

will be spuriously correlated due to the fact that they both increase as the number of attempted steals goes up. This fact underscores the importance of expressing statistics in percentage form.

The following is an example showing the use and computation of the most popular measure of association, the Pearson product-moment correlation. Imagine that we have reason to believe that a pitcher's earned run average is associated with the ratio between his total strikeouts and total walks in a given year. To explore this hypothesis, we choose a sample of ten National League pitchers and their 1982 statistics. The sample and statistics can be found in the accompanying table, with the columns marked X and Y. In order to calculate the correlation coefficient, we must (1) sum these columns, (2) square each individual SO/BB ratio and ERA and sum these squares (columns 3 and 4), and (3) multiply each pitcher's SO/BB ratio and ERA and sum these "cross-products". The results are then plugged into the following formula:

$$\frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

with N equalling sample size. Steps in the computation are shown in the table. The resulting correlation (-.82) is extremely high, and strongly supports the hypothesis (remember as ERA goes down, SO/BB should rise) [2].

Note exactly what we have done. We have shown that SO/BB is strongly associated with ERA. We have not shown that the former causes the latter; from the coefficient itself, we could just as soon conclude that the latter causes the former, or that a third variable (for example, the ratio of strikes and balls) causes both. Any causal claims must be based on arguments independent of statistics, although statistics can and should be used as evidence for them.

Let us continue, however, by claiming that SO/BB causes ERA, or at least is logically prior. We then wish to determine whether we can predict the pitcher's ERA using his SO/BB ratio. A second statistical technique closely related to correlation can be used for this prediction. This technique is called simple regression. The goal of simple regression is to deduce a formula which gives us the best possible estimate of a variable Y, given an associated variable X. Now, many possible interpretations of "best" can be proposed; the one universally adopted is that the mean of the squared differences between the equation and the actual scores is minimized. Consider Diagram 5 of our sample. The predictor is always on the horizontal axis, the to-be-predicted on the vertical. The points lie at the intersection of each of our sample's ERA and SO/BB. The long line represents that equation which comes closest to the points, given our definition of best estimate. This line can be represented by the following formula:

$$Y = a + bX$$

where Y equals ERA, X equals SO/BB, a equals the line's "intercept" (the value of Y where X equals 0), and b equals the line's "slope" (the number of units of change in Y, given a one unit change in X). The slope is the coefficient necessary for predicting Y from X. The lines between the points and the regression line are the errors in the prediction of scores from the equation. The means of their squared vertical distance is smaller than for any other line.

Let us now compute the slope of the regression line which best fits our data. The equation is as follows:

$$\frac{N \Sigma XY - (\Sigma X)(\Sigma Y)}{N \Sigma X^2 - (\Sigma X)^2}$$

Note the similarity between the equations for correlation and simple regression; the numerators are identical, the denominator of the latter is half of the former, additionally without the square root. That half retained must be for the predictor variable, not the predicted. Plugging in the numbers from the third step of the earlier computation:

$$\frac{-57.26}{59.35}$$

yields a slope of -.96. Thus, for every increase of 1 in SO/BB, ERA in our sample is best predicted as decreasing by .96 [3].

To use the slope for prediction, we adopt a method which "transforms" the scores so tthat the regression line moves vertically to the point where the intercept equals zero. We first calculate the mean for each variable (ERA = 3.80, SO/BB = 2.25). By definition, the regression line will always pass through the point representing the intersection of these means. We can then predict a case's Y score using the following formula:

mean of Y + (slope)(X for case - mean of X)

For predicting pitcher's ERA, this would be

mean ERA + (slope)(pitcher's SO/BB - mean SO/BB)

For S eve Rogers, we would plug in

3.80 + (-.96)(2.75 - 2.25) = 3.32

which is what his ERA would have been predicted to be. Thus, the equation overestimated Rogers' ERA by .92. For Steve Carlton, we replace 2.75 with his SO/BB ratio (3.32) and obtain 2.77. In this case, the equation underestimated Carlton's ERA by .33 [4].

Let us wrap this up by looking at correlation and regression from a different vantage point. If one has no other information about a variable Y, one can best predict an individual score on Y by using the variable's mean. If, however, one knows that Y is correlated with variable X, one can make a better prediction of the individual's Y by using information about the corresponding X score. Regression gives us a line representing the best

prediction of Y from X; correlation gives us the amount of scatter (variance) around that line, or the average amount of error in prediction. Squaring the correlation coefficient gives the amount of variance in Y accounted for by the use of the regression equation in predicting Y from X. Subtracting this square from 1.0 gives the amount of variance still unaccounted for; the "standard error of estimate". The main point is that the higher the correlation between X and Y, the greater the amount of variance in Y accounted for when using the regression line, and the lower the standard error. When correlation is perfect, using the regesssion line leads to no error. When correlation is zero, one gains no more advantage in predicting Y scores from X than by using the mean of Y for prediction.

The regression model can be extended in two important ways. First, an equation can be deduced predicting Y from more than one predictor (e.g., mean hits allowed per inning along with SO/BB), permitting us to cut down on the error variance while at the same time uncovering the relative impact of each predictor on Y. Second, nonlinear relationships, such as that imagined between fielder's chances and errors, can be examined. These issues will be among those described in later papers in this series.

FOOTNOTES

1 - The square root of a sample's variance, the "standard deviation" (SD), is a second measure of dispersion. It is used because its values are measured in the same units as the variable whose dispersion it measures, while the variance is measured in squared units. As squared units are often difficult to interpret, the standard deviation is usually a preferable measure. Further, the sizes of both variance and SD are functions of both the sample dispersion and the size of the unit of measure. Due to this fact, the SDs of two measures with "scales" of different size can not be compared without transforming one measure into the same scale as the other. This is true of most statistics, including means. Thus, batting average and slugging percentage (both percentages) are comparable, but batting average and number of hits (a natural number) are not.

2 - The sample chopsen is not representative; extreme cases were chosen to inflate the correlation for expository purposes. Using the entire population of N. L. pitchers who pitched 162 or more innings in 1982, the coefficient was actually -.49.

3 - The actual figure for 1982 N. L. pitchers (162+ innings) was -.44.

4 - The actual means were 3.58 (ERA) and 2.08 (SO/BB); thus Rogers' ERA is predicted as 3.29, Carlton's as 3.03.

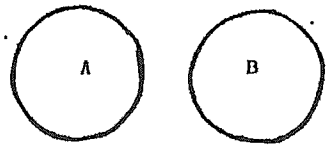


Diagram 1

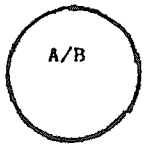


Diagram 2

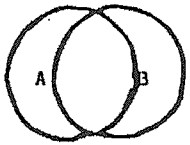


Diagram 3

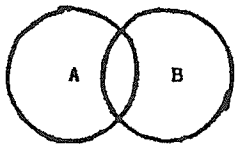
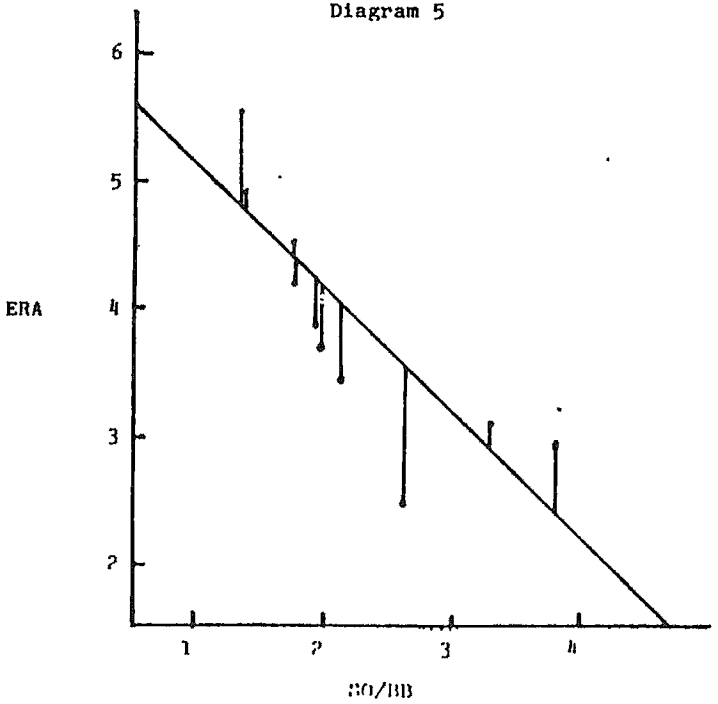


Diagram 4

Diagram 5



	X	Y	X^2	Y^2	XY
Pitcher	SO/BB	ERA	(SO/BB)	ERA	(SO/BB) (ERA)
Carlton	3.32	3.10	11.02	9.61	10.29
Knepper	1.80	4.45	3.24	19.80	8.01
P. Niekro	1.97	3.61	3.88	13.03	7.11
Rhoden	1.83	4.14	3.35	17.14	7.58
Rogers	2.75	2.40	7.56	5.76	6.60
Ruthven	1.95	3.79	3.80	14.36	7.39
Seaver	1.41	5.50	1.99	30.25	7.76
Soto	3.86	2.79	14.90	7.78	10.77
Walk	1.42	4.87	2.02	23.72	6.92
Welch	2.17	3.36	4.71	11.29	7.29
Total (Σ)	22.48	38.01	56.47	152.74	79.72

$$\frac{(10)(79.72) - (22.48)(38.01)}{\sqrt{(10)(56.47) - (22.48)^2}\sqrt{(10)(152.74) - (38.01)^2}} =$$

$$\frac{797.20 - 854.46}{\sqrt{564.70 - 505.35}\sqrt{1527.40 - 1444.76}} =$$

$$\frac{-57.26}{\sqrt{59.35}\sqrt{82.64}} = \frac{-57.26}{(7.70)(9.09)} = -0.82$$

MINOR LEAGUE EFFECTS ON MAJOR LEAGUE PITCHING PERFORMANCES

Terry Bohn

I set out to determine whether or not a major league starting pitcher's success, or lack of it, was influenced by his minor league career. I surveyed 101 major league pitchers, active through the 1981 season, who had at least 100 starts. I then recorded for each his minor league statistics in Triple A, below Triple A and for all of the minor leagues. I looked at two factors. First is the length of time spent in the minors as determined by innings pitched and number of decisions. The second factor is success in each level of the minors as determined by earned run average and winning percentage.

The method I used to determine the effect of the minors on major league success was to select three subgroups from this list of 101 pitchers. The first is a group of nine "super" pitchers who had a career winning percent of .575 or better and an ERA of 3.50 or under. This list is comprised of: Gura, Candelaria, Carlton, Guidry, John, McGregor, Palmer, Richard and Seaver. The second group is a list of the eight "poor" pitchers. Their criteria was a career major league winning % of under .450 and an ERA of over 4.00. These pitchers were: Abbott, Banister, Clancy, Espinosa, Falcone, Honeycutt, Jefferson and Kravec. The third group is the 84 "average" pitchers not included in the other two groups. I compiled the minor league stats for these groups and compared them with each other.

When the nine "super" pitchers were compared to the other two groups in the survey, some interesting conclusions arose. One factor was these nine pitchers consistency, as a group, in ERA and winning %, throughout their minor league careers.The second factor was the number of innings pitched in both Triple A and the other minor leagues. These super nine had more innings in Triple A and fewer in the lower minors than the other pitchers in both categories. They pitched an average of 250 Triple A innings and just 196 in the lower minors. The average pitchers were almost reversed in that they averaged 198 Triple A innings and 261 in the others. The differences were larger with the poor pitchers with only 137 Triple A innings and a total of 343 in the others.

When the eight poor pitchers were compared to the average pitchers in the survey, some other differences came up. First, these pitchers pitched more innings and had a poorer ERA in the lower minorsthan the other pitchers. They averaged 343 innings and a 3.53 ERA as compared to 261 innings and a 3.01 ERA for the average hurlers. They also didn't have as good a won-lost recoed in the lower minors as did the other pitchers. These eight had an average record of 20-22 while the rest of the pitchers were 18-13.

When the minor league statistics for the best and worst pitchers were compared to each other, two large differences arose. The poor pitchers had many more innings pitched and decisions in tje lower minors. They labored for an average of 343 innings as compared to just 196 for the super pitchers. The poor pitchers averaged 42 decisions (20-22) while the super pitchers had only 23 decisions (13-10) in the lower minors. One would think that if you give a pitcher an average of 343 innings in the lower minors and he compiles a mediocre 3.53 ERA, and an average of 42 decision of which he couldn't win half of them, you could predict he is not going to be a big winner in the majors.

I feel three conclusions can be drawn from this survey:

1- The number of innings pitched in various levels of the minor leagues affects major league pitching success.
2- The number of pitching decisions in various levels of the minor leagues affects major league pitching success.
3- Pitching consistency throughout their minor league careers breeds success in the majors.

I'm not sure of the significance of this survey, but I think it can serve as a guideline for fans to predict how their favorite minor leage pitchers will fare in the majors based on thier minor league pitching statistics.

SUBGROUP #1 NINE "SUPER" PITCHERS (Avg. per pitcher in parenthesis)

	IP	(AV)	ER	(AV)	ERA	W	(AV)	L	(AV)	W%
Below AAA	1767	196	644	71	3.28	119	13	91	10	.567
Triple A	2251	250	818	91	3.27	151	17	118	13	.561
Total Minors	4018	446	1462	162	3.27	270	30	209	23	.564

SUBGROUP #2 EIGHT "POOR" PITCHERS (Avg. per pitcher in parenthesis)

	IP	(AV)	ER	(AV)	ERA	W	(AV)	L	(AV)	W%
Below AAA	2742	343	1075	134	3.53	162	20	179	22	.475
Triple A	1098	137	410	51	3.36	75	9	48	6	.610
Total Minors	3840	480	1485	185	3.48	237	29	227	28	.511

SUBGROUP #3 84 "AVERAGE" PITCHERS (Avg. per pitcher in parenthesis)

	IP	(AV)	ER	(AV)	ERA	W	(AV)	L	(AV)	W%
Below AAA	21945	261	7347	87	3.01	1500	18	1072	13	.583
Triple A	16660	198	6711	80	3.63	1134	13	838	10	.575
Total Minors	38605	459	14058	167	3.28	2634	31	1910	23	.580

CHUCK WASELESKI ON THE IMPORTANCE OF GETTING THE LEADOFF BATTER ON BASE

Pitching coaches and baseball announcers (not neccesarily in that order) enjoy expounding on the importance of keeping the leadoff batter off the base paths. But just what is the relationship between the success of the first batter in an inning to his team's offensive production?

The table below contains a breakdown of each half-inning of all 162 Boston Red Sox games in 1983 according to whether or not the leadoff batter reached base. The Red Sox were chosen for this analysis because, even if their hitting, baserunning, and pitching are not typical, their 1983 season was certainly average. More importantly, it is difficult to gather data on 2900 Padre half-innings from here in Massachusetts.

The breakdown indicates that, when the leadoff batter reaches base safely, his team will score in the inning 50 percent of the time, or on average, every other inning. If he fails to reach base, his team can be expected to score 16.8 percent of the time, or about once every six innings. If runs are scored in an inning, the average scoring is 1.91 runs if the leadoff batter has reached base and 1.67 runs if he has not.

Another way of looking at the data indicates that the 35 percent of the time that the leadoff batter reaches safely (1015 times in 2899 innings) accounts for about 65 percent of all runs scored. And, of course, the 65 percent of the time that the leadoff batter is retired will account for about 35 percent of all runs.

It appears reasonable to assume that a successful leadoff batter is as good as a run. The average inning when the leadoff batter reaches base is 0.96 runs (970 runs in 1015 innings). Even when the 85 innings with leadoff home runs are discounted, the average inning only drops to 0.90 runs (836 runs in 930 innings). The average inning when the leadoff batter is retired is only 0.28 runs (527 runs in 1884 innings).

When the leadoff batter reaches safely:

1983	Total Inns.	Runs Scored 0	1	2	3	4	5	6	7	8	Total Runs	Run-Scoring Rate	Pct. of All Runs	Runs per Inning
Red Sox	501	262	113	69	29	19	6	2	0	1	464	47.7%	64.1%	0.93
Opponents	514	245	139	74	23	22	7	1	3	0	506	52.3%	65.5%	0.98
TOTAL	1015	507	252	143	52	41	13	3	3	1	970	50.0%	64.8%	0.96

When the leadoff batter is retired:

1983	Total Inns.	Runs Scored 0	1	2	3	4	5	6	7	8	Total Runs	Run-Scoring Rate	Pct. of All Runs	Runs per Inning
Red Sox	950	793	93	41	13	6	3	0	1	0	260	16.5%	35.9%	0.27
Opponents	934	775	91	44	15	5	3	0	0	1	267	17.0%	34.5%	0.29
TOTAL	1884	1568	184	85	28	11	6	0	1	1	527	16.8%	35.2%	0.28

Based on this information and on the table below, for example, Jerry Remy (batting average .275, slugging percentage .319, on-base percantage .320), the customary leadoff batter in the Red Sox order, would result in 76 first-inning runs (and 41 run-scoring first innings) in 162 games. Moving Wade Boggs into the first position in the batting order (with his .361 batting average, .486 slugging percentage, and .444 on-base percentage) would result in 91 first-inning runs and 49 run-scoring first innings in 162 games. Of course, this comparison avoids the question of where to place Remy in the batting order if it is preferable to hit Boggs leadoff. (As a Red Sox fan, may I suggest that Remy bat in some other team's lineup?)

Run production can be approximated based on how often the leadoff batter reaches base as follows:

Leadoff batter reaches safely	Approximate runs
0 times in 9 innings	2.52 runs
1 time	3.20
2 times	3.88
3 times	4.56
4 times	5.24
5 times	5.92
6 times	6.60
7 times	7.28
8 times	7.96
9 times	8.64

So, if the numbers, and moreover, common sense, show that the success of the first batter in an inning has a significant effect on run production, how important is it to get a leadoff runner on first base over to second? Specifically, is it important enough to attempt to steal?

Breakdown according to base reached:

	Total Inns.	Runs Scored 0	1	2	3	4	5	6	7	8	Total Runs	Run-Scoring Rate	Runs per Inning
First base	768	445	135	112	33	29	9	2	2	1	653	42.1%	0.85
Second base	149	58	47	23	12	6	3	0	0	0	168	61.1%	1.13
Third base	13	4	6	1	1	1	0	0	0	0	15	69.2%	1.15
Home run	85	0	64	7	6	5	1	1	1	0	134	100.0%	1.58
no-out steal of 2nd	38	14	16	8	0	0	0	0	0	0	32	63.2%	0.84
one-out steal	20	13	1	6	0	0	0	0	0	0	9	35.0%	0.45
Successful attempt	58	27	17	14	0	0	0	0	0	0	41	53.4%	0.71
Unsuccessful	25	22	2	0	0	1	0	0	0	0	6	12.0%	0.24
Attempted steal	83	49	19	14	0	1	0	0	0	0	47	41.0%	0.57
No attempt	685	396	116	98	33	28	9	2	2	1	602	42.2%	0.88

The table on the previous page breaks down successful leadoff batters according to the base they were occupying when the second batter in the inning came to the plate. The run-scoring rate with the first batter on first base is 42.1 percent (runs scored 323 times in 768 innings), and the run-scoring rate with the first batter on second base is 61.1 percent (91 times in 149 innings). Therefore, getting the runner to second seems attractive, right?

Maybe not. A successful, no-out steal of second base resulted in a run-scoring inning 63.2 percent of the time, slightly higher than when the runner was already at second base. The difference can probably be attributed to a handful of catcher errors on steal attempts, allowing the runner to go to third. But a caught stealing reduces the run-scoring rate to 12.0 percent, even lower than when the leadoff batter was retired.

Overall, it appears that, on average, attempting to steal second does not affect the frequency of run production. The run-scoring rate for innings when the leadoff batter attempts to steal second is virtually the same as when he does not. In fact, innings with steal attempts result in a slightly lower run-scoring rate (41.0 percent versus 42.2), even with 58 successful steals in 83 attempts (69.9 percent).

So it appears that, as a general strategy, as opposed to a situational strategy to be used when a single run is important enough to risk the consequences, a leadoff runner on first base is at least as valuable as a leadoff runner attempting to get to second base of his own accord. But, from the perspective of scoring multiple runs, it seems to be prefereable for the runner to let someone else move him around. When the leadoff batter reaches first and does not attempt to steal, 2 or more runs resulted 25.3 percent of the time (173 times in 685 innings). Steal attempts reduce this frequency to 18.1 percent (15 times in 83 innings). Only once in those 83 innings did more than two runs score, and that was after an unsuccessful attempt.

In summary, a leadoff runner on first base will score and will ignite multiple-run innings often enough to make a steal attempt too big a gamble in the majority of circumstances. In a late-inning situation where a single run can make a significant difference, the advantage of a stolen base may be worth the risk. However, the stolen base should be returned to its proper place--as a situational strategy, where the element of surprise (shock, in the case of the Red Sox) can increase the chances of success and contribute to a single, meaningful run.

BEST FIELDING 3RD BASEMEN continued

Brooks Robinson is rated as the best fielding third baseman but tops only the Misplay Index. No presently active third baseman is among the leaders on the Misplay Index, a result that is consistent with the data in Table 1. The overall occurrence of errors has dropped sharply over the years. The higher proportion of errors at third base, therefore, becomes an interesting question.

Who is the greatest fielding third baseman since 1925? At second base one name stood out and the answer was obvious. At third base the result is less clear. My vote goes for Brooks Robinson, as I have stated, but the data here leaves room for reasonable doubt. On with the debate!