Select SABR-L Posts on Analytical Topics

For a period of several years, principally around from 1998 to 2003, I was fairly active on SABR-L, the LISTSERV for the Society for American Baseball Research. Though dated, the research for my posts may still be of some general interest, so I have compiled many of them here. Most of these are as they appeared in the original posts. The edits consist mainly of removing names of other posters for confidentiality, fixing typos, and a comment here or there for context.

List of Posts

Comeback Percentage (24 Sep 2005)

I'm a little behind in responding to the discussion of the likelihood of the home team coming back heading into the bottom of the ninth inning. For those still interested, I looked at all games from 1980 through 1983 using Retrosheet and came up with the following percentages.

Runs Behind	Occurrences	ComeBack%
1	813	18.1%
2	729	7.0

3	616	3.2
4	455	1.3
5	335	0.6
6	216	0.0
>=7	409	0.0

Player Versatility (28 May 2005)

I found seven players who have played all 10 positions (including DH) over the course of their career. Halter and Lyons are the only two who have played at least two games at their least common position.

Name	P	C	1B	2B	3B	SS	LF	CF	RF	DH
Campaneris	1	1	1	36	76	2097	68	2	1	8
Halter	2	2	55	68	262	262	28	15	24	14
Lyons	2	4	115	118	229	6	59	237	43	23
Pecota	2	1	28	157	272	177	13	2	19	14
Rojas	1	7	2	1447	46	39	79	124	10	16
Sheldon	1	4	11	13	59	52	4	1	2	2
Tovar	1	1	1	215	227	77	378	471	205	90

Runs and Wins (17 Apr 2004)

The Pythagorean calculation of winning percentage can be algebraically derived from a fairly simple logistic regression model:

ln(w%/(1-w%)) = b x ln(RSpG/RApG)

Thus, the "best" fit exponent for the Pythagorean estimate of winning percentage is simply the coefficient calculated from the logistic regression model.

Running this regression for the team-seasons 1901 through 2002 gives b = 1.86 and an R^2 of .904.

Algebraically solving the regression equation for w% gives the "Pythagorean" formula:

w% = RSpG^1.86/(RSpG^1.86+RSpG^1.86), the Pythagorean winning percentage format.

Win Probability Added (22 Nov 2003)

I am just now catching up on several SABR-L posts and would like to comment on the "new statistic" referred to in the Business Week article a couple of weeks ago. In our book, Paths to Glory, Mark Armour and I formally introduced a nearly identical statistic and labeled it win probability added (WPA). As we discuss in the book, WPA estimates how each plate appearance (for batters) or batter faced (for pitchers) affected his team's probability of winning the game. The method is derived by establishing the probability of a team winning from every possible inning/base/out situation and calculating the change in this probability as a result of the plate appearance. By summing the individual probabilities up over the course of a season, one can generate estimate of the player's value that is situation-dependent, an interesting contrast to the traditional situation-independent sabermetric-type statistics.

As we discuss in our appendix, several others have previously performed

analysis using win probabilities, although the formulation of the actual statistic was slightly different. The research, however, typically suffered from one of two major limitations: the amount of game data was insufficient or was generated from simulations not actual games. In the early 1960s George Lindsay looked game strategies based on win probability data from 782 1958 AL, NL, and IL games. And as has been mentioned in other posts, the Mills Brothers authored a book titled Player Win Averages in 1970. More recently in the American Statistician, Jay Bennett looked at the individual players in the 1919 World Series using a modified version of Lindsay's tables.

I first presented an analysis of win probability added at the spring regional meeting of the Halsey Hall chapter in 2000. For that presentation I made a preliminary analysis of clutch hitting ability by comparing a player's WPA to a traditional situation-independent linear weights calculation. The theory being that a player who's WPA exceeds his linear weights value was a clutch hitter that year. When I compared this difference for batters over a couple of years, I found little correlation, suggesting that batters do not carryover a particular ability to hit well in high-impact situations from one year to the next.

Winning Percentage By Pitchers Per Game (11 May 2003)

In response to the question: Has anyone ever studied the winning percentage of teams who use only one pitcher in a game, vs. the winning percentage of teams who use two pitchers, three pitchers, four pitchers, five pitchers, etc.?

Running the Retrosheet data for the four seasons 1980 through 1983 gives the following result:

Pitchers/Gm	Wins	Total	WinPct
1	2189	2845	76.9%
2	2939	5054	58.2%
3	1617	4128	39.2%
4	675	2382	28.3%
5	229	793	28.9%
6	55	185	29.7%
7	6	34	17.6%
8	4	8	50.0%
9	1	1	100.0%
Total	7715	15430	50.0%

Clemens Run Support (29 Sep 2001)

Given the current Clemens discussion I thought it might be interesting to use this data to see if one could find any patterns in Clemens' run support. For example, if Clemens consistently received greater than average run support, one might attribute this to some intangible impact of Clemens on the offense. However, as the data shows, over the course of his career, Clemens received run support little different than what might be expected based on his teams.

	<------ Clemens ---------->			<------- Team ------------>			
Year	GS	Runs	Inn	R/G	Team	Runs	Inn	R/G
1984	20	123	175.0	6.33	BOS	810	1437.7	5.07

1985	15	61	131.0	4.19	BOS	800	1459.7	4.93
1986	33	201	289.0	6.26	BOS	794	1416.0	5.05
1987	36	200	324.0	5.56	BOS	842	1440.0	5.26
1988	35	153	314.0	4.39	BOS	813	1417.3	5.16
1989	35	159	315.7	4.53	BOS	774	1457.7	4.78
1990	31	131	271.7	4.34	BOS	699	1434.0	4.39
1991	35	157	312.3	4.52	BOS	731	1437.3	4.58
1992	32	128	283.7	4.06	BOS	599	1457.7	3.70
1993	29	94	258.3	3.27	BOS	686	1452.0	4.25
1994	24	93	213.0	3.93	BOS	552	1025.0	4.85
1995	23	130	205.7	5.69	BOS	791	1280.3	5.56
1996	34	150	314.3	4.29	BOS	928	1455.3	5.74
1997	34	154	297.0	4.67	TOR	654	1447.3	4.07
1998	33	151	302.3	4.50	TOR	816	1458.0	5.04
1999	30	134	259.3	4.65	NYA	900	1424.0	5.69
2000	32	154	279.7	4.96	NYA	871	1417.0	5.53
Total	511	2373	4546.0	4.70		13060	23916.3	4.91

Where:
GS = Clemens Starts
Inn = Offensive Innings; for Clemens, it reflects the total innings played by the team's offense in his starts.
R/G = Runs per 9 Inn

James' Win Shares & Garvey (18 Aug 2001)

I used the recent discussion over the Hall of Fame worthiness of Steve Garvey as an excuse to program Bill James' new Win Shares player evaluation methodology. As many of you know, at the Milwaukee convention, James presented a paper outlining the rational and methodology for his Win Shares system. The calculation of Win Shares is an extremely complex process. James provided a short form approach to the calculation which itself requires a fair number of interlocking formulas.

As to first baseman, I count just shy of 100 players who have played at least 1000 games at first base (data through 1997). The top 60 as ranked by their career Win Shares total are listed below. The Win Shares per 150 games played (all games, not just those at first) is also shown for comparison. By the career Win Shares method, Garvey ranks 30th of these players.

Rank	Name	WinShr	/150G	Rank	Name	WinShr	/150G
1	Musial, Stan	700	35	31	Fairly, Ron	268	17
2	Gehrig, Lou	502	35	32	McGriff, Fred	267	26
3	Anson, Cap	486	32	33	Cavarretta, P	264	20
4	Foxx, Jimmie	469	30	34	Cooper, Cecil	260	21
5	Murray, Eddie	451	23	35	Hrbek, Kent	256	22
6	McCovey, Willie	448	26	36	Camilli, Dolph	256	28
7	Carew, Rod	440	27	37	Kuhel, Joe	253	19
8	Brouthers, Dan	429	39	38	Chase, Hal	251	22
9	Connor, Roger	425	34	39	Palmeiro, R	249	23
10	Banks, Ernie	385	23	40	Chambliss, C	248	18
11	Mize, Johnny	377	30	41	Fournier, Jack	247	24
12	Perez, Tony	374	20	42	McInnis, Stuffy	246	17
13	Beckley, Jake	363	24	43	Scott, George	246	18
14	Cash, Norm	352	25	44	White, Bill	244	22
15	Sisler, George	341	25	45	May, Lee	244	18
16	Konetchy, Ed	317	24	46	York, Rudy	242	24

17	Tenney, Fred	316	24	47	Hargrove, Mike	241	22
18	Hernandez, K	311	23	48	Adcock, Joe	237	18
19	Powell, Boog	309	23	49	Mattingly, Don	237	23
20	Daubert, Jake	306	23	50	Buckner, Bill	234	16
21	Greenberg, Hank	304	33	51	Watson, Bob	233	20
22	Cepeda, Orlando	304	25	52	Pipp, Wally	233	19
23	Davis, Harry	304	26	53	Kluszewski, Ted	231	21
24	Bottomley, Jim	303	23	54	Mayberry, John	230	22
25	Vernon, Mickey	301	20	55	Burns, George	225	18
26	Hodges, Gil	300	22	56	Joyner, Wally	221	20
27	Clark, Will	294	27	57	McGwire, Mark	220	26
28	Judge, Joe	294	21	58	Blue, Lu	219	20
29	Terry, Bill	290	25	59	Tucker, Tommy	216	21
30	Garvey, Steve	289	19	60	Grimm, Charlie	216	15

Methodological accuracy check for those interested:

To test the accuracy of my short form approach calculation (WS-SF), I compared my calculation to a set of four seasons of Milwaukee ballplayers James calculated using the long method (WS-LM) and shown in Appendix IV of the handout.

Catcher	Year	WS-LM	WS-SF	First	Year	WS-J	WS-SF
Simmons	1982	19	18	Cooper	1982	29	29
Crand.	1957	13	11	Jaha	1993	15	15
Surhoff	1989	8	9	Brock	1989	10	12
Nilsson	1993	6	9	Torre	1957	10	10

Second	Year	WS-LM	WS-SF	Third	Year	WS-LM	WS-SF
Schoen.	1957	16	12	Mathews	1957	33	34
Gantner	1982	15	14	Molitor	1982	30	29
Gantner	1989	12	12	Molitor	1989	27	28
Spiers	1993	3	5	Surhoff	1993	16	15

Short	Year	WS-LM	WS-SF	Left	Year	WS-LM	WS-SF
Yount	1982	39	38	Vaughn	1993	22	25
Logan	1957	18	16	Oglivie	1982	21	21
Spiers	1989	9	9	Coving.	1957	15	15
Listach	1993	8	8	Braggs	1989	10	12

Center	Year	WS-LM	WS-SF	Right	Year	WS-LM	WS-SF
Yount	1989	34	34	Aaron	1957	37	37
Thomas	1982	25	24	Hamil.	1993	18	18
Bruton	1957	11	10	Deer	1989	12	13
Yount	1993	10	12	Moore	1982	8	7

According to James, the short form approach differs from the long method by 0 - 28% of the time, 1 - 37%, 2 - 16%, 3 - 8%, and 4+ - 11%. The test versus the above 32 players comes out very close to these values but without as many large errors: 0 - 28%, 1 - 38%, 2 - 25%, 3 - 6%, 4+ - 3%.

As a further check, I compared the career totals for several players that James included in his handout with my Win Shares estimate for the first basemen. For the most part, my short form approach seems accurate:

Player	WS-LM	WS-SF

Adcock	236	237
Cooper	241	260
Musial	604	700
Gehrig	489	502
Foxx	435	469
Murray	437	451

The short form approximation works well except in the case of Musial. I suspect here that much of the difference is due to the defensive outfield valuation. In the short form a player is given one Win Share per 48 games in the outfield regardless of L/C/R. I'd guess that in the long method, Win Shares is adjusted based on outfield position; thus, the short form would inflate the value of corner outfielders while underestimating that of centerfielders. I assume that Musial's long career in left led to much of his difference between the long and short form method.

Effect on scoring of new ballparks (14 May 2000)

For some time I have been thinking that one posters recent suggestion was one of the best ways to evaluate the effect of the new ballparks on the recent run increase:

> 4. The way to study this issue is to take a set of parks that haven't
> changed at all and determine if their park factors have changed over
> time. I don't know which parks haven't changed so I have no advice to
> offer. However, if you discovered that the Metrodome's park factor for
> runs was steady for 18 years without any corresponding change in the
> dimensions of the park, then it would indicate to me that ballpark turnover
> is not a factor in the recent offensive increase.

My cursory research suggests that the only unchanged ballpark in the AL over the 1990s is Fenway. The installation of the Stadium Club seats in 1989 was the last material change. In the NL several parks may have remained unchanged during the decade. I decided to use Dodger Stadium as neither the dimensions nor the seating capacity changed over the ten years in question.

The tables below look at the Fenway Park and Dodger Stadium park factors for runs and homeruns over the decade 1990 through 1999.

Park Factor for Runs

Dodger Stadium

Year	GH	GA	RH	RA	PF	3Yr
1990	81	81	674	739	0.91	
1991	81	81	622	608	1.02	
1992	81	81	558	626	0.89	0.94
1993	81	81	648	689	0.94	0.95
1994	55	59	439	602	0.78	0.87
1995	72	72	557	686	0.81	0.84
1996	81	81	597	758	0.79	0.79
1997	73	73	559	664	0.84	0.81
1998	76	73	575	665	0.83	0.82
1999	72	75	677	749	0.94	0.87

Fenway Park

Year	GH	GA	RH	RA	PF	3Yr
1990	81	81	712	651	1.09	
1991	81	81	752	691	1.09	

1992	81	81	669	599	1.12	1.10
1993	81	81	756	628	1.20	1.14
1994	64	51	673	500	1.07	1.13
1995	72	72	747	742	1.01	1.09
1996	81	81	981	868	1.13	1.07
1997	72	75	750	811	0.96	1.03
1998	73	73	731	704	1.04	1.04
1999	72	72	714	669	1.07	1.02

Park Factor for Home Runs

Dodger Stadium

Year	GH	GA	HRH	HRA	HRPF	3Yr
1990	81	81	127	139	0.91	
1991	81	81	103	101	1.02	
1992	81	81	59	95	0.62	0.85
1993	81	81	114	119	0.96	0.87
1994	55	59	96	109	0.94	0.84
1995	72	72	117	144	0.81	0.91
1996	81	81	111	164	0.68	0.81
1997	73	73	136	156	0.87	0.79
1998	76	73	135	140	0.93	0.82
1999	72	75	177	169	1.09	0.96

Fenway Park

Year	GH	GA	HRH	HRA	HRPF	3Yr
1990	81	81	105	93	1.13	
1991	81	81	145	128	1.13	
1992	81	81	91	100	0.91	1.06
1993	81	81	107	134	0.80	0.95
1994	64	51	135	105	1.02	0.91
1995	72	72	133	169	0.79	0.87
1996	81	81	214	180	1.19	1.00
1997	72	75	135	164	0.86	0.94
1998	73	73	161	182	0.88	0.98
1999	72	72	132	162	0.81	0.85

Where:
GH = Games at home
GA = Games away
RH = Runs at home
RA = Runs on road
PF = Park Factor
3yr = Three year moving average park factor
HRH = Homeruns at home
HRA = Homeruns on road
HPF = Homerun park factor

Summary: My take on these tables is that, while clearly not conclusive given the sample size, the reduction in the three year run park factor in Fenway Park suggests that the new ballparks are indeed having an effect in the AL. That is, Fenway appears to be moving from a park which increases runs scoring to one not much above the league average. The same conclusion cannot be drawn for the NL from the Dodger Stadium data. However, there does seem some evidence of a run park factor drop in early/mid 90s coinciding with the entrance of the Colorado franchise.

Methodological Note: Both run and homerun factors are calculated on a per game basis. The homerun factor probably ought to be calculated on a

per at bat basis, but I don't have that data going back to 1990. Given that the tables are comparing one year to another in the same park, I don't this is too serious a problem.

Aparicio & Leadoff Hitting (23 Apr 2000)

In one of the early Baseball Abstracts, Bill James introduced a formula for calculating how many runs a leadoff hitter ought to score based solely on that player's statistics (using H, 2B, 3B, HR, BB, SB & CS) in order to remove the influence of teammate hitting. I thought it might be interesting to apply it to the current Luis Aparicio discussion.

I guessed at the seasonal lead-off hitters for the 1956 - 1971 pennant winners using World Series box scores. The analysis does not include the player or two who had too few seasonal at bats or players from teams who had multiple leadoff hitters in the World Series.

James ranks leadoff hitters by what he calls efficiency: shown in the table below as "effic" and calculated as projected runs per 1000 outs. Aparicio falls in the bottom third of these leadoff hitters although it must be recognized that any analysis looking solely at pennant winners almost surely self-selects for better players.

Also, comparing him solely to the early 60's Yankee leadoff hitters, he looks pretty good. My understanding is that putting "bat-control" middle infielders at the leadoff spot was typical of that time (the early 60's). I'd be intersted in taking another look at the data if anyone has further information on who the leadoff hitters were.

Name	Year	Team_Lg	proR	actR	Outs	Effic
Buford, Don	1971	BAL_A	95	99	312	304
Buford, Don	1970	BAL_A	101	99	359	281
Buford, Don	1969	BAL_A	102	99	375	272
Gilliam, Jim	1956	BRO_N	109	102	407	268
Gilliam, Jim	1959	LA_N	97	91	387	251
Agee, Tommie	1969	NY_N	97	97	403	241
Brock, Lou	1967	STL_N	110	113	465	237
Brock, Lou	1968	STL_N	108	92	464	233
Mcauliffe, Dick	1968	DET_A	97	95	421	230
Versalles, Zoilo	1965	MIN_A	108	126	479	225
Schoendienst, Red	1957	MIL_N	60	56	269	223
Bauer, Hank	1957	NY_A	79	70	353	224
Wills, Maury	1963	LA_N	78	83	349	223
Bauer, Hank	1956	NY_A	90	96	407	221
Virdon, Bill	1960	PIT_N	65	60	299	217
Wills, Maury	1965	LA_N	93	92	433	215
Cash, Dave	1971	PIT_N	72	79	335	215
Flood, Curt	1964	STL_N	96	97	457	210
Bauer, Hank	1958	NY_A	68	62	329	207
Aparicio, Luis	1959	CHI_A	90	98	442	204
Linz, Phil	1964	NY_A	55	63	272	202
Aparicio, Luis	1966	BAL_A	89	97	466	191
Schoendienst, Red	1958	MIL_N	56	47	314	178
Wills, Maury	1966	LA_N	72	60	408	176
Kubek, Tony	1963	NY_A	70	72	412	170
Richardson, Bobby	1961	NY_A	78	80	482	162

where: proR = runs projected by the formula, and actR = atual runs scored.

One final note: In both years shown, Aparicio scores several more runs than projected by the formula. This does not hold true over the course of his career.

Innings with no base runners (22 Feb 2000)

Someone queried:

> Does anyone have a good number (either a hard number or an estimate) of the
> percenatge of innings (actually half innings) in which no runners reach base?

Using the Retrosheet files for the four years 1980 - 1983 indicates a total of 138,726 total half innings. Of those, no runners reached base in 42,471 or 30.6% of them.

Error ratios (31 Jan 2000)

Someone asked:

> I know that there have been several detailed fielding studies posted
> here in past months, but can anyone answer what I think is a
> straightforward question. What percentage of errors result in a runner
> reaching base as opposed advancing a runner who is already on base.

I had a chance to look at this question using the 1980 through 1983 Retrosheet files. Note that all totals below represent four years of data. I broke the data down by looking at the batter's destination under various error occurrences. The two main splits are between those situations in which the batter is charged with an At Bat and those in which he isn't.

Error--Batter Charged with Time At Bat

	No Hit		Single		Double		Triple	
Dest	Count	Pct	Count	Pct	Count	Pct	Count	Pct
DNRB	182	2	25	1	4	1	0	0
First	6235	80	342	16	0	0	0	0
Second	1242	16	1446	67	47	14	0	0
Third	118	2	336	16	254	78	0	0
Home	9	0	17	1	20	6	45	100

Error--No At Bat Charged

	Sac Hit		All Others	
Dest	Count	Pct	Count	Pct
DNRB	6	2	1899	95
First	233	72	92	5
Second	66	20	13	1
Third	15	5	4	0
Home	4	1	0	0

Note that "All Others" includes events such as stolen bases and sacrifice flies.

Totals

	No Hit		All	
Dest	Count	Pct	Count	Pct
DNRB	2087	21	2116	17
First	6560	65	6902	55
Second	1321	13	2814	22
Third	137	1	727	6
Home	13	0	95	0

Where Dest = Batter Destination on the Play; DNRB = Did Not Reach Base. Note that percents may not sum to 100 due to rounding.

Walks & Runs (9 Nov 1999)

I would like to weigh in in support of the idea that walks are relatively less valuable in a low scoring environment. My preliminary analysis suggests that walks have a lower linear weights type run value in a low scoring environment.

First I looked at all 1176 team-seasons between 1946 and 1998 and ran a multiple linear regression analysis for Batting Outs (Outs), 1B, 2B, 3B, HR, and Walks as the dependent variables versus Runs as the independent variable. The resulting values are similar to those in the various linear weights formulas:

Value of Event--All 1176 Team Seasons

Type	Out	1B	2B	3B	HR	Walks	R^2
ALL	-0.10	0.50	0.76	1.20	1.47	0.36	.94

I next sorted the seasons by run scoring to isolate the 100 highest and 100 lowest scoring team-seasons. [Note: High scoring teams ranged from 5.09 up to 6.67 runs per game and low scoring teams ranged from 3.57 down to 2.86] Recalculating the regression equation, unfortunately, results in values that don't seem valid, i.e. triples worth more than homers in the high scoring case, and outs as barely negative in the low scoring case. These anomalies remain even as the team-seasons are increased to two or three hundred. As an aside, this helps illustrate the importance of large sample sizes.

Value of Event--100 High & Low Scoring Only

Type	Out	1B	2B	3B	HR	Walks	R^2
High Scoring	-0.06	0.46	0.56	1.40	1.24	0.34	.93
Low Scoring	-0.02	0.30	0.48	0.90	1.04	0.26	.88

As one way to overcome the above anomalies, I reran the regression equasion for high and low scoring teams requiring all variables, except walks, to be within +/- 10% of the overall value:

Value of Event--100 High & Low Scoring Only--Only Walks Float

Type	Out	1B	2B	3B	HR	Walks	R^2
High Scoring	-0.11	0.45	0.68	1.31	1.33	0.58	.89
Low Scoring	-0.09	0.46	0.68	1.08	1.42	0.39	.81

I realize that other ways exist to view the question, but the above suggests, to me at least, that walks are in fact less valuable in a low run environment.

Additional Bases loaded by IBB discussion (28 Oct 1999)

I took another look at using the intentional walk to load the bases now that I don't have a 7 PM Central time deadline. Table 1 below summarizes the probability of a run scoring in the indicated base-out situations over all occurrences. Table 2 shows the probability of a run scoring given a bases-loaded situation generated from an intentional walk on the previous "play".

Table 1 Probability of Run Scoring--All Occurrences

	<----- AL --------->			<------- NL ------->		
Outs	0	1	2	0	1	2
1st & 3rd	.879	.667	.290	.846	.655	.279
2nd & 3rd	.874	.695	.270	.837	.660	.263
Bases Loaded	.874	.678	.334	.850	.657	.314

Table 2 Probability of Run Scoring--IBB to Load Bases

	<----- AL --------->			<------- NL ------->		
Outs	0	1	2	0	1	2
Bases Loaded	.855	.663	.283	.827	.660	.279

Comment: On average, in most bases loaded situations a run is more likely to score than in the 1st/3rd and 2nd/3rd situations. However, when bases are loaded through a base on balls the probability of a run scoring remains similar to the prior situation. This suggests that loading the bases with an IBB in a single run environment is not necessarily a poor strategy.

Bases loaded vs Runner on third only (26 Oct 1999)

A week or so ago in response to some of the strategies in the LCS, SABR-L had some discussion regarding the wisdom of intentionally walking the bases full in the final inning of a tie game. Expected run and probability tables, previously researched on SABR-L by Tom Ruane among others, is one way to examine the issue.

The table below looks at the probability of scoring at least one run from third when only third base is occupied as compared to the bases being loaded. All games over the period 1980 - 1983 are included.

Runner on Third only	<----- AL --------->			<------- NL ------->		
Outs	0	1	2	0	1	2
Times Run Scores	945	2351	1393	1011	2344	1259
Occurances	1114	3506	5047	1242	3602	4760
Probability	.848	.671	.276	.814	.651	.264
Bases Loaded						
Outs	0	1	2	0	1	2
Times Run Scores	974	1784	1051	708	1499	888
Occurances	1115	2631	3150	833	2280	2825
Probability	.873	.678	.334	.845	.657	.314

Although, on average, the probability of scoring is greater with the bases loaded, the probabilities are close enough that given the wide range of possible pitcher/batter matchups, it seems likely that in particular offence/defense situations the probability of scoring would be less when the bases are loaded. In other words, it may make sense to load the bases with intentional walks in certain situations.

Out of curiosity I then recalculated the above probability tables only in the situation in which the score was tied and the game was in the ninth inning or later.

Runner on Third only	<----- AL --------->			<------- NL ------->		
Outs	0	1	2	0	1	2
Times Run Scores	19	40	19	19	50	37
Occurances	27	78	100	26	84	120
Probability	.704	.513	.190	.731	.595	.308
Bases Loaded						
Outs	0	1	2	0	1	2
Times Run Scores	45	73	43	37	95	50
Occurances	53	120	127	49	148	167
Probability	.849	.608	.339	.755	.642	.299

Unfortunately the sample sizes are really too small to draw any definitive conclusions. In at least one instance, two out in the NL--although of no statistical significance--a run was more likely to score in the runner on third only situation.

Klein and Fictional Questions (11 Sept 1999)

Several days ago someone pointed out that the Klein home/road discussion really incorporates two distinct questions. I think this needs reiterating and expanding.

(1) The first question is one of value. If one accepts the principles of sabermetrics, then one can objectively determine, with more or less accuracy, how valuable a player was to his team. By calculating the run value of a player's statistics and comparing those to the context in which those statistics were accumulated, i.e. the runs per game, a player's worth can, again within some arguable degree of accuracy, be objectively known.

This value is empirically derived and verifiable at the team level. Whether due to some unique ability to take additional advantage of an already advantageous situation (e.g. Klein in the Baker Bowl) is irrelevant. Obviously, other factors go into a players overall value such as fielding and intangible contributions, but I don't think these are pertinent to the current debate.

(2) The question of what Klein would hit in a "statistically neutral" park falls into what I would call an alternative universe scenario. Like what would have happened if the South won the Civil War, or Hitler invaded Britain, or Kennedy hadn't been assassinated, the question of what Klein would have hit in a "statistically neutral" park is a fictional question. Fictional questions can be fun to debate, they can yield insight into a problem, and clearly some responses to the problem are more thought-out and reasonable than others but, and this is the key, they are not empirical problems. A fictional question is not empirically verifiable.

I have a certain sympathy for those who argue Klein's stats are what they are and we shouldn't try to change them. They don't need to be changed, they need to be put into context, and the two are not the

same. No matter how well reasoned, the fact remains that we really don't know what Klein would have hit if he played his career somewhere else because there is no way to prove one hypothesis versus another. It can be fun and enlightening to debate how Klein would have hit elsewhere and, again, some hypotheses are more likely than others, but it is not empirically provable one way or another.

More Klein Perspective (4 Sept 1999)

Several days ago I tried to put Klein's home and road performance in some additional context by examining his offensive winning percentage (a Bill James devised statistic that calculates a player's value by comparing his runs created per game with the actual runs scored in those games) both at home and away. I have tried to add additional perspective by looking at several contemporaries on the same basis over the same 1928 to 1933 time frame.

	RC	OW%	OffGms	OW	OL
Klein					
Home	608	.864	38.4	32.8	5.6
Road	299	.665	46.2	30.1	16.1
Total	907	.743	84.6	62.9	21.7
Ott					
Home	338	.750	42.2	31.0	11.1
Road	394	.788	42.3	33.0	9.3
Total	731	.758	84.5	64.0	20.5
Foxx					
Home	483	.857	37.1	31.4	5.7
Road	411	.768	44.6	33.6	11.0
Total	894	.795	81.7	65.0	16.7
Gehrig					
Home	430	.774	44.5	34.0	10.6
Road	579	.864	45.3	38.5	6.8
Total	1009	.806	89.8	72.4	17.4

The raw data is form TB1, and I decided not to relist the fairly lengthy column descriptions. Please contact me if you want the methodological specifics.

While I suppose it would be nice to be able to evaluate a few more NL contemporaries such as Waner, Terry, or Hartnett (for whom I have not been able to find home/away splits), I'd guess based on their full season statistics and Ott's home/road Offensive W/L pct that a legitimate case could be made for Klein as the best player in the NL over the 1929 to 1933 time frame. On the other hand, he falls short of the Gehrig, Ruth, or Hornsby peaks.

Chuck Klein--Home & Away (28 Aug 1999)

The recent discussion surrounding the significance of Chuck Klein's home statistics in evaluating his ability brings to mind the Sandy Koufax debate. Like then, I still don't understand the logic of deeply discounting a players home statistics. Assuming teams play half their games at home, a player's ability to help his team win at home is no

more or less valuable than on the road. As long as his achievements are placed in the proper context, a player should be evaluated by the whole of his accomplishments.

Home/Road Breakdown For Chuck Klein's Initial Years as a Phillie
(Statistics calculated from data in TB I)
Home

Year	RC	RC/G	R_BB	G_BB	R/G_BB	OW%
1928	40	11.39	881	75	5.87	.790
1929	97	13.35	1083	76	7.13	.778
1930	124	17.27	1187	77	7.71	.834
1931	105	15.18	814	76	5.36	.889
1932	127	16.55	936	77	6.08	.881
1933	116	19.47	792	72	5.50	.926
Total	608	15.82	5693	453	6.28	.864

Road

Year	RC	RC/G	R_Lg	G_Lg	R/G_Lg	OW%
1928	18	6.41	4888	539	4.53	.666
1929	72	8.67	5526	540	5.12	.742
1930	72	8.55	5838	541	5.40	.715
1931	40	4.65	4723	542	4.36	.533
1932	50	5.55	4754	541	4.39	.615
1933	47	5.20	4116	546	3.77	.655
Total	299	6.48	29845	3249	4.59	.665

Where:
RC = Runs Created using basic formula: ((H + W)* TB) /(AB + W)
RC/G = Runs Created per game (game = 25.5 outs)
R_BB = Runs scored in Phillie home games, both teams
G_BB = Phillie home games
R/G_BB = The run context for Phillie home games, i.e.Runs per game in Phillie home games (R_BB / G_BB*2)
OW% = Bill James calculation of offensive winning percentage:
RC/G^2 / (RC/G^2 + R_BB^2)
R_Lg = Total runs in NL less runs scored in Phillie home games
G_Lg = Total games in NL less Phillie home games
R/G_Lg = The run context for Phillie away games, i.e. runs per game in NL games exclusive of Phillie home games

Despite the fact that games in the Baker Bowl averaged 1 to 2 more runs per game than elsewhere, Klein's performance was extremely valuable as his runs created rose significantly above the average runs scored in those home games. I don't see how one's home performance is any less valid in terms of helping his team win, than one's road performance. The reasons for his excellence at home, while interesting, are irrelevant when evaluating, within the proper context, the level of that excellence.

As an aside, Klein wasn't all that bad on the road. Only once did he have an offensive winning percentage below .600.

Hornsby's Fielding (22 Aug 1999)

Rogers Hornsby's fielding ability remains a topic on which little consensus seems to be developing. After reading his biography by Charles Alexander and looking at the statistics I have a hard time viewing his

defensive ability, at least until late in his career, as much worse than average. In fact, early in his career he appears to have been quite a good fielder.

Hornsby came up as a nineteen year old shortstop in September 1915, played creditably in the field and in 1916 won the starting shortstop job in spring training. For the last half of the season, however, manager Miller Huggins moved Hornsby over to third. Huggins moved Hornsby back to short for the 1917 season where he appeared to have a pretty good season in the field: he led the league in double plays, and had a better than average fielding percentage and range factor. By 1919 Branch Rickey was managing the team, and he decided to move Hornsby to second. Despite working out at second in spring training, Hornsby spent most of the year at third and only played 25 games at second.

Although he apparently had decent fielding stats at short, I'm not arguing he was a major league shortstop--two of the best baseball minds of the time both attempted to move him. But, the very fact that he could play shortstop at least adequately suggests he had some defensive ability.

Playing second base as a 24 to 26 year old from 1920 through 1922, Hornsby had a range factor above the league average every year and twice led in double plays. In May 1923, he tore his knee while making a throw, came back too quickly, and was in a cast for two weeks. It obviously cost him some range that year, and I would argue it could very well have affected the rest of his career in the field.

In 1926 his real health problems started. In May, in a collision at second base he displaced two vertebrae which caused back problems for the rest of his career. Additionally, a long term battle with painful carbuncles began shortly thereafter. A heel bruise in August 1928 and its secondary effects which caused him to miss most of the 1930 season robbed Hornsby of most of whatever range he had left.

While no defensive whiz, he appears slightly above average as a youngster. Later as injuries took their toll, he fell below average, but not egregiously so except in years in which he tried to play through injuries such as 1923 or 1931. All in all, I have trouble reconciling his fielding career as I understand it with some of the more negative analysis of his fielding.

Year	Pos	G	Rng	LgRng	Pct	LgPct	DP
1915	SS	18	5.22	5.21	.922	.931	
1916	SS	45	4.93	5.17	.910	.916	
1916	3B	83	3.08	3.00	.928	.938	
1917	SS	144	5.52	5.15	.939	.931	X
1918	SS	109	5.89	5.64	.933	.934	
1919	3B	72	3.11	3.10	.933	.946	
	(1B-5 Games; 2B-25; SS-37)						
1920	2B	149	5.82	5.43	.962	.963	X
1921	2B	142	5.51	5.45	.969	.962	
1922	2B	154	5.66	5.58	.967	.961	X
1923	2B	96	4.95	5.57	.962	.959	
1924	2B	143	5.72	5.62	.965	.961	
1925	2B	136	5.17	5.61	.954	.967	
1926	2B	134	5.06	5.29	.962	.965	

1927	2B	155	5.68	5.63	.972	.966	
1928	2B	140	5.32	5.74	.973	.970	
1929	2B	156	5.34	5.47	.973	.968	X
1930	INJ						
1931	2B	69	4.52	5.22	.961	.964	
Tot	2B	1561	5.36	5.51	.965	.964	
Tot	SS	356	5.49	5.32	.932	.933	
Tot	3B	192	2.98	2.99	.924	.942	

Notes on Table: X in DP column means lead league in double plays.

Sources: Statistics from STATS All-Time Major League Handbook. Biographical information from Rogers Hornsby, a Biography, by Charles Alexander.

More on Top 2Bs (21 Aug 1999)

Given the recent debates on Ryne Sandberg and the top second basemen, I thought it might be interesting to see how they ranked by Offensive Wins Above Replacement (OWAR)-- the sabermetric methodology detailed and used in the Bill James Historical Abstract. The OWAR is based on a replacement player W/L Pct of .350. To compare with TPR from Total Baseball, I have also included Offensive Wins Above Average (.500). I have listed below the top 25 who had more than 1000 games at 2B. Note the statistics are from the whole career (although seasons before 1894 are excluded), not only those games at second.

Name	AB	OW	OL	Pct	OWAR	OWAAv	G at 2B
Collins, Eddie	9949	209	75	.735	109	67	2650
Morgan, Joe	9277	187	79	.702	94	54	2527
Hornsby, Rogers	8173	167	46	.785	92	61	1561
Lajoie, Nap	9589	181	77	.700	90	52	1949
Carew, Rod	9315	168	86	.660	79	41	1130
Gehringer, Charlie	8860	156	84	.650	72	36	2206
Sandberg, Ryne	8385	144	92	.611	61	26	1995
Frisch, Frankie	9112	148	103	.590	60	23	1762
Whitaker, Lou	8570	144	100	.591	59	22	2308
Grich, Bobby	6890	127	75	.629	56	26	1765
Doyle, Larry	6509	122	66	.648	56	28	1719
Herman, Billy	7707	128	85	.601	53	22	1636
Biggio, Craig	5750	108	52	.675	52	28	1054
Fox, Nellie	9232	140	122	.533	48	9	2295
Alomar, Roberto	6048	107	62	.632	48	22	1526
Pratt, Del	6826	118	82	.589	48	18	1688
Randolph, Willie	8018	128	105	.549	46	11	2068
Doerr, Bobby	7093	116	85	.577	46	15	1852
Evers, Johnny	6137	111	75	.596	46	18	1686
Lopes, Davey	6354	111	76	.594	46	18	1416
Gordon, Joe	5707	102	63	.619	44	20	1519
Myer, Buddy	7038	112	86	.565	43	13	1340
Lazzeri, Tony	6297	104	75	.581	41	14	1456
Huggins, Miller	5558	99	69	.590	40	15	1530
Schoendienst, Red	8479	123	113	.521	40	5	1657

My own subjective take on this list is that the top four are clearly the

best, followed by Gehringer and Carew. The next ten to twelve can be placed in almost any order depending upon (1) how one analyzes their defense and (2) how one views peak versus career value.

Baserunner advancement the value of batting events (19 Jul 1999)

The recent debate on measuring the value of baserunner advancement centers on whether or not advancement of baserunners by hitters is random and unrelated to any specific batter characteristic or whether systematic differences exist between players. Unfortunately, this interesting topic has become confused with whether or not clutch hitting exists. I suspect significant systematic differences between hitters exist that have nothing to do with clutch hitting.

I remember Bill James arguing that a double by Rickey Henderson likely has less baseruuenr advancement potential than a double by a lumbering slugger (I forget who he used as an example) because Henderson's speed would get him some doubles on balls that might not be that well hit; conversely a slow player would really need to hit the ball to get to second. Other differences between players that affect baserunner advancement may also exist: maybe there's a differential between right handed hitters and left handed hitters; maybe there's a difference between flyball and groundball hitters--it wouldn't surprise me if groundball hitters were more likely to advance runners on an out but less likely on a hit.

The value difference between a single and a walk in the runs created formulas rests mainly in the baserunner advancement differential--the batter ends up on first in both cases. But is it the case that no variation exists in the singles between players? Did Willie Wilson's singles, which probably included a number of infield hits, have the same advancement potential as Willie Aikens' singles, one of the slowest players in baseball, who was probably hitting singles off the right field fence? Using the 1980 - 1983 Retrosheet data I took a look at this question.

The table below compares the baserunner advancement on singles by Willie Wilson and Willie Aikens for 1980 - 1983.

Batter	Year	BR_1B	BA_1B	BA_1B/S	BR_2B	BA_2B	BA_2B/S
Aikens, W	1980	43	71	1.65	27	50	1.85
Aikens, W	1981	21	36	1.71	9	17	1.89
Aikens, W	1982	25	37	1.48	21	37	1.76
Aikens, W	1983	22	27	1.23	13	23	1.77
Aikens, W	Total	111	171	1.54	70	127	1.81
Wilson, W	1980	48	60	1.25	32	47	1.47
Wilson, W	1981	22	31	1.41	15	25	1.67
Wilson, W	1982	40	53	1.33	29	43	1.48
Wilson, W	1983	18	22	1.22	22	32	1.45
Wilson, W	Total	128	166	1.30	98	147	1.50

Where:
BR_1B = Number of times a single was hit with a runner on first base.
BA_1B = Total number of base advanced by those runners. For example, a runner ending up at third equals two bases advanced. Note that for the

few times a runner was thrown out on the basepaths I credited one base advanced.
BA_1B/S = BR_1B / BA_1B; that is, bases advanced from first per single.
BR_2B = Number of times a single was hit with a runner on second base.
BA_2B = Total number of bases advanced by those runners.
BA_2B/S = BR_2B / BA_2B; that is, bases advanced from second per single.

I'm not claiming the above offers concrete proof of differences between hitters. After all, Wilson never got to hit with himself on base, and the table only looks at two players over four years.

But I think the data strongly supports the idea that different players have systematic differences in value of the same hitting event such as a single. In all eight comparisons, i.e. each year and from both first and second, runners average a greater advance on an Aikens' single than on one by Wilson. It may only be of marginal importance, but this would suggest a single by Aikens might be more valuable than one by Wilson [Although this probably should lead to a base running comparison as Wilson as a baserunner is surely more valuable than Aikens].

My point is that different hitters may have different overall values from the same batting event based on systematic differences between them. This has nothing to do with clutch hitting. I think, though, that it is a mistake to simply dismiss these potential systematic differences without significantly more research--which I hope to get to in the not to distant future.

Fielding opportunities by position based on Pitcher Hand (17 Jul 1999)

The recent discussion over fielding opportunities by position based on the whether the pitcher throws right or left led me again to the 1980 - 1983 Retrosheet files and some intriguing obsrvations.

Table 1 -- First "Out" By Position

Pos	<--LHP---->		<---RHP---->	
	"Outs"	Pct	"Outs"	Pct
1	5420	5.8%	12144	5.7%
2	1711	1.8%	4025	1.9%
3	7337	7.9%	22091	10.3%
4	14348	15.4%	39037	18.2%
5	14027	15.0%	26518	12.4%
6	18232	19.5%	37957	17.7%
7	9006	9.6%	23598	11.0%
8	12841	13.8%	28630	13.3%
9	10466	11.2%	20636	9.6%
Total	93388	100.0%	214636	100.0%

Where the first out is defined as the sum of (1) the first putout if there was no strikeout or infield assist on the play and (2) the first assist if the first putout was made by an infielder. All table data consists of four year totals.

The infield results are intuitive: with a left-hander pitching 24.3% of the first outs are made by the 1B or 2B. When a righty takes the mound, this increases to 28.5%. Conversely, a third-baseman's out percentage with a lefty pitcher is 15.0% and only 12.4% when a righty pitches. I assume the main reason for the differential is that managers work to get

additional right handed batters to face left handed pitchers and vice versa.

The outfield results on the other hand seem counter-intuitive. A greater proportion of outs to the left fielder against righties?

My first thought was that I must have erred in generating the data. Hence, I ran a table for which position fielded the ball for all batted balls including hits.

Table 2 -- Total Balls in Play Fielded By Posiiton

Pos	<--LHP---->		<---RHP---->	
	Fielded	Pct	Fielded	Pct
1	7082	5.3%	15535	5.1%
2	2007	1.5%	4682	1.5%
3	8227	6.1%	24543	8.0%
4	15233	11.3%	41490	13.5%
5	16325	12.1%	30703	10.0%
6	20270	15.1%	42038	13.7%
7	22436	16.7%	49201	16.0%
8	23574	17.5%	53779	17.5%
9	19270	14.3%	44691	14.6%
Total	134424	100.0%	306662	100.0%

Again, the infield results are as expected, but now the outfield results exhibit little difference between left and right handed pitchers.

To complete the verification of Table 1, I recalculated the Table 2 for hits only (net of homeruns).

Table 3 -- Hits Fielded By Posiiton

Pos	<--LHP---->		<---RHP---->	
	Fielded	Pct	Fielded	Pct
1	501	1.4%	1088	1.3%
2	42	0.1%	81	0.1%
3	294	0.8%	696	0.9%
4	572	1.6%	1594	2.0%
5	1225	3.4%	2358	2.9%
6	1289	3.6%	2679	3.3%
7	13397	37.0%	25195	30.9%
8	10322	28.5%	24108	29.6%
9	8544	23.6%	23666	29.1%
Total	36186	100.0%	81465	100.0%

Notes: Hits plus outs fall slightly short of the totals (Table 2) due to non-covered events such as errors. The hit total will not total the the four year hit totals because not all hits have a "fielded by" position identified.

Subtracting Table 3 from Table 2 yields results similar to Table 1. In other words, the seemingly backward outfield values for outs are generated from two separate fields in the database: which position had the first "out", and which position fielded the ball. I must say the results still seem so counter-intuitive that I won't completely rule out making an error, but the calculation methodology is the same for all positions.

I have only been fiddling with this for a day or two and, as yet, have no good idea why hits follow the expected pattern and outfield outs fall the opposite way. Does anyone have any thoughts on this distribution, or suggestions on other ways to look at the question?

Lack of AL Lefties (9 Jul 1999)

Twins manager Tom Kelly was recently asked the most surprising things he's seen this year in the American League (outside of events surrounding his own team). He responded (1) the poor play of the Baltimore Orioles and (2) the lack of left-handed starting pitchers. As to the latter, Kelly noted that his switch hitters rarely hit from the right side. Below is a table of games started by lefties over the past 10 years.

Year	Total GS	GS By Lefties	Pct
1990	2266	699	30.8
1991	2268	619	27.3
1992	2268	601	26.5
1993	2268	681	30.0
1994	1594	482	30.2
1995	2025	622	30.7
1996	2266	651	28.7
1997	2264	643	28.4
1998	2268	608	26.8
1999YTD	1179	230	19.5

Note that the 1999 YTD numbers may not be perfect: I think players who moved from the AL to NL during the season (only a few, if any) may not show up in these numbers, additionally I entered the handedness manually. Nevertheless, Kelly's observation seems accurate based on the evidence.

More Correlation (8 Jul 1999)

Someone asked for a correlation of OBP plus Isolated Power?

Here's the Expansion era correlations including OBP plus Isolated Power (OPI).

Statistic	Correlation to runs scored	
Avg	.828	
OBP	.881	
SLG	.924	
OPS	.959	
OxS	.960	[OBP x SLG]
OIP	.943	

Correlation of batting stats to runs by decade (5 Jul 1999)

I thought it might be interesting to take the suggestion and look at correlation between run scoring and batting statistics by decade.

<--------- Batting Statistic ----->

Decade	Avg	OBP	SLG	OPS	OxS
1870s	.916	.882	.898	.904	.912
1880s	.863	.809	.858	.875	.867
1890s	.754	.872	.844	.874	.879
1893-1899 Only	.865	.922	.915	.940	.940
1900s	.930	.901	.917	.952	.953
1910s	.842	.887	.862	.924	.928
1920s	.869	.909	.915	.963	.964
1930s	.833	.930	.919	.956	.959
1940s	.843	.886	.898	.945	.947
1950s	.824	.852	.842	.937	.943
1960s	.818	.909	.923	.962	.964
1970s	.831	.894	.904	.955	.956
1980s	.760	.841	.890	.934	.936
1990s	.823	.863	.921	.953	.954

Technical notes: Run scoring was defined as runs per game, i.e. team runs scored divided by team games. Decades defined as years ending in 0 through 9 (not the more technically accurate 1 through 0). Decade of 1870s includes NL team seasons only; the NA is not included.

Conclusion: When the NL elected to use batting average as its benchmark statistic in the 1870s, this seemed a reasonable choice. BA was in fact more highly correlated to run scoring than any other batting stat during these initial seasons (it must be noted that NL in the 1870s consisted of only 28 team seasons). BA pretty much held its own through 1910 (although the 1890s appear as a bit of an anomaly), and by the thirties clearly was not as well correlated as the others. I also find it interesting that over the last four decades, slugging average is consistently higher correlated to run scoring than on base percent. I would have thought the opposite before I started this analysis.

More OPS vs BA (4 Jul 1999)

In my haste to correct and embarrassment over posting incorrect correlation data for the expansion era, I failed to accurately express my thoughts regarding the relative value of batting average.

BA is obviously less correlated to runs scored than OPS, but, and this is my point, not enough to render it valueless. The .83 correlation between BA and scoring indicates a strong relationship between BA and scoring. If the correlation was, say, .30 then, by all means lets throw it out, but the correlation is strong enough that it imparts meaningful information.

Additionally BA is widely understood and intuitive. Hits per At Bat (no, I don't want to get back into that calculation) is directly derived from on field events. Something like OPS which is the sum of two rate stats has no intrinsic meaning, e.g. an OPS of 1.015 does not directly tie to 1.015 anything. My opinion is that if we're going to start with mathematical combinations of rate statistics which then lose any direct relationship to the underlying events, we ought to use a measure, e.g. Runs Created or Linear Weights, which is expressed in runs or wins. I just don't see any statistic that isn't expressed as some specific on field event--such as H/AB, TB/AB, times on base per PA, runs, wins, etc.--ever gaining widespread acceptance.

Correlation to runs scored--Constant over time? (3 Jul 1999)

Looking at the correlation between batting statistics and run scoring made me curious over whether the relationships have remained constant over time. I thus ran the correlations between batting statistics and runs scored for the period 1901 - 1919 as compared to the modern expansion era (61 - 98).

Statistic	---Correlation to runs scored---	
	Expansion Era	Dead Ball Era
Avg	.83	.88
OBP	.88	.83
SLG	.92	.86
OPS	.96	.89
OxS	.96	.90 [OBP x SLG]

I find these results fascinating. In the dead ball era, batting average was more highly correlated to run scoring than either on base percent or slugging average. In fact, it was almost as highly correlated to runs scored as the other two combined.

Correlation to runs scored—Correction (3 Jul 1999)

After recalculating the correlations I get the following table:

Statistic	Correlation to runs scored
Avg	.83
OBP	.88
SLG	.92
OPS	.96
OxS	.96 [OBP x SLG]

First baseman defense (26 Mar 1999)

Several months ago on this List we had a discussion regarding the fielding of Bill Buckner in particular and first basemen in general. One of the knocks against using assists per game as a measure of Buckner's fielding ability was the suspicion that his slowness afoot led to more 3 to 1 assists and fewer unassisted putouts, thus inflating his ability as measured by assists.

As I now have the computing power to load all the publicly available Retrosheet files (1980 through 1983) I thought I'd take deeper look at the data. I apologize in advance if this topic is too out of date.

The analysis looks at three types of first baseman assists and putouts:

3ua: Plays on which the first baseman made the putout and no assist was credited.
3to1: Plays on which the first baseman made the first assist and the pitcher the first putout.
3toX: Plays on which the first baseman made the first assist and the putout was made by a fielder other than the pitcher.

The table below indicates the MLB statistics for the starters (defined as >=700 [400 in 1981] defensive innings) along with the individual totals for Buckner, Garvey, Hernandez, and Murray for comparison. All

data is per 9 innings.

		3ua	3to1	3toX	Total
Max	1980	1.89	0.62	0.39	
Avg	1980	1.42	0.39	0.24	2.05
Min	1980	1.00	0.21	0.14	
Max	1981	1.75	0.60	0.38	
Avg	1981	1.36	0.43	0.25	2.03
Min	1981	1.03	0.28	0.08	
Max	1982	1.91	0.70	0.35	
Avg	1982	1.46	0.43	0.24	2.13
Min	1982	1.10	0.21	0.14	
Max	1983	1.83	0.72	0.39	
Avg	1983	1.40	0.46	0.25	2.11
Min	1983	1.10	0.27	0.15	
4 Year Avg		1.41	0.43	0.24	2.09
Buckner	1980	1.19	0.62	0.17	1.98
Buckner	1981	1.26	0.48	0.33	2.07
Buckner	1982	1.64	0.70	0.25	2.58
Buckner	1983	1.41	0.72	0.39	2.52
Buckner	Avg	1.41	0.64	0.29	2.35
Garvey	1980	1.34	0.51	0.14	1.99
Garvey	1981	1.69	0.38	0.11	2.17
Garvey	1982	1.50	0.44	0.20	2.14
Garvey	1983	1.50	0.29	0.15	1.93
Garvey	Avg	1.49	0.42	0.15	2.06
Her'dez	1980	1.66	0.31	0.38	2.35
Her'dez	1981	1.52	0.42	0.35	2.29
Her'dez	1982	1.56	0.48	0.35	2.39
Her'dez	1983	1.61	0.65	0.37	2.64
Her'dez	Avg	1.59	0.47	0.37	2.43
Murray	1980	1.24	0.26	0.18	1.67
Murray	1981	1.25	0.57	0.19	2.01
Murray	1982	1.30	0.39	0.23	1.92
Murray	1983	1.56	0.52	0.18	2.26
Murray	Avg	1.34	0.42	0.19	1.96

Conclusion: Buckner's advantage in assists does not carry over into unassisted putouts where his figures are average. The tables above also further support Hernandez' fielding reputation.

Let me caveat all this by saying I realize a dichotomy exists on this List between those who have a high degree of confidence in range factors as a measure of fielding ability and those who feel that they are much too dependent on the pitching staff and other aspects that do not even out the number of balls hit to any area over a season. I do not mean to reopen that debate; only present the above data for those who might be interested.

Home/Road & Koufax (10 Mar 1999)

As one reader suggested, using only road stats to evaluate Sandy Koufax ignores Koufax' real advantage over opposing pitchers in Dodger Stadium. Looking only at Koufax' road performance appears to diminish the glitter of his stats. But his tremendous home record relative to the rest of the league in Dodger Stadium is not irrelevant.

Relative to the opposition, it doesn't matter where a pitcher saves his team runs. A pitcher uniquely better than the competition in his home park is no more or less valuable in terms of saving runs and winning baseball games over the course of a season than one's relative value in the rest of the league's parks.

The table below summarizes Koufax' record in terms of runs allowed per game relative to the rest of the NL pitchers in both his home and road parks.

Year	1960	1961
Runs/G at Memorial Stadium Except Koufax Pitching	4.57	4.72
Koufax R/G At MS	5.46	5.02
Koufax % of Others at MS	126%	106%
R/G at 9 Other NL Parks	4.19	4.49
Koufax R/G at 9 other NL Parks	3.28	3.25
Koufax % of Lg other NL Pks	78%	72%
Combined Home/Road	97%	89%

Year	1962	1963	1964	1965	1966
Runs/G at Dodger Stadium Except Koufax Pitching	4.36	3.57	3.41	3.11	3.30
Koufax R/G At DS	2.05	1.44	0.97	1.63	1.81
Koufax % of Others at DS	47%	40%	28%	53%	55%
R/G at 9 Other NL Parks	4.51	3.87	4.11	4.15	4.20
Koufax R/G at 9 other NL Parks	4.16	2.42	3.34	3.22	2.34
Koufax % of Lg other NL Pks	92%	63%	81%	78%	56%
Combined Home/Road	67%	52%	51%	65%	55%

Speed Scores and Reaching on Errors (20 Jan 1999)

The last issue of "By the Numbers" included my analysis of the influence of team speed on opposition errors. I intended to update that analysis for a future issue by looking at individual players. Given the recent discussion on this board on what influences batters reaching base on error, I thought I'd share the analysis here on the SABR-L.

I examined the effect of speed on reaching base on error by looking at all 219 batters who had at least 300 at bats in 1980 using the 1980 Retrosheet files. As a proxy for player speed I used the Bill James

Speed Score (excluding the fielding range factor). The table below summarizes the data by breaking it into quintiles. I also looked at correlations based on the 219 individual player records.

Quint	SpdScr	RBoE	Opp	/Opp	RBoE.GB	Opp.GB	/Opp.GB
1	3.67	259	11498	.0225	252	5751	.0438
2	4.84	277	11370	.0244	264	5620	.0470
3	5.49	276	11943	.0231	267	6073	.0440
4	6.82	300	12559	.0239	289	6519	.0443
5	10.95	370	13690	.0270	359	7596	.0473

Quint: Speed Score Quintile
SpdScr: Top Bill James Speed Score of the Quintile
RBoE: Reached base on error
Opp: Opportunities--At bats that put the ball in play but did not result in a hit.
/Opp: RBoE per opportunity
RBoE.GB: Reached base on error on ground ball
Opp.GB: Opportunities on ground balls only
/Opp.GB: RBoE.GB per GB opportunity

Summary: A very modest correlation of .14 exists between reaching base on error per opportunity and a player's speed score. Because nearly all times first base is reached on error comes on a ground ball, much of this correlation can be explained by the relationship between speed scores and the propensity to hit ground balls (correlation = .30), i.e. faster players tend to be ground ball hitters. When one looks at times reached base on error as a percentage of non-base hit ground balls the correlation between speed and reaching base on error evaporates to basically zero (correlation = .04).

One poster suggested looking at whether a batter is hitting left or right.

Quint	Bats	RBoE	Opp	/Opp	RBoE.GB	Opp.GB	/Opp.GB
4	L	133	6143	.0217	129	3113	.0414
	R	167	6416	.0260	160	3406	.0470
5	L	156	6139	.0254	152	3571	.0426
	R	214	7551	.0283	207	4025	.0514
All	L	538	25291	.0213	522	13501	.0387
	R	944	35769	.0264	909	18058	.0503

Bats: The side of the plate the at bat took place from. Switch hitter at bats are therefore split.
All other columns are the same as above.

Summary: Here it seems fairly conclusive that those hitting from the right side are much more likely to reach base on error than those hitting from the left side. I would suppose righties are more likely to reach base on error because they are more likely to hit to the SS or 3B who typically make more errors (mainly due to the longer throw) than first of second basemen.

Expansion and Gould Thesis (30 Aug 1998)

One poster commented in regards to Gould's thesis:

> Gould's thesis about competitive evolution and variation in baseball
> is true, I am sure. But I am sure only because I am sure that
> variation increases with each expansion. (Of course Rod Carew had
> that big year in '77, McCovey in '69, Killebrew in '61 . . . .)
> A few examples do not prove anything, but only because they do not
> measure league-wide variation. If Schell and others find no increase
> in league-wide variation in '61, '69, '77, and '93, then the Gould
> thesis is in big trouble as a principal expanation of an important
> phenomenon in baseball.

To recap: Gould hypothesized that the declining variation in batting averages over the course of baseball history evidenced an increase in the overall level of play. He noted that this declining variation plateaued around 1940.

While I intuitively believe in the merits of Gould thesis, I previously on this list raised two potential problem areas. The suggestion above notes a third area to investigate.

The table below calculates the standard deviation of batting average for all players with at least 300 at bats (the criteria Gould uses in his 1983 Vanity Fair article) over the past 45 years.

Year	StdDev	Year	StdDev	Year	StdDev
1953	.030	1968	.028	1983	.028
1954	.033	*1969	.029*	1984	.030
1955	.029	1970	.032	1985	.026
1956	.031	1971	.030	1986	.028
1957	.031	1972	.030	1987	.030
1958	.029	1973	.029	1988	.027
1959	.029	1974	.028	1989	.028
1960	.024	1975	.031	1990	.027
*1961	.028*	1976	.030	1991	.029
*1962	.027*	*1977	.029*	1992	.028
1963	.028	1978	.025	*1993	.030*
1964	.028	1979	.028	1994	.033
1965	.028	1980	.029	1995	.030
1966	.027	1981	.029	1996	.029
1967	.032	1982	.025	1997	.028

Average standard deviation 45 year period: .029
Average standard deviation 1961, 1962, 1969, 1977 & 1993: .029

In other words, no increase in league-wide variation exists in the expansion years.

Gould's Extinction of .400 hitting (2 Aug 1998)

Given the recent discussion on the evolution of the baseball talent level, I thought it might be interesting to review the well published comments of SABR member and famous paleontologist Stephen Jay Gould. He has written extensively over the last several years how the "extinction of .400 hitting really measures the general improvement of play". In his recent book _Full House_ he concludes several convincing and eloquent chapters as follows:

"In a quick summary of a long and detailed argument, symmetrically

shrinking variation in batting averages must record general improvement of play (including hitting, of course) for two reasons--the first (expressed in terms of the history of institutions) because systems manned by best performers in competition, and working under the same rules through time, slowly discover optimal procedures and reduce their variation as all personnel learn and master the best ways; the second (expressed in terms of performers and human limits) because the mean moves toward the right wall [i.e. human limits], thus leaving less room for the spread of variation. Hitting .400 is not a *thing* [Gould's emphasis], but the right tail [of the normal distribution] of the full house for variation in batting averages. As variation shrinks because general play improves, .400 hitting disappears as a consequence of increasing excellence in play."

While not necessarily disagreeing with his analysis I have two difficulties.

(1) He notes that the increase in ability (i.e. this decrease in variation) reached a plateau about 1940. Now this doesn't quite make intuitive sense to me. As has been previously discussed on this List, the post-war integration of baseball and the increased acceptance of Latin American ballplayers very likely raised the overall level of play. I would suggest this improvement in overall ability of the best players (as measured by variation in batting average) is masked by an offsetting increase in variation from expansion. Thus an analysis based on variation in batting average looses much of its validity after WWII.

(2) As the table below indicates, pitchers do not (or only extremely weakly) follow this trend. It seems an explanation that addresses only half the batter/pitcher equation should not be viewed as universal.

Annual Standard Deviation of Opponents Batting Average and Batter Batting Average (averaged over decades)

Year	Pitchers	Batters
1890s	.022	.038
1900s	.022	.034
1910s	.022	.034
1920s	.021	.034
1930s	.021	.031
1940s	.020	.029
1950s	.021	.028
1960s	.022	.027
1970s	.022	.028
1980s	.024	.028
1990s	.022	.028

Source: Sean Lahman's Database
Notes: Regulars only; defined as 150 IP or 400 AB
1890s start in 1893; 1990s through 1997

Note that my hitter calculations differ slightly from Gould's (probably due to the definition of a regular) and show the variation attaining a minimum in the 1960s. This, of course, slightly weakens my argument in (1) but I still think a case could be made for the factors mentioned above applying past the 1960s.

I'm curious as to how Gould's thoughts on this matter are regarded by baseball researchers. Is his theory widely accepted? Has there been debate on his theories over the past few years? I would enjoy any comments on his analysis.

Garvey's Fielding (29 Mar 1998)

In a debate a poster asked who else should have won the Gold Glove from 1974 - 1977 if Steve Garvey did not deserve it has continued an intersting debate. Looking at who played first base at that time in the NL indicates the GG voters may not have been as far off as has been suggested. Below are the NL first basemen other than Garvey with at least 130 games at first.

1974: Joe Torrre, Willie Montanez, John Milner, Tony Perez, and Lee May. Only May and Perez over 140 games.
1975: Mike Jorgenson, Perez, and Montanez. Only Montanez over 140 games.
1976: Perez, Bob Watson, and Mike Ivie. Only Watson over 140 games.
1977: Keith Hernandez, Perez, Dan Driessen, Watson, Willie McCovey, and Montanez. Four were over 140 games.

With the exception of 1977 when Hernandez probably deserved the GG, it's not obvious that Garvey was the wrong choice.

For comparison sake, I ran the Bill James Defensive W/L Pct (which incorporates Fielding Pct, Assists/G, estimated 3-6-3 DPs, and errors by shortstops and third basemen) for the six 1974 players with over 130 Games:
Torre: 59%, Perez: 57%, Milner: 57%, Garvey: 51%, May: 47%, Montanez: 41%.

I realize the above in no way proves Garvey was a good player; his awards may have been deserved only because no great fielding first baseman was playing.

But is it possible that Garvey has become so universally regared by us sabermetricians as overrated that he might possibly now be underrated?

Mazeroski's Defense (22 Mar 1998)

The recent debate on Mazerowski's HOF credentials and by inference the value of his defense has been fun. One poster suggested the value of Maz' defense over and above an average second baseman (Fielding Runs in Total Baseball speak) at 51 runs per 162 games over the course of his career. I don't find it possible that Maz's defense (or any player's) is anywhere near that valuable.

1. Based on Pete Palmer's batting linear weights, the suggested defensive calculation values PO and A above the league average at .72 runs and DP above league average at .50 runs. Even accounting for the adjustment to avoid double counting DPs, these values intuitively seem too high. Palmer himself does not try to translate from batting linear weights to fielding; he uses much lower values in calculating FR: .20 for DP and PO and .40 for A.

2. TB IV gives Maz the second highest career fielding rating of all

time, 362 FR (behind Lajoie). Even this rating of 27 FR/162G seems high as no other exclusively 20th century player has more than 265 FR. For comparison the ranking is 1.53x Ozzie Smith's 10th place ranking (236 FR) and 2.29x Rabbit Maranville's 34th place ranking (158 FR). And both played in more games than Maz.

3. In the 1983 Baseball Abstract, Bill James analyzes the claim that Ozzie Smith saves 100 runs per season. In the essay he concludes Smith saves maybe 25 to 35 runs a season. Part of his logic assumes that of the plays Smith makes beyond an average shortstop, 80% have the value of a negative hit at 1/3 to 2/5 of a run each.

4. James created a defensive won/lost ranking system. I ran the system for 2B/SS/3B seasons from 1957 (the first Gold Glove Award) through the present. Each player is rated on 40/30/20/10 point scale; I had to adjust out the 10 point scale (team Defensive efficiency rating) due to data limitations. The following are the seven players who calculate to 10 wins above average (.500):

Name	W	L	Pct	+.500
B. Robinson	73	31	.698	21
Maz	67	31	.685	18
O. Smith	164	100	.609	18
L. Aparicio	93	74	.558	10
B. Grich	55	35	.609	10
B. Bell	75	47	.628	10
C. Boyer	36	16	.695	10

At the 10 runs per win rule of thumb, this give Maz 180 (13.5/162G) defensive runs above average. Note this ranking system probably underrates the greatness of Maz as it places a ceiling on the total runs a second baseman is responsible for.

5. For Maz to be worth 51 FR/162G over the course of his career requires him to be the greatest fielder of all time by a large margin. Yet in twelve seasons of over 125 G Maz won "only" eight gold gloves. Now I realize the gold glove award isn't perfect but it certainly is relevant as to what his contemporaries felt at the time. Bench for example caught over 120 games 12 times: 68-77 and 79. He won the gold glove every year except the last. B. Robinson had 143+ games at third from 60-75--he won the GG every year. K. Hernandez played 90+ games in the 13 years 76-88, he won GG the last 11 of those. Other players such as Ozzie Smith also had marginally better GG award voting records.

6. Finally, sorry if this is too long but I find the analysis of baseball players fascinating, to believe Maz is worth 51 FR/162G is put his defensive value on par with the era's greatest pitchers. For example, by TB IV, Koufax from 62-66 averaged 49 pitching runs (runs above what an average pitcher would prevent) per season and three times led the league. From 62-69 Marichal averaged 32 PR/yr. Gibson's 1968 season was worth 63 PR--in other words Maz' *average* season with the glove was worth nearly as much Gibson's 1968 season.

The above points can obviously be expanded and debated. But to quote Bogey--as well as I remember--from the end of the Maltese Falcon: "some may be unimportant, I won't argue that, but look at the number of them"

Bill Mazeroski was clearly a great defensive player, maybe even the

greatest of all time, be he wasn't saving 51 runs per season more than an average second baseman.