WHEN WILL

COOPERSTOWN

OPEN ITS

DOORS TO

ALIENS?

BASEBALL Analyst

No. 22 2/86

December 6, 1985

Dear Friends:

At the end of the 1985 calendar year I will be leaving the Texas Rangers. This is a decision that began to form back in August and I informed the Rangers of my intentions in mid-September. I discovered that I was no longer enjoying my work, a very disturbing thought as there was no other passion in my life. It is my desire to generate new passions in my life and perhaps stoke up the old flame for this work. Sometimes I feel all I need is to get back to more original research and I will begin to enjoy this endeavor again. Whatever the case, I knew I could not remain in this job as it stands now and be happy. I lose a little financial security, but personally I cannot afford to be unhappy.

At present I do not see myself severing myself from baseball or sabermetrics at all. I will do one book on baseball and probably a second. I research and write most of the brief radio spots, "A Page from Baseball's Past." Those of you who live around Texas, Detroit, Kansas City, Baltimore, Toronto and Milwaukee will have a good chance to hear some of these in the 1986 season. I encourage you to write the radio station and tell them how you love the show, that it inspires you to dash out and buy the sponsor's product.

I also am involved in designing a baseball game that will be based on the ongoing season like a Rotisserie League but with much greater realism. Plan on it for the 1987 season. I have also been approached about some work for a consultant to two clubs, one in each league. And after a while I may return to work with a major league organization with the job designed more to suit my sanity.

Still, I may just vanish from the baseball world after a year or so and find my feet and happiness on a totally different path. Whatever, I would like to express my thanks for all the support and assistance I have been given by the sabermetric community. For those who need to get in touch, you might write me at 2201 Violet Lane, Arlington, Texas 76006; Telephone (817) 649-0293. If you find you are getting no response, try me at 1943 Strait Tow Road, Crystal, Michigan 48818.

The one thing I ask is freedom from questions about how I got my job with the Rangers, what did I do for the Rangers, and what you could do to get a job in baseball. I have answered every one of those letters for four years now, and it is time to stop. I will cover that in my second book, okay?

In closing, I would like to share a favorite poem which has been on my bulletin board for a couple of years since it first appeared in The Christian Science Monitor. Seen in this present light it may explain why I am compelled to make these changes. But I offer it to you my friends in the spirit that it first hit me. As unlikely as it may seem, in this poem is the spirit at the heart of all scientific inquiry. Can you see it? This is something I have tired to bring to my work in sabermetrics, and it is what I wish to leave with you.

Craig R. Wright

The transposition

Before I used to strive
for intellectual acquisition—
a thin man bearing stones.
But now I have thrown out
small reason
and the thoughts of the troubled mind.
Now I am happy to confess
that I do not know.
And being ignorant, thus,
I am like water,
transparent as the spirit
nurturing the phenomenal world.

Sometimes I am like water
swirling in quiet pools,
in hidden recesses.
And sometimes like the meandering
water of rivers,
feeding the lands.
Sometimes like the raindrops
and the congealed water
of the mountain snow.
And sometimes the tempest is my person,
raging with the waters
of the mighty sea.

What water I am is decided solely
by the breath that stirs me—
the God I love.
And being, in profound plainness
given to the simplicity
of guiding motion,
I am myself now and find
that I have cast the stones
of yesterday
into the stream
of perpetual tomorrow.

David Sparenburg

EDITOR: Jim Baker

FOUNDER/PUBLISHER: Bill James

BUSINESS MANAGER: Susie McCarthy

STATISTICAL CONSULTANT: John Borkowski

EDITORIAL OFFICE

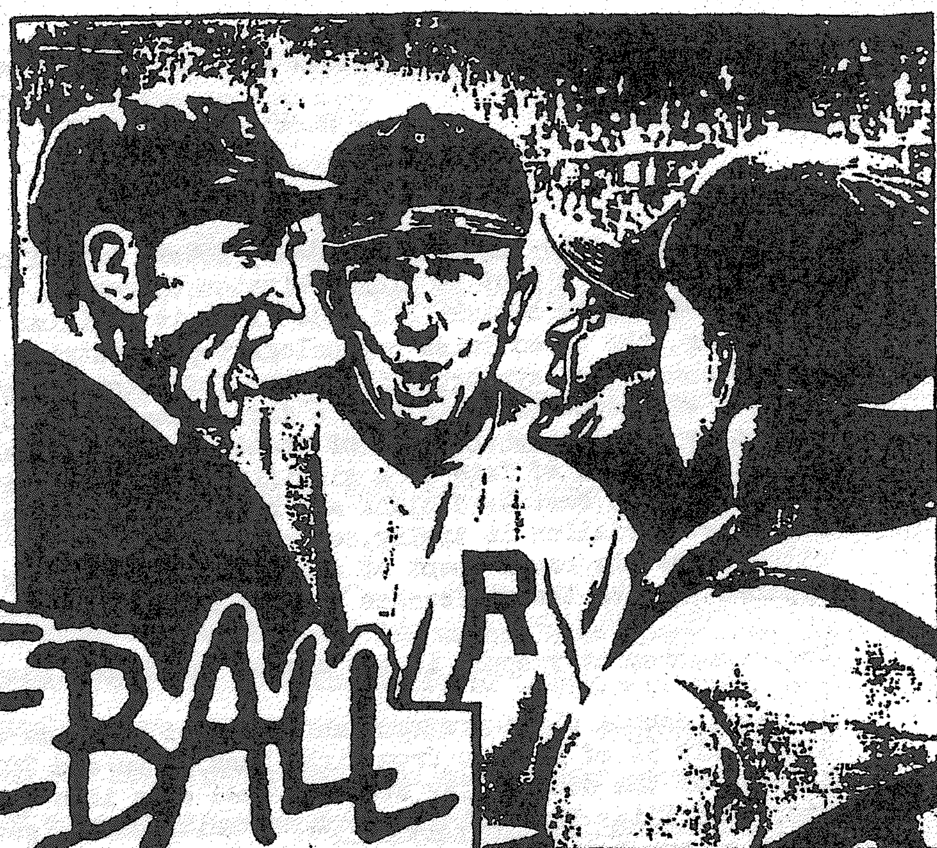
20 Lincoln Street
Jersey City, New Jersey 07307
(201)656-3892

BUSINESS OFFICE

Box 171
Winchester, Kansas 66097

SUBSCRIPTION RATE

$12.00 per annum



BASEBALL Analyst

February 1986

Issue No. 22

IN THIS ISSUE: Obviously the preceding page represents a major event in the world of sabermetrics. Craig was the one man who was on the "inside." Now that's changed. I suppose many of you will ask of Craig's decision, "how can he give up such a neat job with a major league baseball team for nothing but uncertainty?--I'd quit my dull, everyday job to take his place in a minute!" Well, keep this in mind; nothing is what it seems to be from the outside. Sometimes you have to get right in the middle of what you might consider perfection to see it for what it really is. Don't judge Craig by your own experience. Remember, he's been there. Only he really knows what he wants. The ANALYST wishes Craig the best in whatever he decides his new passion is. Sanity is everything.

Of course now that Craig will no longer be submitting, the rest of you are going to have to pick up the slack. He has done a lion's share of the work over the past eight issues and now its somebody else's turn. Craig's article OBSERVATIONS ON THE STOLEN BASE appears on page 5. We also have Dick O'Brien on page 4 with SYNCHRONICITY AND THE DOUBLE PLAY. Bill James is on page 10 with PERIPHERAL AND IDENTIFYING CHARACTERISTICS OF GROUND-BALL AND FLY-BALL TYPE PITCHERS. Mark Pankin adds RUN ESTIMATION BY GAME-LINE ASSEMBLY on page 17. A pretty heavy line-up folks. Next time we'll have pieces from Mike Kopf and Jess Boronico.

ABOUT THE COVER: Another Joe Schwind original. Joe is especially incensed about the failure of the Hall of Fame to recognize many of the intergalactic players who have added so much to the game. Many have been subjects of his drawings.

SYNCHRONICITY AND THE DOUBLE PLAY

Dick O'Brien

In the movie, "The Eagle Has Landed", Robert Duvall plays a German colonel who masterminds a plot to kidnap Winston Churchill. During the early moments of the film he asks his aide if he is familiar with the works of Dr. Carl Jung, the noted Swiss psychologist-psychiatrist. He goes on to mention Jung's concept of synchronicity - the fortuitous coincidence of either propitious or inopportune timing - a factor that can and frequently does either ab ort or catalyze a sequence of events that can impact history or personal lives. The idea of fate as a significant factor is inherent within the concept. All the best laid plans may be either consummated or negated by the slightest variation in timing, either by the planner or his opponent. In either case the requirement of accurately determining the reaction of your opponent may be the difference between success and failure.

Does synchronicity apply in the use of the double play strategy?

At best, double plays are successful 24% of the time when attempted.* Since only 5% of runners from first base score and 20% from second base, the use of the double play strategy must come into question. (See pp. 224, 266, 1985 Elias Baseball Analyst). Traditionally, getting two outs for the price of one batted ball has mezmerised baseball mavens into thinking its reward is worth whatever its cost. If the standard managerial cliche is, "We just take one game at a time," why don't they believe in taking one out at a time? Because the appeal of two for the price of one is greater than one at a time. It's simply a matter of greed. Proportionately, more second-basemen have been put out of action attempting the double play than going for a single out. And while a double play usually requires the ball being handled by three players rather than two, the possibility of mishandling the ball is further increased.

One suspects that far too often, managers are obsessed irrationally, by the presence of a runner on first base to the exclusion of their primary concern; getting the batter to make an out according to his pattern of predictability. Getting two outs for the price of one is strictly speaking, a bonus situation to be sought, but only as a residual effect of the primary effort - getting one man out for certain. How many times have we seen a shortstop throw to second before he has the ball or the secondbaseman bungle a pivot in his haste, drop a throw or throw it into the middle of next week. If any one or more of these occur we end up with no one out and runners at first and second.

If we face a contact, high-average spray hitter or an upper-cutting, low-average slugger in a double play situation, are we going to play for one out or two? Employing the usual defensive strategy for each of the above type

* A 24% double play success rate is a ballpark guestimate. The contention is based on 2197 DPs in the AL with 9107 runners on first base with less than two outs; 1729 DPs in the NL with 7117 runners. Admittedly, not all double plays occur with runners on first, but - well, you understand the thought process. Also, one can strike out into a double play, but this alternative does nothing to vitiate the thrust of the argument.

(Continued on page 20)

CRAIG WRIGHT:

OBSERVATIONS ON THE STOLEN BASE

DATA: Steal attempts as well as straight pick-offs in the American and National League games covered by Project Scoresheet in 1984 (thanks guys).

TREATMENT: Leagues studied seperately, broken down between RHP and LHP as well as breaking into thirds the steal attempts according to non-intentional walk averages of the pitchers. Also noted were each pitcher's base runners in the form of hits plus hit-by-pitches plus walks.

In regard to the differences between LHP and RHP we now have clear evidence that the LHP is tougher to steal on. In both the National League and American League the steal success rate against RHP was about 12% greater than against LHP. The number of steal attempts per base runner was also much higher against the LHP. Overall there were about 23% more steal attempts against the RHP versus the LHP.

What was especially interesting to note is that the edge enjoyed by the LHP is 99% their ability to pick runners off after they break for second (pick-off caught stealings). In both leagues the steal success rate without the pick-off caught stealings was nearly identical between RHP and LHP.

STEAL SUCCESS RATE WITHOUT PICK-OFFS

	AL	NL
LHP	.6875	.7481
RHP	.6882	.7545

That surprised me as I assumed the runners would be a little slower taking off against a lefty as they would realize the lefty passes that point of absolute committment to home plate over first base a fraction of a second later than the right-hander. That apparently is not so. They seem to generally run off the same cue regardless of whether a righty or lefty which gives no edge either way to the catcher but does lead to more than a few runners taking off too early against LHP. The pick-off caught stealing rate per base runner is <u>127%</u> greater for LHP compared to RHP (10.9 per 1000 base runners versus 4.8).

Do base stealers run less against lefties because they are likely to be less successful or because it is harder to get a safe jump off a lefty? I am not sure how we will answer that one, but I expect it is a combination of the two. If one needs to be more selective against a lefty, that would mean the hit-and-run play and ordered steal attempt should be more dangerous with a lefty.

What factors influence the frequency of stolen base attempts? Certainly the leagues make a difference due to a number of factors including (1) higher steal success rates on the artificial turf surfaces which are more prevalent in the National League (2) the defensive speed that is especially attractive in artificial turf is easily converted to offensive speed (3) the lack of the DH and fewer offensive parks in the National Leagues creates a higher relative value for the stolen base and a reduced cost for the caught stealing.

Certainly, the steal success rate influences how often the steals are attempted, not just out of the fear of being caught but that a lower steal success rate would also suggest that generally the pitcher is mechanically tougher to get a jump on. Even within the groups of lefties and righties the attempt rate seems to follow inversely the success rate.

As to the theory that stealers run more against wild pitchers because there are more opportunities (pitches) that continues to appear to be no more than a minor influence. The theory appears stronger than it is because the wilder pitchers tend to throw harder and generally have more exaggerated mechanics that are easier to get a jump on. In both leagues, for both righties and lefties, the pitchers with the higher walk averages had higher steal success rates against them to go along with the increased attempts per base runner.

Besides that overlap of poor control and higher success rates, the "opportunity theory" sinks even further when you look at the exceptions in the high and low walk groups. Invariably you will find the exception in the good control group is someone like Tom Seaver who in 1984, threw fewer pitches per batter as his non-intentional walk average was only 2.21 and he averaged fewer strike outs

than the league norm. Yet Tom has a very slow delivery, allowed a high success rate (.867) and was run on a lot, almost 50% more per base runner than the American League average for his low-walk group.

The exception in the high-walk group would be someone like Steve McCatty who was wild in 1984, averaging 61% more non-intentional walks per inning than Seaver, but McCatty, who does not throw hard, has simple quick mechanics and very little leg kick from the stretch. The result is only 11 steals in 18 tries and only 65 attempts per 1000 base runners compared to Seaver's 108.

There is one more note of interest from this study. The question has arisen more than once whether straight pick-offs should count as a caught stealing. After looking at this data I would say, no, that it should simply remain a category credited to the pitcher but not against the runner as a thwarted steal attempt. With the National League attempting 104 steals per 1000 base runners and the American League just 69, I would argue for a caught stealing designation for straight pick-off only if the National League averaged significantly more straight pick-offs per base runner. Actually they averaged slightly less than the American League (4.69 and 4.91). If anything, it would appear that the base stealers are being very slightly penalized with the present pick-off caught stealing definition. Still, it is a just definition as we see that there is indeed a jump in the number of pick-off caught stealing attempts in the running league, 9.32 per 1000 runners versus only 4.58 in the AL. The magnitude of that gap also suggests to me that part of the aggressiveness of the National League base stealer involves an anticipation of the best pitch to go on rather than actually waiting for the pitcher to commit to the plate. It could also be that the pitchers have developed better moves to first, but I still lean more to the former theory as the best move in the world cannot pick off a runner who knows the pitcher's motion —— unless the runner tries to "cheat" by anticipating the pitcher's committment to the plate.

AMERICAN LEAGUE

Left-Handed Pitchers:

Walk Avgs.	SB-CS (Not Pick-Off)	P.O.CS	Straight Pick-Off	Steal%	Steal Attempts Per 1000 Base Runners
LOW	105-51	25	17	.580	58
MIDDLE	103-47	26	16	.585	54
HIGH	111-47	24	12	.610	78
TOTALS	319-145	75	45	.591	61

Right-Handed Pitchers:

Walk Avgs.	SB-CS (Not Pick-Off)	P.O.CS	Straight Pick-Off	Steal%	Steal Attempts Per 1000 Base Runners
LOW	297-131	15	28	.670	74
MIDDLE	281-143	20	26	.632	66
HIGH	307-127	14	34	.685	80
TOTALS	885-401	49	88	.663	73

PER 1000 BASE RUNNERS	LHP	RHP
Straight Pick-Offs	5.11	4.81
Pick-Off Caught Stealings	8.52	2.68

NATIONAL LEAGUE

Left-Handed Pitchers:

Walk Avgs	SB-CS (Not Pick-Off)	P.O.CS	Straight Pick-Offs	Steal%	Steal Attempts Per 1000 Base Runners
LOW	130-42	44	10	.602	85
MIDDLE	136-51	30	11	.627	83
HIGH	135-42	25	4	.668	104
TOTAL	401-135	99	25	.631	89

Right-Handed Pitchers:

Walk Avgs	SB-CS (Not-Pick-Offs)	P.O.CS	Straight Pick-Offs	Steal%	Steal Attempts Per 1000 Base Runners
LOW	419-142	39	36	.698	98
MIDDLE	415-119	48	13	.713	123
HIGH	426-149	31	25	.703	114
TOTAL	1260-410	118	74	.704	111

PER 1000 BASE RUNNERS	LHP	RHP
Straight Pick-Offs	3.51	4.58
Pick-Off Caught Stealings	13.91	7.30

AMERICAN AND NATIONAL LEAGUES

Left-Handed Pitchers:

Walk Avgs	SB-CS (Not Pick-Offs)	P.O.CS	Straight Pick-Offs	Steal%	Steal Attempts Per 1000 Base Runners
LOW	235-93	69	27	.592	70
MIDDLE	239-98	56	27	.608	67
HIGH	246-89	49	16	.640	88
TOTAL	720-280	174	70	.613	74

Right-Handed Pitchers:

Walk Avgs	SB-CS (Not Pick-Off)	P.O.CS	Straight Pick-Offs	Steal%	Steal Attempts Per 1000 Base Runners
LOW	716-273	54	64	.686	86
MIDDLE	696-262	68	39	.678	90
HIGH	733-276	45	59	.695	97
TOTAL	2145-811	167	162	.687	91

PER 1000 BASE RUNNERS	LHP	RHP
Straight Pick-Offs	4.4	4.7
Pick-Off Caught Stealing	10.9	4.8

COMBINE LHP AND RHP:

Walk Avgs	SB-CS (Not Pick-Offs)	P.O.CS	Straight Pick-Offs	Steal%	Steal Attempts Per 1000 Base Runners
LOW	951-366	123	91	.660	81
MIDDLE	935-360	124	66	.659	82
HIGH	979-365	94	75	.680	94

Bill James

PERIPHERAL AND IDENTIFYING CHARACTERISTICS OF GROUND-BALL AND FLY-BALL TYPE PITCHERS

One of the important bits of information released by the Elias Baseball Analyst was the groundout/flyout ratios of all pitchers. This is information that analysts of the game have needed for many years, and which we still need for many pitchers not covered in the present work, including minor league pitchers and those from the past. Perhaps more importantly, we also need to identify team groundball/flyball tendencies, which are essential for the interpretation of defensive statistics, and which are not carried in the Elias book.

Since we have good information for the 1984 season, it is now possible to develop better ways to identify and classify groundball pitchers. What are the characteristics of such a pitcher? Can one study the record of a pitcher and estimate accurately the percentage of his outs which come on ground balls and the percentage which come on fly balls? What is the most accurate way to do that? This article is an effort, a first effort with the good data available, to address those questions.

PART I
DATA AND DISTINCTIONS

I began by forming five groups of 20 pitchers each. Three were from the American League two from the National. I'll discuss the American League study first.

A. AMERICAN LEAGUE

The American League groups considered all pitchers without any innings cutoff. The groups were formed by the 20 with the highest groundout/flyout ratios, the 20 with the lowest groundout/flyout ratios, and the 20 who were in the exact center of the league. The groundout/flyout ratios of the top group ranged from 1.57-1, for Dave Rozema, to 2.75-1, for Bob Stanley. The central group ranges from Mark Clear, at .99-1, to Jay Howell, at 1.08-1. The lower group ranges from Bill Caudill, at .36-1, to Ken Schrom, at .71-1.

Having formed the groups, I simply recorded all the pitching statistics for the pitchers in the group, plus their assists total, taken from their fielding records, and began looking for any distinctions between them. I found several systematic differences, most of which could have been expected:

1) The groundball pitchers allowed fewer home runs. The high groundout group allowed 195 home runs while facing 9,968 batters, a rate of 20 HR/1000 batters. The low groundout group allowed 269 home runs while facing 10,464 batters, a rate of 26 HR/1000 batters. The central group was, as expected, in the middle in this respect.

2) The groundball pitchers allowed fewer sacrifice flies. The high groundball group allowed 72 sacrifice flies; the low groundball group, in 5% more at bats, allowed 92 sacrifice

flies. The central group was in the center in this respect.

3) The groundball pitchers allowed more sacrifice hits (bunts). The high groundball group allowed 87 sacrifice bunts; the low groundball group allowed only 54 bunts. The central group, which pitched quite a few more innings than the other two, allowed 90, so that they were, again, in the middle in this respect (on a per inning or per batter basis.)

4) The groundball pitchers had many more assists. The groundball pitchers were credited with 387 assists, which is 39 assists per 1000 batters faced. The low groundball group was credited with 222 assists, which is 21 per 1000 batters faced. The median group was credited with 35 assists per 1000 batters.

5) The groundball pitchers were charged with more unearned runs, confirming a finding previously put forward by Craig Wright. The groundball group was charged with 154 unearned runs, which accounted for 14% of their total runs. The flyout group was charged with 102 unearned runs, which is 9% of their total runs. The median group was, again, in the middle, with 10% of their runs being scored un-earned.

6) The groundball pitchers issued more intentional walks. I hadn't anticipated this one, but the reason for it is obvious: the groundball pitcher is more likely to issue an intentional walk to try to set up a double play than is a flyball-type pitcher. The groundball pitchers issued 88 intentional walks; the flyball group issued 70.

The median group, which had fewer relief pitchers than the other two groups, issued fewer intentional walks than either extreme group.

7) The flyball pitchers also tended to be strikeout pitchers. The groundball pitchers struck out 117 batters per 1000 batters faced. The flyball group struck out 142 per 1000 batters. The median group was in the center in this respect.

8) The groundball pitchers allowed more hits. The groundball pitchers allowed 242 per 1000 batters. The flyball group 227 hits per 1000 batters.

Those are the distinctions which I take to be significant; you may find others. The most effective pitchers, by far, were the central group, which had a composite record of 189-144 (.568) with a 3.64 ERA. The extreme groups were comparable in performance, the groundball pitchers going 138-141 (.495) with a 3.77 ERA, and the flyball pitchers 144-134 (.518) with a 3.82 ERA. Counting unearned runs, the flyball group was more effective, allowing fewer runs per 9 innings and having a better winning percentage; however, the difference is small.

Both groups had more relievers than starters, while the median group did not:

	G	GS	CG	GF	ShO	Sv
GROUNDBALL GROUP	755	244	36	295	9	88
MEDIAN GROUP	658	360	71	144	12	22
FLYBALL GROUP	793	247	31	332	9	103

The abbreviations, for those of you who have stumbled on this article in your dentist office or something, are for "Games pitched, Games Started, Complete Games, Games Finished, Shuouts, and Saves." The key point is that the data for the extreme groups is similar. The rest of the data will be presented in the same form, in case it is of some use:

	IP	H	TBF	R	ER	HR	SH	SF
GROUNDBALL GROUP	2326	2408	9968	1128	974	195	87	72
MEDIAN GROUP	2911.2	2799	12240	1307	1178	255	90	102
FLYBALL GROUP	2452	2380	10464	1162	1060	269	54	92

	HB	TBB	IBB	SO	WP	Bk	Assists
GROUNDBALL GROUP	51	828	88	1162	74	11	397
MEDIAN GROUP	59	931	56	1710	76	10	425
FLYBALL GROUP	50	910	70	1490	65	16	222

B. NATIONAL LEAGUE

From doing the American League study, it was apparent that extreme ratios tend to be most common in small groups of innings, and thus that there was a measure of random chance in them. When moving the study to the National League, I decided it would be desirable to mitigate this effect by setting a 100-inning minimum for inclusion in the study. In the National League, I dispensed with the central group, which tends essentially to confirm the trends, and formed only the two extreme groups.

The upper groundball group consists of twenty pitchers with groundball ratios ranging from 1.40-1 (Larry McWilliams) to 3.68-1 (Steve Trout). The low groundball group ranges from 0.45-1 (Sid Fernandez) to 0.87-1 (Jose DeLeon). I will begin here by presenting the data:

	G	GS	CG	GF	ShO	Sv
GROUNDBALL GROUP	841	374	53	249	17	99
FLYBALL GROUP	787	437	49	185	13	61

	IP	H	TBF	R	ER	HR	SH	SF
GROUNDBALL GROUP	3106.2	3042	13063	1351	1189	207	144	88
FLYBALL GROUP	3295.1	3124	13834	1516	1352	299	142	111

	HB	TBB	IBB	SO	WP	Bk	Assists
GROUNDBALL GROUP	40	1038	139	1811	90	33	566
FLYBALL GROUP	51	1061	88	1980	68	27	382

It should be noted that the groundball pitchers were somewhat more effective in the National League. The groundball group had a composite won/lost record of 170-165 with a 3.44 ERA (4.00 Total Runs/9 innings), while the flyball group had a composite record of 180-198 with a 3.69 ERA (4.23 total runs). This would be consistent with Craig Wright's argument that groundball pitchers tend to be more effective on artificial turf; however, it is very weak evidence in that regard.

In general, all of the same effects which were observed in the American League are also apparent here. The groundball group allows fewer home runs and sacrifice flies, has more assists, allows more hits per inning, more sacrifice bunts and unearned runs, and gives more intentional walks. The flyball group strikes out more batters.

There are some differences. Perhaps the largest, and perhaps the most intriguing discovery of the study--it will be interesting to see if it holds up with more research--is that the sacrifice hit totals, which reflect a large and consistent spread in the American League, are almost even in the two National League groups. That means, if it holds up, that the groundball tendency of the pitcher is a major determinant of when a bunt is ordered in the American League, but not in the National League. One can easily imagine why that might be so, but it is a large and significant finding, and should not be embraced too readily or trumpeted too loudly until it can be confirmed.

Another point of note is that most of the indicators--six of the eight--would also point toward there being more ground balls in the National League than in the American League. The National League has fewer home runs and fewer sacrifice flies, but more sacrifice bunts, more intentional walks. more pitcher's assists and more un-earned runs. The two exceptions are the weakest indicators--the National League has more strikeouts and fewer hits/batter.

PART II
DISTINCTIONS NOT OBSERVED

There were a couple of things that I expected to find in this study, but did not.

1) I expected to find the distinction between the strikeout tendencies of the different types of pitchers to be much more extreme, and thus much more useful for identification, than it was.

2) I would have expected that the ground ball pitchers could have been identified by the ratio between their innings pitched and their outs accounted for--that is, that because the ground ball pitchers would get more double plays, these "excess outs" could be isolated and identified in the study.

Not so. The distinction does exist, but it is not strong enough to be very useful. What I mean by outs/innings ratio is this:

$$\frac{\text{TOTAL BATTERS FACED - HITS - WALKS - HIT BATSMEN}}{\text{INNINGS PITCHED TIMES THREE}}$$

If all outs were accounted for this would be 1.00; each double play causes the ratio to drop lower.

It is lower. In the American League, the flyout group has an outs/innings ratio of .968; the groundout group of .957. In the National League, the ratios are .971 and .960. The lowest outs/inning ratio for any pitcher was for Juan Agosto, a groundball pitcher in the American League, who had a ratio of

.916 (152/166). But the variation between individuals is too large, and the overall distinction too small, to be useful.

3) I might have expected there to be more left-handed pitchers in the groundball group; this, at least, has been suggested. Not so at all within this study; in fact, the concentration of left-handers was higher in the flyout group than in the groundout group in both leagues. In the American League, the groundball group contained 5 left-handers, the central group contained 5, and the flyball group contained 6. In the National League, the groundball group contained 6 lefthanders; the flyball group contained 8.

PART III
CONSTRUCTION OF A PREDICTIVE FUNCTION FOR GROUNDBALL TENDENCIES

I had hoped, on beginning this research, that I would be able to develop a way to estimate accurately what percentage of a pitcher's outs resulted from ground balls. It seemed to me that it should be possible--and it may still be possible--to develop a method that would take a pitcher's statistical line, look at the home runs, look at the sacrifice hits and the sacrifice flies, look at the assists, and say "this fellow gets 41% of his outs on ground balls."

I am somewhat disappointed with what I have. The best formula that I have developed takes an unimaginative form and yields decent, but not reliable, results. The fomrula is:

GROUND BALL PERCENTAGE EQUALS .307 + 1.84X
where X equals:

National League:
(Pitcher's Assists - (Park Adjusted) Home Runs Allowed - Sacrifice Flies Allowed + Intentional Walks + Un-Earned Runs) divided by innings pitched.

American League:
(Pitcher's Assists - (Park Adjusted) Home Runs Allowed - Sacrifice Flies Allowed + Intentional Walks + Un-Earned Runs + Sacrifice Flies) divided by innings pitched.

Before I discuss the accuracy of this formula, I have to explain something. The way that Elias has chosen to state their data (I would probably have made the same choice for their purpose) tends to exaggerate the distinctions between pitchers. Comparing a pitcher with a ratio of .50-1 to one with a ratio of 2.00-1 it looks like one pitcher has four times the ground ball tendency of the other--but it's really only two times. The one pitcher gets 33% ground outs, the other 67%. This is a consistent problem with the ratio-stated method; it always makes the differences look a lot bigger than they really are. Besides that, when estimating the percentage, you can't really estimate the ratio; you have to estimate the percentage and turn it into a

ratio. I'm going to deal mostly with "percentages" here, not ratios; a 1.25-1 ratio is 56% ground balls.

For the groups of pitchers discussed above, this formula is acceptably accurate in assessing their groundout tendencies. I'll present the data in chart form; that is, the American League high-groundball group can be estimated by the formula above to have gotten 65.9% of their non-strikeout outs on ground balls, whereas they actually got 65.1%:

	Estimated	Actual
AMERICAN LEAGUE GROUNDOUT	.659	.651
AMERICAN LEAGUE CENTRAL	.517	.507
AMERICAN LEAGUE FLYOUT	.372	.377
NATIONAL LEAGUE GROUNDOUT	.646	.668
NATIONAL LEAGUE FLYOUT	.432	.435
AMERICAN LEAGUE, TOTAL	.509	.509
NATIONAL LEAGUE, TOTAL	.559	.543

For many <u>individuals</u>, it is also accurate. For <u>most</u> individuals, it is reasonably accurate. For example, Jose DeLeon (estimated 40%, actual 46%), Dave Dravecky (37%/46%), Andy Hawkins (50%/46%), John Candelaria (41%/45%), Ed Whitson (48%/45%), Pete Falcone (44%/45%), Mark Davis (39%/44%), Ed Lynch (48%/44%), Tim Lollar (37%/40%), Craig Lefferts (40%/39%), Bill Laskey (42%/39%), Joe Price (39%/38%), Rich Gossage (38%/35) and Steve Trout (83%/79%). The formula yields below-average estimates for 17 of the 20 NL pitchers who were among the league low in groundout/flyout ratio, and above-average estimates for 18 of the 20 at the top of the scale, although two of them screw us up by going over the top--over 100%.

It must be assumed--and it is apparent--that these estimates are more accurate for any group than for the individual seasons within a group. For an analysis of a <u>team</u>, then, the formula might be good enough. But for an analysis of an individual, the possibility of an error is still too large.

Take, for example, Jeff Russell, the young pitcher recently acquired by the Rangers.

Estimated: 68%
Actual: 46%

The formula is completely off, and there is nothing apparent that we can do about it. If you look at his stats, you would swear he was a groundball pitcher. He had 34 assists while pitching to 787 batters, a higher assists rate than most of the National League's top ground ball pitchers. He allowed only 3 sacrifice flies (an extremely low rate) and walked 8 batters intentionally (a high rate). He was a fraction worse-than-league in strikeouts and hits allowed per batter faced. He was near the middle of the league in home run rate and un-earned runs allowed. Nothing about the record suggests that he would have one of the lowest Ground Ball rates in the National League--but he did.

Craig Wright has said that the pitcher's assist total is

the best indicator of his ground-ball tendency, and he would seem to be right about that--but there are anomalies. Russell is one; Bob McClure is another. Comparing David Palmer and Frank Williams of the National League, I note that they have similar records (9-4, 7-3) and reasonably similar ERAs. Palmer faced 444 batters and had a ground outs-to- air out ratio of 2.25-1; Williams faced 454 batters with a ratio of 2.19-1--yet Palmer had only 13 assists, and Williams had 35.

Home Runs allowed (with park adjustments) and Sacrifice Flies allowed would probably be as meaningful as assists, except that the numbers are smaller, and thus require more time to become stable indicators. All three of those seem to be basically straight-line indicators; that is, the percentage of ground ball outs that will become pitchers' assists seems to be about the same for pitchers who throw 65% ground balls as it is for those who throw 35%. The percentage of fly balls which leave the yard appears to be about the same for pitchers who throw 65% fly balls as it is for those who throw 35%. This would be very useful knowledge--were it not for the anomalies, were it not for the fact that both categories are subject to other factors, such as pickoffs at first base (1-3), which Bob McClure gets in abundance, and the fact that there are simply more hard-hit balls (hence home runs) off some pitchers than there are off others.

The other major indicators--intentional walks, unearned runs, possibly sacrifice hits--are also subject to influence from several others factors, and consequently don't show the same straight-line relationship to the groundball tendency. Perhaps some adjustments could be made for other things which have an impact on the column, and thus the relationship could be made more valuable. The remaining relationships--strikeout tendencies, hits/inning records--are so vague as to be of questionable value in the identification of groundball pitchers.

In any case, I am certain that it can be done better than I have done it here. The method that I have developed here is probably adequate for classification purposes. I noted early on in the study that one can make a surprisingly accurate classification of groundball/flyball tendencies in the American League by focusing on only two categories--sacrifice hits allowed, and sacrifice flies. Of the 20 American League pitchers with the highest ground ball rates, only 5 allowed more sacrifice flies than sacrifice hits. Of the twenty with the lowest ground ball rates, 15 allowed more sac flies than sac hits.] That gives us a pretty good start on a classification right there, and it leaves the two major categories--assists and home runs allowed--untouched.

So if you can improve on this system, great. That's what it's here for.

RUN ESTIMATION BY GAME-LINE ASSEMBLY

by Mark D. Pankin

In a recent Analyst article, Bill James explained a computer implemented technique to estimate runs from offensive contributions based on team total AB, H, 2B, 3B, HR, BB, and R in individual games ("Technique of Run Estimation by Game-Line Assembly", Issue 17, April, 1985). Bill thought that the technique could be a significant alternative runs estimator, particularly for atypical cases, but that it is not practical. In this article, I report on some follow-up work I have done. It is almost like a "good news/bad news" story. The good news is that the technique can be implemented in a practical way, but the bad news is that there are questions about the accuracy of the method. However, the bad news is tempered because the method may work best for exceptionally high offensive performance, where there is the greatest need for additional run estimating techniques. Also, with a practical computer implementation, it may be possible to modify the technique to work better for average performances.

This article first describes the computer implementation, including what I think are corrections to Bill's article, and makes it available to others who have the appropriate computer and software. Then, I discuss the results of the technique and some of the problems with it.

Computer Implementation. Bill developed a spreadsheet to carry out the game-line estimation technique on his Kaypro. I believe that there are two errors in his article. (No doubt that he will let us know if I am wrong.) First in Section 3--The assessment, the five formulas in cells n4-r4 should use the absolute value of the differences rather than the differences themselves. Thus, the formula for n4 should be 2*abs(h51-h4). I base this on Bill's idea of penalizing large differences. Without the absolute values, two large differences, one postive and one negative, can cancel each other. Also, negative differences increase the suitability score, which does not make sense in light of Section 4. The second error is in Section 6--Re-Assessing the needs is in line 51. Bill shows formulas that use line 49, such as b49*(a51/a49) for b51, in order to express the player's performance as if he had the number of at bats in a51 (1500 in his example). But line 49 contains the totals of the games selected, not the player's record. He really used line 1 to compute the numbers in the example at the bottom of the page. The correct formula for b51 is b49*(a51/a1), and the formulas for the rest of line 51 are corrected in a similar manner.

These corrections are provided in case anyone else wants to duplicate Bill's work. (I would not be surprised to find out I was the only one who was "crazy" enough to dig into the article. Thanks for writing it!) As he pointed out, the Kaypro implementation is not practical. The problem is not the fundamental technique, but the computer and spreadsheet software. Fortunately, the IBM PC and compatibles, which have much more memory, along with an advanced spreadsheet such as Lotus 1-2-3 can handle the game-line estimation technique in a resonable manner. I have a Compaq portable with 256K of RAM, and I use Lotus 1-2-3 to implement this run estimation method. (This article, however, was written using an Apple Macintosh, so please don't ask me how to get the different type faces on an IBM PC.) 1-2-3 and a few other spreadsheet programs have several features that are extremely useful in this case. Powerful database functions permit the streamlining of Bill's section 4, selection of most desirable games, and the elimination of section 5, storing several totals of selected games. With this elimination, there is no need to limit the number of iterations to a specified fixed number. Instead, my set up has a target number of at bats, and games are selected until that target is reached or exceeded.

An even more useful feature is the "Macro" language built into 1-2-3. At a minimum it can be used to assign repetitive sequences of keystrokes to a single key, which greatly facilitates issuing the commands needed to run the spreadsheet. However, I carried this further and used the Macro language to automate

the entire process. All the input player or team lines are entered in one area of the spreadsheet, and then a single key is pressed. The spreadsheet takes over and one at a time loads the data for an individual player or team into line 1 as described in Bill's article and carries out the game-line selection, re-assessment, and accumulation until the target number of at bats is reached. Then the final results are copied back to the input area and the numbers for the next team or player are loaded so the estimation process can repeat. This goes on until all teams and players have been estimated with no additional attention needed. The time required for an estimation depends on the number of game-lines, how strict the selection standards are, and the target number of at bats. I have loaded 158 game lines (all major league games from a recent week). When choosing the most suitable game only and using 2000 target AB, it takes about five to seven minutes for each estimation. Loosening the selection criterion to (4*best + avg.)/5 reduces the computation time by about half. Because I can start the program and leave it alone while I do other things, this is a practical method for doing game-line based run estimations.

I am willing to make this spreadsheet available to anyone who can use it. You must have the Lotus 1-2-3 spreadsheet software and a computer with at least 256K RAM that can run it. Also, you must know how to perform basic spreadsheet operations and procedures such as loading and saving files in 1-2-3. I will provide a disk file with some instructions to help you, but this spreadsheet is not a fancy "turn-key" product like those you can pay quite a bit for. On the other hand, the price is right -- FREE if you send me a formatted floppy disk in a reusable disk mailer (if it gets to me, I probably can get it back to you in the same mailer). If this sounds like too much trouble, send me $5 to cover the cost of the disk, mailer, and postage, and I will send you the spreadsheet. Please specify if you want the disk formatted in a particular way. Otherwise, I will format it as 8 tracks per sector, double sided. My address is 1018 N. Cleveland St., Arlington VA 22201.

Run Estimation Results. First, here are the results for the three players in Bill's articles. The first line shows the actual data for the player, the second shows the game totals from my spreadsheet prorated to the same number of AB, and the third has the corresponding data that Bill reported.

Player (Year)	AB	H	2B	3B	HR	BB	Runs
Enos Cabell (84)	436	135	17	3	8	21	62 (Runs Created)
--Pankin Estimates	436	135	17	3	8	24	61
--James Estimates	436	136	17	4	8	21	64
Gus Bell (54)	619	185	38	7	17	48	101 (Runs Created)
--Pankin Estimates	619	185	38	6	17	49	83
--James Estimates	619	185	38	6	17	51	98
Eddie Murray (84)	558	180	26	3	29	107	129 (Runs Created)
--Pankin Estimates	558	179	26	3	29	105	125
--James Estimates	558	179	25	5	28	67	121

My estimates are based on the 158 game-lines from all major league games played during the week of May 12-18 as taken from The Sporting News box scores, one line for each team or two per game. I used 2000 target at bats, and on each iteration the most suitable game only was selected to add into the total. These comparisons are quite bothersome. Enos Cabell comes out about the same no matter how he is evaluated. [However, when I used games taken from June 23-29, his estimated runs dropped to 39.] With Gus Bell, both of us match his actual data very closely, but Bill's estimate is much higher and closer to the basic runs created formula. This difference is caused by Bill's use of only 20 game-lines, which contain one or two Gus Bell type games with more runs scored than the Gus Bell type game-lines in my set

of 158. [I get 85 estimated runs for Bell using June 23-29 games.] For Eddie Murray, Bill misses badly on the walks, but his estimated runs is close to mine and the runs created. Again, this is probably caused by one or two of his game-lines being high in runs and home runs, but low in walks. [My June 23-29 game-lines estimate Murray's runs at 124.] What all of this shows is that the game-line estimation method is sensitive to the particular game lines used. Obviously, 20 game-lines are not enough, which Bill pointed out, but even with two sets of 158 game-lines (combining them requires more than the 256K I have in my Compaq), there are troubling results. One is the huge drop (61 to 39) in Cabell's estimated runs between my two sets. Another is that the estimate for Bell is so far below the runs created value while the estimate for Murray is not. The problem, if any, with runs created is that it tends to overestimate exceptionally high performance, which would lead me to expect a greater difference from runs estimated by game-lines in Murray's case than Bell's IF the game-line estimation method is accurate.

One test often used on run estimating techniques is to see how they work for teams. After all, we know how many runs a team produced in a given season, so any method or formula worth its salt ought to estimate team runs well. Unfortunately, game-line estimation falls down badly on this score. I tested the method on the 1984 season totals for all major league teams and for the American and National League totals. In general the estimates are well below the actual runs scored, over 100 runs too low for some teams. This was very disturbing, and my first thought was that there was something wrong with the spreadsheet set up or I had chosen an unusual week. The results did not change significantly when I used a different week's games (also 158 game-lines). I tried adjusting the penalty function, number of target at bats, game selection criterion among other things, but nothing changed the fundamental underestimation. Then I decided to see if I could find a reason for the low run estimates. After analyzing the game-line data in several ways, I believe the problem is that the run estimates are based on the overall average runs per at bat of the selected games. Team season totals represent more or less average performance. Hence the game-lines selected for team will tend to be taken from games near the center of the spectrum. However, as runs scored in a single game grows, the runs per at bat does not grow as rapidly. The following table summarizes the data from the 316 game-lines in the two weeks used:

Runs	# of games	Avg. R/AB	Low AB	High AB
0	26	0.000	27	34
1	36	0.032	21	40
2	42	0.061	27	43
3	49	0.092	24	48
4	31	0.118	27	49
5	34	0.139	27	55
6	33	0.171	30	49
7	27	0.201	31	50
8	14	0.224	31	41
9	7	0.262	31	38
10	8	0.275	32	42
11	5	0.282	36	43
12	1	0.324	37	37
14	1	0.341	41	41
16	1	0.363	44	44
19	1	0.413	46	46

Up until about 9 runs per game, each additional run per game results in about 0.030 more runs per at bat, with the curious exceptions of going from 4 to 5 and from 7 to 8. Above 9 runs per game the growth in runs per at bat is much slower. Granted there are few games in the sample at that level, it seems safe to say that the approximate relationship $R/AB = 0.030*(R/G)$ that holds up to 9 R/G grossly overestimates the R/AB for games with 10 or more runs. To see why this causes underestimation consider two sets of three games each. In one, 4 runs are scored in each game, and in the second there is one game with 12 runs and two games with no runs. Thus each set averages 4 runs per game, but in the first set the average runs per at bat is 0.118, while in the second set the average runs per at bat is $0.324/3 = 0.108$. If the hits and walks data for the two sets of games have the same relationship as the runs per at bat, then the game-line run estimate for the second set will be low. I think that something along these lines happens when considering a team for an entire season, but I don't know how to verify this without an extensive study and whether this presumed effect explains the large estimated vs. actual underestimates.

Obviously there a lot of loose ends in the above analysis, but it does shed some light on the evaluation of the game-line run estimation technique. Presently, there seem to be two major problems. One is that the estimates are very sensitive to the game-lines used. A sufficiently large set of game-lines should correct this problem. However, as the number of game-lines grows, the computer requires more memory and calculation time increases. The second problem is that the accuracy of the method is poor when applied to team data. The method may be accurate for extremely high or low levels of performance, but it is not clear how this can be verified. One possible solution to both problems is using groups of game-lines selected to measure performance in a given range. For example, only high run game lines might be used to evaluate outstanding offensive seasons. However, accumulating the appropriate sets of game-lines might be quite a chore. The whole question of whether the game-line estimation method can be a useful tool is up in the air. The comments and ideas of others are welcomed. Even if this goes nowhere, I thank Bill for presenting the method and providing me with a fascinating Lotus 1-2-3 exercise.

* * * *

DICK O'BRIEN continued from page 4

batters without regard to a runner on first base will, I submit, pay richer dividends than the automatic relapse into double play thinking. Or if it's early in the game and our pitcher can't throw low strikes today, are we going to the bullpen for one who can?

Granted, I may be overstating the case against the use of the double play strategy, but I'm equally certain that its blanket use by certain managers whenever there are less than two outs and a runner occupies first base is equally specious reasoning. When you come down to the question of whether or not to go for the double play the most important factor to consider is the batter's predicability patter against the man on the mound or his possible successor.

Davis Jackson's statistical summary of the 1984 season, The Last Word, shows (p.60) that both leagues convert 40% of double play opportunities (from a low of 32% - Mariners and Phillies - to a high of 52% - Expos, Orioles and Cards). Thus when everything occurs according to the predictability pattern and in the right sequence a 52% success rate is the most one can hope for. Is the game worth the candle?

I'm certainly not arguing for the abandonement of the use of the double play. There are many time it should and must be tried. All that's being suggested is that its questionable overall success rate be examined to determine its best use. Defensive strategy should be determined on the best way of getting the batter out, not on getting the batter and the base runner out on the same play. This should be true often enough to discourage the automatic double play strategy syndrome.