SABR XXIII
SAN DIEGO

JUNE 25, 1993

SUBTLE ASPECTS OF THE GAME

MARK D. PANKIN

CLASSIFICATION OF BATTERS

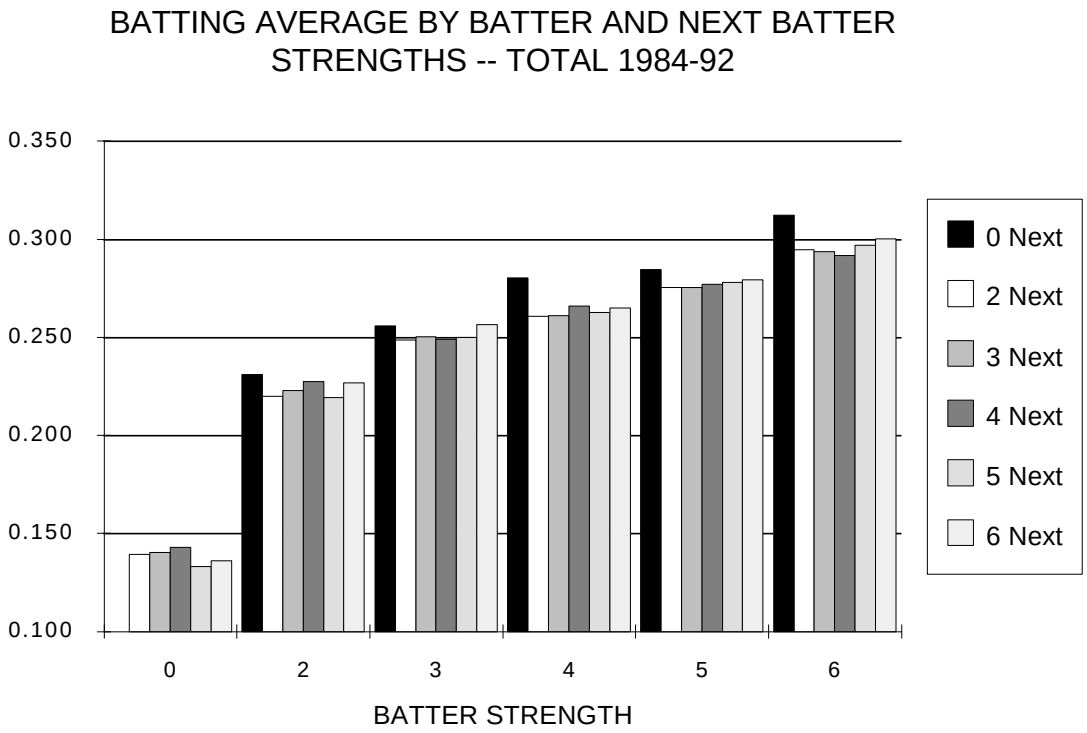
AVERAGES, DIVIDING POINTS

RATING	OBP	SLUGGING	
3	0.372	0.477	AVERAGE
	0.346	0.427	DIVIDING POINT
2	0.329	0.397	AVERAGE
	0.313	0.367	DIVIDING POINT
1	0.284	0.316	AVERAGE
0 = PITCHERS	0.173	0.174	AVERAGE

- ADD OBP AND SLUGGING RATINGS TO FORM FIVE GROUPING PLUS PITCHERS:

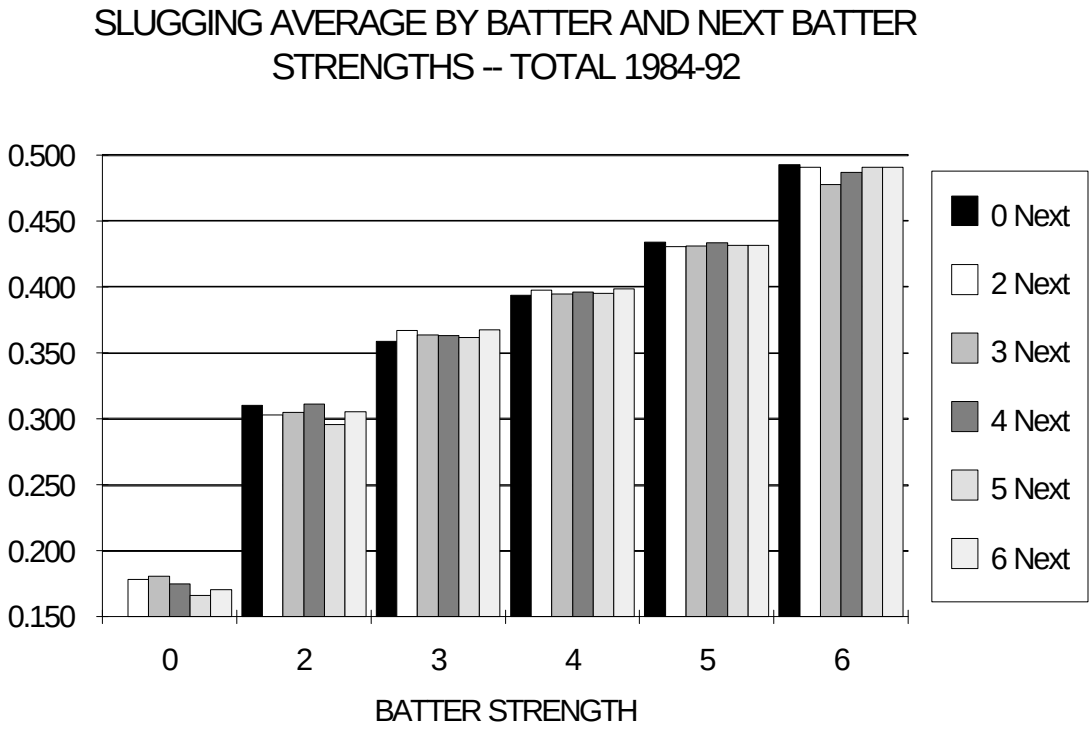
  6 = BEST HITTERS
  5 = ABOVE AVERAGE HITTERS
  4 = AVERAGE HITTERS
  3 = BELOW AVERAGE HITTERS
  2 = WORST HITTERS (EXCLUDING PITCHERS)
  0 = PITCHERS

BATTING AVERAGE



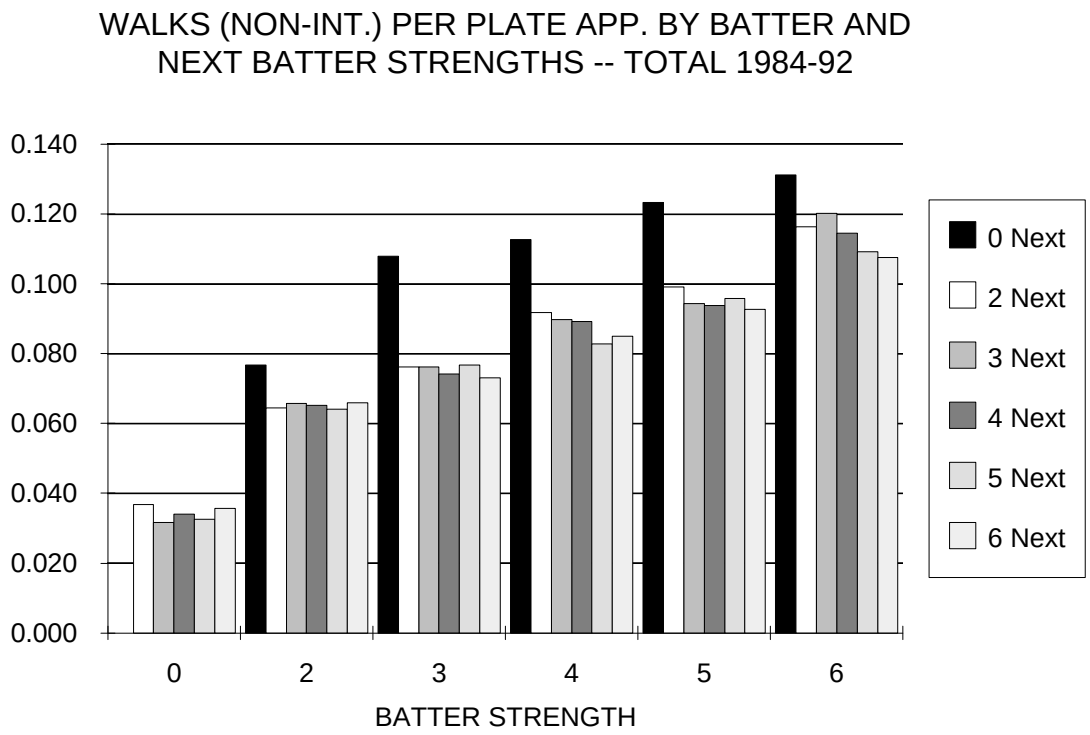
- HIGHER BAs WHEN PITCHER BATS NEXT ARE SIGNIFICANT
- OTHER DIFFERENCES ARE NOT SIGNIFICANT

SLUGGING AVERAGE



- DIFFERENCES BY NEXT BATTER STRENGTH ARE NOT STATISTICALLY SIGNIFICANT

NON-INTENTIONAL WALKS PER PA



- HIGHER LEVELS WHEN PITCHER BATS NEXT ARE SIGNIFICANT EXCEPT WHEN 6's BAT (TOO FEW PLAYS)

- TENDENCY FOR MORE WALKS WHEN A WEAKER BATTER FOLLOWS (MANY OF THE DIFFERENCES SHOWN ARE STATISTICALLY SIGNIFICANT)

CLASSIFICATION OF RUNNERS

- BASED ON (SB + CS)/(1B + BB + HBP)
- MEASURES FREQUENCY OF STOLEN BASE ATTEMPTS WHEN PLAYER REACHES FIRST BASE
- COMPUTABLE FROM STANDARD DATA
- THREE WAY CLASSIFICATION:

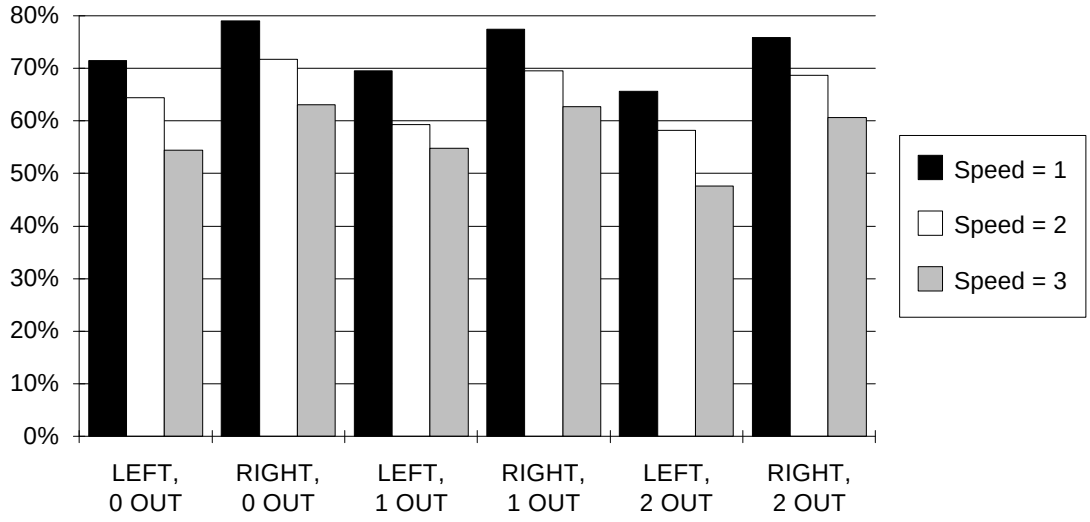
  3 = FASTEST: 14.1% AND ABOVE
  2 = MIDDLE: BETWEEN 4.8% AND 14.1%
  1 = SLOWEST: LESS THAN 4.8% AND ALL PITCHERS

ADVANCEMENT ON SINGLES (1)

- GRAPH SHOWS PERCENT FIRST TO SECOND (WHEN NO RUNNER ON SECOND), SO LOWER IS BETTER

- DIFFERENCES ARE SIGNIFICANT

PCT. OF RUNNERS ON 1ST STOPPING AT 2ND ON SINGLES
1ST, 1ST & 3RD ONLY, 1984-92 MAJOR LEAGUE DATA

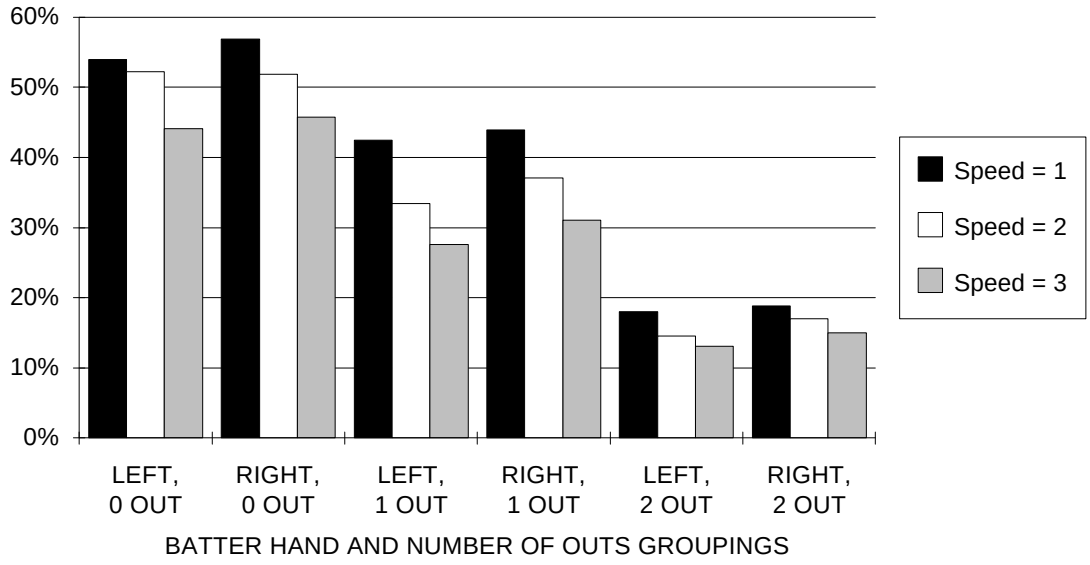


BATTER HAND AND NUMBER OF OUTS GROUPINGS

ADVANCEMENT ON SINGLES (2)

- GRAPH SHOWS PERCENT SECOND TO THIRD, SO LOWER IS BETTER
- DIFFERENCES ARE SIGNIFICANT

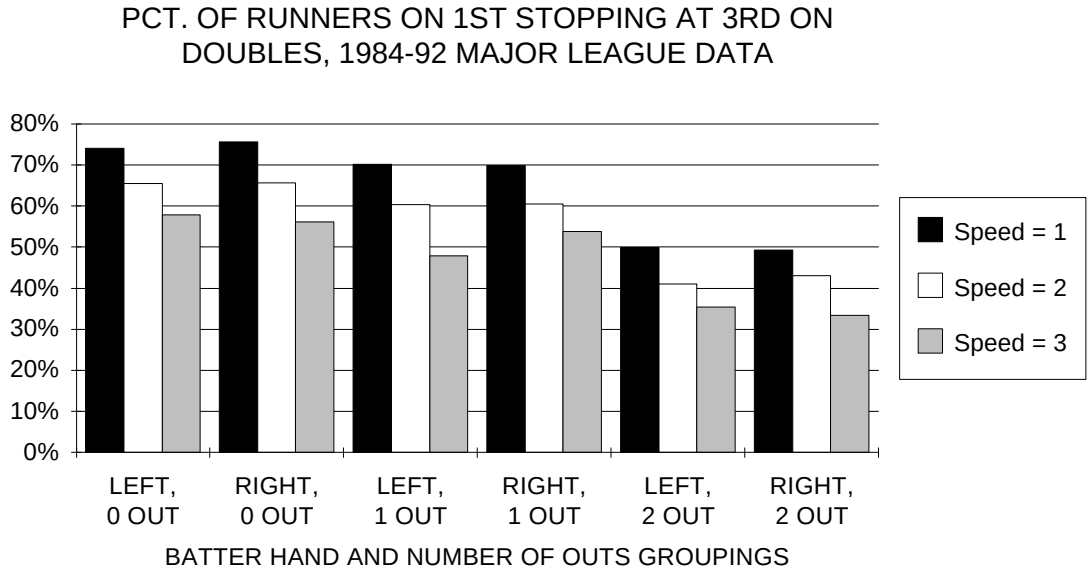
PCT. OF RUNNERS ON 2ND STOPPING AT 3RD ON SINGLES
1984-92 MAJOR LEAGUE DATA



BATTER HAND AND NUMBER OF OUTS GROUPINGS

ADVANCEMENT ON DOUBLES

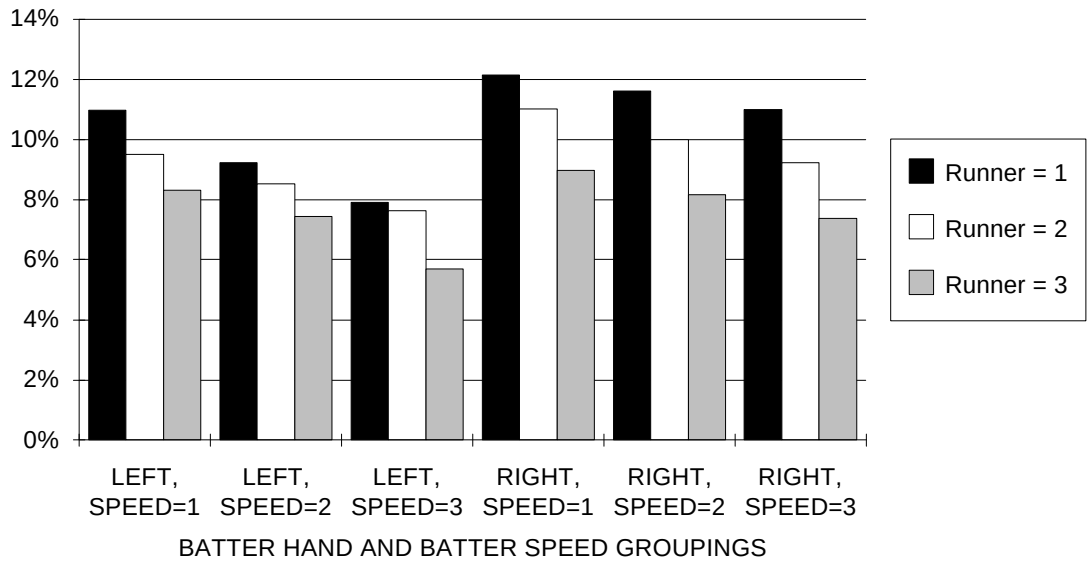
- GRAPH SHOWS PERCENT FIRST TO THIRD, SO LOWER IS BETTER
- DIFFERENCES ARE SIGNIFICANT



AVOIDING DOUBLE PLAYS

- MOST DIFFERENCES ARE SIGNIFICANT: BOTH FASTER BATTERS AND FASTER RUNNERS ARE INVOLVED IN FEWER DOUBLE PLAYS

GIDP PERCENTAGE, RUNNER ON FIRST ONLY
1984-92 MAJOR LEAGUE DATA

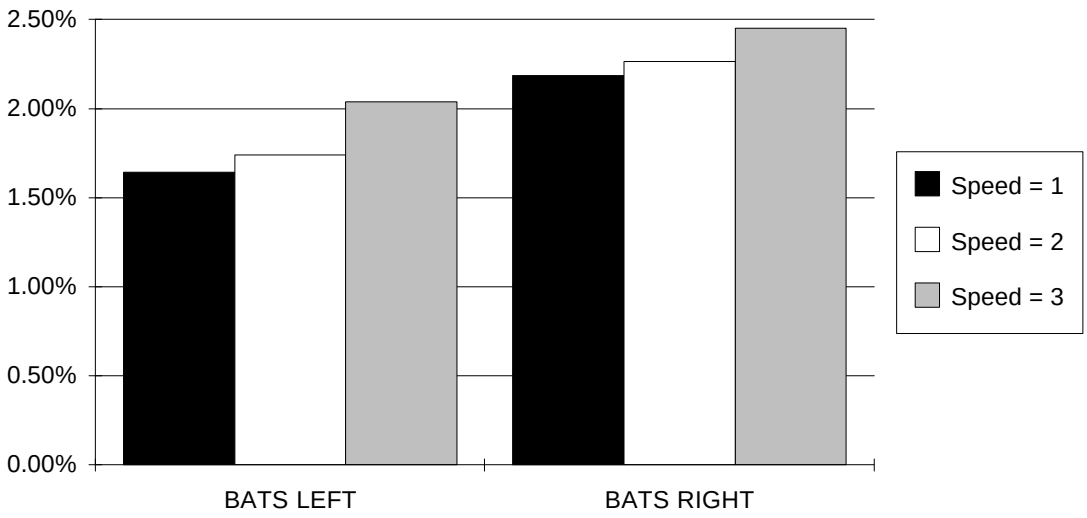


BATTER SAFE ON ERROR

- DIFFERENCES ARE SIGNIFICANT: FASTER BATTERS REACH BASE MORE OFTEN ON ERRORS

- BATTING HAND IS MORE IMPORTANT: SLOWEST RIGHT REACH MORE ON ERRORS THAN FASTEST LEFT

SAFE ON ERROR % BY BATTER HAND AND SPEED RATING
NO RUNNERS ON BASE, 1984-92 MAJOR LEAGUE DATA



CONCLUSIONS

- STRENGTH OF FOLLOWING HITTER HAS ONLY SLIGHT EFFECT ON BATTING PERFORMANCE

  - BATTING AND SLUGGING AVERAGES NOT AFFECTED
  - SLIGHTLY MORE WALKS WHEN WEAKER HITTERS NEXT
  - HIGHER BATTING AVERAGES AND MORE NON-INTENTIONAL WALKS WHEN PITCHER FOLLOWS

- FASTER BATTERS AND RUNNERS HAVE STATISTICALLY SIGNIFICANT ADVANTAGES

  - MORE LIKELY TO ADVANCE FURTHER ON HITS
  - AVOID SOME DOUBLE PLAYS
  - REACH BASE MORE OFTEN ON ERRORS

- KEY QUESTION: WHAT ARE THE EFFECTS ON RUN SCORING AND CONSTRUCTION OF BATTING ORDERS?

NOTES AND COMMENTS

Introduction. This section provides explanations and additional information to accompany the overhead slides. In a sense, it is an article organized those slides.

For several years, I have been developing a so-called "Markov Chain" model of major league baseball. [I have given several talks and written several articles on the subject, so no further explanation is provided here.] One product of the model is a method for determining how many runs a lineup will score on the average over a large number of games and an associated model that finds the highest scoring batting order given the nine hitters.[1]

Any mathematical model of a complex process incorporates simplifying assumptions. The most important one in my model is the assumption that players' batting performance is not affected by where in the order they hit or the lineup positions of the other players. Baseball lore contains statements like "if you put a weak hitter after a strong one, the strong hitter won't do well because he won't get anything to hit." I decided to do a study to see if that is indeed the case. Except for individual player egos and psychology (e.g. Barry Bonds did not want to lead off because he figured that a high RBI total would increase his pay, and it is hard to argue with him at this point), the possible effects caused by the strength of the next batter seemed to me to be the most likely reason my assumption on batting performance might not be correct.

Another simplifying assumption in my model is that runner advancement on hits and outs, double plays, and batters reaching on errors all take place according to major league averages. Although, I doubt that this assumption has much of an effect when comparing two batting orders, there are obviously differences among players according to their speeds and base running abilities. In order to improve the model, I decided to collect the necessary data to account for these differences.

This talk reports on the primary results of these two investigations. The source of the data is the Project Scoresheet database, which currently contains play-by-play data for every major league game in the 1984-92 seasons. My objective was to obtain statistically meaningful results, which requires large numbers of plays, which in turn means broad groupings of players.[2]

The page numbers referenced below are those of the slide copies that appear earlier.

Page 1: Classification of Batters. Many studies have shown that run creation can be modeled as a function of two types of baseball events: 1) getting on base, and 2) advancing on the bases. The first is measured by on base percentage (OBP), and the second corresponds to slugging average (SA). Consequently, I decided to use these two

---

[1] These models are implemented in the Draft module of the APBA computer baseball game, which is published by Miller Associates (1-800-654-5472)

[2] Those who wish to find out about statistically meaningless comparisons can watch almost any telecast or read Elias. We don't take seriously the batting leaders after the first 10 games, approximately 40 at bats, so why should we care that Jose Canseco hit .300 last year in "late inning pressure situations" with runners on (9 for 30)?

measures when classifying batters by ability. Each player is classified by his one season performance, so the same player may have different ratings in different years. In order to get broad groupings, I divided all regular and semi-regular (at least 200 plate appearances) position players into three approximately equal groups for each of OBP and SA. Players with fewer than 200 plate appearances probably are, on the average, weaker hitters, so the overall distribution of players may have more 1s. However, better hitters tend to get more playing time, so the number of plate appearances will be distributed more uniformly.

Obviously, there will be hitters near the dividing points with different ratings and similar abilities. Keep in mind the ratings are for the purposes of forming broad groupings, not for evaluating individual players. The averages for the rating groups show that considered as groups, there are distinct differences in hitting ability among the groups.

The 3,2,1 ratings serve to provide labels. No claim is made that the 3s are 50% better than the 2s. Pitchers are put into their own separate group (rating = 0).

There are nine groups if we use all possible combinations of the OBP and SA ratings. This seems to be too many, so I decided to add the OBP and SA ratings to obtain five groups (2,3,4,5,6) plus the pitchers. I do not claim that a 3 in OBP is somehow equal to a 3 in SA in batting ability. The addition is done to obtain an appropriate number of groups. It should be noted that high OBP and high SA are not independent. There are exceptions, but players who are above average in one category are often above average in the other. I think the qualitative descriptions in the overhead slide are justified.

Definition of Next Batter. In most cases this is obvious, but two special cases are worth discussing. If a player is the last batter of the game for his team, the next scheduled batter is used as the next batter for the purposes of this study. If a batter is the last batter in an inning, the first batter of the next inning is considered the next batter, whether or not he was in the lineup at the end of the previous inning (i.e. he may be a pinch hitter or have entered the game on defense). There are relatively few such cases, so this decision is probably not crucial. In effect, I am assuming that the opposition and batter acted as if they knew there would be a substitution. While this is not likely to be true in all cases, making the opposite assumption that they acted as if the scheduled next batter would actually hit suffers from the same problem. No definition can be perfect since in some cases, who bats next depends on what the previous batter does.

Total Plate Appearances. The tables show the distribution of plate appearances by batter strength and next batter strength for each league and in total.

AMERICAN LEAGUE

Batter Strength	Next Batter 0	2	3	4	5	6	Total
0		6	4	1	6	5	22
2	3	51908	32092	32012	27108	23067	166190
3	8	37344	26295	28661	20543	27748	140599
4	5	34946	28742	32803	31071	32806	160373
5	3	25159	26309	33838	30433	33970	149712
6	3	18243	27039	32610	39254	46225	163374
Total	22	167606	140481	159925	148415	163821	780270

NATIONAL LEAGUE

Batter Strength	Next Batter 0	2	3	4	5	6	Total
0		8402	8837	11042	8479	6979	43739
2	26594	42494	22186	24970	14943	16553	147740
3	8638	27670	17438	22954	15831	20098	112629
4	6547	31831	24049	28659	21449	23428	135963
5	1983	20067	19851	24694	19702	21204	107501
6	721	18417	19790	22884	26405	28324	116541
Total	44483	148881	112151	135203	106809	116586	664113

TOTAL MAJOR LEAGUES: 1984-92

Batter Strength	Next Batter 0	2	3	4	5	6	Total
0		8408	8841	11043	8485	6984	43761
2	26597	94402	54278	56982	42051	39620	313930
3	8646	65014	43733	51615	36374	47846	253228
4	6552	66777	52791	61462	52520	56234	296336
5	1986	45226	46160	58532	50135	55174	257213
6	724	36660	46829	55494	65659	74549	279915
Total	44505	316487	252632	295128	255224	280407	1444383

The few cases where strong hitters (5,6) are followed by a pitcher are most likely due to part-time players or late season call-ups who had very good performance statistics for the year. We see that in general weak hitters tend to be followed by weak hitters and strong hitters by strong hitters. Of course, there only so many strong hitters to bat, so there are plenty of cases where weak and strong hitters are adjacent in the lineup. Note that over half the time pitchers are preceded by 2s.

Page 2: Batting Average. The best way to compare quickly the differences and similarities of performance levels is by a graph. The six batting strength groupings appear along the x-axis. Each batter strength group has six bars, one for the batting average of

batters of the indicated strength when followed by each of the six strengths. (There are only five bars for the pitchers because pitchers never follow pitchers.)

In each case, the bar for the batting average when the pitcher bats next is higher than the other five bars for the same batting strength. These differences are statistically significant.[3] One possible explanation for this difference is that pitchers don't want to walk the number eight hitter, who is usually a weak hitter, and give the pitcher a chance to bunt. Consequently, they throw more in the middle of the strike zone. However, as we shall see, *non-intentional* walks are also higher when the pitcher bats next, and slugging average is not affected. Another possible explanation is that the number eight hitters are more selective and more willing to walk.

The heights of the bars in each group when non-pitchers bat next are about the same and there are no patterns to the differences. Statistically, the differences are not significant. Hence, we conclude that batting average is not affected by the strength of the next batter, except when the pitcher bats next.

Page 3: Slugging Average. The heights of the bars in each batter strength group are not much different, nor are there patterns such as the bars getting taller as the next hitter gets stronger. In general, the differences in slugging averages graphed are not statistically significant.

Page 4: Non-intentional Walks per Plate Appearance. There are two significant effects for non-intentional walks.[4] The first is that they are significantly more likely when the pitcher bats next. This is likely due to a combination of "unintentionally intentionally" walking the number eight hitter and greater selectivity on the part of the number eight hitters.

The second effect is a noticeable pattern of fewer walks of batters rated 4 and higher in front of stronger hitters. This effect becomes more pronounced as batter strength increases. Not all of the differences are statistically significant and the pattern is not perfect, but the effect is clear from the graph.

Page 5: Classification of Runners. Now we turn to the effects of faster and better runners. As was the case for batting ability, we need a measure of running ability in order to make the classification. There are several sources of speed ratings, but I wanted to use "standard" data, the type that can be found in the Baseball Guide, for example. That pretty much means using stolen bases. One commonly computed and discussed statistic is stolen base percentage: SB/(SB+CS). Its drawback is that some players have high percentages, but very few steal tries (e.g. 4 of 5, 2 of 2). Also, many fast runners attempt a lot of stolen bases, but are not particularly good at it. My solution is to compute an approximation to how often a player tries to steal when he has a chance (i.e. reaches first):

---

[3] Statistical significance, in this case, means that the differences in batting averages are highly unlikely (less than 5% probability) be due to random fluctuations if, in fact, there is no true difference between the two cases. The calculations depend on the amount of the difference and the numbers of at bats.

[4] *Intentional* walks show the expected patterns: they are much more likely when the pitcher bats next, and more likely when weaker hitters bat next, which to a greater extent, is the pattern shown by non-intentional walks

(SB+CS)/(1B+BB+HBP). This statistic is far from perfect. Not all steal tries are of second, and the runner may be blocked by a runner on second. Also, some fast and good runners just don't try to steal very much. My contention is that the top group consists, on the whole, of much better runners than the middle group, which is turn contains much better runners than the bottom group. Examination of the players in each group (not shown here) bears out this assertion.

Page 6: Advancement on Singles (1). The graphs shown on this and the following pages are somewhat different from those illustrating batting effects. It is more convenient to show the percent of runners who stop at second, so a lower percentage indicates better base running. Hit location affects the chances of advancing beyond second, but the Project Scoresheet database does not have hit location for all plays. Instead, I used batter handedness as a surrogate for hit location. The number of outs can also affect whether the runner tries for an extra base, so that is also part of the grouping. The x-axis shows six groupings by batter hand and number of outs. Each has three bars, corresponding to runner on first speed. Since a slow runner on second can prevent a fast runner on first from going to third, only the cases where second base is open are tabulated.

The differences in the graph are both significant and expected. There is better advancement on singles by left handed batters, faster runners advance to third more often, and the number of outs affects the advancement: runners are slightly more likely to go to third as the number of outs increases. What may be surprising is that in all cases but one, more than half of the time the runner stops at second. (Infield singles are included in the tabulation.)

Page 7: Advancement on Singles (2). This graph is similar to the previous one, and the conclusions are much the same. It is interesting to note that more than 80% of time, runners score from second on two-out singles. Also, it is slightly harder to score when a right handed batter singles, probably due to having to wait to see if some hits will go through into left field.

Page 8: Advancement on Doubles. This graph and the effects are similar to those for advancement on singles. However, there is virtually no difference between left and right handed batters, which is not surprising.

The advancement on hits effects shown probably are no surprise to most of you. I doubt that the next two topics have been quantified as they are here.

Page 9: Avoiding Double Plays. The graph shows the percent of time a ground into double play (GIDP) occurs when there is a runner on first only, which is the purest situation to analyze, with none or one out. The x-axis groups are determined by batter handedness and batter speed. Hitting into double plays probably depends most on whether the batter tends to hit the ball in the air or on the ground or not all (i.e. strikes out), but this information is not part of the standard data in the Project Scoresheet database. I think

it is fair to assume the distribution of flyball vs. groundball hitters is more or less the same by batter hand and batter speed.[5]

Almost all of the differences in the graph are statistically significant. Not surprisingly, we see that right handed batters ground into DPs more frequently than lefties. For this reason, runner speed makes a greater difference when the hitter bats right. Except for the fastest left handed batters, the reductions in GIDPs due to faster batters are not as great those due to faster runners. This suggests that more double plays are foiled by being broken up at second than by the batter just beating the throw.

Page 10: Batter Safe on Error. To avoid complications due to base runners, only situations with no one on are considered. The graph shows the percentages of these in which the batter reaches due to an error (no hit is scored on the play) by batter hand and batter speed. We see that faster batters do reach more frequently on errors. While this may not be a surprise, before doing this study, I wasn't so sure. Infielders tend to play a little closer in for faster batters, so they may not get to as many balls and have slightly fewer chances to make errors. Also, official scorers might base the hit/error decision on a close call on the speed of the batter. Perhaps the advantage results from infielders rushing their throws when the batter is speedy.

Note, however, that batter handedness is more important than batter speed. The slowest right handed batters reached on errors more frequently than the fastest lefties. This shows that the longer throws from the left side of the infield provide are more significant than the advantage the left handed batter has getting to first.

Page 11: Conclusions. I am in the process of revising my basic run scoring Markov model, and the data shown above will be incorporated.

My conclusion is that the basic assumption that batting performance is not affected by the strength of the following hitter is still valid. Leaving aside the effects when the pitcher bats next (which affects one batter per team in one league and usually not for the whole game), we see that batting and slugging averages are not affected, but strong hitters draw slightly more walks when followed by weak hitters. Additional walks lead to additional runs even when weaker hitters follow. (Note that Barry Bonds is one of the league leaders in runs scored while batting fifth. Getting on base, which means not making outs, is critical to scoring.) My revised model will account for these additional walks.

There is no doubt that the advantages of faster runners are real, which is hardly a surprise. What is new, is that some of these have now been quantified and can be incorporated into mathematical models.

Statistical significance is useful for telling us that something not due to random fluctuations is taking place, but it is far from the whole story. Many of the statistically significant differences are small, especially the chances of reaching on an error. The important issue is how much these differences effect run scoring. My revised models will be able to answer that question. I hope to present some of the answers next year in Arlington, Texas. See you there!

---

[5] If this is not so, it could well be because fast players, especially left handed batters, may try to hit the ball on the ground to get more infield hits. Such behavior would lessen the advantage fast batters have in avoiding GIDPs, which would make the batter speed effects shown even more significant.

I am always looking for feedback, comments, methods for improving my analysis and models, and research ideas. Please feel free to contact me.

Mark Pankin
1018 N. Cleveland St.
Arlington, VA 22201
(703) 524-0937