On Why Teams Don't Repeat

Phil Birnbaum

This article originally appeared in "Baseball Analyst," February, 1989.
birnbaum@sympatico.ca, *www.philbirnbaum.com*

1. Introduction

The assumption that luck is a factor in baseball performance allows us to deduce one reason pennant-winning teams tend not to repeat: they usually won their pennant with the help of a good deal of luck, and luck tends not to hold from year to year.

While random chance is a factor in helping any team to the record it posts, it is *more* likely, in retrospect, to have helped teams which posted good records than teams which posted bad records: and, conversely, teams which performed badly are more likely to have been hurt by bad luck than teams which finished .500 or better.

We can phrase this effect as follows:

> The majority of extreme achievements (good or bad) have been substantially aided by luck; that is, the talent that produced the achievement is not as good (or bad) as the achievement itself.

Rephrasing that as an example, of teams that win more than 100 games (an extreme achievement), most are teams that were not talented enough to win 100 games, but did so by luck. The same applies to players; of players who hit 40 home runs in a season, most were players who were not legitimate 40-home-run hitters, but just had a lucky year, and of players who have had ERAs of less than 2.00 in a season, most achieved that performance by a fair deal of luck: their talent was not alone enough to carry them to that mark.

In this essay, we'll deduce and quantify the above effect, thus giving a major reason why teams and players don't repeat: they just weren't good enough to achieve what they did in the first place.

2. Why

For this study, we assume that for every team (or player), there exists a talent level for that team which represents its probability of success. For example, we assume that a team with 100-game talent will, independently for every game, win that game with probability 100/162, and a player with 35-home-run talent will hit a home run in a particular at-bat with probability, say, 35/600. While these assumptions are not completely accurate -- a 100-game team will have a better than 100/162 chance of winning against a poor team and less than 100/162 chance of winning against a good team -- they are close enough and reasonable enough to not affect the conclusions that will follow. Also, when we refer to a lucky team, we are talking about one which has exceeded its talent (say, a 100-game talent that wins 102 games); when we say a team is unlucky, we mean it failed to realize its talent (a 100-game talent winning only 96 games).

Now, suppose a team wins exactly 100 games in a season. What is the expected talent level of that team?

Because of the effects of luck, that team is probably not exactly a 100-game talent (almost certainly not, considering that in theory, team talent can be something other than a whole number of games, such as 99.94 or 100.12). We're therefore looking at a team that is either (1) a less-than-100 game talent that got lucky, or (2) a more-than-100-game talent that got unlucky. But it is a fact that there are many, many, more less-than-100-game talents than there are more-than-100-game talents. There are probably several times as many teams capable of playing 95-game baseball than there are 105-game teams, for instance. So of teams that wind up with records of 100-62, there will be many more who were 95-game talents who got lucky than there will be 105-game talents who got unlucky, and the average talent of 100-62 will be less than 100 games.

The same reasoning applies to hitters; a .333 hitter, say, is either a .300-.332 talent who got lucky, or a .334+ talent who got unlucky. But the distribution of batting talent follows the tail-end of a normal curve; there are so many more talents worse than .333 than better that it is virtually certain that our .333 hitter would normally hit worse than .333.

Another way to look at it is that good luck moves teams towards the top, while bad luck moves teams towards the bottom. So, if you sample teams at the top, you'll find for the most part that they were lucky, because the teams that weren't moved to the bottom.

You can try an experiment with real data. Take a season's worth of actual runs / runs created data. Sort the teams by runs created, which, in a very rough way, represents the team's talent. Assign each team a "+" if it exceeded its RC estimate, and a "-" if it scored fewer runs than the prediction. The +'s and -'s will probably be distributed pretty evenly. Then, sort the teams by actual runs scored, which represents the achievement. Since the +'s move up and the -'s move down, you should notice the +'s are concentrated among the teams in the top half of the league. Again, the reason: luck moves mediocre teams up to the top among the genuinely good teams, while (bad) luck moves talented teams down to the ranks of the less talented.

3. The calculation

We have seen, so far, that the typical 100-game-winning team is actually a team of lesser talent than 100 games. But in order to conclude that this effect is the reason teams don't repeat, we need to determine what the expected talent of such a team is. If the average 100-62 team is actually of 90-72 talent, then we have a plausible explanation for its failure to repeat. If, however, such a team has average talent of 98-64, that doesn't explain why it fails to win even those 98 games the following year. To determine how responsible the effect is for failure to repeat, we need to get some kind of estimate of the magnitude of the effect, although we have already shown why it must exist.

Graph #1 (next page) shows the distribution of team wins in the American League over the years 1961-1984 (1981 excluded). What may be hard to see from the graph, because of the irregularity between 75-80 wins, is that the distribution is roughly bell-shaped, but with a peak at around 85 wins rather than the peak at 81 we would find if the distribution were strictly normal. The peak occurs in the mid-80's because of a tendency by teams to stabilize at a level where they have a chance at 90-95 games and a pennant (see 1985 Abstract, Blue Jays comment). The exact number of seasons per win level are listed in the table on the last page. All team-seasons of more or less than 162 games (including 1972 seasons) were expressed in wins per 162 games.

The above distribution is that of actual wins, or achievement; it follows that the distribution of team talent is different. That's because the effect of luck is to spread out the achievements away from the talents. A 35-home-run hitter may hit 40 or 45 in a season if he has a lucky year, even if there isn't anyone in the league with strict 40 homer talent.

We can estimate the distribution of talent by finding a talent function that would theoretically produce the above team-win distribution when the seasons were actually played out (mathematical details are in the notes). I did this by taking a first guess, calculating, and then adjusting my guess until what I got looked like it approximated the above team-win distribution. What I finally settled on looks like graph 2 (superimposed on the actual team win distribution above).

A few things are immediately apparent from the graph: primarily, the talent distribution is clustered around the centre much more than the actual-win distribution. There are many teams who actually won more than 100 games in a season (far right of actual-win curve), but almost none who actually had the talent to do it. Also, there are many more teams whose talent is around 81-90 games than there are teams who actually won that many; again, random chance spreads the achievements away from the talent. The talent curve here is not perfect, though; there are probably theoretical reasons why the graph should look smoother and more bell-shaped. The conclusions of this study, however, should not be affected by this failing.

Again, the reason I settled on this talent distribution was that it predicted the actual distribution fairly well. Graph 3 is the actual-win graph superimposed on the predicted-win (predicted by the talent distribution, that is) graph. It seems to fit fairly well, considering that under our assumptions, the predicted-win curve must be round and bell-shaped. The numerous peaks and valleys in the actual-win curve are almost certainly caused by chance.

If there is significant error in the curves, it probably occurs in the far left (50-65 wins) and far right (97+ wins) portion. Because there have been so few teams to finish with very low or very high winning percentages, the actual-win curve is very jagged at these points, and no theoretical curve can match it very well. You might move my talent curve up or down at either end if you feel my theoretical-win curve lies too low or high with respect to the actual. Such changes, though, won't substantially change the results presented here, I think.

Again, numeric values are presented in the table. Just to read off a few figures, of 264 American League teams between 1961 and 1984, 5 finished with seasons of exactly 97 wins (after normalizing to 162 games and rounding to the nearest game). The talent curve says that of these same teams, only 2.4 had talent of exactly 97 wins, and the theoretical distribution predicts that 3.9 teams "should have" finished with exactly 97 wins. Also, 3 teams finished with records of 60-102, while only an estimated 1.9 teams had exactly that talent. And so on.

The talent distribution shown above and in the table is the main result of this study; all results and conclusions to follow are based on this talent distribution.

---

4. The results

Having obtained an estimate of the talent distribution, we can derive the average talent of a team with a given record. Some of the results (full results in the table):

Teams winning 85 games have average talent of about 85 games.
Teams winning 90 games have average talent of about 88 games.
Teams winning 100 games have average talent of about 93 games.
Teams winning 108 games have average talent of about 98 games.

The results substantially support the premise that teams with extreme records are lucky -- the average 100-game-winning team is actually only good enough to have won 93, while the average 108-game-winning team is only, on average, a 98-game team. What this really means -- I haven't stated it explicitly yet -- is that if the same 108-game-winning team, with the same players, were to play the same season over again, it would, on average, win only 98 games, 10 less than it actually did. It means that the team exceeded its talent by 10 games not because of leadership, nor heart, nor managerial skill, but only because of random luck.

Of course, when we say that the average 108-game-winning team is actually a 98-game-winning talent, we're not saying that's true for every team that wins 108 games. Some 108-game winners may actually be 108-game talents, some may be 95-game talents, and some may even be 85-game talents that were extremely lucky. All we're saying is that if you took a large number of teams that won 108 games, their average talent would be about 98 games. The actual calculation of how many 108-game winners are actually 108-game, 98-game, 85-game, etc. winners appears below (full data in table, again):

40-86	87-89	90-92	93-95	96-98	99-101	102-104	105-107	108-110	111+
0.9%	3.7%	11.0%	16.2%	24.3%	22.7%	14.6%	4.9%	1.8%	0%

Note that the chance a 108-game-winning team is actually a legitimate 108-game-talent is less than 2%, while a rather large 32% of such teams weren't even good enough to win 96 games. This result, again, strongly suggests that teams don't repeat because they were just lucky the first time.

5. Making a conclusion

We have established that the average, say, 100-game-winning team is around a 93 game talent, although it could actually be more talented than 100 games or less talented than 90 games. But for a particular team, how can you tell which it is? Were the '84 Tigers really a 104 game talent, or were they a 100-game talent or a 93-game talent?

Well, when a team wins more games than its talent indicates, there are at least five reasons why:

1. It produced more offense than its talent should have; that is, individual players got lucky and hit more home runs than they "should have", or more doubles, or more walks, or had a higher steal percentage, etc. This is the "career year" effect.

2. It produced more runs from its offensive components than it should have; that is, it exceeded its runs created estimate.

3. Same as #1, but for pitching: the team's *opponents* produced fewer offensive accomplishments than they should have. That is, the pitchers of the team in question had "career years".

4. Same as #2, but for pitching: the team's opponents scored fewer runs than their runs-created estimate predicts.

5. The team produced more wins than it should have given its number of runs scored and runs allowed; that is, it exceeded its pythagorean projection.

Numbers 2, 4, and 5 should theoretically be easy to check: just line up the particular team and see if it exceeded its projections, and by how much. You should be able to line up all teams in any period of years who won more than 100 games, and find that on average, (a) they exceeded their runs created estimates, (b) their opposition scored less than their runs created estimates, and (c) they exceeded their pythagorean projection. I did this for (c), and found that the AL 100-game-winners

since 1961 beat their estimates by an average 3.5 games. (You have to use the 1.83 coefficient in the pythagorean formula, because it's otherwise inaccurate for very good or bad teams.)
For the '84 Tigers, (a) they actually *undershot* their runs produced estimate by 14; (b) their pitching staff allowed 24 *more* runs than their estimate; but (c) they overshot their pythagorean projection (exponent 1.83) by *five-and-a-half* games. So far, then, we estimate their "luck" as having been about a game and a half (5.5 from pythagoras minus a games for the 38 runs of bad luck).

For point 1 and 3, whether the players had career years, you have to sit down and figure out. I would work it by taking each player's age and career record prior to 1984, and estimating what a reasonable expected 1984 performance would look like (in fact, since this is 1988, we have post-1984 data to use in projecting estimated '84 performance) and comparing it to the player's actual 1984 performance. I won't go through that, because I'm not very good at that kind of thing, but you might look at the bench, Willie Hernandez, etc. Note that the effect is not limited to huge career year effects -- if everyone on the team gets lucky by even just two home runs, you've got maybe three extra wins right there.

So, get your best estimate for #1 and 3, get your numbers for #2, 4, and 5, and add them up. That's your estimate of how lucky the team was. If you do that for a few different teams, you'll get an estimate of, when a team gets lucky, how the luck manifests itself. Does the luck show up, on average, as, say, 40% in runs created, 20% by Pythagoras, and 30% by the players, or what? Actually, it's probably roughly proportional to the standard deviations of the deviances from expected, so you could figure out which deviates most, expected pythagoras minus actual, expected runs minus actual, or expected performance minus actual. I suspect player performance would be the biggest, but that's just a guess.

6. Principles

The effect we have been discussing here -- for the purposes of this paragraph, we'll call it the Extremes Effect -- is similar to two other effects discussed by Bill James: the Whirlpool effect, and the Plexiglass principle.

The Whirlpool Principle (1983 *Abstract*, p. 220) states that "all teams are drawn forcefully towards the centre"; winning teams will tend to decline towards .500 in subsequent seasons, and losing teams will tend to improve towards .500.

The Whirlpool Principle is caused by two factors: One, the Extremes effect, which states that any team is closer in talent to .500 than its record indicates; and second, the *talent* of teams tends to approach .500 as time goes on. The reason for the latter is obvious; good teams tend to have good players, who, if they stay in the lineup, will eventually decline in talent, and who, if they don't stay in the lineup, will have to be replaced by worse players (If Wade Boggs gets hurt, there's nobody at his level of talent to replace him). Bad teams, on the other hand, will get rid of mediocre players and replace them with (better) young talent.

We can see how much of the Whirlpool principle is caused by the first (extremes) effect and how much by the second, as follows: take all AL teams from 1961-84 (excluding 1980-81) and look at how they did in the following year (teams were grouped in 5-win groups starting at 53-57 wins -- the first row below is the average of the group):

Wins in 1st year:	56	61	65	70	76	80	85	90	94	99	104	109
Predicted talent:	62	65	69	73	78	81	85	88	90	92	95	98
Wins in next year:	69	70	72	74	76	83	83	88	90	88	93	103

These results pretty much agree with the Effect; the weak teams improved beyond their expected talent, while the strong teams declined past theirs. The pattern is not quite as strong for the good teams as for the bad, but it still holds (except for the 109 group, which consists of only three teams). If we are to trust the theoretical values, it appears that most of the whirlpool effect is caused by the extremes effect at high or low levels of achievement.

The Plexiglass Principle states that any team that has a large gain (decline) one season will tend to have a decline (gain) the next. There are three reasons for this. Two are the ones described above in the explanation of the whirlpool principle: that the team was probably lucky and talent changes tend to move towards .500. The third is that the average team that has a large gain (decline) over only one season has often done so out of luck, since personnel changes tend normally not to produce a large season-to-season fluctuation: a team's talent normally doesn't change drastically in only one year. Teams that show a large single-season fluctuation are thus *particularly* lucky, and since that luck doesn't necessarily carry on to the next season, many teams bounce back. (It would seem, then, that if there is an obvious reason for the large fluctuation, such as a free-agent gain or loss, or two, the bounce-back should be smaller. I haven't done the necessary research to check that, though.)

---

7. A few notes

- The basic effect here applies to individual players, as well as teams. Mark McGwire hit 49 homers last year. That's an extreme event, and there aren't very many 49-home-run hitters in existence, so, knowing nothing else about Mark McGwire, we would assume that his achievement was helped by a substantial amount of luck. I say "knowing nothing else" because for an individual, you have a much better idea of his level of talent than you do for a team. For Mark McGwire, we have previous major and minor league batting records, scouting reports of his talent, etc, that suggest that before 1987, nobody thought McGwire was an extraordinary power hitter at all. We might therefore conclude that his 49-homer performance was the result of luck based on the prior evidence of his talent, which suggests a figure in the range of 30 home runs, perhaps. For a team, on the other hand, since it is probably in its current state for only a year, what with trades, retirements, injuries, etc., we haven't had time to form independent notions of how good the team is, and we must usually resort to such analytical methods as described in section 4 to make a conclusion about the team's actual talent.

- The effect applies in reverse to bad teams, or bad players. Just as a 100-game winning team is not as good as its record indicates, neither is a 60-game winning team as bad. Note that the numerical results in the table may not be applicable because expansion does strange things with the distribution of teams at the low end. I think they're reasonably accurate, nonetheless. Anyway, since the principle applies at both ends, we could rephrase the effect as "On average, a team's (player's) talent is closer to .500 than its record indicates, and the difference is due to luck."

- The effect provides a suitable explanation of the sophomore jinx. Since very seldom does anyone talk about the sophomore jinx with respect to a rookie of modest first-year stats, the "disappointing" second seasons match up with extreme rookie accomplishments which were probably caused, in part, by luck.

- The following table shows the chance of a team repeating or improving its record next year, assuming its talent stays the same as this year. Note that because a team's talent will almost always change, these numbers understate the actual chance for the bad teams, and overstate the actual chance for the good teams. It might therefore be better to think of this as the chance a team would have a better record if it could play the *current* season over again.

The top number is the number of actual wins in the season; the bottom is the probability of reaching or exceeding that number.

50	55	60	65	70	75	80	85	90	95	100	105	110
.88	.81	.73	.68	.64	.62	.58	.52	.41	.29	.21	.14	.08

Table

I'll explain how to read the table by explaining one row of the table, the row that starts off "97":

Column A: The average team that won 97 games actually had talent of 91.32 games.

Column B: Of the 264 AL teams from 1961-84 (excluding 1981), exactly 5 finished with exactly 97 wins (after normalizing to 162 games and rounding).

Column C: The theoretical talent distribution predicts that of 264 teams, 3.86 teams should finish with 97 wins.

Column D: Of 264 teams, 2.36 should have talent of exactly 97 games (actually, talents of 96.5 - 97.5 games).

Column E: In theory, of teams that win 108 games, 7.96% of them have talent of exactly 97 games.

Of course, none of the numbers are really accurate to anything approaching two decimals.

Wins	A	B	C	D	E
40	54.12	0	0.00	0.00	0.00
41	54.47	0	0.00	0.00	0.00
42	54.84	0	0.00	0.00	0.00
43	55.23	0	0.00	0.00	0.00
44	55.63	0	0.00	0.00	0.00
45	56.05	0	0.11	0.00	0.00
46	56.49	0	0.15	0.00	0.00
47	56.95	0	0.19	0.00	0.00
48	57.43	0	0.25	0.02	0.00
49	57.92	0	0.32	0.04	0.00
50	58.43	0	0.41	0.09	0.00
51	58.97	0	0.51	0.14	0.00
52	59.52	0	0.62	0.19	0.00
53	60.09	1	0.76	0.34	0.00
54	60.68	2	0.91	0.49	0.00
55	61.29	0	1.07	0.63	0.00
56	61.92	2	1.25	0.86	0.00
57	62.57	4	1.45	1.04	0.00
58	63.24	1	1.67	1.24	0.00
59	63.94	3	1.89	1.57	0.00
60	64.65	3	2.13	1.90	0.00
61	65.39	0	2.39	2.17	0.00
62	66.15	4	2.65	2.45	0.00
63	66.92	2	2.93	2.76	0.00
64	67.72	5	3.22	3.08	0.00
65	68.53	3	3.52	3.32	0.00
66	69.36	2	3.82	3.58	0.00
67	70.20	6	4.14	3.88	0.00
68	71.05	5	4.46	4.17	0.00
69	71.91	4	4.79	4.53	0.00
70	72.78	7	5.13	4.87	0.00
71	73.65	6	5.48	5.26	0.00
72	74.52	7	5.83	5.60	0.00
73	75.38	2	6.18	6.11	0.00

74	76.25	7	6.54	6.61	0.00
75	77.11	8	6.89	6.94	0.00
76	77.95	12	7.24	7.30	0.00
77	78.79	11	7.57	7.64	0.00
78	79.61	2	7.89	7.96	0.00
79	80.40	10	8.18	8.34	0.00
80	81.18	5	8.45	8.70	0.01
81	81.93	9	8.67	9.24	0.02
82	82.66	6	8.85	9.79	0.03
83	83.37	7	8.96	10.24	0.06
84	84.04	6	9.01	10.87	0.12
85	84.69	10	8.99	11.23	0.23
86	85.32	11	8.89	11.51	0.41
87	85.92	10	8.71	11.69	0.71
88	86.51	5	8.45	11.78	1.18
89	87.07	11	8.11	11.33	1.85
90	87.63	6	7.71	10.33	2.68
91	88.16	9	7.24	9.06	3.64
92	88.69	7	6.72	7.68	4.66
93	89.22	3	6.16	6.16	5.53
94	89.74	5	5.58	4.17	5.40
95	90.26	2	5.00	2.90	5.29
96	90.79	3	4.42	2.61	6.55
97	91.32	5	3.86	2.36	7.96
98	91.86	8	3.34	2.21	9.76
99	92.40	4	2.85	1.48	8.36
100	92.96	2	2.40	1.09	7.69
101	93.52	0	2.00	0.77	6.61
102	94.09	3	1.65	0.53	5.40
103	94.68	2	1.34	0.43	5.07
104	95.26	3	1.08	0.31	4.12
105	95.85	1	0.86	0.17	2.48
106	96.45	0	0.68	0.09	1.41
107	97.04	0	0.52	0.06	0.98
108	97.63	1	0.40	0.05	0.83
109	98.22	1	0.31	0.04	0.65
110	98.79	0	0.23	0.02	0.31
111	99.36	0	0.17	0.00	0.00
112	99.92	0	0.12	0.00	0.00
113	100.47	0	0.09	0.00	0.00
114	101.01	0	0.06	0.00	0.00
115	101.53	0	0.04	0.00	0.00
116	102.03	0	0.03	0.00	0.00
117	102.53	0	0.02	0.00	0.00
118	103.00	0	0.01	0.00	0.00
119	103.46	0	0.01	0.00	0.00
120	103.90	0	0.01	0.00	0.00

How I got most of this stuff

The first thing I had to do was come up with a talent distribution, t(x), where t(x) is the number of teams out of 264 with talent x games. For example, if we set t(80)=8.4, we're guessing that 8.4 out of 264 teams have talent of exactly 80 games.

Now, let g(x,n) = out of 264 teams, # of teams of talent x winning exactly n games. By binomial theorem,

$$g(x,n) = t(x)C(162,n)\left(\frac{x}{162}\right)^{n}\left(\frac{162-x}{162}\right)^{162-n}$$

And so the total number of teams winning n games, which we'll call a(n), is given by:

$$a(n) = \sum_{x=0}^{162} g(x,n)$$

I chose t(x) in such a way that the above numbers a(n) matched the actual number of teams winning n games. (Compare columns B and C of chart.)

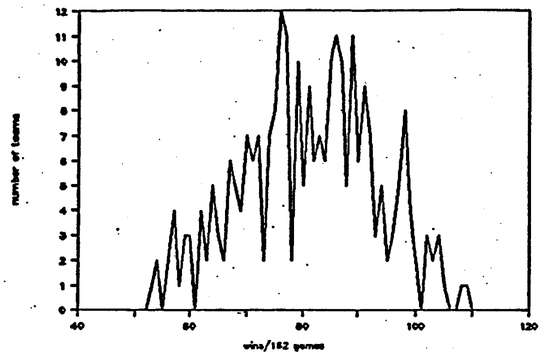
Then, to find the average talent of n-game-winning-teams, we just average g(x,n) weighted by x:

Average talent of teams that win n games =

$$\frac{0*g(0,n)+1*g(1,n)+...+161*g(161,n)+162*g(162,n)}{g(0,n)+g(1,n)+...+g(161,n)+g(162,n)}$$

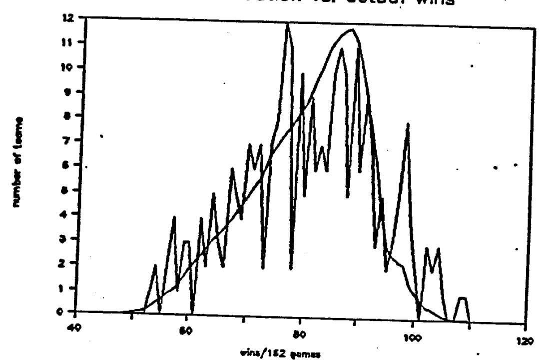
1

Actual win distribution: AL, 1961–84



2

Talent distribution vs. actual wins



3

Actual win distribution vs. projected