A Bullpen is Not Just for Protecting Leads: Relief Strategy Reexamined.

Bob Sawyer (Baseballnut570@hotmail.com)

In his 2004 SABR34 presentation, "Coming from Behind: Patterns of Scoring and Relation to Winning"[i], David W Smith drew the far-reaching conclusion that the advent of the closer had very minimal impact on late inning success. This conclusion is based upon a study of lead-retention percentages over 73 seasons dating back to 1901. Graphing victory percentage does not show a trend of greater success in recent years for the front-running team than in the 1950s. Based on this data, Smith and his disciples on this issue have advocated that closers and other aspects of the modern bullpen represent stratagems without adequate payoffs.

A more sophisticated look at the graph Smith presented to support these claims shows that each of the three sets of data points has some weak, yet discernible trend. The overall trends shown in figure 3: Winning percentage for teams leading after 1, 4, or 8 innings (all margins combined) from 1901 – 2003[ii] (fn1) are for reduced percentages of victory for teams holding leads after the 1st and 4th innings, and a very gentle arc that in recent years favors improved victory percentage by teams with leads at the end of the 8th inning. Smith did not publish the numerical data for a least-squares error minimization calculation, but his published Figure 3 graphs makes these trends fairly obvious to anyone with experience in curve-fitting

The upward arc indicates that today's game finishers are slightly more successful in the 9th than pitchers of the past, but the difference is so small that it not unfair to say that any concerns about closer quality are much ado about next to nothing. The downward trends on the other two portions of figure 3 present us with a mystery. Given the increasing amount of attention devoted to bullpens in post-war Baseball, why would taking the lead be a less reliable way to win now than it was in 1901?

Other sports provide a ready answer to this puzzle. Scoring first is an extremely sure way to win a World Cup soccer game, but is of very little importance in an NBA Championship game. Scoring first is very useful in Football and Hockey, but not nearly as much as in Soccer at the highest level. The differences in importance between the four sports largely results from differences in the frequency of scoring plays. Nonetheless, until scoring opportunities have all but disappeared, both basketball teams and soccer teams attempting to comeback will always be told by their coach to put more effort into defense. Regardless of the sport involved, the chance for a comeback win diminishes toward zero if the team with the lead keeps on scoring regularly.

At the dawn of the 20th century Baseball teams would lose 80 percent of the time when they trailed in a game. Today they lose only 78% of the time. Are baseball teams doing something differently today that might account for that? Well, pinch hitting became steadily more common over the decades, as did mid-inning pitching changes. And of course there is the increasing standard practice of using the bullpen to protect leads. One of the greatest changes has been reduced reliance on starting pitching , especially by the team which is trailing. This change tended to get overlooked in the midst of those mentioned above because it was already well under way by the mid-1920s. When was the last time you saw a starting pitcher sent out to work the 9th inning when his team was trailing by more than one run?

Early in the 20th Century more than 75 percent of all starts resulted in complete games. Yet no more than 50% of starts can result in team victories. Therefore the 1900s was a period when at least 50% of all defeated teams were getting complete games from their pitcher. We see in the box scores of World Series games that sometimes a starting pitcher returned to the mound for another inning even though his team was behind by three runs. We may be reasonably sure without checking that many, many times a starting pitcher stayed on the mound during regular season even though his team trailed by two runs or more.

The precepts Game and Decision theory tells us that such a practice can sometimes be the optimal strategy, but only if one of two conditions is met. If there is absolutely no substitute pitcher from whom better performance can be expected, then it makes sense to leave the current pitcher in the game. Tabletop games such as this author are quite familiar with this situation. It can also make good to leave the starter or current reliever out there to get pounded when the advantage from bringing in someone else would be too small to justify "using up" further pitchers. These are games that must be treated as being either "already in the bag" or "as good as lost" and the managers attentions turn to how to maximize victory in subsequent games. Under all other circumstances, sticking with a pitcher who is demonstrably ineffective is a suboptimal strategy.[iii]

Aside from game theory arguments, how could we tell whether a pattern of pitcher usage makes sense? David W Smith believed that team won-lost records provide the ultimate evidence of pitching strategy effectiveness. Smith was wrong about that. His investigative approach only works when there is no opportunity for a counter strategy by the opponent. As we will see, even for the bottom of the 9th inning situations, the other team's defensive strategy can change the numbers on which Smith focused.

It is readily apparent when playing a tabletop simulation game that employing relief specialists does change the percentage of runs scored in the later innings of a game. When the relief pitchers are higher quality than a tiring starter then scoring by the opponent is reduced. On the other hand, if real world relief pitchers are actually less efficient than the pitchers they replace, then more of the scoring would be taking place in the later innings.

Therefore I elected to investigate whether the inning-by inning scoring pattern given in Smith's 2004 presentation was stable from decade to decade. My findings were that it was not. There has been ample variation in come from behind wins between five and six year periods. And furthermore, there is a very clear historical trend toward reduced scoring in the late innings.

SECTION II

At the start of his article and presentation Smith presents a table taken from the conglomeration of data from 122,906 official games spread over 73 seasons. Just under one percent of those games terminated before the 9th inning and another 11,513 games (9.37%) went into extra innings. During innings one through eight, the home team scored between 7% and 18% more frequently than the visitors did—depending on the inning. During extra innings, the home team scored only 73% as many runs as the visitors scored.[iv]

This last phenomena is explained entirely by the victory conditions of Baseball. Except in the case of over-the-fence home runs after 1920, the team batting last wins in a walk-off as soon as the go-ahead run scores. Smith's sample included 11,513 extra-inning games, of which 1.624% officially ended in a tie. The visitor won by a single run in 25.7% of extra inning games and by more than one run in 21.3%. Home teams won 47.2% of extra-inning games by a single run and just 4.11% of the time by more than one run.[v] Smith's sample is missing some high scoring seasons just prior to WWII. If these were added in we would expect a higher percentage of wins by more than one run. Adding only seasons prior to 1920 would shift these ratios in the opposite direction.

The walk-off rule likewise pollutes the data for home team scoring in the 9th inning. Home teams dropped from a collective .518 per previous inning all the way to .415 per inning whereas visiting teams were held to .450 runs per inning (down from .470). The Home team bats in the 9th only when tied or trailing so part of the explanation could the quality of the opposing pitchers. On the other hand, as indicated by the gap during extra innings (.457 for visitors to .334 for home,) at least part of the difference will be due to mid-inning walk-off victories. Whatever the explanation, it is beyond denial that home team's offense is far more effective in the 9th inning than is indicated by percentage of runs being scored in the bottom of the 9th. This consideration shaped the methodology of my study.

Smith's posted Figure 3 is presumably taken from a slide from his 2004 Convention presentation. The vertical scale is the percent chance of ultimately winning the game by a team with any kind of lead. All the data points for an 8th inning lead fall between 92 and 97 percent. Some teams, such as the 1954 Indians never lost a game in which they led after 8 innings. The worst seasonal performance in Smith's sample was 41 wins and 10 losses for the 1978 Seattle Mariners.[vi]

The graph for 1st inning leads has a spread from 66 percent to 73 percent. This total was right on 70% from the late thirties to the early sixties but clearly dropped below 70 percent for more recent decades. The graph for 4th inning leads is difficult to read correctly. Based on the peaks of the triangles marking data points the spread seems to run from 76% to 81%. Using the bases of the triangles yields a range from slightly more than 75% to just over 80%. Either way, over the decades there has been a gradual yet undeniable long term decline in victory frequency on the order of two percent.

David W Smith does not directly say so in his paper, but he encourages his readers to draw the conclusion that major league teams appear to be either mishandling their pitching staffs, or devoting a great deal of thought to something that makes little difference. In a direct remark to this author at SABR40 Smith said his study showed that bullpens of 1992 or 2002 were no more effective than those of 1952. In his paper, Smith expounds as follows.

"The pattern is striking. Teams which lead after one inning win nearly 70% of the time and the winning percentage gets consistently better with each passing inning. A lead after 8 innings will hold up 95% of the time. This is a point worth keeping in mind the next time you hear an announcer praise a team for the way its great bullpen gets the job done.[vii]

A lead of four runs at the end of 8 has been converted to a win 99.5% of the time. The raw numbers are 44324 wins and only 213 losses in the 73 seasons examined.
Of course, a modern reality is bullpen specialization with every team feeling obligated to find a closer to handle these situations. I, therefore, analyzed this question across the last century to look for any obvious patterns. As Figure 3 shows, there are no discernible differences from 1901 through 2003 (data shown for leads of all sizes after one, four, and eight innings). What this says about the necessity of a closer is interesting, but not our main topic here.[viii]

Conclusions
(1st three conclusions skipped)
• The advent of the closer has not changed late inning success[ix]

It is important to realize that Smith's graph is not showing league-wide blown save percentage. For one thing, when scoring levels are high, a large number of leads will be too large to count as save opportunities. For another, the winning team may have its pitchers blow the save more than once and yet emerge victorious. While we may be certain that a lead of any particular size would be easier to make stand up when runs are scarce. Smith's figure 2 is not about protecting leads as this concept is normally understood when talking baseball. The data points that make up the graphs have to do with the frequency at which the team with the lead at a given point emerged ultimately victorious.

Smith inferred from the lack of change in ultimate victory percentage that relief pitchers as a group were not accomplishing anything of significance. We know from experience and calculations by Markov chains that extra runs do matter—(at least until the lead is more than 10 runs). So which is right? Both are. Imagine a see-saw with people on both sides of the center shifting outward with the intent of stranding the other up in the air. Smith's Figure 3 records the angle of the see-saw for each season and finds little change. This study measures the distance from the center for the two parties and finds that the distance has changed. And from the small trend in Smith Figure 3, we learn that the trailing team has made more effective adjustments than the team with the lead.

PART III. Analysis of Scoring Patterns

Smith's Table 2 gave the conglomeration of 73 seasons of data; seasons that did not include the early twenties. I did my study to establish whether or not the shape of scoring was the same in specific time periods as in Smith's Table 2. When measured on a percentage basis, there has been a steady decline in scoring for the 7th, 8th and 9th innings.

My study of scoring patterns utilized some shortcuts because I had no ready access to how many times a team batted or pitched into extra innings. I excluded both these innings and the runs that were scored in them from the sample. Rather than guesstimate how many innings were involved and then deal with the effects of the walk-off rule, I opted to exclude the bottom of the 9th inning as well. Rather than try to cut and paste for every single team season I selected four franchises from each league and followed them for decades or half decades. Runs scored and allowed were totaled for each inning and then divided by the grand total for the 17 half innings not effected by the walk-off rule. This division should have the happy effect of normalizing to the scoring level of the time period.[x]

TABLE ONE. Scoring distributions over time: portion of runs from the 1st through top of 9th

PERIOD	# TEAMS	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
Random		.1176	.1176	.1176	.1176	.1176	.1176	.1176	.1176	.0588
1871-1875	23	**.1349**	.1009	.1104	.1156	.1161	**.1274**	**.1272**	.1150	.0535
1894-1902AL	40	**.1502**	.1111	.1291	.1110	.1141	.1157	.1138	.1131	.0419
1901N-1910	40	**.1505**	.1038	.1163	.1138	.1134	**.1159**	**.1120**	.1146	**.0598**
1911-1914	40	**.1469**	.1035	.1167	**.1115**	**.1085**	**.1208**	.1168	.1174	.0579
1915-1919	40	**.1424**	.1033	.1142	.1141	.1115	.1198	.1115	**.1275**	.0556
1920-1924	40	**.134**1	.1014	.1130	.1198	.1141	**.1200**	.1193	**.1204**	.0579
1925-1930	48	**.1304**	.1002	.1184	.1152	.1179	.1198	.1175	**.1240**	.0564
1931-1935	40	**.1367**	.1058	.1159	.1128	.1125	**.123**7	.1161	.1218	.0547
1936-1941	48	**.1333**	.1028	.1147	.1177	.1143	.1164	.1196	.1125	.0587

1942-1946	40	**.1370**	.1006	**.1123**	**.1214**	.1180	**.1204**	.1179	.1173	.0552
1947-1952	40	**.1339**	**.0990**	.1154	.1175	.1191	.1173	**.1223**	**.1215**	.0541
1953-1957	40	**.1337**	.1026	.1124	**.1222**	.1145	.1211	**.1203**	.1196	.0536
1958-1962	40	**.1305**	.1016	**.1210**	.1196	.1151	.1240	.1182	.1157	.0543
1963-1968	48	**.1336**	.0993	.1138	.1176	.1176	**.1240**	.1182	.1187	.0573
1969-1973	40	**.1362**	.1026	.1135	**.1207**	.1124	**.1242**	.1178	.1176	.0548
1974-1978	40	**.1359**	.1084	.1157	.1171	.1189	.1176	.1181	.1163	.0521
1979-1982	32	**.1338**	.1081	.1173	.1171	**.1217**	.1185	.1157	.1155	.0524
1983-1987	40	**.1294**	.1044	.1195	**.1211**	.1192	**.1216**	.1143	.1175	.0530
1988-1992	40	**.1383**	.1032	.1181	**.1256**	**.1214**	**.1204**	.1131	.1101	.0499
1993-1997	40	**.1321**	.1019	**.1202**	**.1209**	**.1213**	**.1233**	.1145	.1119	.0539
1998-2002	40	**.1276**	.1084	**.1203**	**.122**5	.1183	**.1254**	.1178	.1102	**.0495**
2002-2006	40	**.1331**	.1063	.1196	.1197	.1169	**.1250**	.1174	.1103	.0517
2007-2013	48	**.1292**	**.1097**	.1157	.1233	**.1243**	**.1256**	.1140	**.1076**	.0508

The highs and lows have red or green print and value well above random are in bold face. This made it somewhat easier to identify trends. While the entries represent percentages of the sample, they are best read as effectiveness ratings. The lower the number the more effective were the pitchers collectively working that inning. The 9th inning numbers are much lower than half of the other columns because only scoring by the visitors was included, and visitors are typically outscored by the home team on a per out basis. Hence the 9th inning numbers are best understood as using a different scale

As expected the first inning saw the most scoring and the second inning saw the least, no matter what time period was involved. What has changed is the difference between the two and whether the sum of the 1st and 2nd is greater or less than the average of other inning.

My methodology's implicit assumption that the number of 8th innings for official games always equals the number of 1st innings is not strictly true. Yet Smith's data shows this to be true to within one percent. Of the 122910 games in Smith's study only 0.98% had no 9th inning. Even fewer official games had no 8th inning, even fewer still no 6th inning and all official games had at least a top of the 5th inning. As an experiment and for the benefit of readers, I applied my methodology to Smith's aggregate data with and without compensating for missing innings. The correction barely changes the numbers. I also realized that because I did not know how many times the home team did not take final at bats in shortened games, the correction was just educated guessing.

The numbers in Table One do not represent percentages of all scoring; for example, one cannot add two numbers from the same row together to get 25.01 and then deduce from that sum that a quarter of all runs of that time period were scored in those innings. On the other hand, if one makes the same addition for two different rows, getting 25.5 for one row and 24.7 for the other, one can very reasonably conclude that a greater proportion of all runs was scored in those innings during the former period than during the latter. Table One tracks whether pitching+Defense for later innings has improved or degraded relative to early innings.

With only 32 to 48 team-seasons in the groups, there was bound to be some degree of random fluctuation between adjacent innings. Along with the anomaly of World War II, these fluctuations somewhat obscure the main pattern. The peculiar results for the 9th inning in the 19th century turn out to be related to the practice of the host team opting and in some cases being required to bat first. Prior to 1950 Home teams had the option of choosing to bat first. This option was exercised only 18 times after 1900, but it was used in 40.5% of games during 1894.[xi] The 1899 Cleveland Indians played 35 of their scheduled home games in the parks of their opponents. Given that Retrosheet.org lumps these games with their opponent's normal home games the 40 team sample for the 1890s must be regarded as unsuitable for the shortcuts of my methodology[xii] The very low 9th inning numbers for the 1870s National Association of Professional Base Ball Players quite possibly also results from home teams choosing not to bat last.

Table two uses the same data as table one. Adjacent innings are lumped together and the time periods are consolidated differently. A clear pattern emerges starting with the 1920 to 1941 Golden Age. World War II brought a shortage of Major League-quality players at all positions, but it is clear that starting pitching was temporarily at an all-time low relative to potential relief pitchers. We also see that Smith's conglomerate sample appears to be a good representation for the scoring pattern of 1969 to 1987, but is not indicative of the percentages of scoring in other time periods. In particular, it bears only a passing resemblance to the percentages from 19th or 21st Century baseball.

Table Two: Average results over larger time periods

	1st and 2nd	3rd and 4th	5th and 6th	7th and 8th	top of 9th
1870s NA	23.58	22.60	24.35	24.24	5.25
1890s to 1902AL	26.13	24.01	22.98	22.69	4.19
1901NL to 1919	**25.01**	**22.89**	**23.00**	**23.33**	**5.78**
1920 to 1941	**23.62**	**23.19**	**23.47**	**24.03**	**5.69**
1942-1946	24.19	23.44	24.02	23.14	5.24
1947 to 1968	**23.36**	**23.49**	**23.82**	**23.86**	**5.48**
1969 to 1987	**23.97**	**23.55**	**23.85**	**23.33**	**5.31**
1988 to 2007	**23.77**	**24.17**	**24.29**	**22.64**	**5.13**
2008 to 2013	23.89	23.90	24.99	22.16	5.08
Smith's Data	23.72	23.64	23.86	23.42	5.38
With adjustment	**23.76**	**23.68**	**23.89**	**23.33**	**5.34**
Random	23.53	23.53	23.53	23.53	5.88

.

Section III

How does this data link to Smith's results? The explanation will take us to the shoreline of Expected Victory Theory, but fortunately we will only have to get our feet wet. We won't need any Markov chain because we there is a simple way to identify whether a change in expected future scoring will raise or lower the percentage chance to win. I call it the innings of scoring-method. In order to avoid dealing with too many variables such as "x"," y", and "z", let us start with a very concrete example.

The sixth inning has just ended, so both teams have 3 more half innings of at bats if needed. To keep things simple let us say that is the home team that trails and the margin is two runs. To keep the numbers concrete we shall set normal scoring as 0.5 runs per half inning and specify for the moment that it applies to both sides Thus the 2-run lead represents exactly four innings of scoring. Who is going to win? We don't know, but since the home team needs to outscore the visitors by the results from 4 typical half innings while coming to bat on three more times, they probably will wind up on the short end of the score when all is said and done. Teams with this lead in this inning have historically won 70% of the time, which matches the predictions of the Markov approach. That the innings of scoring method shows a 7 to 3 advantage for the visitors in the example is simply a lucky coincidence.

Now suppose that the 7th inning goes by and the lead remains two runs. Now the trailing team needs to pack 6 innings of scoring into two. The ratio of 6 to 2 is more favorable for the leading team than 7 to 3, and so the innings of scoring method declares that there is a reduced chance of a come-from-behind win than in the previous scenario, and that is what Smith's Figure 3 shows. Again, the historical ratio is in accord with the theoretical prediction.

Now what if scoring were only 0.4 runs per inning? In that case the 2 run lead represents five innings of effort, and so the likelihood of a come from behind victory has gone down because the innings of scoring ratio is 8 to 3. And 3 of 11 is less than 3 of 10. Therefore, if the leading team had some stratagem that would lower the scoring for both sides from 0.5 runs per innings to 0.4 runs per inning it would help its cause.—that is, it would raise its expected victory percentage from the historical 70% to something better.

The initial example assumed that all innings were equal as far as scoring was concerned. Yet Table One and Table Two illustrate that in some periods the 7th inning was a bigger than normal scoring inning and in others it was not. Over time, the 9th inning has become increasingly a low scoring inning. Collectively, the 7th through 9th inning scoring began the century barely below average, with the 7th inning being the easiest of the game and (top of) the 9th inning apparently providing something of a problem relative to the rest of game. During the Golden age, or lively ball period collective scoring rose above the overall norm during the late innings. Whatever the reasons, pitchers as a group were more effective in the first six innings of 1920 to 1941 than in the final innings of regulation.

During WWII this was no longer the case. Maybe it was the relief pitchers who shown brilliantly rather than being a case of starting pitchers who were subpar, but either way professional Baseball's collective assumption that a tiring starter was better than a fresh second line reliever received a hard jolt. Which brings us to the question of how the prevailing strategy of the Golden Age (call this "STRATEGY-30") compares with post-war pitching stratagems; ("STRATEGY-60".) Were managers right to abandon STRATEGY-30 and start using STRATEGY-60 instead?

Table two sums the last three innings for STRATEGY-30 at 29.72, which is would be 1.516 runs over three innings when the scoring level is 4.5 per nine innings. STRATEGY-60 sums to 29.34, which is just under the random 29.41, and so expected scoring for three innings would be 1.496 runs.

3 innings times 29.72/29.41 times 0.5 runs per inning = 1.516 expected runs

3 innings times 29.34/29.41 times 0.5 runs per inning = 1.496 expected runs

Switching back to our "fool proof" technique: 3.496 to 1.516 is better for the trailing team than 3.516 to 1.516; thus the Trailing team increases its chance for a come-from-behind win if it adopts STRATEGY-60 while the leading team pursues STRATEGY-30. The increase is not large, but it is quite real none the less.[xiii] The .02 runs represents one 9th of an out's worth of offense, so that every 27 games of using the inferior late inning strategy concedes a full inning's scoring to the opponent. Someone who managed that way for 2000 games against opponents using the superior STRATEGY-60 would be disadvantaging his team by 77.5 innings of scoring! And you cannot do that without losing some games that otherwise could have been won. A parallel argument produces results that are slightly more strongly in favor of the leading team's going to STRATEGY-60.

This argument is more general than the specific numbers used above, because the direction of the advantage is the same whether overall offense is 0.333 runs per inning, 0.45 R/inning, or 0.56 R/inning. In each case the advantage remains one ninth of an out. The bigger the lead in runs, the less important this advantage becomes, but the advantage does not disappear. At worst, the advantage becomes so small that one's primary concern becomes making sure that subsequent potential victories are not jeopardized by our actions today. (eg a five run lead with one inning to go would be 5.5 to 0.5 or 11 to 1.) We see these clearly when we realize that a five run lead represents approximately 30 outs of offense; two hundred seventy times the importance of the difference in strategies.

But notice what has happened. If both stuck with STRATEGY-30 (then given the posited 0.5 runs per inning) the relevant numbers in runs were 3.516 to 1.516. With both using STRATEGY-60, the same situation evaluates as 3.496 to 1.496. The difference is the same and the ratio is very nearly the same. And that means that any gain in expected victory percentage that either side might have gained has been mostly canceled by the opponents selection of a counter strategy of pitching improvement. We have assumed that the leading team and the trailing team made equal improvements; in the next section the reader can find reasons why it is normally the trailing team that can maker bigger improvements to its pitching + defense.

We didn't specify whether our hypothetical game took place during the Golden Age or sometime after WWII. We simply postulated a given level of offense. If we postulate a lower level, then the .02 run advantage shrinks, yet remains the output of 1/9th of a typical out. Thus the separation in Expected Victory Percentage from STRATEGY-30 and STRATEGY-60 remains intact.

We did not consider what the actual percentage chances of victory actually were. The conclusion of the argument was that if one side did X and the other side did not X then Probability shifted in a particular direction. The line of reasoning also demonstrated that when both sides did X then any change from the situation where both sides did non-X would be smaller than the change if only one side did X. If you change the size of the lead, the size of the shift would change but not the direction of the change. If the level of offense changes the argument still holds. This assures us that the above argument for what to do when down two runs will also work for different size leads and for tie games.

We are assuming in the argument that our hypothetical manager has the resources to employ either STRATEGY-30 or STRATEGY-60. This assumption is completely reasonable for the real world time period in which Relief Aces first emerged, 1947-1960 Managers could have used their pitchers the way they were in the late 1930s just as some managers in the 1920s used their staffs the same as in the previous decade. Whenever his manager did not have resources to make the adaptation to STRATEGY-60, then it behooved his general manager to strive to get the needed players on the roster.

The tactics developed from 1969 to 1987, which we will call "STRATEGY-80" cannot be compared directly to STRATEGY-30 or to the tactics of today because the pitching staffs have changed. Can you imagine a manager with a contemporary staff feeling comfortable using STRATEGY-60, let alone STRATEGY-30 on a regular basis? If asked, the manager would say either that he doesn't have the arms, or that the older strategy would misuse the arms he does have. Nor are there many pitchers today who could remain effective is used in the manner of Mel Harder, Lefty Grove, Carl Hubbell or Dizzy Dean.

What Table Two tells us is that not only was STRATEGY-80 was an improvement upon STRATEGY-60, STRATEGY-80 is likewise trumped by the practices of 1988 to 2007. We see that late inning scoring further declined with each twenty year period and thus the bullpen evolution that took place was not without positive results. The data is only preliminary, but Major league teams appear to have taken yet another small step forward by their current practices, {a.k.a. STRATEGY-2100}.

These arguments for the increasing practice of getting tiring pitchers out of the game in favor of fresh arms covers ground only vaguely related to what is the optimal time to employ one's best relief pitcher. In fact, each of our strategy sets could in principle broken down into three parts; what to do when trailing, what to do when tied, and what to do when leading. The reason I did not break things down this way is that managers have until fairly recently been astute enough to realize that the best defensive tactics when down a run is the same as for when the game is tied, which is exactly the same as when leading by a single run. For managers of the 1950s and 1970s the partition of strategies was (I) what to do when way ahead, (II) what to do when far behind, and (III) what to do while this game remain up for grabs. STRATEGY-30, STRATEGY-60, and so forth may be thought of as packages of tactics which were in vogue during different time periods.

Readers should note that because the chance to win equals one minus the opponent's probability of victory, any combinations that work in favor of come from behind win are working against the winning percentage of the team with the lead, and conversely. We noted above that if the leading team could somehow impose an overall reduction in scoring that affected both teams equally then its lead had a better chance of standing up. What if the home team could maintain its own scoring while holding down the opponent's?

Is this not exactly what the use of a 1960's style relief ace was intended to do? Let us suppose that the relief ace could reduce opponents scoring from .5 runs per inning to .4 runs per inning. Expected future runs would be 1.5 and 1.2 in this case with the trailing team needing 8.75 innings worth of expected scoring to force a tie when down by two runs. (3.5 expected runs needed divided by 0.4 runs per inning =8.75 innings of expected scoring.) A 7 to 3 advantage in innings of scoring has been increased to 8.75 to 3.

However, the trailing team has a counter-strategy available. It could bring its own relief ace into the game. Again assuming that the effect is to reduce enemy scoring to 0.4 runs per game, the home team advantage shifts back the other way to 8 to 3. This is still better than the original 7 to 3 ratio, so regardless of what the trailing team elects to do, the leading team's optimal strategy is to try to upgrade its late inning pitching.

Just as the leading team's hope is that the opponent will stand pat with its pitching via strategy-30, a trailing team can gain a substantial advantage if the home team keeps its relief ace on the bench. (7 to 3 shifts to 6.4 to 3.) Hence regardless of what the team with the lead elects to do, the trailing team improves its chances by using its better relievers. We see in these two parallel arguments that if a team has only one reliever who is substantially better than the weakest starting pitcher then over the course of the season there will be game after game for which that reliever's services will be highly desirable.

The difference in opponents scoring between STRATEGY-30 and STRATEGY-60 are smaller than .5 to .4 runs per inning. But the direction of the advantage remains in favor of STRATEGY-60. Failure to adopt the more modern approach during the final three innings concedes a potential advantage to the opponent of one out's worth of runs for each nine games. And the argument works right up to the point where the lead is so large that the Manager would be wasting a portion of tomorrow's chance for victory if he brought in one of his best available pitchers.

The same structure of argument can show that STRATEGY-60 is untenable in the face of STRATEGY-80, which is itself suboptimal once resources for STRATEGY-2100 are available.[xiv] No matter who has the lead, cutting down scoring by the opposition will increase the chance of victory, and table Two indicates that the best strategies to do that are the recent one.

Because the advantage between the progressions of strategy was so modest, there was historical disagreement over what was optimal use of pitchers. Thus we should be wary of drawing any conclusions about specific tactics. All Table Two tells is that the long run evolution of defensive strategy and tactics has lead to a smaller and smaller percentage of scoring after the 6th inning. Within these 20-year time periods we may be assured that some managers tried to use their pitchers the way they had managed them 10 years before. But the hard facts working against practitioners of the old ways were that with more and more power threats in opposing lineups and without trick pitches to which to resort, starting pitchers were losing effectiveness by the time the 8th inning arrived. These facts have driven the evolution of staff usage during each subsequent decade.

Section IV Wrap up

The specific data points for Smith's Figure 2 and Figure 3 are the product of two percentages---the leading team's probability of avoiding what amounts to a blown save---and the percentage of blown leads that still result in wins. The first of these is greatly dependent on pitching by the team with the lead. The second is highly sensitive to the quality of the pitching the former trailing team can put on the mound should it catch up or go ahead.

According to the Stats Baseball Scoreboard 35.65 % of blown saves in 1992 and 1997 resulted in victories: With 1049 blown saves in 3297 tries,[xv] the expected victory value of a generic save opportunity was .7944 victories. Because lead greater than three runs are harder to lose and do not count as save opportunities this data is entirely consistent with Smith's tables. I do not have the precise numbers, so suppose for the sake of argument that for 2012 only 32.1% blown saves subsequently led to victories. Now suppose also that that the success rate has risen to 70.1% (which it has.) What is the value of a save opportunity under these assumptions? It is .79698, which is barely different from the previous .7944. The save percentage changed by a very significant 2%, but Smith's methodology will find a change in victory percentage of only 0.3 percent. When we throw in the leads lost by starting pitchers who remained in the game, it is not hard to envision how it is possible for the real world percentages in the paragraph above to change significantly, yet with the resulting final product being nearly or completely unchanged.

Which brings us back to the negative slope that was deemed a mystery. Why might the trailing team be able to improve its pitching more than the team with the lead? That is because a pitcher who has allowed the other team to build a two run lead by the 6th inning is in most cases a pitcher who is not having a good day. That is, these games are days in which he is not pitching as well as he would in a randomly selected start. Under these circumstances, a manager of a post-war team could usually find at least one substitute pitcher from which he could expect better performance. This realization ought to and eventually did trigger a trend that steadily improved 8th and ninth inning performance by the trailing team.

Meanwhile, the invention of the save as a semi-formal statistic made managers increasingly aware that their starting pitcher might not be the best choice for preserving a lead. What we find historically is that the net effect of reduced reliance upon starting pitchers has been to make 1st inning and 4th inning leads slightly less reliable as an precursor of ultimate victory while 8th inning leads are very slightly more secure than a century ago. This is not because there is no discernible benefit to modern day bull pen strategies by the team with the lead, but because the trailing side resorted to a defensive counter-strategy which diminished or even entirely eliminated the victory-percentage payoff from use of LOOGies, setup men and closers. Because it looked at wins rather than runs, Smith's methodology had no way to detect that relievers were being used with increasing effectiveness by both sides.

Suggested Further Research

The scoring pattern identified in my sample consolidated data from distinct scoring conditions to keep the sample large without having to add team-seasons. A more specific study using data from additional teams could separate the innings by whether the team at bat began the inning ahead, behind or tied. This would definitively answer three questions for which my methodology assumed answers.

Does the home team have different scoring pattern than the visiting team? I assumed the two paralleled each other and did not monitor the two rates separately.

Which came first historically, improved late inning pitching performance by the trailing team or by the leading team? I assumed these two evolved simultaneously.

Have there been time periods in which performance by the leading team improved more than performance by the trailing team? Based on Smith's Figure 3, my initial conclusion is that the trailing team's pitching performance has improved more in the long term, but this is only speculation at this point.

APPENDIX

Given the way the relevant Retrosheet pages are linked, it is much less time consuming copy and paste sample data from the same franchise year after year than to collect data for each team in a ten-team league-season. Unfortunately, there are still a great many seasons not covered for this data. Improvisation was needed to keep the sample size from shrinking or expanding to the point where recent data overwhelmed data from years such as 1899 when there were only 12 teams in all. Imbalanced schedules for the seasons 1969 to 1992 in the NL and 1969 to 1976 the AL also presented cause for concern. As a work around, I opted for 8 franchises balanced between the divisions of post 1969 baseball.

The franchises I chose were the Pirates, Phillies, Dodgers and Reds from the NL and the Orioles-Browns Indians, Athletics and White Sox from the AL. None of these franchises were part of the National Association and so for the 1870s my data base was every team from the three years for which data appears in Retrosheet.org.[xvi] Data from the 1890s covered only 1894 and1899, so I used data from all twelve teams plus data from all eight teams of the 1901 and 1902 American league (which was using similar rules and equipment.)

Doing this gave me sample of 40 team-seasons of pre-foul strike baseball, but left me short of team-season data from 1901 to 1909. Retrosheet covered only the 1901, 1902, 1904, and 1909 seasons. To get up to forty team-seasons for the extreme dead-ball years I included 1910 in this group and used the full National league for the 1901 and 1902 seasons. This meant that the next two groups of seasons covered only 9 years (1911-1919). To compensate, the 1914 Federal League was included in the 1911 to 1914 grouping. From that year onward, data was taken solely from the eight sample franchises.

How seasons should be grouped is not self-evident, but neither can it be a matter of whim. It is very clear that scoring conditions of 1918 were radically different from those of 1921, but whether 1918, 1919 or 1920 represents the end of the dead ball era is not determinable by events or by the changes in scoring. The 1919 season had livelier baseballs than the years of WWI. The 1920 season saw official limits on spit balls and related trick pitches. Starting in 1921, new balls began to be brought into play, which greatly enhanced the effect of the 1919 and 1920 changes.

Conditions by 1943 were clearly unlike those of 1940. And 1947 was nothing like 1945. But should 1941, 1942, or 1946 be grouped with 1943-45? I opted to use 1942 to 1946 as the "War years" because the inferior baseballs of 1942-to 1945 were still in widespread use during 1946.

Rule changes in 1963, 1969, and 1988 clearly created new scoring levels. And something happened in 1993 that raised hitting for both average and power in both leagues. This leaves post-war sequences of 16 years, 6 years, 19 years, five years, and 21 years to be subdivided in some way. Dividing into groups of exactly five seasons involves a risk of grouping of apples with oranges. Table 2 highlights the long term trend by grouping entirely by temporal proximity. Thus periods with different levels of offense are grouped together. The advantage from meshing time periods was that a graph with steady increases resulted.

---

[i] "Coming from Behind: Patterns of Scoring and Relation to Winning" David W Smith. Presentation for SAB34, with written version found Retrosheet.org/Research . Subsequent page number citations reference this document.

[ii] Smith," Figure 3, page 4.

[iii] In recent years, baseball managers have begun to make pitching changes on the assumption that even though their ace starting pitcher is still getting the batters out, there are players in the bullpen who would perform at least as well if not better. However much we fans might bewail the delays brought on by such substitutions, as a table-top game designer I can attest to the logic behind many (though not all,) of these machinations.

[iv] Smith Table 2, page 2

[v] Smith Tables 2,3 and 4. Tables three gives the number of extra inning decisions. And table 4 gives a break down. The

count of 10th Innings (11513) exceeds extra-inning decisions (11326). The difference of 187 is the count of tie games.

[vi] Smith Page 5

[vii] Smith page 3

[viii] Smith page 4

[ix] Smith pg 8

[x] Parks effects should be likewise eliminated, since the home and road totals for each inning are added together. Since the 9th inning numbers include only runs scored by visiting teams the numbers in Table One might be knocked off kilter if the sample franchises are not collectively neutral.

[xi] "How large is the Advantage of batting last?", Mark Pankin found at Retrosheet.org/Research.

[xii] The number half innings for the visiting team in the 9th is normally within 2% of the number of official 1st Innings, 2nd Innings, and so forth. My practice of simply summing all the runs from each inning or half inning tacitly assumed that for each team-season in the

sample there would be equal numbers of the 34 half innings being included. But this breaks down whenever the visiting team bats last, as happened a lot in 1894 and 1899. I thereby conclude that while Table 1 and Table 2 give the correct proportions among the first 8 innings for each sample period, that the 9th inning data for the 1890s is out of whack with the rest of the table. In order to get proper data, one would need to sum runs scored in the top of the ninth regardless of which team is batting in that inning.

[xiii] For those who play table-top games, there is no conceptual difference between a small edge and a large edge provided that the drawbacks of chasing that edge are the same. Think of it this way, a bet on 4 numbers of a roulette wheel will probably lose. The same is true of a bet that covers 3 numbers. But if the total cost and total payoff of the bets were the same, would it not be crazy to bet on only three numbers rather than four? Managing a baseball game is like being forced to play roulette for a set number of spins. Playing the percentages in Baseball is analogous to choosing the available bets with the highest expected profits. Bill McKechnie was an optimistic stoic. He said "Take care of the percentages and they will take care of you." The version used by players of simulations games is "Take care of the percentages or be prepared to get your butt kicked by someone who does."

[xiv] I refrained from direct comparison between STRATEGY-30 and STRATEGY-2100 even though the differences would be of far greater significance than the difference between Strategy-30 and STATEGY-60. One reason is that the pitching staffs of the golden age were too small to support Strategy-2100. There is also a concern that transporting players and strategies forward or backward in time might not achieve comparable results in terms of changing the innings in which runs are scored.

[xv] STATS Baseball Scoreboard 1998 p152& p164 . STATS 1993 Baseball Scoreboard p 31 & p 212.

[xvi] As of August 1, 2014 these were 1871, 1872 and 1874.