By the Numbers

Volume 17, Number 2 The Newsletter of the SABR Statistical Analysis Committee May, 2007

Review

Academic Research: Do Countries Specialize in Certain Types of Ballplayers?

Charlie Pavitt

In trade, some countries specialize in certain types of products, and there are models in economics that attempt to explain how and why that happens. Do those hypotheses apply to baseball players? Do countries specialize in certain positions or types of players? Here, the author reviews an academic study that investigates this question.

This is one of a series of reviews of sabermetric articles published in academic journals. It is part of a project of mine to collect and catalog sabermetric research, and I would appreciate learning of and receiving copies of any studies of which I am unaware. Please visit the Statistical Baseball Research Bibliography at www.udel.edu/communication/pavitt/biblioexplan.htm. Use it for your research, and let me know what is missing.

Evan Osborne, Baseball's International Division of Labor, Journal of Sports Economics, Volume 7 Number 2, pp. 150-167

I've not found many recent academic publications of statistical baseball research interesting, but here's a welcome exception. Osborne describes two models of international trade, comparative advantage and product lifecycle. The comparative advantage model implies that countries entering the international trade system specialize in exporting one form of product indefinitely, whereas the product-lifecycle model implies that countries begin by specializing but, over time, the proportion among different exports comes to approximate the proportions world-wide. Osborne's goal was to compare the ability of these models to represent the extent to which different countries' "export" of major league baseball players has specialized among positions and skills across time.

Osborne concentrated on six nations that have been producing major leaguers for quite some time; Canada, Cuba, the Dominican Republic, Mexico, Puerto Rico (not technically a nation, but certainly functioning as one in this context), and Venezuela. He divided time into three eras, 1940 to 1959, 1960 to 1979, and 1980 to 2002; Canada, Cuba, and Puerto Rico contributed enough players to qualify for each of these eras, whereas the Dominican, Mexico, and Venezuela did not for 1940-1959, and so conclusions about the latter three countries are based on only the later two periods. Using an array of differing and mostly simple statistical tests, Osborne reached the following conclusions: Canada and Mexico have tended to specialize in pitchers, whereas the Dominican, Puerto Rico, and Venezuela have produced relatively few. Of these, the Dominican has specialized in righthanders and Mexico in lefties (upon seeing this, Valenzuela and Higuera came readily to my mind). There was no evidence of bias toward or away from strikeout or control pitchers.

In this issue

Turning to position players, the hitters produced by the two pitching-rich countries have tended to have relatively poor batting averages whereas two of the three pitching-poor countries (Puerto Rico and Venezuela) have tended to excel here. The Dominican and Venezuela not surprisingly have relatively overproduced middle infielders (and Venezuela catchers) and underproduced at the infield corners and outfield. Puerto Rico has also been high in middle infielders and catchers, although not noticeably low elsewhere. Canada has been short on both corner and middle infielders. All of these findings support the comparative advantage model. Home run hitting is the exception; consistent with the product-lifecycle model, in general major leaguers from these nations have increased in home runs per at bat over time. While many of these results are not surprising and none are earth-shattering, I found it useful to see the extent to which our impressions about player specialization find support here.

Charlie Pavitt, chazzq@udel.edu ♦

Informal Peer Review

The following committee members have volunteered to be contacted by other members for informal peer review of articles.

Please contact any of our volunteers on an as-needed basis – that is, if you want someone to look over your manuscript in advance, these people are willing. Of course, I'll be doing a bit of that too, but, as much as I'd like to, I don't have time to contact every contributor with detailed comments on their work. (I will get back to you on more serious issues, like if I don't understand part of your method or results.)

If you'd like to be added to the list, send your name, e-mail address, and areas of expertise (don't worry if you don't have any – I certainly don't), and you'll see your name in print next issue.

Expertise in "Statistics" below means "real" statistics, as opposed to baseball statistics: confidence intervals, testing, sampling, and so on.

Member	E-mail	Expertise
Shelly Appleton	slappleton@sbcglobal.net	Statistics
Ben Baumer	bbaumer@nymets.com	Statistics
Jim Box	jim.box@duke.edu	Statistics
Keith Carlson	kcsqrd@charter.net	General
Dan Evans	devans@seattlemariners.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	Han60Man@aol.com	Statistics
John Heer	jheer@walterhav.com	Proofreading
Dan Heisman	danheisman@comcast.net	General
Bill Johnson	firebee02@hotmail.com	Statistics
Mark E. Johnson	maejohns@yahoo.com	General
David Kaplan	dkaplan@UDel.Edu	Statistics (regression)
Keith Karcher	karcherk@earthlink.net	Statistics
Chris Leach	chrisleach@yahoo.com	General
Chris Long	clong@padres.com	Statistics
John Matthew IV	john.matthew@rogers.com	Apostrophes
Nicholas Miceli	nsmiceli@yahoo.com	Statistics
John Stryker	john.stryker@gmail.com	General
Tom Thress	TomThress@aol.com	Statistics (regression)
Joel Tscherne	Joel@tscherne.org	General
Dick Unruh	runruhjr@iw.net	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

Reply to "Are Career Home Run Trends Changing?"

JP Caillault

A previous article in BTN found that recent players hit home run peaks later in their career than players of the more-distant past. Here, the author replies to that article and its conclusions.

I would like to comment on "Are Career Home Run Trends Changing in the Steroid Era?" by Yoshio Muki and John Hanks, which appeared in the August, 2006 issue of *By The Numbers*.

There are (at least) three major flaws in the work.

First, 19 of the 36 "modern era" players were active in 2006 and so have not yet had a chance for their careers to decline (or improve). The records of those players are not complete, so it's unfair to compare their stats to players whose careers are over. In footnote 2 the authors acknowledge the importance of this, but only state that "a number" of active players are included in their study, neglecting to mention that that "number" comprises more than half of their "modern era" population of players. The authors also state in footnote 2 that they performed a "side study" which excluded all active players and that "no significant differences were found" between the numbers presented in the paper and in the side study. However, a paper which relies exclusively on statistics and significance should include some numbers to support the claim that "no significant differences were found," especially given the fact that more than half of one group's data disappears in the side study (making the number of "Classical Era" players = 64 and "Modern Era" players = only 17). In fact, it's hard even to believe that "no significant differences were found," since a glance at Figure 4 shows that if the 19 active players (all of whom have positive slopes) are removed from the plot, the distribution of slopes seems as random as the authors claim it is in Figure 3.

Second, the authors attribute equal significance to every season of a player's career, irrespective of how much the player played. For example, the slope they calculate for Harmon Killebrew is +0.42 (it's actually +0.00042, but the authors have used HR/1000AB without ever telling the reader). However, this slope includes Killebrew's first season in the majors, when he was 18 years old, in which he batted only 13 times (and didn't hit a HR). If that one season is removed from the study, then Killebrew's slope changes to -0.37. In fact, Killebrew didn't reach even 100 AB in any of his first five seasons. If they are all removed, then his slope plummets to a value of -2.71, the lowest of any of the 100 players included in the study. There are many other examples like this, two of the most prominent being Babe Ruth and Jimmie Foxx, whose first seasons consisted of 10 ABs and 9 ABs, respectively. Those types of seasons shouldn't be completely removed, but a proper weighing of each season is required for the final results to have any meaning.

Third, the authors say nothing about correcting the players' statistics for the times in which they played, despite the fact that comparing eras is essentially the point of their paper. A player who would have played in the NL from 1986-2006 (the same duration as Barry Bonds' career thus far), and hit HRs at exactly the same rate as the NL as a whole, would have a slope of +0.59. The authors indicate that they would consider this player to have demonstrated increased HR efficiency, despite the fact that he would have been exactly average during his entire career. Similarly, a player whose career span exactly mimicked Hank Aaron's (1954-74 NL, 1975-76 AL) and who hit HRs at exactly the same rate as the league would be considered by the authors as having decreased HR efficiency as time went on, since his slope would be –0.43. If the authors instead calculated the ratio of a player's HR/AB rate to that of the league, then plotting the slope of that ratio would be a better indicator of the change in the player's HR prowess. In that case, one who hits HRs at exactly the same rate as the league throughout his career would have a slope = 0, as expected.

Given these problems with the paper, its content from Figure 6 onward is probably meaningless.

JP Caillault, jpc1957@msn.com ♦

Get Your Own Copy

If you're not a member of the Statistical Analysis Committee, you're probably reading a friend's copy of this issue of BTN, or perhaps you paid for a copy through the SABR office.

If that's the case, you might want to consider joining the Committee, which will get you an automatic subscription to BTN. There are no extra charges (besides the regular SABR membership fee) or obligations – just an interest in the statistical analysis of baseball.

The easiest way to join the committee is to visit http://members.sabr.org, click on "my SABR," then "committees and regionals," then "add new" committee. Add the Statistical Analysis Committee, and you're done. You will be informed when new issues are available for downloading from the internet.

If you would like more information, send an e-mail (preferably with your snail mail address for our records) to Neal Traven, at beisbol@alumni.pitt.edu. If you don't have internet access, we will send you BTN by mail; write to Neal at
4317 Dayton Ave. N. #201, Seattle, WA, 98103-7154.

Submissions

Phil Birnbaum, Editor

Submissions to *By the Numbers* are, of course, encouraged. Articles should be concise (though not necessarily short), and pertain to statistical analysis of baseball. Letters to the Editor, original research, opinions, summaries of existing research, criticism, and reviews of other work are all welcome.

Articles should be submitted in electronic form, either by e-mail or on CD. I can read most word processor formats. If you send charts, please send them in word processor form rather than in spreadsheet. Unless you specify otherwise, I may send your work to others for comment (i.e., informal peer review).

If your submission discusses a previous BTN article, the author of that article may be asked to reply briefly in the same issue in which your letter or article appears.

I usually edit for spelling and grammar. If you can (and I understand it isn't always possible), try to format your article roughly the same way BTN does.

I will acknowledge all articles upon receipt, and will try, within a reasonable time, to let you know if your submission is accepted.

Send submissions to:
Phil Birnbaum
88 Westpointe Cres., Nepean, ON, Canada, K2G 5Y8
birnbaum@sympatico.ca

Quantifying the Impact of Opponent Quality

David Roher

Does good pitching beat good hitting? Previous studies suggest no such effect exists overall. But could an effect exist for certain players? Is it possible that pitcher X is unusually good at beating good hitters, while hitter Y is unusually good at hammering the best pitchers?

"Good pitching beats good hitting" is a common mantra among baseball insiders. Unlike many other old baseball ideas, it has partially stood the test of deeper inquiry. In its 2006 book *Baseball Between the Numbers,* Baseball Prospectus found a strong correlation between pitching and playoff success, and was unable to link hitting with October victories.[1] However, in the August, 2001 edition of *By the Numbers,* Tom Hanrahan showed that, at least in the regular season, no trend for good pitching against good hitting appears to exist. But no one has calculated the precise degree to which this idea is true for individual players. Do some players exhibit the ability to perform equally well regardless of opponent quality?

To answer this question, I took Retrosheet's 2006 play-by-play data and parsed through each individual at-bat. Then, for each player, I made a coordinate graph of at-bats: on the y-axis was a numerical value of each batting result, using Tom Tango's linear weights,[2] and on the x-axis was the quality of the opponent for that particular at-bat. For hitters, opponent quality was measured in Baseball Prospectus' Fair RA pitching statistic, and for pitchers, it was measured in Baseball Prospectus' EqA hitting statistic. After I plotted the points, I then ran a linear regression. The final statistic, which I call OQE, or Opponent Quality Effect, is the slope of this best-fit line (hOQE for hitters, pOQE for pitchers). The greater the slope, the more that player is affected by the quality of his opponent.

Here are the 25 hitters (minimum 3.1 PA/G) most affected and the 25 least affected by the quality of their opponent in 2006, left and right respectively:

	hOQE	EqA		hOQE	EqA
Ray Durham	0.853	0.300	Nick Punto	-0.313	0.267
A.J. Pierzynski	0.802	0.266	Conor Jackson	-0.209	0.278
Jeff Conine	0.698	0.259	David Ortiz	-0.195	0.344
Jamey Carroll	0.643	0.264	Kevin Millar	-0.185	0.291
Robinson Cano	0.615	0.307	Adrian Gonzalez	-0.171	0.300
Melky Cabrera	0.603	0.273	Craig Biggio	-0.156	0.250
Angel Berroa	0.580	0.209	Geoff Jenkins	-0.156	0.277
Lyle Overbay	0.570	0.299	Austin Kearns	-0.128	0.286
Craig Monroe	0.554	0.268	Alex Rodriguez	-0.120	0.318
Jermaine Dye	0.552	0.328	Joe Crede	-0.075	0.278
Pedro Feliz	0.551	0.242	Freddy Sanchez	-0.051	0.293
Omar Vizquel	0.539	0.270	Adam Kennedy	-0.049	0.261
David Eckstein	0.536	0.250	David Wright	-0.049	0.313
Ichiro Suzuki	0.522	0.296	Scott Podsednik	-0.033	0.250
Carl Crawford	0.511	0.296	Andruw Jones	-0.018	0.302
Ramon Hernandez	0.500	0.288	Todd Helton	-0.013	0.296
Brandon Inge	0.498	0.269	Mark Loretta	-0.011	0.255
Carlos Beltran	0.493	0.328	Jacque Jones	-0.010	0.280
Kenji Johjima	0.492	0.279	Garret Anderson	-0.007	0.269
Manny Ramirez	0.491	0.352	Yuniesky Betancourt	-0.004	0.254
Jose Reyes	0.483	0.294	Adam Dunn	0.002	0.289
Brian Roberts	0.469	0.280	Hanley Ramirez	0.015	0.294
Pat Burrell	0.468	0.301	Luis Gonzalez	0.020	0.271
Jose Castillo	0.467	0.238	Ryan Freel	0.025	0.271
Adam Everett	0.464	0.226	Marcus Giles	0.025	0.259

[1] Nate Silver and Dayn Perry, *Baseball Between the Numbers Why Everything You Know About the Game is Wrong*, 1st ed. New York: Basic Books, 2006, p. 361.

[2] http://www.tangotiger.net/bsrexpl.html

What is most surprising about the set of data is the apparent lack of correlation relative to EqA, at least to the eye. Each set of 25 has their share of both excellent and poor hitters. Although it may appear after a brief look at the top and bottom 25 hitters that low OQE would favor a higher EqA, this is untrue over the sample space of the entire major leagues: a measure of the correlation between EQA and OQE produced an R^2 value of just .001, indicating an insignificant to non-existent correlation between the two statistics. There is no evidence that as hitters get better overall, their ability to hit better pitching in comparison to worse pitching increases or decreases. In addition, measures of the correlation between OQE and over 100 other hitting statistics revealed nothing statistically significant.

However, it is not hitting that improves as part of the old mantra. After completing the data for hitters, I expected to see OQE for pitchers inversely correlate with their quality. Here are the 25 pitchers (minimum 75 IP) most affected and 25 least affected by the quality of their opponent in 2006, left and right respectively (note that hitter and pitcher OQE are not to scale and thus not comparable):

	pOQE	Fair RA		pOQE	Fair RA
Kevin Gregg	0.012	4.89	Tony Armas	-0.056	5.76
Vicente Padilla	0.004	4.79	John Maine	-0.055	4.14
Ruddy Lugo	0.004	3.90	Doug Davis	-0.055	5.43
Aaron Heilman	0.004	4.14	Michael O'Connor	-0.054	4.84
Jon Lester	0.004	4.81	Brett Myers	-0.053	4.21
Jake Woods	0.002	5.13	Anibal Sanchez	-0.050	2.99
Enrique Gonzalez	0.001	5.97	Paul Maholm	-0.049	5.07
Ryan Franklin	0.001	5.09	Aaron Cook	-0.048	4.44
Scott Baker	0.000	6.76	Ian Snell	-0.046	5.01
Scott Linebrink	0.000	3.44	Ben Sheets	-0.045	3.82
Juan Cruz	0.000	4.22	Aaron Harang	-0.045	4.11
Randy Johnson	0.000	5.48	Byung-Hyun Kim	-0.044	6.06
Brad Hennessey	-0.001	5.07	Chris Carpenter	-0.043	3.30
Ryan Dempster	-0.001	5.96	Jason Marquis	-0.042	6.28
Scott Kazmir	-0.001	3.67	Brandon Claussen	-0.041	6.36
Brandon McCarthy	-0.001	4.52	Josh Johnson	-0.040	3.75
Kelvim Escobar	-0.002	4.24	Taylor Buchholz	-0.040	6.20
Brad Radke	-0.002	4.90	Cole Hamels	-0.040	4.57
Seth McClung	-0.002	6.74	Anthony Reyes	-0.040	5.11
Zach Miner	-0.002	5.35	Ramon Ortiz	-0.039	5.92
Felix Hernandez	-0.002	4.99	Horacio Ramirez	-0.037	4.93
Boof Bonser	-0.002	4.52	Rich Hill	-0.037	4.61
Jon Garland	-0.003	4.86	Wandy Rodriguez	-0.036	6.21
Scott Proctor	-0.003	3.54	Clay Hensley	-0.036	4.02

Surprisingly, there is still no correlation, even among hundreds of other statistics. The R^2 between OQE and Fair RA was a mere .0007.

But if no such correlation exists, then what accounts for the success in the playoffs among better-pitching teams? Even without a correlation to another statistic, there is still a disparity in the OQEs of different players. One possible explanation is that out of sheer luck, playoff teams in the past 35 years have had low OQE pitchers. Another, and my hypothesis, is that the average major league pitcher is less affected by a change in opponent quality than the average hitter. There is still more research to be done on that subject, however.

From the study, it appears that OQE's randomness does not lend itself to any practical applications. But what if it could? To answer that question, one must look at the situations in which having a player with a low OQE would be advantageous: in summary, the situations in which a player is likely to face a high quality opponent. The main regular-season situation for hitters is late-inning at-bats against strong relief pitchers. Since National League teams often depend on pinch-hitting in these situations, it would be advantageous for them to acquire a hitter that had a very low OQE, since a disproportionate amount of at-bats would be against high-quality pitchers. For pitchers, there isn't as clear cut a situation – the most likely application would be for lefty specialists and set-up men. Because they are the only pitchers in the current bullpen structure for most teams that come in specifically in tough situation, they are more likely to face higher quality opponents.

However, I think the most important application would be for the playoffs. Many times, teams have a very good idea of their playoff chances before the season. If a team like the Yankees were looking for a final piece that would help them win the World Series, it wouldn't hurt to use OQE to break a tie in a personnel decision. Conversely, if a team knew that one aspect of their team was going to be very weak, they could acquire players with very high OQEs, because the only time that they would succeed would be against a poor quality opponent in the first place.

It is important to remember that because OQE is completely independent of player quality, it cannot possibly be useful on the scale of other sabermetric statistics. However, there is still a small chance that a player's OQE from year to year could be as constant and predictable as overall production. Results from this study are inconclusive towards that hypothesis, but further research on the topic could prove or disprove it.

David Roher, david.roher@gmail.com ♦

Do You Have Any Idea How Fast You Were Going?

Russell A. Carleton

Years ago, Bill James came up with the "speed score," an estimate of a player's speed on the basepaths, which was based on the player's yearly statistics. Here, the author comes up with an alternative, this time using play-by-play statistics, such as number of times taking an extra base on a hit. In addition, he investigates whether traditional speed-related stats indeed measure a single attribute, or whether what we call "speed" is actually comprised of more than one basic skill.

Baseball is, at its heart, a game of running. After all, the object of the game is to run from one base to another before being tagged (after, of course, hitting the ball). The points assigned in the game are called “runs.” Baseball aficionados often have soft spots in their heart for guys who can run like an antelope. In fact, there are a few baseball players here and there, who despite lacking any ability to get on base, hit the ball for power, or other hitting abilities, seem to make their living as pinch-runners. Come to think of it, teams will occasionally bat these players in the leadoff spot (!) if only for the fact that they are fast.

A strange thing in baseball statistics is that we have plenty of statistics to describe the act of hitting, but relatively few that address a player’s abilities on the basepaths. Attempts have been made, from the easy-to-calculate (stolen base percentage) to the more esoteric (times advancing from first to third on a single), although these are limited to simply describing specific situations. Bill James has given us a statistic called “speed score,” which attempts to pin one number on a player, based mostly on seasonal stats. The formula is moderately easy to calculate and creates a range from roughly 3 (painfully slow) to 10 (Vince Coleman), although in some odd cases, I’ve seen numbers like 27 show up (usually due to small sample sizes). The fine folks at *Baseball Prospectus* have taken the formula and made it so that the scores are mathematically restricted to a range between 0 and 10. Outside of the James method, I’m not familiar with any other empirically-based speed ratings out there. That’s not to say that they’re not out there, only that I haven’t come across them.

I don’t know what process James went through to construct the five factors used in his formula. *En face*, they appear to be based on most of the right things that would be evidence for speed: stolen base percentage, double plays grounded into, triples, etc.[3] My problem is that from a statistical/methodological point of view, it’s not clear if they form a statistically coherent factor. I propose to use some advanced statistical methodology, primarily a factor analytic approach, to investigate if a better speed score (or scores) might be built from the data that are available and to investigate how the James formula shakes out. A warning to those who are still reading: Some major numerical nerdiness follows.

First things first. What events in baseball involve speed and/or running ability? Or at least what events are more likely to happen with a fast runner than with a slow runner? A few things surely spring to mind. The most obvious is the stolen base, although even stolen base raw totals can be deceiving. A player who steals 50 bases, but is caught 100 times isn’t fast, just reckless and lucky 1/3 of the time. So, stolen base success rate is preferred as a better metric of base stealing ability, and by extension, speed.

But even then, stolen base percentage has its shortcomings. Not all stolen base attempts end in a caught stealing or a stolen base. Some (probably those that are the “run” part of a hit and run) are negated by a foul ball or a ball in play. It makes sense that a manager would only want a runner with a decent chance to make it to be running on a play and that a faster runner would get the “steal” sign more often. So, I will calculate the percentage of times that a runner was on first (with second base open) that he ran for second, excepting situations on a two-out 3-2 count. (Everybody runs then.)

On a related note, there’s another hidden variable that can tell us a little something about a runner’s speed. It’s not a secret to the opposing manager if there is a fast runner on first base. So, he’ll tell the pitcher to keep an eye on the runner by throwing over to first base. In 2006, the ten runners who drew the throws in the greatest percentage of times when they were on first (while second was unoccupied) were Ryan Freel, Dave Roberts, B. J. Upton, Nook Logan, Chris Duffy, Juan Pierre, Jose Reyes, Alfonso Soriano, Curtis Granderson, and Alfredo

[3] James’s five factors that go into his speed score are as follows. All stats are seasonal.

```
Sp1: ((SB+3)/(SB+CS+7)-0.4)*20,
Sp2: SQRT((SB+CS)/((H-2B-3B-HR)+BB+HP))/0.07
Sp3: 3B/(AB-HR-K)/0.02*10
Sp4: ((R-HR)/(H+BB-HR-HP)-0.1)/0.04
Sp5: (0.055-GDP/(AB-HR-K))/0.005
```

After calculating all five for each player, the lowest is dropped, and the remaining four are averaged to determine the player’s speed score.

Amezaga. Sounds like a pretty good list of speedsters. I calculated this percentage for all players. On a methodological note, if a batter led off the inning with a single, drew a throw, and then watched his three teammates strike out, that only counts as one situation in which he was on first, not three.

There are also other types of "stolen" bases. A runner who goes from first to third on a single has essentially stolen an extra base, as has a runner who goes from second to home on a single or first to home on a double, but some "station to station" runners are not particularly good at this particular skill. It's easy enough to calculate what percentage of the time, given these circumstances, each player is able to take the extra base. But what about success rates? I did calculate the success rates for each of the three situations, but generally, they were pretty high. In fact, I've done research elsewhere that suggests that when runners try for that extra base, they usually make it, around 93-95% of the time. Many players had perfect scores on the success rates, which doesn't help us much in selecting out the good from the bad. Runners, or perhaps third base coaches, want to be sure when trying for that extra base and they will likely be more sure with a fast runner. So, the variable of interest is how often the runner is allowed to try, rather than how often he is to make it, in each of the three "extra" base situations.

One minor note of methodology is that for runners going from first to third on a single, I did not count situations in which the runner found himself on second after the single, but could not go to third due to a runner already being there. This was an attempt to make sure that I did not punish a runner for having a slow gentleman on his team that he often got stuck running behind. However, there is the possibility that in some cases, with a runner on first and second, the runner on second may not have been sure to go home, and so the runner on first may have had to hesitate, costing him his chance to go to third.

Then there's the issue of triples. Fast players are usually the ones who are able to leg out triples. Most often, these are hits to the (right-center field) gap which would have been easy doubles for slower players. So, some form of triple to double ratio seems to make sense as a measure of speed. In this case, I will use the player's percentage of "in play" extra base hits which are triples. Essentially, the formula is 3B / (2B + 3B).

The other thing that speed helps with is the ability to make it to first safely on a ground ball in the infield. This can either take the form of an infield hit or beating out a potential double play ball. So, I selected for all ground balls hit that were then fielded by an infielder. To check for infield hits, I looked at all ground balls where a fielder's choice was not recorded, so that I did not reward a player for reaching first before the throw when the fielder clearly had other ideas. I found the percentage of these ground balls in which the batter was credited with a single. For the avoidance of double plays, I selected for situations in which there were less than two outs with a runner on first where a ground ball was hit to an infielder and an out was recorded at second base. Whether or not the defense completed the double play was the variable of interest. The percentage of times that the batter was able to avoid the double play was obtained.

Many of the things I have listed are present in some form in the original James speed scores, although James used only seasonal stats to calculate his speed scores. Thanks to Retrosheet, the general public has more complete play-by-play data available for such projects. It seems a shame not to take advantage of it. I used the play-by-play data from 2003-2006 and for each player-year, I calculated all of the variables listed above. I restricted the sample to those with at least 100 PA in the season under consideration.

Following the calculation of the percentages, a little bit of numerical gymnastics was needed. Probability variables are not normally distributed, so a statistical conversion needs to be applied to normalize them. I used the natural log of the odds ratio method, which first calculates the odds ratio of a given probability (probability / (1 – probability)) and then takes the natural log of that odds ratio. This leads to a distribution that roughly approaches normality, and indeed, a check of the skew statistics of the newly calculated data points showed that all nine had skew less than an absolute value of 1. Three is generally considered the standard cutoff point for skew which will violate the assumption of normality too greatly to proceed.

Another problem arose that while all of the variables being considered were now in log-odds ratio form, they were on slightly different scales. For example, the best runners might draw a throw 70% of the time they reached first, while for attempting to reach third from first on a single, the best runners might attempt it 40% of the time. A score of 50% might place a runner in the bottom of one category, but the top of another. The variables needed re-scaling. Thankfully, since they had just recently been normalized, the variables could easily be turned into z-scores, using the yearly mean and standard deviation as their basis. Now, the score of 0.7 meant the same thing relative across all the variables.

These z-scores served as the basis for an exploratory factor analysis. For those unfamiliar with factor analysis, the best way to explain it is that it's a way to sort out what variables go together. To take an example from my day job (I'm a child clinical psychologist by training), I often use a certain measure to look at how likely it is that a child has a particular problem. There are some questions that ask about symptoms of depression (i.e., feeling sad, crying, lack of energy) and some that ask about behavioral problems (i.e. getting into fights, vandalizing property). Suppose that the child in front of me has depression, but is well-behaved. I would expect that the answers to most or all of the "depression" questions would be high, while the behavioral problems questions would be low. Further, I would expect that if the rating of feeling sad is high, then the rating for crying will also be high, because those two symptoms generally go together. Now, suppose that I didn't know off-hand which symptoms tended to go together, but I wanted to find out. Exploratory factor analysis will tell me exactly that.

In this case, I have several measures of events in a game that involve baserunning. Are these all measurements that seem to go together? Do they all represent some underlying property (speed) or are they all measuring something completely separate to where knowing one doesn't mean a lot for any of the others? The computer program will look at the data and create new factors and tell me which of my old variables go in which of the new factors (and by extension, which of my old variables go together).

For those readers with knowledge of factor analysis, I submitted the nine variables (z-scores derived from the logged odds-ratio of the various events) in the data to a Varimax rotation with principle components analysis extraction. I requested that the computer save all factors with Eigenvalues over 1. My goal was to create new factors as orthogonal as possible to one another (hence Varimax, not Oblimin), although as we will soon see, I was only somewhat successful.

If the preceding paragraph made your head spin, don't worry. It means that I asked the computer to try to sort out what variables went with what in such a way that they really formed separate concepts, so that if there were two or three different pieces to speed/baserunning, I would know that.

Indeed, two factors emerged from the data, accounting for 43.9 percent of the original variance. I've reproduced the factor loading table below, with blanks denoting factor loadings below .20. For those who aren't familiar, a higher number (1.0 is the maximum) means that the old variable is very related to the new factor that's been created by the program. In general, we hope to see each variable have a high number on one factor and a low number on all the rest.

	Factor 1	Factor 2
Avoiding DP on ground ball percentage	.696	
Percentage of times on first drawing a throw	.694	.440
Percentage of times on first attempting to steal	.653	.483
3B / (2B + 3B)	.624	
Infield hits per grounder	.573	
Stolen base success rate	.443	
Percentage of time attempting to go 2nd to home on a single		.691
Percentage of time attempting to go 1st to 3rd on a single		.624
Percentage of time attempting to go 1st to home on a double		.497

The two factors appear to shake out fairly nicely. The two variables dealing with drawing a throw and attempting to steal from first cross-load on both factors, although they do so with fairly high factor loadings on both. As such, I allowed them to remain cross-loaded. Conceptually, the first factor appears to encompass acts that involve actually running an extra 90 feet within a short amount of time, along with the two variables looking at drawing throws at first and attempting to steal, so I will label this factor as "speed." The other factor shapes up nicely, as it is a collection of stats that show how often a player goes for that extra base. This might be the work of the manager or the third base coach, or a measure of how much of a risk taker the player is as a runner, or some of both. We'll call this factor "green light." Despite the Varimax rotation, the cross-loadings produced intercorrelation between the two factors, with a correlation coefficient of .659.

A few more minor details and then we're ready to calculate the actual speed scores. Factor analysis is designed to create factors with high levels of internal consistency (if a runner is fast according to one of the variables in the scale, he's generally going to be fast according to all the other ones), and internal consistency is most often measured by a statistic called Cronbach's alpha. "Speed" and "Green Light" have Cronbach's alphas of .70 and .71 respectively. .70 is generally considered the cutoff for an appropriate level of internal consistency. This means that they can be added together and that the results will form a coherent factor. I allowed that because of missing data or hiccups in the data set, one of the z-scores might be missing for a player, and as such, calculated the mean score (dividing by the constant does nothing to the distribution statistically) for each variable, which owing to their all being z-scores were all on the same numerical scale. I looked to see what the AR(1) inter-class correlation was over the four years in the data set for stability over time. Speed had an ICC of .77, while green light checked in at .70.

So, who were the fastest runners of 2006 according to the speed score, minimum of 100 PA? Who were the ones given the green light the most often?

Ten Fastest Runners	Ten Slowest Runners
1) Ichiro Suzuki (1.76)	1) Mike Jacobs (-2.05)
2) Chris Duffy (1.68)	2) Adrian Gonzalez (-1.86)
3) Bernie Castro (1.59)	3) Yadier Molina (-1.43)
4) Carl Crawford (1.53)	4) Manny Ramirez (-1.33)
5) B.J. Upton (1.51)	5) Mike Piazza (-1.31)
6) Dave Roberts (1.51)	6) Adam Laroche (-1.28)
7) Corey Patterson (1.49)	7) Carlos Delgado (-1.27)
8) Juan Pierre (1.32)	8) Jonny Gomez (-1.24)
9) Willy Taveras (1.31)	9) Bengie Molina (-1.19)
10) Jody Gathright (1.30)	10) Javy Lopez (-1.11)

Ten Most "Green Lights"	Ten Most "Red Lights"
1) Corey Patterson (1.64)	1) Paul Konerko (-2.05)
2) Alfredo Amezaga (1.63)	2) Mike Piazza (-1.99)
3) Richie Weeks (1.50)	3) Ryan Garko (-1.87)
4) Jason Repko (1.44)	4) Adrian Gonzalez (-1.71)
5) Willy Taveras (1.43)	5) Dan Johnson (-1.69)
6) B. J. Upton (1.35)	6) Victor Martinez (-1.65)
7) Chone Figgins (1.31)	7) Josh Bard (-1.62)
8) Carl Crawford (1.17)	8) Kevin Millar (-1.61)
9) Brian Roberts (1.13)	9) Jim Thome (-1.60)
10) John McDonald (1.11)	10) Toby Hall (-1.55)

Not shockingly, the list of the fast includes gentlemen who mostly ply their trade in center field, and the list of the slow includes those who catch, play first base, and are designated to hit (and apparently, not designated to run). The list of those who are high on "green light" ranking contains a pretty good roster of guys who have a reputation (some deservedly so) as fast runners and risk-takers, and several folks from the fastest runner list make an appearance on the green light list. The list of "red light" players similarly contains some big lumbering guys who have reputations of being station-to-station runners.

With all of this done, how does my method compare to the Bill James version? The five component speed scores from the James method have a Cronbach's alpha of .69, although James drops the lowest of the scores. His overall speed score method has an ICC of .70 over the four years in the study (2003-2006) among those with 100 PA or more in the season in question. My method is more stable over time, although the difference is not overwhelming. I calculated the correlations between each of my two factors and James' speed scores. My speed score had a correlation of .807 with James's. The green light score correlated at .718. Also, James's method has minimal skew (.455), once the sample is restricted to those with 100 PA or more.

The James method has the added benefit of being infinitely more easily calculated. The method that I have used here employs the type of statistical rigor that I would use in constructing a brand new scale, but even after all that, the James method does just about as well and it has appropriate scale properties. To that end, I would recommend, unless the reader has a particular masochistic streak about him, that he use the James method and that he use it with confidence.

But even if my scale isn't a vast improvement, we have learned a few things here. Speed and base running have two components. One is actual leg speed to get from one base to the next. The other is whether the manager and third base coach have confidence in the poor chap to actually make that attempt. Because different managers have different appetites for risk, runners will be sent at different rates if they switch teams (or managers). But then, knowing that actual speed and perception of speed are two separate factors, and that they are not perfectly correlated, then it makes sense that some players are given the green light more (or less) than their speed score would indicate they should be. For example, Ichiro has a speed score of 1.76, but a "green light" score of .62, for a difference of 1.14. He was the third most under-sent runner in baseball in 2006, behind Steven Drew and Joe Inglett. On the other side of the coin, Jonny Gomez had a speed score of -1.24, yet a green light score of .05. Meaning that despite being, on average, one and a quarter standard deviations below the league average on speed, he was sent at roughly an average rate. Indeed, I may have created a metric more helpful in evaluating third base coaches than roster players!

Russell Carleton, RCARLETO@depaul.edu ♦

ERA By Innings Pitched: "No Wonder He Didn't Pitch Much"

Fred Worth

In a previous issue of BTN, an article showed that, in general, the more at-bats a player had, the better his performance. Here, the author does a similar analysis for pitchers, and, just for fun, looks at how often each IP total occurred in baseball history.

I was amused by the article by Abbott Katz in the November 2006 issue of "By The Numbers." Mr. Katz looked at all batters with n at bats in a given season to see what trends might be deduced from the data. I decided to do a similar investigation, but this time for pitchers.

I used Lee Sinins' "Sabermetric Baseball Encyclopedia" to obtain my data.

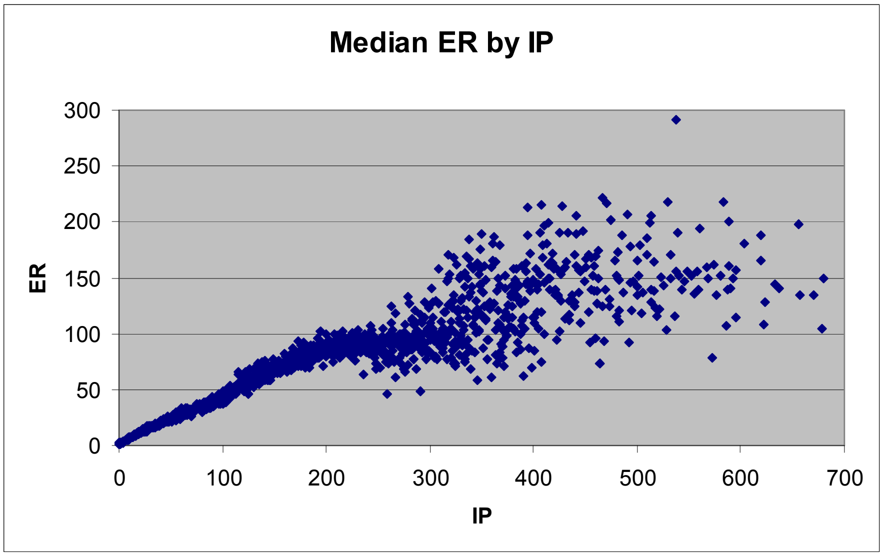
Before going to the data, let me just mention a couple of amusing observations.

- The fewest innings pitched that has never been accomplished in a season is 278-2/3. And since no one has thrown more than that in any season since Charlie Hough in 1987, it seems likely to stay that way.
- The only one to ever pitch 258-1/3 is Luis Tiant in 1968.
- Chappie McFarland had 270-1/3 in 1904, the only pitcher with that total.
- Wayne Garland had 282-2/3 in 1977, the only pitcher with that total.
- Table 1 shows the pitchers who allowed at least 4 ER with 0 IP.

Table 1 – Players who allowed at least 4 ER in zero IP

	year	ER	IP
Bob Kammeyer	1979	8	0
Dennis Tankersley	2003	7	0
William Childers	1895	6	0
Lino Urdaneta	2004	6	0
Doc Hamann	1922	6	0
Tom Qualters	1953	6	0
Kirtley Baker	1894	5	0
Bob Uhl	1940	4	0
Paul Stuffel	1953	4	0
Bob McGraw	1918	4	0
Nick Altrock	1919	4	0

Let me proceed by showing the median earned runs by inning pitched:

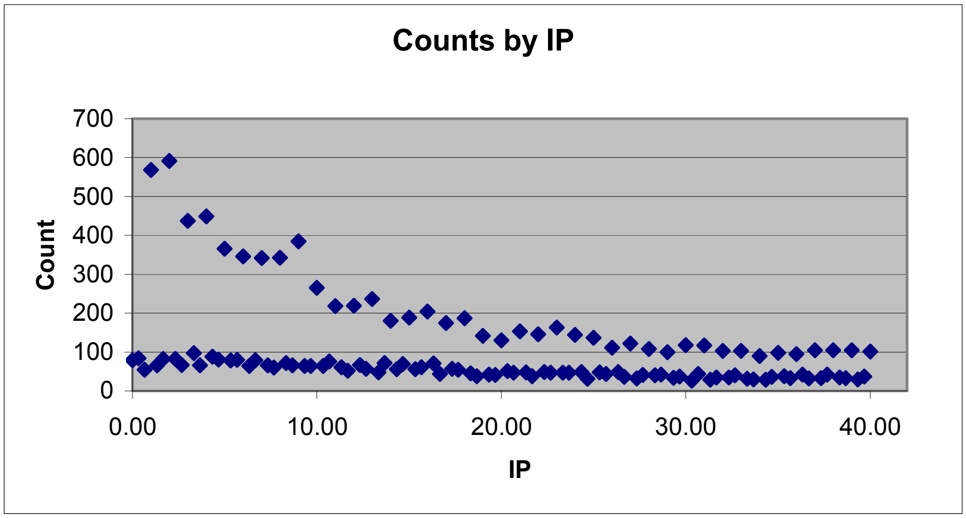


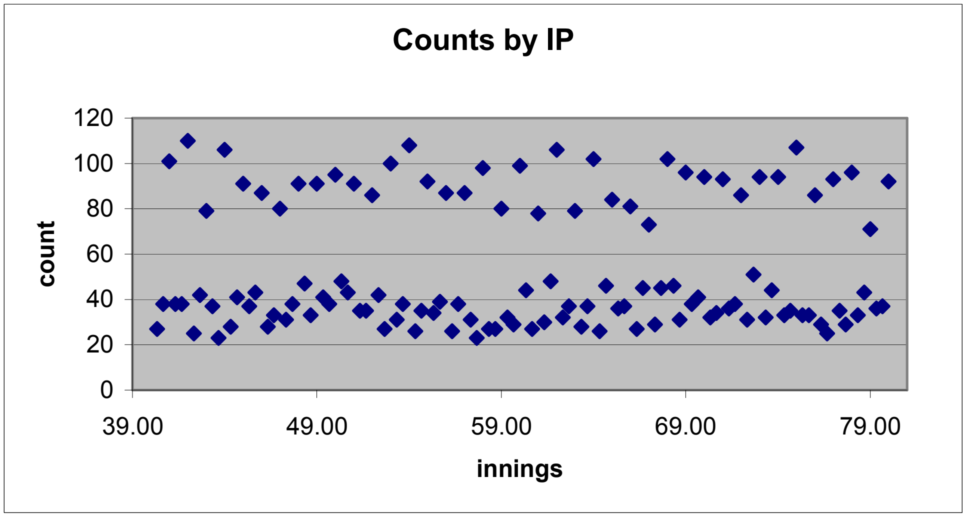
Surely there is no surprise in the generally upward trend since this refers to earned runs, not ERAs. The widely scattered data for 300+ IP is due simply to the fact that, for those innings, n is generally rather small. In light of that observation, let me now present a chart showing n for each IP. For obvious reasons, I will omit all IP with n=0.

The chart takes up the next three pages.

One immediate observation is that, for small numbers of IP, n is biggest for non-fractional IPs. In fact, 311 IP is the smallest n for which n is not greater than for all n ± 2/3. Surely all manner of conjectures are possible for this fact.

Here are two charts showing n for various numbers of IP:





It is not hard to see the tendency to whole number IP.

(Article continues on page 17)

IP	n
0.0	79
0.3	84
0.7	54
1.0	568
1.3	66
1.7	82
2.0	591
2.3	82
2.7	67
3.0	437
3.3	97
3.7	66
4.0	449
4.3	88
4.7	81
5.0	366
5.3	78
5.7	80
6.0	346
6.3	64
6.7	79
7.0	342
7.3	66
7.7	60
8.0	343
8.3	72
8.7	66
9.0	385
9.3	64
9.7	64
10.0	265
10.3	64
10.7	76
11.0	218
11.3	61
11.7	52
12.0	219
12.3	67
12.7	57
13.0	236
13.3	48
13.7	72
14.0	180
14.3	56
14.7	69
15.0	189
15.3	56
15.7	61
16.0	204
16.3	71
16.7	44
17.0	175
17.3	57
17.7	54
18.0	187
18.3	45
18.7	38
19.0	142
19.3	42
19.7	41
20.0	130
20.3	51
20.7	47
21.0	153
21.3	48
21.7	38
22.0	146
22.3	49
22.7	47
23.0	163
23.3	47
23.7	47
24.0	144
24.3	49
24.7	31
25.0	137
25.3	48
25.7	44
26.0	111
26.3	49
26.7	36
27.0	122
27.3	32
27.7	41
28.0	108
28.3	40
28.7	42
29.0	100
29.3	34
29.7	37
30.0	118
30.3	26
30.7	44
31.0	117
31.3	29
31.7	35
32.0	103
32.3	35
32.7	40
33.0	103
33.3	31
33.7	30
34.0	90
34.3	29
34.7	36
35.0	98
35.3	38
35.7	33
36.0	95
36.3	42
36.7	32
37.0	105
37.3	33
37.7	42
38.0	105
38.3	35
38.7	33
39.0	105
39.3	30
39.7	37
40.0	101
40.3	27
40.7	38
41.0	101
41.3	38
41.7	38
42.0	110
42.3	25
42.7	42
43.0	79
43.3	37
43.7	23
44.0	106
44.3	28
44.7	41
45.0	91
45.3	37
45.7	43
46.0	87
46.3	28
46.7	33
47.0	80
47.3	31
47.7	38
48.0	91
48.3	47
48.7	33
49.0	91
49.3	41
49.7	38
50.0	95
50.3	48
50.7	43
51.0	91
51.3	35
51.7	35
52.0	86
52.3	42
52.7	27
53.0	100
53.3	31
53.7	38
54.0	108
54.3	26
54.7	35
55.0	92
55.3	34
55.7	39
56.0	87
56.3	26
56.7	38
57.0	87
57.3	31
57.7	23
58.0	98
58.3	27
58.7	27
59.0	80
59.3	32
59.7	29
60.0	99
60.3	44
60.7	27
61.0	78
61.3	30
61.7	48
62.0	106
62.3	32
62.7	37
63.0	79
63.3	28
63.7	37
64.0	102
64.3	26
64.7	46
65.0	84
65.3	36
65.7	37
66.0	81
66.3	27
66.7	45
67.0	73
67.3	29
67.7	45
68.0	102
68.3	46
68.7	31
69.0	96
69.3	38
69.7	41
70.0	94
70.3	32
70.7	34
71.0	93
71.3	36
71.7	38
72.0	86
72.3	31
72.7	51
73.0	94
73.3	32
73.7	44
74.0	94
74.3	33
74.7	35
75.0	107
75.3	33
75.7	33
76.0	86
76.3	29
76.7	25
77.0	93
77.3	35
77.7	29
78.0	96
78.3	33
78.7	43
79.0	71
79.3	36
79.7	37
80.0	92
80.3	38
80.7	27
81.0	104
81.3	27
81.7	28
82.0	82
82.3	33
82.7	35
83.0	75
83.3	32
83.7	28
84.0	92
84.3	32
84.7	27
85.0	80
85.3	27
85.7	34
86.0	81
86.3	24
86.7	19
87.0	82
87.3	24
87.7	25
88.0	83
88.3	27
88.7	32
89.0	73
89.3	20
89.7	20
90.0	95
90.3	31
90.7	18
91.0	62
91.3	24
91.7	26
92.0	97
92.3	28
92.7	19
93.0	72
93.3	11
93.7	20
94.0	72
94.3	33
94.7	25
95.0	79
95.3	22
95.7	18
96.0	80
96.3	19
96.7	23
97.0	75
97.3	23
97.7	21
98.0	71
98.3	16
98.7	25
99.0	72
99.3	29
99.7	17
100.0	62
100.3	26
100.7	20
101.0	55
101.3	20
101.7	22
102.0	55
102.3	18
102.7	12
103.0	59
103.3	19
103.7	12
104.0	63
104.3	12
104.7	12
105.0	66
105.3	17
105.7	15
106.0	59
106.3	23
106.7	19
107.0	61
107.3	18
107.7	21
108.0	53
108.3	16
108.7	18
109.0	62
109.3	13
109.7	10
110.0	44
110.3	14
110.7	11
111.0	55
111.3	24
111.7	21
112.0	65
112.3	17
112.7	16
113.0	61
113.3	16
113.7	13
114.0	61
114.3	17
114.7	16
115.0	51
115.3	18
115.7	14
116.0	58
116.3	13
116.7	12
117.0	64
117.3	20
117.7	16
118.0	62
118.3	11
118.7	11
119.0	53
119.3	13
119.7	11
120.0	47
120.3	12
120.7	17
121.0	46
121.3	10
121.7	6
122.0	60
122.3	10
122.7	14
123.0	48
123.3	7
123.7	16
124.0	59
124.3	7
124.7	12
125.0	66
125.3	7
125.7	7
126.0	58
126.3	10
126.7	15
127.0	55
127.3	13
127.7	10
128.0	57
128.3	17
128.7	15
129.0	59
129.3	26
129.7	10
130.0	61
130.3	13
130.7	7
131.0	60
131.3	10
131.7	4
132.0	37
132.3	10
132.7	10
133.0	51
133.3	16
133.7	14
134.0	61
134.3	15
134.7	14
135.0	44
135.3	12
135.7	16
136.0	56
136.3	9
136.7	15
137.0	45
137.3	14
137.7	7
138.0	48
138.3	13
138.7	9
139.0	45
139.3	10
139.7	14
140.0	58
140.3	15
140.7	13
141.0	58
141.3	18
141.7	16
142.0	48
142.3	9
142.7	10
143.0	46
143.3	8
143.7	7
144.0	55
144.3	12
144.7	13
145.0	50
145.3	11
145.7	12
146.0	45
146.3	12
146.7	11
147.0	53
147.3	8
147.7	12
148.0	47
148.3	11
148.7	9
149.0	40
149.3	17
149.7	8
150.0	51
150.3	11
150.7	10
151.0	44
151.3	20
151.7	11
152.0	51
152.3	6
152.7	12
153.0	55
153.3	7
153.7	5
154.0	59
154.3	13
154.7	10
155.0	40
155.3	11
155.7	11
156.0	54
156.3	10
156.7	15
157.0	39
157.3	10
157.7	13
158.0	53
158.3	8
158.7	12
159.0	46
159.3	12
159.7	12
160.0	54
160.3	8
160.7	20
161.0	44
161.3	10
161.7	10
162.0	43
162.3	12
162.7	9
163.0	46
163.3	14
163.7	7
164.0	45
164.3	11
164.7	16
165.0	43
165.3	12
165.7	11
166.0	33
166.3	15
166.7	17
167.0	41
167.3	13
167.7	9

IP	n
168.0	57
168.3	12
168.7	12
169.0	36
169.3	21
169.7	13
170.0	53
170.3	18
170.7	9
171.0	52
171.3	8
171.7	16
172.0	42
172.3	18
172.7	7
173.0	51
173.3	10
173.7	15
174.0	39
174.3	10
174.7	15
175.0	53
175.3	10
175.7	14
176.0	44
176.3	16
176.7	14
177.0	39
177.3	13
177.7	9
178.0	40
178.3	13
178.7	15
179.0	57
179.3	12
179.7	9
180.0	48
180.3	11
180.7	14
181.0	50
181.3	9
181.7	14
182.0	43
182.3	12
182.7	8
183.0	48
183.3	12
183.7	16
184.0	55
184.3	13
184.7	10
185.0	41
185.3	11
185.7	8
186.0	47
186.3	14
186.7	12
187.0	47
187.3	8
187.7	8
188.0	47
188.3	17
188.7	9
189.0	43
189.3	12
189.7	9
190.0	43
190.3	13
190.7	18
191.0	60
191.3	7
191.7	9
192.0	40
192.3	13
192.7	16
193.0	43
193.3	17
193.7	9
194.0	48
194.3	13
194.7	18
195.0	39
195.3	14
195.7	18
196.0	63
196.3	22
196.7	18
197.0	41
197.3	10
197.7	17
198.0	54
198.3	14
198.7	13
199.0	55
199.3	21
199.7	12
200.0	35
200.3	11
200.7	12
201.0	54
201.3	15
201.7	11
202.0	45
202.3	12
202.7	10
203.0	51
203.3	12
203.7	10
204.0	58
204.3	16
204.7	14
205.0	55
205.3	12
205.7	14
206.0	45
206.3	15
206.7	20
207.0	45
207.3	16
207.7	17
208.0	67
208.3	19
208.7	14
209.0	41
209.3	11
209.7	12
210.0	37
210.3	13
210.7	13
211.0	51
211.3	16
211.7	25
212.0	52
212.3	16
212.7	14
213.0	42
213.3	8
213.7	13
214.0	47
214.3	17
214.7	10
215.0	54
215.3	12
215.7	13
216.0	42
216.3	12
216.7	12
217.0	59
217.3	16
217.7	13
218.0	42
218.3	15
218.7	14
219.0	42
219.3	14
219.7	10
220.0	47
220.3	14
220.7	12
221.0	36
221.3	12
221.7	14
222.0	47
222.3	17
222.7	9
223.0	42
223.3	14
223.7	9
224.0	42
224.3	10
224.7	9
225.0	41
225.3	8
225.7	7
226.0	50
226.3	8
226.7	10
227.0	42
227.3	6
227.7	6
228.0	45
228.3	11
228.7	8
229.0	35
229.3	16
229.7	8
230.0	33
230.3	12
230.7	7
231.0	42
231.3	16
231.7	12
232.0	35
232.3	4
232.7	12
233.0	39
233.3	10
233.7	11
234.0	40
234.3	9
234.7	15
235.0	40
235.3	8
235.7	6
236.0	35
236.3	6
236.7	8
237.0	39
237.3	10
237.7	7
238.0	33
238.3	9
238.7	9
239.0	37
239.3	8
239.7	8
240.0	41
240.3	6
240.7	4
241.0	33
241.3	7
241.7	7
242.0	34
242.3	5
242.7	8
243.0	33
243.3	6
243.7	4
244.0	40
244.3	3
244.7	12
245.0	43
245.3	2
245.7	5
246.0	27
246.3	10
246.7	11
247.0	28
247.3	6
247.7	4
248.0	27
248.3	3
248.7	3
249.0	29
249.3	4
249.7	13
250.0	50
250.3	8
250.7	8
251.0	29
251.3	8
251.7	5
252.0	35
252.3	10
252.7	3
253.0	24
253.3	3
253.7	8
254.0	31
254.3	5
254.7	4
255.0	31
255.3	6
255.7	6
256.0	25
256.3	5
256.7	10
257.0	33
257.3	3
257.7	9
258.0	27
258.3	1
258.7	3
259.0	29
259.3	2
259.7	4
260.0	34
260.3	4
260.7	10
261.0	23
261.3	10
261.7	5
262.0	27
262.3	4
262.7	3
263.0	21
263.3	2
263.7	6
264.0	23
264.3	6
264.7	2
265.0	20
265.3	7
265.7	5
266.0	23
266.3	4
266.7	4
267.0	26
267.3	3
267.7	4
268.0	23
268.3	5
268.7	5
269.0	24
269.3	7
269.7	5
270.0	28
270.3	1
270.7	4
271.0	21
271.3	3
271.7	2
272.0	23
272.3	5
272.7	4
273.0	18
273.3	5
273.7	3
274.0	23
274.3	2
274.7	2
275.0	23
275.3	7
275.7	12
276.0	20
276.3	7
276.7	6
277.0	19
277.3	5
277.7	5
278.0	19
278.3	2
279.0	15
279.3	2
279.7	4
280.0	16
280.3	6
280.7	4
281.0	18
281.3	5
281.7	6
282.0	17
282.3	7
282.7	1
283.0	13
283.3	4
283.7	6
284.0	12
284.3	6
284.7	2
285.0	15
285.3	3
285.7	3
286.0	13
286.3	3
286.7	2
287.0	11
287.3	2
287.7	3
288.0	17
288.3	5
288.7	4
289.0	14
289.3	2
289.7	1
290.0	11
290.3	1
290.7	7
291.0	22
291.3	2
291.7	3
292.0	13
292.3	5
292.7	3
293.0	15
293.3	5
293.7	6
294.0	10
294.3	1
294.7	4
295.0	11
295.3	3
295.7	4
296.0	11
296.3	5
297.0	13
297.3	1
297.7	6
298.0	10
298.3	2
299.0	18
299.3	1
299.7	5
300.0	11
300.7	2
301.0	9
301.3	1
301.7	2
302.0	12
302.3	3
302.7	2
303.0	9
303.3	2
303.7	3
304.0	5
304.3	2
304.7	1
305.0	15
305.3	1
305.7	2
306.0	7
306.3	1
307.0	10
307.3	1
308.0	8
308.3	2
308.7	2
309.0	11
309.3	3
309.7	2
310.0	10
311.0	6
311.3	7
312.0	10
312.3	2
312.7	2
313.0	9
313.3	2
314.0	10
314.3	1
315.0	15
315.3	1
315.7	3
316.0	10
316.3	3
316.7	2
317.0	6
317.3	1
317.7	4
318.0	7
318.3	1
319.0	3
319.3	1
319.7	3
320.0	5
320.3	2
320.7	4
321.0	10
321.3	2
322.0	7
322.3	4
322.7	1
323.0	4
323.3	1
323.7	4
324.0	5
324.3	3
324.7	2
325.0	7
325.3	3
325.7	2
326.0	7
326.3	1
326.7	2
327.0	6
328.0	2
328.3	2
328.7	2
329.0	3
329.3	1
329.7	2
330.0	8
330.3	3
330.7	1
331.0	4
331.7	1
332.0	6
332.3	2
332.7	1
333.0	4
333.3	1
333.7	1
334.0	6
334.3	3
334.7	1
335.0	5
335.7	1
336.0	7
336.3	1
336.7	2
337.0	3
337.3	1
338.0	1
338.3	1
338.7	3
339.3	2
339.7	4
340.0	3
341.0	4
341.7	2
342.0	6
342.3	2
342.7	4
343.0	3

IP	n
343.3	3
343.7	3
344.0	2
344.7	1
345.3	2
346.0	3
346.3	2
346.7	4
347.0	2
347.7	2
348.0	4
348.3	2
348.7	3
349.0	1
349.3	1
349.7	1
350.0	4
351.0	2
351.7	3
352.0	3
352.3	2
353.0	4
353.3	1
353.7	1
354.3	1
354.7	1
355.0	1
355.7	1
356.7	1
357.3	2
357.7	1
358.0	4
359.0	4
360.0	1
360.7	3

IP	n
361.0	1
361.7	1
362.0	3
363.0	2
363.7	1
364.0	2
364.3	2
365.0	1
365.7	1
366.3	3
366.7	3
367.0	2
367.3	1
367.7	1
368.0	2
368.7	1
369.0	2
369.3	1
369.7	3
370.0	2
371.0	2
371.3	2
371.7	1
372.0	1
372.3	2
372.7	1
373.0	2
374.0	1
375.0	3
375.3	1
376.0	4
376.7	1
377.3	2
377.7	1
378.0	2

IP	n
378.7	2
379.0	2
379.7	2
380.0	2
380.3	2
381.0	1
381.3	1
382.0	2
382.3	1
382.7	4
383.0	3
383.3	2
384.3	1
384.7	1
385.7	1
386.7	1
387.3	1
387.7	2
388.0	2
388.3	1
389.0	3
389.3	2
390.0	2
390.3	2
390.7	1
391.0	2
391.7	1
392.7	2
393.0	2
393.3	1
393.7	1
394.0	1
394.3	1
394.7	1
395.7	1

IP	n
396.7	2
397.0	1
397.3	1
398.0	1
399.7	1
400.0	1
400.3	1
401.0	1
401.3	2
402.0	1
402.3	1
403.0	1
404.0	2
404.7	1
405.0	2
405.7	1
406.3	1
407.0	1
407.7	1
408.0	2
408.7	1
409.0	1
410.0	1
410.3	1
411.0	2
413.0	1
414.0	1
414.3	1
415.0	2
415.7	1
416.3	1
416.7	1
418.0	1
418.7	1
419.3	2

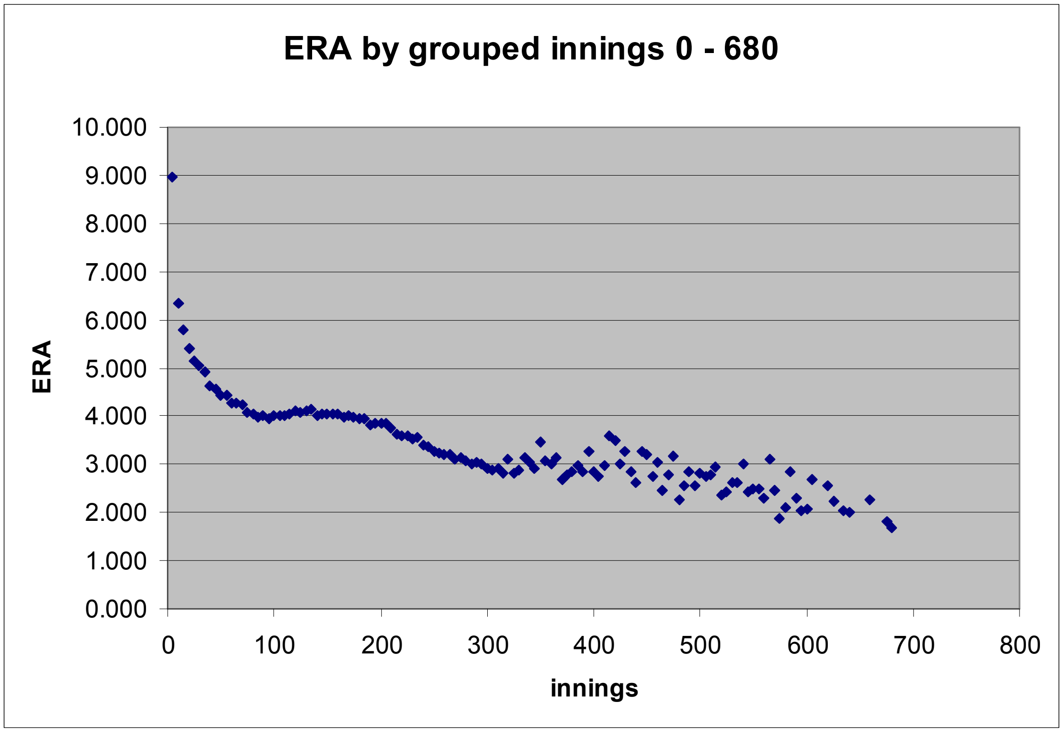
IP	n
420.0	2
420.7	1
421.7	1
422.3	2
422.7	1
423.7	1
424.0	2
425.0	2
425.3	2
426.0	1
426.3	1
427.7	1
429.7	2
430.7	1
431.0	1
432.3	1
432.7	1
434.0	1
434.3	2
434.7	1
437.3	1
437.7	1
439.3	1
440.0	2
440.3	1
440.7	1
441.0	1
441.3	2
444.0	1
444.3	1
445.0	1
445.3	1
445.7	1
447.3	2
449.0	1

IP	n
450.0	1
451.0	1
453.0	2
453.3	1
454.7	1
455.3	1
456.0	1
457.3	1
458.0	1
458.7	1
459.7	1
460.3	1
460.7	2
461.0	1
462.0	1
462.3	1
464.0	1
466.3	1
466.7	1
468.0	1
469.7	1
470.7	1
473.7	1
474.0	3
474.3	1
478.7	1
480.0	3
480.7	1
481.7	1
482.0	1
482.3	1
482.7	1
483.3	1
485.0	1
487.0	2

IP	n
491.3	1
492.0	2
493.0	1
493.3	1
494.7	1
496.0	1
500.0	1
500.3	1
500.7	1
501.0	2
503.0	2
504.0	1
506.0	2
509.0	1
509.3	1
512.0	1
513.3	1
513.7	1
514.0	1
516.0	1
516.7	1
517.3	1
518.7	1
521.0	1
523.0	2
526.0	1
528.7	1
529.7	1
532.0	1
532.7	2
536.3	1
538.0	1
538.3	1
539.3	1
540.0	1

IP	n
543.3	1
546.3	1
548.7	1
552.3	1
555.3	2
557.3	1
559.0	1
561.0	2
567.0	1
569.0	1
573.0	1
574.0	1
577.0	1
581.0	1
583.0	1
585.7	1
587.0	2
588.7	1
589.3	1
590.7	1
593.0	1
595.0	1
595.7	1
603.0	1
619.0	1
620.0	1
622.0	2
623.0	1
632.3	1
636.3	1
656.3	1
657.7	1
670.7	1
678.7	1
680.0	1

One item that is not likely to surprise anyone is the following chart, showing ERA by IP. In order to not have too much clutter, I've done it with grouped IP (0-5, 5.3-10, etc.)



Again, the wild scattering above IP=300 is due to small n for those IP. But surely no one is surprised that ERA goes down as IP goes up. One item that is interesting is the temporary increase at around 100 IP. The data are shown on the next page.

Notice the ERAs are strictly decreasing until 80 < IP ≤ 85. ERAs generally increase until 130 < IP ≤135 and then they generally decrease again. Obviously better pitchers get more IP. Relief pitcher usage, and its changes over time, is a confounding variable here. Perhaps 80-100 IP is around the point where managers finally realize that a pitcher, who had been very effective, has finally lost it (eg., Steve Blass in 1973).

There are two possible occurrences of the Law of Large Numbers in this data. The question is which it is. Is it that major league pitchers have a tendency to a certain ERA and as the number of IP increases they tend to that? Or is it that more and more pitchers with a particular number of IP show the tendency of pitchers with that IP to have a particular ERA?

As in the essay by Mr. Katz, there are too many unknowns to be able to draw any clear conclusions. The one conclusion that seems clear is "good pitchers pitch more."

from	to	ERA
0	5	8.949
5.3	10	6.349
10.3	15	5.783
15.3	20	5.411
20.3	25	5.153
25.3	30	5.038
30.3	35	4.931
35.3	40	4.628
40.3	45	4.560
45.3	50	4.441
50.3	55	4.425
55.3	60	4.285
60.3	65	4.264
65.3	70	4.229
70.3	75	4.072
75.3	80	4.061
80.3	85	3.988
85.3	90	4.011
90.3	95	3.963
95.3	100	4.019
100.3	105	4.010
105.3	110	4.024
110.3	115	4.049
115.3	120	4.117
120.3	125	4.065
125.3	130	4.096
130.3	135	4.151
135.3	140	4.022
140.3	145	4.030
145.3	150	4.037
150.3	155	4.037
155.3	160	4.049
160.3	165	3.983
165.3	170	4.006
170.3	175	3.970
175.3	180	3.949
180.3	185	3.941
185.3	190	3.833
190.3	195	3.851
195.3	200	3.844
200.3	205	3.848
205.3	210	3.742
210.3	215	3.630
215.3	220	3.592
220.3	225	3.579
225.3	230	3.530
230.3	235	3.560
235.3	240	3.407
240.3	245	3.363
245.3	250	3.271
250.3	255	3.239
255.3	260	3.191
260.3	265	3.214
265.3	270	3.100
270.3	275	3.154
275.3	280	3.078
280.3	285	3.024
285.3	290	3.029
290.3	295	2.998
295.3	300	2.915
300.3	305	2.896
305.3	310	2.910
310.3	315	2.831
315.3	320	3.111
320.3	325	2.816
325.3	330	2.888
330.3	335	3.129
335.3	340	3.055
340.3	345	2.928
345.3	350	3.466
350.3	355	3.061
355.3	360	3.021
360.3	365	3.155
365.3	370	2.697
370.3	375	2.775
375.3	380	2.839
380.3	385	2.990
385.3	390	2.855
390.3	395	3.281
395.3	400	2.858
400.3	405	2.750
405.3	410	2.977
410.3	415	3.582
415.3	420	3.505
420.3	425	2.996
425.3	430	3.274
430.3	435	2.845
435.3	440	2.617
440.3	445	3.284
445.3	450	3.195
450.3	455	2.748
455.3	460	3.036
460.3	465	2.468
465.3	470	2.781
470.3	475	3.178
475.3	480	2.280
480.3	485	2.546
485.3	490	2.856
490.3	495	2.545
495.3	500	2.828
500.3	505	2.741
505.3	510	2.788
510.3	515	2.946
515.3	520	2.354
520.3	525	2.429
525.3	530	2.636
530.3	535	2.637
535.3	540	3.026
540.3	545	2.418
545.3	550	2.490
550.3	555	2.477
555.3	560	2.291
560.3	565	3.120
565.3	570	2.448
570.3	575	1.883
575.3	580	2.090
580.3	585	2.861
585.3	590	2.292
590.3	595	2.054
595.3	600	2.056
600.3	605	2.701
605.3	610	------
610.3	615	------
615.3	620	2.571
620.3	625	2.226
625.3	630	------
630.3	635	2.050
635.3	640	1.994
640.3	645	------
645.3	650	------
650.3	655	------
655.3	660	2.281
660.3	665	------
665.3	670	------
670.3	675	1.798
675.3	680	1.683

Fred Worth, WORTHF@hsu.edu ♦

The Effects of Travel on Home Field Advantage

Andrew Boslett, Matt Hoover, Thomas J. Pfaff

One hypothesis for the cause of home field advantage is that visiting teams need to adjust to travel and to the park characteristics of the home team. If that's the case, it seems reasonable that there should be more of an effect for the earliest games in a series or road trip, than for the later games. Here, the authors examine the 2006 season to investigate.

In sports, home-field advantage is the difference between winning percentage at home and winning percentage on the road. In the 2006 major leagues, the visiting team only won 45.4% of the time**.** There are other possible explanations besides which team bats last on why there is such an advantage to being the home team. First, it is important to note that baseball fields do not have a standard configuration. This is in contrast to most other sports, such as basketball and football, which have set criteria on the playing arena's exact dimensions and characteristics. Because each baseball park has a different arrangement, it stands to reason that an away team could face a learning curve in adjusting to another park's surroundings and that a team might tailor its roster to the advantage. Furthermore, flying into a city and living out of a hotel could possibly have an effect on how an away team performs. It can be assumed that living in one's own home is more advantageous than living in a hotel. Moreover, a team could possibly tire over a long period of consecutive road series.

If travel affects player performance or if players must adjust to new ballparks, as opposed to just living out of a hotel, we would expect that visiting team winning-percentages would increase over the course of a series. That is, one would expect there to be a gradual improvement in winning percentage from 1st games of series to 3rd games of series.

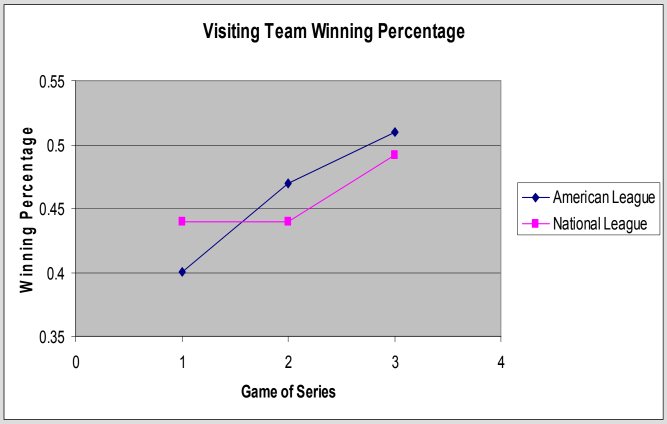
Our study's original focus was to analyze and compare visiting-team winning percentages of intra-league games in 2006. That year, American League teams exhibited a much greater improvement over the course of a series than did the National League teams. Due to these findings, our investigation shifted to trying to find out possible explanations behind the difference between the American and National Leagues.

Results – Series

The data set used was the set of intraleague games from the 2006 season. The winning percentage of visiting teams was calculated for each of the first three games of each series. (We elected to ignore the few fourth games of series due to a small sample size.) The number of games sampled for the American League for games one, two, and three was 325, 321, and 293, respectively, and for the National League it was 376, 373, and 347, respectively. The general trend in winning percentage for the visiting teams can be clearly seen in the data, from [4], in figure 1.

From the graph, you can see an upward trend in winning percentage for visiting teams in both leagues. In the American League the trend is more pronounced: visiting teams on average only win 2 of every 5 first games of a series, but by the third game of the series the winning percentage is over 50%, meaning that home field advantage has effectively evaporated. The trend in the National league is not as extreme: winning percentage starts at about 45%, remains constant during the second game, and jumps up to 49% for the final game of the series.

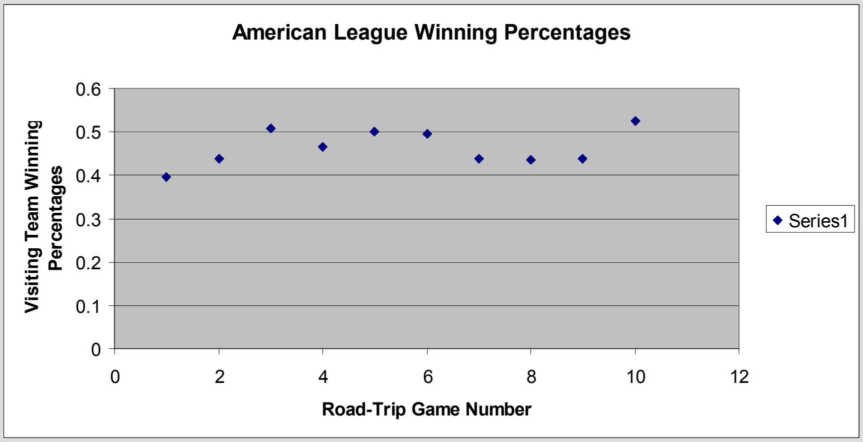
Figure 1 – Winning Percentage by Game of Series



To determine whether the distribution of wins and losses was statistically different for each game of the series we performed a chi-squared test to test for equal winning percentage over the three games. For the American League there is strong statistical evidence that the likelihood of the visiting team winning does not remain constant over the course of the series. The chi-squared value for the American league was 7.269 giving a p-value of 0.026. For the National league the chi-squared value was only 2.021 giving a p-value of 0.364. While the differences in winning percentage were significant at the 5% critical level for the American League, the National League results were not significant.

A two-proportion test performed on the first and third game of the series provided similar results. The American League had a p-value of 0.007, while the National League had an insignificant 0.213 p-value.

Figure 2 – Visitor Winning % by Game of Road Trip, AL



Results – Road Trips

Figure 2, data from [4], gives the American League visiting team winning percentage over the course of a road trip. We see general upturns from games 1 to 3, 4 to 6, and 8 to 10, signifying probable series. The number of games sampled for each of the ten games is as follows: 149, 148, 148, 135, 128, 123, 73, 39, 32, and 19. As the road-trip progresses, however, we do not see a general decline in winning percentage. The p-value, 0.73, is from a chi-squared test comparing the winning percentages over the course of a road-trip. It is not statistically significant and, thus, there is no detectable impact of road-trip game number on winning percentage. Further, a linear regression test yields a p-value of 0.370 (r^2 = 0.101) with a slope of 0.004.

Interestingly, if we use only the first six games (since there is a notable drop in the size of the data set from six to seven) we get a p-value of 0.059 (r^2 = 0.632) with a slope of 0.018. This suggests an improvement in the visitors' winning percentage on a road trip, at least for the first six games on the road.

Figure 3 – Visitor Winning % by Game of Road Trip, NL

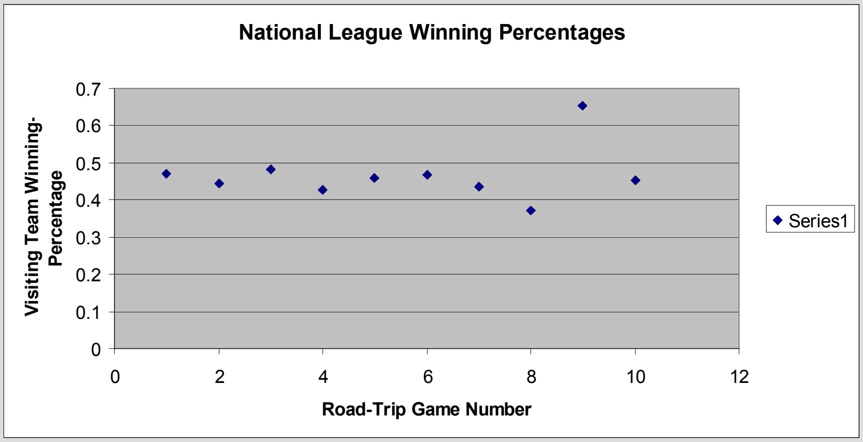


Figure 3, data from [4], gives the National League's winning percentage over the course of a road trip. The number of games sampled for each of the ten games is as follows: 172, 171, 168, 152, 146, 139, 85, 54, 46, and 22. The data for the National League is much more sporadic than the data for the American League. We see a drop in winning percentage between 1st and 2nd games of road trips. There is also a tremendous increase in winning percentage between 8th games and 9th games (from under 40% for 8th games to approximately 65% in 9th games of series). Furthermore, there is a general decrease in winning percentage (besides the 9th game outliers) as the road trip proceeds. The p-value, 0.347, is from a chi-square test comparing the winning percentages over different games of a road trip. It is not statistically significant at a 5% critical level and thus there is no discernible difference in winning percentage in the National League over the course of a road-trip. A linear regression test with the game nine outlier removed yields a p-value of 0.252 ($r^2 = 0.182$) with a slope is –0.005. As with the American League, we looked at just the first six games to get a p-value of 0.888 ($r^2 = 0.006$) with a slope of –0.001. This is very different from the results of the first six games of a road trip for the American League. It appears that AL teams may improve over the early part of a road trip, but NL teams did not.

Park Factor and Travel Distances

Figure 4, data from [1], illustrates the variance and standard deviation for ESPN's Park Factor statistics. The Park Factor measures the effects of a baseball stadium on offensive production as compared to other parks. It is a weight that can be applied to batting average to indicate how a batter should perform in a given ballpark. If the Park Factor is equal to 1.0, there is no effect on batting performance. If it is greater than 1, then the batter is expected to perform better against the pitcher in that park as opposed to a park that has a smaller Park Factor. If it is less than 1, then the pitcher has the advantage in the given park. Comparing the variance of ballparks in the American and National Leagues should indicate if there is a difference between the structural characteristics of the ballparks in each league. To find out if there is a statistically significant difference between the variances of the American League and the National League, we performed an F-Test and obtained a p-value of 0.574. Although the difference is not statistically significant at a .05 level, this does not necessarily mean that it is not significant in the real world.

Table 4 – SD and Variance for ESPN's Park Factors

League	Standard Deviation	Variance
American	0.076	0.005706
National	0.088	0.007824

Table 5, data from [3], gives the average travel distance in miles between all cities of each league. We acquired the distances by compiling the distances between each league's teams. These results support the hypothesis that American League teams travel more than National League teams. It may be that American League teams to have a steeper assimilation curve; on the other hand, the 62 mile difference may not have any real life significance.

Table 5 – Average Travel Distance (in miles) Between League Cities

League	Distance
American	1438
National	1376

Conclusion

The trends observed support the hypothesis that visiting team performance improves over the course of the series as teams adjust to the new ballparks and rest after travel. However, if this was indicative of the true cause of home-field advantage, we would expect similar results for the National and American Leagues. This is not what happened. Because of this, we proceeded to check if there was any difference in the way teams adjusted to new ballparks or traveled in the American and National Leagues.

We examined visiting team winning percentage over the course of a road trip to see if the scheduling differences between the leagues could explain our results. Our tests show that there is no statistical evidence that the game of the road trip affects visiting team winning percentage. While American league ballparks are marginally farther away from each other on average, we were not able to directly look at the effect of distance traveled on visiting team winning percentage. This would be a promising area for future research.

Finally, we looked at the differences in the variation of the Park Factors between American League and National League parks. The idea behind the Park Factor was to capture all of the structural characteristics of a ball park that affect team performance into one statistic. And, while it is clearly an imperfect measure, there does not seem to be a significant difference between the required adjustments for American and National League ballparks.

None of the data we analyzed provided us with evidence that contradicted our hypothesis. Though analysis of park factors and travel distances produced no statistically significant effects the results for the National and American league showed roughly the correct relationship. A look at these trends over the course of many seasons would be needed to see if the effects are truly statistically significant.

Bibliography

[1] ESPN.com, http://www.espn.com.
[2] Major League Baseball, http://mlb.mlb.com.
[3] Rand McNally Road Atlas 07. U.S/Canada/Mexico Copyright 2007 by Rand McNally & Company Chicago, Illinois.
[4] Retrosheet, http//www.retrosheet.org.

Andrew Boslett, 58 Corning Blvd, Corning, New York, aboslet1@ithaca.edu
Matthew Hoover, 619 Washington St, Wellesley MA 02482 mhoover1@ithaca.edu
Thomas J. Pfaff, Department of Mathematics, Ithaca College, Ithaca, NY 14850, tpfaff@ithaca.edu ♦