DOI 10.1515/jqas-2012-0034 — JQAS 2013; 9(3): 271–283

David Beaudoin*

Various applications to a more realistic baseball simulator

Abstract: This paper develops a simulator for matches in Major League Baseball (MLB). Aspects of the approach that are studied include the introduction of base-running probabilities which were obtained through a large data set, and the simulation of nine possible outcomes for each at-bat. Various applications to the simulator are investigated, such as the definition of a measure of the ability of a batter/pitcher, in-play strategy and the determination of the optimal batting order for a given team.

Keywords: analysis of variance; in-play strategy; in-play probabilities; logistic regression; major league baseball; measure of performance; optimal batting order; simulation.

***Corresponding author: David Beaudoin,** Associate Professor, Département Opérations et Systèmes de Décision, Faculté des Sciences de l'Administration, Pavillon Palasis-Prince, Bureau 2636, Université Laval, Québec (Québec), G1V0A6 Canada, e-mail: david.beaudoin@osd.ulaval.ca

1 Introduction

Professional sports teams are always looking for ways to gain an edge over their opponents. In baseball, here are methods to do so: optimizing in-play strategy (e.g., should a team attempt stealing a base or not?), finding the batting order that maximizes a team's expected number of runs, and measuring accurately the true ability of players (which may lead to signing underrated free agents or drafting more efficiently). This paper aims to contribute in this regard through a very realistic game simulator.

The book "Moneyball" by Michael Lewis (2003) opened a lot of people's eyes regarding the usefulness of statistics in baseball. It describes different approaches for fielding a competitive team (new prospect selection methods, the use of more meaningful statistics when gauging players, etc.). Since the publication of that book, more and more teams have been inclined to hire statisticians or "Sabermetric" analysts. The term "Sabermetric" is derived from the acronym SABR, which stands for the Society for American Baseball Research (www.sabr.org). The main aim of this group is to study baseball history, but some "Sabermetric" people analyze baseball through objective analyses, rather than beliefs (which may be biased). The main purpose of this work is to provide tools for team management to increase their team's success on the field through a handful of applications.

Several authors have employed baseball game simulations in their work. A common trait to the procedures used in the past is the set of rules for runner advancement which are quite simplified and do not reflect accurately what is observed in Major League Baseball (MLB) games. For example, a single with a runner on second base may always score a run in simulated games. Moreover, outs are treated such that runners never advance following the play, which is clearly not the case in real games. Strike outs, ground outs and fly outs are pooled together and viewed as a single possible outcome.

The lack of appropriate data with respect to the way runners advance in all kinds of situations explains the application of such simple rules. More detailed data are now available through the source www.retrosheet.org and some authors have gone on to publish entire books of empirical data, such as Tango et al. (2006). Prior to that, authors had to rely on simple baserunning rules. We now list some references to projects using such methods. Please note that there exist enormous non-academic literature on this topic, but we focus mainly on academic work.

D'Esopo and Lefkowitz (1977) develop an interesting statistic called the "scoring index." Basically, they calculate the number of runs a player would generate if he batted in all nine spots in their modelling approach. In other words, the authors assume independent and identically distributed at-bats according to the batter's hitting distribution. This paper adheres to a simple model to determine the location of runners after the occurrence of any at-bat.

Cover and Keilers (1977) came up with a similar idea to evaluate a batter's performance, but this time through play-by-play computations. The latter consist of following the sequence of outcomes for a given player and calculating its resulting number of runs per game. The statistic is named the "offensive earned-run average" (OERA). Ano (2001) expands this measure to account for stolen bases, a measure the author calls the "modified offensive

earned-run average" (MOERA). Both papers follow conventions for runner advancement which are slightly different from the ones considered in D'Esopo and Lefkowitz (1977). Cover and Keilers (1977) and Ano (2001) assume singles and doubles to be long, i.e., a single moves all runners two bases, and a double scores all runners.

Mills and Mills (1970) also look for a measure of a batter's ability. These authors estimate the average contribution per at-bat to the probability of winning a game; this statistic is called the "player win percentage." Similarly, Lackritz (1990) evaluates players in terms of their effects on their team winning proportion.

Kinoshita (1987) applies the same concept to evaluate pitching abilities. This work simulates at-bats according to probabilities suggested by the pitcher's number of singles, doubles, triples, home runs and walks allowed in the season.

The main objective of the paper by Hirotsu and Wright (2005) is to optimize pitcher substitution strategies incorporating handedness of opposing batters. The technique is illustrated via a fictitious game between the San Francisco Giants and the Colorado Rockies. The authors simulate games complying with the rules of runner advancement proposed by D'Esopo and Lefkowitz (1977).

Bukiet, Harold, and Palacios (1997) introduce a Markov chain method, which is then applied to achieve several goals such as finding optimal batting orders and predicting the number of games a team might win. Finally, we mention the work of Sueyoshi et al. (1999) which proposes an approach for baseball evaluation referred to as "a benchmark approach." This paper combines the OERA defined above with data envelopment analysis (DEA). The authors consider the offensive results of 30 players belonging to the Central League in Japan.

More recently, Baumer (2009) examines the impact of baserunning ability on a team's number of runs scored over a complete season. The method makes player-specific estimation, but considers only eight ways to "take the extra base" and does not account for the number of outs, a factor that turns out to be statistically significant (see Section 2). Here is a quote that shows the author recognized the potential of incorporating this factor: "Since runners are more likely to be moving with the pitch with two outs, a revision of the baserunning probabilities to take the number of outs into account could be fruitful."

We also mention the work of Beaudoin and Swartz (2010), who develop a hockey simulator in a similar way. Based on a vast data collection process, these authors estimate several parameters and simulate games in order to assess strategies for pulling the goalie late in games. Their simulation program has several applications, just like the one suggested in this paper with respect to baseball.

This paper promotes a much more lifelike baseball simulator, which is based on a broad data collection process. The data grants accurate estimation of runner movement on the bases as a function of the batting outcome and the current situation (position of runners on the bases and number of outs), as well as the simulation of throwing errors. The possible outcomes are more elaborate than previous models, as outs are subcategorized as strike outs, ground outs and fly outs. Runners may now advance following an out, a feature that was non existent in prior work, despite its very common occurrence in real games. A batter reaching on an error is also made possible via probabilities obtained through data. Another key element related to our simulator concerns the estimation of the probability of each possible outcome for any given batter-versus-pitcher confrontation. This crucial step is made in part by virtue of a logistic regression analysis.

The simulation model is detailed in Section 2. We describe an at-bat's possible outcomes, as well as the estimation of their respective probabilities. The set of rules for runner advancement is also presented in this portion. The following three sections relate to numerous applications of the simulator. More precisely, Section 3 defines a measure of the ability of a batter/pitcher called the number of runs generated per game (NRGG). This statistic is calculated for all batters and pitchers in MLB based on their 2009 achievements, and the top 20 is shown for both categories of players. A very important aspect of baseball is studied in Section 4: optimizing in-play strategy. The methodology is illustrated via two scenarios encountered during the 2009 ALCS (American League Championship Series). The last application of the simulation program studied in this project concerns optimal batting orders (see Section 5). The technique is exemplified by finding the best ordering of New York Yankees players. We conclude with a short discussion in Section 6, which exhibits some potential improvements to the current version of the baseball simulator.

2 Simulation model

Each at-bat is simulated according to the multinomial distribution with parameters n=1 and $\mathbf{p}=(p_1, \ldots, p_9)$. In other words, every at-bat can yield nine possible outcomes, which are presented below:

1. Single (1B)
2. Double (2B)
3. Triple (3B)
4. Home run (HR)

5. Base on balls (BB)
6. Hit by pitch (HBP)
7. Strike out (SO)
8. Ground out (GO)
9. Fly out (FO)

Notice that a player reaching the bases via error is handled in the GO outcome, as will be discussed later. Double plays can also occur following a GO or a FO, as this play depends on the rules for runner advancement. This topic is also covered a bit later.

The estimation of the parameters p_1=Pr(single), p_2=Pr(double),..., p_9=Pr(fly out) is vital to the realism of the simulator. How do you estimate those nine parameters for a matchup between a given batter *B* and a given pitcher *P*? Hirotsu and Wright (2004) estimate a parameter ω to calibrate the batting probabilities. A different approach is taken here.

The log5 approach by Bill James (1981) also studies the batter-pitcher matchup. It does account for the opposing pitcher's ability, but does not allow the simulation of the various outcomes of an at-bat. It is based on batting averages, which do not account for walks and HBP.

Clearly, p_1,..., p_9 depend on $p_i^{(b)}$ and $p_i^{(p)}$, where:

- $p_i^{(b)}$ =probability of occurrence of outcome *i* for batter *B* facing an average pitcher (for *i*=1, 2,...,9)
- $p_i^{(p)}$ =probability of occurrence of outcome *i* for pitcher *P* facing an average batter (for *i*=1, 2,...,9)

The statistics on Major League Baseball's official website (www.mlb.com) enable us to easily estimate the parameters $p_i^{(b)}$ and $p_i^{(p)}$ described above for any batter and any pitcher (those estimators will be called $\hat{p}_i^{(b)}$ and $\hat{p}_i^{(p)}$ from now on). Indeed, the statistics of a certain batter were obtained against a wide variety of pitchers, whose abilities differ greatly from one another. It seems reasonable to assume that the statistics of a batter reflect his performance against an average pitcher. A similar argument could be made regarding pitching statistics.

The goal is therefore to estimate the probability of each possible outcome for a given matchup based on the batter's and pitcher's statistics. This was done by first carrying out the following logistic regression model:

$$log(PPO/(1-PPO))=\beta_0+\beta_1*PPOB+\beta_2*PPOP+\epsilon,$$

where:

- PPO=$p_1+p_2+p_3+p_4+p_5+p_6$=P(positive outcome) for the current matchup between batter *B* and pitcher *P*
- PPOB= $p_1^{(b)}+p_2^{(b)}+p_3^{(b)}+p_4^{(b)}+p_5^{(b)}+p_6^{(b)}$ =P(positive outcome) for batter *B* facing an average pitcher
- PPOP= $p_1^{(p)}+p_2^{(p)}+p_3^{(p)}+p_4^{(p)}+p_5^{(p)}+p_6^{(p)}$ =P(positive outcome) for pitcher *P* facing an average batter

From the definitions above, a "positive outcome" corresponds to any of the first six events described earlier (1B, 2B, 3B, HR, BB, HBP). The logistic model thus incorporates the information on the batter and the pitcher to come up with a probability that the batter will reach the bases safely.

Data was collected on over 43,000 at-bats from the MLB 2009 regular season. For each matchup between batter *B* and pitcher *P*, the following three variables were calculated:

- $\widehat{PPOB}=\hat{p}_1^{(b)}+\hat{p}_2^{(b)}+\hat{p}_3^{(b)}+\hat{p}_4^{(b)}+\hat{p}_5^{(b)}+\hat{p}_6^{(b)}$
- $\widehat{PPOP}=\hat{p}_1^{(p)}+\hat{p}_2^{(p)}+\hat{p}_3^{(p)}+\hat{p}_4^{(p)}+\hat{p}_5^{(p)}+\hat{p}_6^{(p)}$
- *Outcome* =1 if the at-bat resulted in 1B, 2B, 3B, HR, BB or HBP, 0 if the at-bat resulted in SO, GO or FO

The logistic regression procedure with *Outcome* as the dependent variable and $\widehat{PPOB}$ and $\widehat{PPOP}$ as the independent variables provide the estimated model below:

$$log\left(\frac{\widehat{PPO}}{1-\widehat{PPO}}\right)=-3.6538+4.3932*\widehat{PPOB}+4.5242*\widehat{PPOP},$$

Simple algebra yields:

$$\widehat{PPO}=\exp(z)/(1+\exp(z)),$$

where $z=-3.6538+4.3932*\widehat{PPOB}+4.5242*\widehat{PPOP}$

All three parameter estimates are largely significant (<1×10e-16), with the standard errors of β_0, β_1 and β_2 taking values 0.12, 0.22 and 0.27, respectively. Both parameters associated with the quality of the batter and the pitcher (β_1 and β_2) turn out to be >0. This was expected since a batter whose probability of a positive outcome is very high (i.e. one of the best batters in baseball) should increase the value of the occurrence of a positive outcome for the current matchup between himself and any pitcher (similarly for a pitcher whose value of $\widehat{PPOP}$ is large, i.e., one of the worst pitchers).

We motivate the model by discussing the interpretation of the resulting parameter estimates. First, we check whether a matchup between an average batter and an average pitcher yields a probability of positive outcome which makes sense. Combining all MLB players' 2009 statistics together, we obtain a probability of occurrence of a positive outcome to be 0.327. Plugging this value in the estimated logistic regression model (i.e., replacing $\widehat{PPOB}$

and $\widehat{PPOP}$ by this value) we obtain $\widehat{PPO}$ =0.323, which turns out to be pretty close to 0.327. Also, one could expect $\beta_1=\beta_2$. The estimates, 4.39 and 4.52, are fairly close and well within each other's confidence interval.

Once the probability of a positive outcome, PPO, has been estimated by $\widehat{PPO}$ for a given matchup between batter B and pitcher P, the values of $\hat{p}_1^{(b)},\ldots,\hat{p}_6^{(b)}$ are adjusted so that their sum equals $\widehat{PPO}$. In other words, we modify the probability of occurrence of each of the six positive outcomes to account for the strength of the pitcher currently on the mound. A similar strategy is used in order to estimate the parameters associated with negative outcomes, i.e. p_1, p_8 and p_9. More precisely, the final estimates for the matchup of interest are provided by the following equations:

– $\hat{p}_i=\hat{p}_i^{(b)}*\widehat{PPO}/\widehat{PPOB}$, for i=1, 2, 3, 4, 5, 6

– $\hat{p}_i=\hat{p}_i^{(b)}*(1-\widehat{PPO})/(1-\widehat{PPOB})$, for i=7, 8, 9

We are thus assuming that the probabilities will scale proportionally to only the batter's statistics. McCracken (2001) shows that both batter and pitcher tendencies are responsible for issuing walks and homeruns, but the frequency with which singles, doubles and triples are hit is primarily a function of the batter's ability. That would be a possible improvement to the current method.

It can be easily verified that, as it should:

$$\sum_{i=1}^{6}\hat{p}_i=\widehat{PPO}$$
$$\sum_{i=7}^{9}\hat{p}_i=1-\widehat{PPO}$$
$$\sum_{i=1}^{9}\hat{p}_i=1$$

Let us demonstrate the method through a matchup between Ryan Howard from the Philadelphia Phillies and Tim Lincecum from the San Francisco Giants. Their values of $\hat{p}_i^{(b)}$ and $\hat{p}_i^{(p)}$, for i=1,...,9 are presented in Table 1, based on their 2009 statistics.

It can be inferred from Table 1 that $\widehat{PPOB}$ =0.1211+0.0521+0.0056+0.0634+0.1056+0.0085=0.3563, while $\widehat{PPOP}$ = 0.1301+0.0395+0.0043+0.0110+0.0748+0.0066=0.2663. From the estimated logistic regression model, we obtain $\widehat{PPO}$ =0.2924. The probability of each of the six positive outcomes for Howard (the first six probabilities in the first row of Table 1) must be lowered to account for the fact that he is facing one of the best pitchers in the game. In fact, from the equations seen earlier, $\hat{p}_i=\hat{p}_i^{(b)}$ *0.2924/0.3563=0.8207* $\hat{p}_i^{(b)}$, which means that the probability of every positive outcome for Howard must be lowered by 17.93%. On the other hand, it can be shown that each of the three negative outcomes for Howard must have their respective probabilities increased by 9.93% ($\hat{p}_i=\hat{p}_i^{(b)}$ *(1-0.2924)/(1-0.3563)=1.0993* $\hat{p}_i^{(b)}$). The final estimates that are used by the program to simulate an at-bat between Ryan Howard and Tim Lincecum are displayed in Table 2. For example, it can be seen that Howard's probability of striking out, which was 26.2% versus an average pitcher, is raised to 28.8% against Lincecum. The chances of hitting a single went from 12.1% (vs average) to 9.9% (vs Lincecum).

A batter with too few at-bats, or a pitcher with a small number of batters faced may yield probabilities which do not make much sense (e.g., a batter with 3 home runs in only 10 at-bats). Special care must be taken in such cases, and several options may be considered like assigning average (or below-average) statistics to players not having sufficient data. Or perhaps the user may choose to include data from previous seasons.

Notice that a batter reaching first, second or third base on a defensive error was accounted for in the outcome GO. Data on errors were collected for all 2430 games from the 2009 season. The information gathered allowed for the estimation of Pr(batter reaches base i following an error | the batter hit a ground ball), for i=1, 2, 3. When simulating one of the nine possible outcomes, if the random number points towards the occurrence of outcome GO, a second random number is generated to check if the batter indeed grounds out, or if he reaches one of the three bases on an error. The data suggest that Pr(batter reaches first base on error | GO)=3.083%, Pr(batter reaches second base on error|GO)=0.501% and Pr(batter reaches third base on error | GO)=0.058%.

The rules of runner advancement have a large impact on the quality of the simulations, especially when

Table 1 Estimated probabilities of each of the nine outcomes for a given batter and a given pitcher (when facing average opposition).

Player	1B	2B	3B	HR	BB	HBP	SO	GO	FO
Ryan Howard	0.1211	0.0521	0.0056	0.0634	0.1056	0.0085	0.2620	0.1916	0.1901
Tim Lincecum	0.1301	0.0395	0.0043	0.0110	0.0748	0.0066	0.2871	0.2442	0.2024

Table 2 Final probability estimates of each of the nine outcomes for a simulated at-bat between Ryan Howard and Tim Lincecum.

Matchup	1B	2B	3B	HR	BB	HBP	SO	GO	FO
Ryan Howard vs Tim Lincecum	0.0994	0.0427	0.0046	0.0520	0.0867	0.0070	0.2880	0.2106	0.2090

assessing in-play strategies. Most authors whose research required the use of a baseball simulator adopted the model developed by D'Esopo and Lefkowitz (1977). This simple model is as follows:

- In the case of a single: The batter and all runners advance one base, except for a runner on second base who scores automatically.
- In the case of a double: The batter and all runners advance two bases.
- In the case of a triple: The batter and all runners advance three bases.
- In the case of a home run: The batter and all runners score.
- In the case of a walk: The batter reaches first base, and all runners who are forced to move advance one base.
- In the case of an out: The batter is out, and all runners stay put.

Clearly, this set of rules does not reflect accurately what happens in real games. For example, a runner on second base does not always score following a single. He may hold at third base, or he may be thrown out. Also, the batter may advance an additional base while the team on defense is trying to retire the runner. Using a more sophisticated set of rules increases the realism of the simulator, and improves the quality of the results. For instance, in-play probabilities can be significantly biased if using rules of runner advancement which are inappropriate, especially late in games. Suppose the road team is trailing by a single run in the top of the 9th inning with a runner on second base. A simulator based on the model elaborated by D'Esopo and Lefkowitz (1977) overestimates the probability of the road team winning the game, since a single invariably scores the tying run (which may not necessarily be the case in reality).

One of the biggest difference between the model of D'Esopo and Lefkowitz (1977) and ours concerns the way runners are handled in the case of outs. The previous model simply does not allow runners to advance in such situations. This work splits "outs" into three distinct categories, namely strike outs, ground outs and fly outs. This discrimination is important as runners may advance following a ground out or a fly out. Therefore, a batter who has a tendency to strike out often will get penalized in our simulator (rightfully so). A runner on third base may now have the opportunity to score on a GO or a FO. Also, our more realistic model allows for double plays, or a batter reaching first base on a fielder's choice.

There are cases where the advancement of runners, if any, is obvious. If a player hits a home run, the batter and all runners score. In the case of a walk or a batter getting hit by a pitch, there is no doubt as to where the runners end up following the play. Let us call them "obvious situations." On the other hand, there are several circumstances where one cannot tell for sure the location of each runner following the at-bat. We shall call them "non obvious situations" and we list them here:

- The batter singles with at least one runner on base, except if there was a runner on third base only (in which case the runner almost always scores and the batter holds at first base).
- The batter hits a double. There had to be a runner on first base prior to the play.
- The batter grounds out with less than two outs with at least one runner on base.
- The batter flies out with less than two outs with at least one runner on base, except if there was a runner on first base only (in which case the runner almost never advances).

Each of the four situations above can be further subcategorized by specifying exactly where runners stood before the play occurred. A total of 23 "non obvious situations" are then obtained.

Data was collected (www.mlb.com) from the 2009 season on over 18,000 plays belonging to any of those "non obvious situations." The location of the batter and all runners before and after the play was recorded in an Excel file. For each "non obvious situation," an analysis of variance was achieved in order to test whether the number of outs had any impact on the runner advancement or not. Indeed, runners may be more aggressive as the number of outs increases. Tables 3 and 4 present an exhaustive list of all 23 "non obvious situations," as well as the conclusion provided by the F-test of every analysis of variance, at the 5% level, regarding the potential effect of the number of outs on the way runners react.

Table 3 "Non obvious situations" associated with outcomes 1B and 2B, specifying whether the runner advancement probabilities vary as a function of the number of outs.

Outcome	Location of runners prior to the play	Significance of the number of outs
1B	R1	Yes
1B	R2	Yes
1B	R1-R2	Yes
1B	R1-R3	Yes
1B	R2-R3	Yes
1B	R1-R2-R3	Yes
2B	R1	Yes
2B	R1-R2	Yes
2B	R1-R3	No
2B	R1-R2-R3	Yes

R1 means that a runner was on first base prior to the play. Similar definitions apply for R2 and R3 regarding second and third base prior to the play, respectively.

It can be seen from Table 3 that the number of outs turns out to be significant at the 5% level for all six subcategories related to outcome 1B. As a result, the simulator uses different sets of probabilities for runner advancement as a function of the current number of outs for these cases. When the analysis of variance concludes that the number of outs has no impact on runner advancement, the probabilities are pooled together over all possible numbers of outs.

Tables 13–19 in the supplementary reading paper show the probabilities pertaining to the rules for runner advancement in all "non obvious situations." Those are the probabilities used in the simulator. One important note: only situations which occurred in the data are presented in those tables. We are aware that other (rare) cases might happen in real life, but their probabilities were estimated as 0. Example: with runners on 2nd and 3rd base, a batter flies out and it turns out that the runner on 3rd base stays put, while the runner on 2nd base gets thrown out (perhaps following a spectacular catch in the outfield).

Table 4 "Non obvious situations" associated with outcomes GO and FO, specifying whether the runner advancement probabilities vary as a function of the number of outs.

Outcome	Location of runners prior to the play	Significance of the number of outs
GO	R1	Yes
GO	R2	No
GO	R3	No
GO	R1-R2	Yes
GO	R1-R3	No
GO	R2-R3	Yes
GO	R1-R2-R3	No
FO	R2	Yes
FO	R3	No
FO	R1-R2	Yes
FO	R1-R3	No
FO	R2-R3	No
FO	R1-R2-R3	No

R1 means that a runner was on first base prior to the play. Similar definitions apply for R2 and R3 regarding second and third base prior to the play, respectively.

The number of outs is significant for all six cases where the outcome is a single (see Table 3). A careful look at the proportions pertaining to these cases (see Tables 13 and 14 in the supplementary reading) exhibits the fact that runners are definitely more aggressive as the number of outs increase. A runner on second base prior to the occurrence of a single scores much more often with two outs than with none. The same comment could be made regarding a runner on first base trying to reach third base on a single.

Based on Table 4, the "non obvious situations" where the number of outs has a significant impact on the advancement of runners following a GO occurs when the location of the runners prior to the play was either R1, R1-R2 or R2-R3. The logical explanation with respect to the first two states may come from sacrifice bunts (which count as ground outs). The batter being retired with the runner(s) advancing was observed much more often with 0 out than with only one. As for the R2-R3 situation, we note that runners were much more conservative (both staying put on the ground ball) with no outs compared to one out.

Two out of the six subcategories related to outcome FO are significant with respect to the number of outs. They correspond to the only cases where a runner was located on second base with nobody on third base prior to the play (R2 and R1-R2). The results show that the runner on second base tried much more frequently to reach third base on the fly out when there were no outs (hoping to be 90 feet away from home plate with one out, increasing substantially the chances of scoring without having anyone getting a hit).

Stolen bases (SB) are omitted from the simulator, but may be included in a future version. The biggest obstacle faced by the programmer when trying to add this characteristic into the simulations is to determine a player's probability of *attempting* a stolen base. This probability can be estimated by the number of SB attempts by the given player divided by the number of times he was "in a position to attempt one." The latter is difficult to evaluate. Determining the number of times a given player stood on first base over the whole season is not that easy (adding

the singles, the walks and the HBP is not enough, since he may have reached first base on a fielder's choice, or on an error). Also, a player being on first base is not sufficient to qualify as being "in a position to attempt a SB" since there may be a runner on second base preventing him from trying to steal second base. To avoid having to spend a huge amount of time overcoming this hurdle it was decided to drop stolen bases from this version of the simulation program.

An additional feature was incorporated in the simulator: throwing errors. Out of the nine possible outcomes defined earlier, only five can lead to a throwing error during the play: 1B, 2B, 3B, GO and FO. Indeed, it is impossible (or highly unlikely) to observe a throwing error committed by the defense on a HR, a BB, a HBP or a SO. There exist two more common ways that a throwing error may occur during a game: on a stolen base attempt and on a pickoff attempt. We need not to worry about the first since SB are omitted from the simulator (as discussed previously). We thus end up with six possible plays where a throwing error may occur: 1B, 2B, 3B, GO, FO and on a pickoff attempt.

It was mentioned earlier that data on errors were collected over all games from the 2009 regular season. This data set was not restricted to errors leading to the batter reaching the bases, but also included throwing errors. The file specifies the number of bases the runners advanced following the throwing error, as well as the play that caused it. Proprietary restrictions prevent the publication of all estimated probabilities, but we reveal below the value of two estimates:

- Pr(runners advance ONE base on a throwing error|1B)=0.0142
- Pr(runners advance TWO bases on a throwing error|1B)=0.0014

As a consequence, every time the simulator determines that a player hits a single, the program first settles where all runners end up following the play (according to our rules of runner advancement), and then allows the possibility that all runners advance an additional base with a 1.42% chance, or an extra two bases with a 0.14% percentage.

Let us now summarize the various steps executed by the simulator for each at-bat:

- If there are runners on the bases prior to the play, simulate whether a throwing error occurs on a pickoff attempt or not.
- Calculate the values of $\hat{p}_i$ that prescribe the likelihood of each of the nine outcomes, based on the batter's and pitcher's statistics. Simulate the occurrence of one of those nine possible outcomes accordingly.
- If the combination of outcome, runners on base prior to the play and number of outs corresponds to an "obvious situation," move the runners in a suitable way.
- In the case of a "non obvious situation," simulate the advancement of the batter and the runners in accordance with the rules specified by our model.
- If the selected outcome was either 1B, 2B, 3B, GO or FO, simulate the occurrence of a throwing error (and the number of bases by which runners advance).

The simulator has been coded in the R programming language. The speed of simulations depends, of course, on the machine on which the program is executed, but just to give a general idea to the reader, let's mention that simulating one million games takes roughly 3 h to run.

3 Measure of the ability of a batter/pitcher

When a batter steps up to the plate, most baseball broadcasters present the following three statistics: the batting average (BA), the number of home runs and the number of runs batted in (RBI). In the case of a pitcher, the majority of television stations show his win-loss record, his earned run average (ERA), as well as the number of strike outs and walks. There exist several such statistics being amassed on every player in MLB. How do you compare two batters or two pitchers based on such a huge amount of variables? For example, is a player with a 0.250 BA and 30 HR better than a player with a 0.300 BA and 15 HR? Would not it be simpler if the quality of a player was measured in terms of a single statistic that incorporated all of the important information?

Perhaps the most intuitive statistic that can be developed for any given batter is its number of runs scored per game if that player filled all nine spots of the lineup. The interpretation of such measure is quite simple and its units, the number of runs scored per game, is easy to understand. The simulator described in Section 2 allows for the estimation of this statistic for all batters by simulating a large number of games. We simply let $\hat{p}_i=\hat{p}_i^{(b)}$, which translates into assuming that the batter constantly faces an average pitcher, and we calculate the average number of runs generated per game. For pitchers, we fix $\hat{p}_i=\hat{p}_i^{(p)}$, that amounts to postulating that the pitcher of interest keeps dealing with an average batter. We shall call this statistic the NRGG.

Similar methods were used in the development of a statistic like this one in the past like the scoring index by D'Esopo and Lefkowitz (1977) and the offensive earned-run average (OERA) by Cover and Keilers (1977). We also mention the work of Lindsey (1977), Pankin (1978), Bennett and Flueck (1983) and Bukiet et al. (1997). The most important distinctions between these approaches and this one lies in the rules of runner advancement which are far more accurate here, and also the number of possible outcomes (various authors consider strike outs, ground outs and fly outs as being equivalent). Moreover, the method for estimating the probability of each at-bat's outcome is novel (see Section 2).

A total of 2,500,000 games were simulated for each batter and each pitcher in order to calculate their NRGG statistic. This choice leads to a margin of error of up to 0.006 runs per game. Indeed, simulations show that an upper bound on the standard error of the runs per game is 5, so that the maximal value for the half length of the confidence interval, E_{max} is:

$$E = s * z_{\alpha/2} / \sqrt{n}$$
$$E_{max} = 5 * 1.96 / \sqrt{2{,}500{,}000}$$
$$E_{max} = 0.006$$

The NRGG statistic was calculated only for batters with at least 200 at-bats and for pitchers who faced at least 200 opposing batters. The average value of the statistic turns out to be 4.87 for batters and 4.56 for pitchers. Theoretically, those numbers should be identical, but differ here because the number of batters and pitchers considered in the calculation is different (334 batters with at least 200 at-bats versus 369 pitchers who faced at least 200 batters). Also, for those numbers to match, one would need to do a weighted average (by at-bats or plate appearances). Tables 5 and 6 present the top 20 batters and *starting* pitchers, respectively, in MLB during the 2009 regular season.

Table 5 NRGG for the top 20 batters in the MLB 2009 regular season.

Rank	Batter	Team	NRGG	Rank	Batter	Team	NRGG
1	A. Pujols	StL	9.81	11	A. Gonzalez	SD	7.56
2	J. Mauer	Min	9.29	12	R. Braun	Mil	7.48
3	P. Fielder	Mil	8.48	13	A. Dunn	Was	7.43
4	J. Votto	Cin	8.40	14	P. Sandoval	SF	7.43
5	K. Youkilis	Bos	8.06	15	A. Rodriguez	NYY	7.41
6	M. Ramirez	LAD	7.98	16	M. Teixeira	NYY	7.36
7	D. Lee	ChC	7.98	17	C. Utley	Phi	7.35
8	H. Ramirez	Fla	7.92	18	T. Helton	Col	7.32
9	B. Zobrist	TB	7.73	19	M. Cabrera	Det	7.31
10	C. Beltran	NYM	7.57	20	J. Bay	Bos	7.22

Table 6 NRGG for the top 20 starting pitchers in the MLB 2009 regular season.

Rank	Pitcher	Team	NRGG	Rank	Batter	Team	NRGG
1	T. Lincecum	SF	2.54	11	U. Jimenez	Col	3.35
2	C. Carpenter	StL	2.61	12	T. Lilly	ChC	3.35
3	Z. Greinke	KC	2.81	13	T. Hanson	Atl	3.39
4	C. Kershaw	LAD	2.85	14	R. Wolf	LAD	3.43
5	J. Vazquez	Atl	2.99	15	J. Jurrjens	Atl	3.46
6	J. Peavy	CWS	3.02	16	E. Bedard	Sea	3.47
7	F. Hernandez	Sea	3.13	17	J. Pineiro	StL	3.48
8	D. Haren	Ari	3.19	18	A. Wainwright	StL	3.48
9	J. Johnson	Fla	3.27	19	J. Verlander	Det	3.53
10	C. Sabathia	NYY	3.33	20	J. Outman	Oak	3.53

Albert Pujols and Joe Mauer, who take the top 2 spots in our batter rankings, were named National League (NL) and American League (AL) most valuable players in 2009, respectively. The NRGG statistic therefore seems to be in agreement with the votes from the *Baseball Writers Association of America*. The Cy Young award is given annually to the best pitchers in baseball, one each for the NL and AL. The winners in 2009 were Tim Lincecum from the San Francisco Giants (NL) and Zack Greinke from the Kansas City Royals (AL). These pitchers rank at the top of their respective leagues according to the NRGG statistic (first and third overall).

4 In-play strategy

The simulator can easily estimate in-play probabilities at any point during a particular game. The user has to input both lineups (including which batter is due up next), each team's current pitcher on the mound and the current state of the game (score, inning, number of outs, runner(s) on base, if any). It is then possible to simulate a very large number of games, where every single one starts at the state that was inputted by the user, and to determine the fraction of games won by each team.

In-play probabilities can be very useful to a team's manager as it opens the way to the comparison of different strategies. The simulator can be used to estimate winning percentages under a few strategies, helping a team to optimize its chances of beating their opponents. This can be done in real time: all that is needed is someone running simulations on a laptop, and letting the coach know what the suggested strategy is.

We consider two games from the 2009 ALCS (American League Championship Series) between the Los Angels Angels and the New York Yankees, a series that was eventually won by the Yankees in six games. We first turn our attention to Game 3, which was held on October 19th. The Angels led 4–3 going into the top of the 8th inning. The leadoff hitter, Hideki Matsui, walked on five pitches before being replaced by Brett Gardner as a pinch runner. Jorge Posada was up next against the Angels' relief pitcher Kevin Jepsen. The Yankees manager, Joe Girardi, had several options to consider:

- Ask Gardner to try stealing second base.
- Instruct Posada to hit a sacrifice bunt.
- Let Posada hit, while asking Gardner to stay put at first base.

Those three strategies are compared through the simulator. Other variations of those tactics could have been commanded by the manager (such as a hit-and-run), but they are omitted here. We simply focus on the three plans of action described above. The success rate of the stolen base attempt is set at 26/31=0.839 since Gardner stole 26 bases in 31 attempts over the 2009 season. As for the sacrifice bunt, we consider three different success rates: 80%, 90% and 100%.

On October 19th, Girardi opted for the stolen base attempt, but Gardner was caught stealing by the catcher Jeff Mathis. Posada homered a few pitches later to tie the game up at 4 apiece. The game went in extra innings, where the Angels won 5–4 in 11 innings. On the Angels side, Kevin Jepsen threw the 8th inning, Brian Fuentes took care of the 9th, while Jason Bulger and Ervin Santana were on the mound for the 10th and 11th innings, respectively. That is the information we input in the simulation program with respect to which pitchers are on the mound for Los Angeles for the remainder of the game. If a simulated game lasted more than 11 innings, it was assumed that Ervin Santana kept pitching until the game finished.

The same concept was used regarding New York's pitchers. In other words, the management of Yankees hurlers during Game 3 of the ALCS is replicated in simulated games, and the last pitcher used in that game stays on the mound indefinitely every time a simulated game goes beyond 11 innings. Specifically, Phil Coke throws the first third of the 8th inning, Phil Hughes performs the remainder of that inning as well as the whole 9th inning, the ace Mariano Rivera takes care of the 10th, whereas David Robertson pitches the first two thirds of the 11th inning and Alfredo Aceves stays on the mound until the game ends.

Team lineups are also inputted in the simulation program, stipulating that Jorge Posada and Bobby Abreu are up next for the Yankees and the Angels, respectively. Each simulated game starts with a 4–3 score in favor of Los Angeles with no outs in the top of the 8th inning with a runner on first base. A total of N=4 million games are simulated under each possible strategy, and the fraction of games won by New York is calculated. Such choice of the value of N guarantees precision up to three decimal places on the winning percentage (the half length of the confidence interval becomes 0.00049 at most).

The estimated probabilities of New York winning the game are displayed in Table 7 for all strategies discussed above. The results suggest that the best strategy consists of attempting a stolen base, which is exactly what the Yankees did in that game. This approach did not yield a positive outcome on October 19th as Gardner was retired on the attempt (and to make matters even worse, Posada followed that up with a home run, which would have given New York the lead, instead of tying the game), but it was still the best plan of attack. Based on our procedure, the sacrifice bunt turns out to be the worst strategy, even if its success rate is guaranteed!

It is interesting to note that the Yankees roughly have one chance out of two of beating the Angels, despite currently trailing by a run, when adopting the stolen base strategy. The latter increases New York's chances of winning the game by about 7% compared to the sacrifice bunt strategy, and 2% versus the "stay put" strategy (no SB, no sacrifice bunt).

The simulator was used once again to evaluate the intentional base on balls strategy which was employed by the Angels manager, Mike Scioscia, on October 22th 2009 for game 5 of the ALCS. Los Angeles led 7–6 with two outs and the bases empty in the top of the 9th inning. One of the most feared hitters in the game, Alex Rodriguez, was up next against pitcher Brian Fuentes. Los Angeles's

Table 7 Estimated winning percentages of the New York Yankees against the Los Angeles Angels on October 19th using various strategies in the top of the 8th inning with no outs, Brett Gardner on first base and Jorge Posada at the plate with the Angels leading 4–3.

Strategy	Estimate of Pr(Yankees win)
Stolen base attempt	48.35%
No stolen base attempt, no sacrifice bunt	46.03%
Sacrifice bunt (100% success rate)	41.33%
Sacrifice bunt (90% success rate)	40.88%
Sacrifice bunt (80% success rate)	40.36%

manager elected to intentionally walk Rodriguez, even though such a decision put the potential winning run at the plate in Hideki Matsui. Matsui eventually walked, followed by Robinson Cano being hit by a pitch to load the bases. The tension was palpable in Angel Stadium as Nick Swisher came to the plate, but Fuentes forced him to fly out to the delight of the fans in Los Angeles, as the Angels won 7–6. But was the intentional walk the strategy that maximized LA's chances of winning the game? Or should Fuentes have faced Rodriguez?

Before moving on to the simulation results, let us first discuss the management of pitchers (as we did for the October 19th game). The Yankees' best relief pitcher, Mariano Rivera, threw the last two thirds of the bottom of the 8th inning. Since the actual game ended halfway through the ninth, we cannot tell for sure whether Rivera would have pitched the 9th inning or not. Therefore, we simulate 4,000,000 games using a specific strategy (walking Rodriguez intentionally or not) assuming that Rivera would have pitched the ninth, and we simulate another 4,000,000 games taking for granted that he was done for the night. The Yankees relief pitchers who had not yet played in that game were Chad Gaudin, Phil Coke, David Robertson and Alfredo Aceves (we chose to exclude those who did not participate in the ALCS at all, presuming it was a sign that their manager did not trust them enough to use them in such crucial games). Prior to each half inning where the Angels are on offense in simulated games, the program therefore picks randomly one of those four relief pitchers to go on the mound for the entire half inning.

On the Angels side, it is conjectured that Brian Fuentes pitches the 9th inning until it is over (no matter what happens), since this is what was observed in Game 5. The relief pitchers who had not played so far were Jason Bulger, Ervin Santana, Matt Palmer and Scott Kazmir (again, excluding those who did not make a single pitch during the ALCS). Accordingly, for each simulated game that goes in extra innings the simulator picks randomly (without replacement) one of those four hurlers to take care of an entire half inning.

We input each team's lineup, noting that Alex Rodriguez and Torii Hunter are up next for the Yankees and the Angels, respectively. Each simulated game starts with Los Angeles leading 7–6 in the top of the 9th inning with the bases empty and two outs. Again, a total of 4 million games are simulated under each scenario (IBB to Rodriguez or not, Rivera pitches the 9th inning or not). The fraction of simulated games won by the Angels is presented in Table 8.

The results suggest that walking Alex Rodriguez intentionally is *not* the optimal strategy. In fact, the numbers from Table 8 indicate that facing Rodriguez increases the Angels' chances of winning the game by roughly 5%, which is quite large. Also, having Mariano Rivera on the mound during the 9th inning inflates New York's winning percentage by about 1%. Notice, however, that fatigue is not accounted for by the simulator, so it is up to the manager to determine whether it is best to remove a pitcher.

Table 8 Estimated winning percentages of the New York Yankees against the Los Angeles Angels on October 19th using various strategies in the top of the 8th inning with no outs, Brett Gardner on first base and Jorge Posada at the plate with the Angels leading 4–3.

Strategy	Estimate of Pr(Angels win)
Face A. Rodriguez, M. Rivera does not pitch in 9th	93.14%
Face A. Rodriguez, M. Rivera pitches in 9th	92.50%
Walk A. Rodriguez intentionally, M. Rivera does not pitch in 9th	88.62%
Walk A. Rodriguez intentionally, M. Rivera pitches in 9th	87.16%

5 Optimal batting order

Another application to the simulation program is the determination of the optimal batting order for any given team. We consider the 2009 World Champions, the New York Yankees. The nine players with the most at-bats during the 2009 regular season were Robinson Cano, Derek Jeter, Mark Teixeira, Johnny Damon, Nick Swisher, Melky Cabrera, Hideki Matsui, Alex Rodriguez and Jorge Posada. The natural defensive positions of these players are suitable for the manager Joe Girardi to put them in the same lineup. Therefore, we are looking for the best possible lineup which includes all nine players mentioned above.

There exist 9!=362,880 ways to order nine batters in a lineup. We could simulate a large number of games under each possible lineup and calculate the average number of runs per game to determine the batting order that maximizes offensive performance. This procedure takes a lot of time, so we make a few assumptions to reduce the number of simulations. The first one consists of assuming that Alex Rodriguez and Derek Jeter should be placed in either one of the first four spots in the lineup, whereas the second premise suggests to put Melky Cabrera among the bottom four. The other six players may bat at any remaining spot.

Those assertions are based on the high level of offensive production from Rodriguez and Jeter, and the mediocre performance by Cabrera. We are left with 12*4*720=34,560 possible permutations, a significant cutback compared to the initial 362,880 potential lineups.

As a first step, a total of 25,000 games are simulated under each of the 34,560 possible lineups. The values of $\hat{p}_i$ are set at $\hat{p}_i^{(b)}$ for every player, which means it is assumed that an average pitcher is on the mound throughout the simulations. Once the run of simulations is completed for a given lineup, a confidence interval is calculated with respect to the average number of runs scored. The largest lower bound of all 34,560 confidence intervals turns out to be 6.603672. All lineups whose *upper* bound on the average number of runs is *smaller* than this number were dropped from the second step, which leaves us with only 1423 lineups.

During that second step, a total of 1,250,000 games are simulated for each of the remaining 1423 lineups. Once again, a confidence interval on the number of runs per game is calculated for every combination of players. The largest lower bound obtained is 6.369681 and we find 40 lineups whose upper bound exceeds this number (no significant difference with the optimal order). We view those 40 lineups as being the optimal ones: they are shown in Tables 11 and 12 (see supplementary reading paper) along with their respective average runs per game and their associated confidence interval. We summarize in Table 9 the number of times each player is spotted in rank *i* among those fourty optimal lineups. The value of the statistic NRGG described earlier is also presented for those players.

It can be seen from Table 9 that Derek Jeter should definitely be the leadoff hitter, while Robinson Cano and Melky Cabrera, the two worst players based on the statistic NRGG, should bat in the last two spots. Jeter's suggested number one batting position may very well be explained by his high on-base percentage (he finished the 2009 season first among all Yankees players) and his relatively low slugging percentage (8th out of the nine players considered here). The best players according to the NRGG statistic, Alex Rodriguez and Mark Teixeira, should be placed in the number 2, 3 or 4 spots.

We verify whether the Yankees manager, Joe Girardi, used a "good" batting order during the 2009 World Series or not. Table 10 depicts each of the nine players' spot(s) in New York's lineup during that series versus the optimal spots according to our top 40 orders.

We formulate the following conclusions based on Table 10:

- Johnny Damon should bat at a much lower spot.
- Robinson Cano should bat at a slighly lower spot.
- Alex Rodriguez and Hideki Matsui should bat at a slightly upper spot.
- Joe Girardi placed Derek Jeter, Mark Teixeira, Nick Swisher, Melky Cabrera and Jorge Posada at their optimal spots.

Recall that the optimal batting order generates, on average, 6.374252 runs per game with a (6.369681, 6.378823) confidence interval (see Table 11 in the supplementary reading). Girardi's most common lineup during the World Series (Jeter, Damon, Teixeira, Rodriguez, Matsui, Posada, Cano, Swisher, Cabrera) averages 6.350583 runs per game. Its confidence interval, (6.346015, 6.355152), does not overlap with the optimal, so we conclude there is a significant difference, although pretty minimal in practice. As a comparison, the worst possible lineup composed of those nine players scores an average of 6.24 runs per game (approximately). In other words, Girardi's batting order was not very far from being optimal.

Table 9 Ranks of nine Yankees players in the 40 optimal batting orders, as well as their NRGG.

Player	Rank in the lineup									NRGG
	1	2	3	4	5	6	7	8	9	
R. Cano						1		39		5.96
D. Jeter	40									6.62
M. Teixeira		5	11	19	5					7.36
J.Damon		1	4	2	12	9	12			6.21
N. Swisher		6	3	3	3	13	11	1		6.18
M. Cabrera									40	4.59
H. Matsui		4	8	9	10	5	4			6.58
A. Rodriguez		24	13	3						7.41
J. Posada			1	4	10	12	13			6.20

Table 10 Batting position of nine Yankees players in the lineup during the 2009 World Series, as well as their suggested position according to the top 40 batting orders (based on our simulator).

Player	Batting position(s) World Series	Batting position(s) in 40 optimal orders
R. Cano	6–7	8
D. Jeter	1	1
M. Teixeira	3	4–3
J. Damon	2	5–7–6
N. Swisher	5–7–8	6–7–2
M. Cabrera	8–9	9
H. Matsui	5–6	5–4–3
A. Rodriguez	4	2–3
J. Posada	5–6	7–6–5

6 Concluding remarks

This paper develops a baseball simulator, whose main assets over previous simulation programs are its more realistic rules of runner advancement and its discrimination of different types of outs (not viewing strike outs, ground outs and fly outs as being equivalent). We investigate various applications to this simulation procedure, such as the introduction of a single statistic (called NRGG) that incorporates all of the relevant information regarding the performance of any given batter/pitcher. According to this new criterion, the best batter in MLB for the 2009 season was Albert Pujols from the St. Louis Cardinals, whereas the best starting pitcher in the game was Tim Lincecum from the San Francisco Giants. Two more applications of the simulator are examined, namely the determination of the in-game strategy that maximizes a team's chances of winning a game at any given moment, as well as the establishment of a team's optimal batting order.

Before we discuss possible improvements to the procedure detailed in this paper, let us highlight the pros and cons of using a modeling approach versus simulations. The latter can include additional variables more easily and could take into account some factors like the current score and who is on base (which may not be possible in a reasonable way under modeling approaches). Also, extra innings can be dealt with more easily and realistically with simulations. On the other hand, models yield definitive results, as opposed to sample averages (confidence intervals need to be calculated to determine if two averages are significantly different from one another).

There are several improvements that could be made to the current simulation method. Future work may add stolen bases into the simulator. Also, any baseball fan is aware of the importance of handedness in pitcher-versus-batter matchups. As a matter of fact, most left-handed pitchers perform better against players batting from the left side of the plate, and vice-versa. As a consequence, if a left-handed pitcher is on the mound it may be more astute to simulate at-bats using the batters' statistics *against lefties only*. This small adjustment is very easy to implement.

Also, a player's speed is only partially taken into account by the simulator. A slower player does not get as many doubles and triples as others, which is taken care of in the estimation of parameters p_2 and p_3, but the use of universal rules for runner advancement prevent us from distinguishing faster from slower players. A future version of the simulation program may create three categories of runners such as "fast," "average" and "slow," where the numerous probabilities guiding the way runners advance in a simulated game could be estimated for each category. Sugano (2008) and Baumer (2009) take into account the effect of player speed on the bases, whereas James (1987) puts forth a framework for measuring player speed.

It may be desirable to include a ballpark and a home field advantage effect. Also, recent player performance may be weighted more heavily than more dated statistics. Some people may also argue that the probabilities of each of the nine outcomes defined earlier could possibly depend on the game situation (runs ahead/behind, inning, outs), but having few data on certain situations may cause trouble.

A possible improvement regarding optimal batting orders lies in the criterion for declaring one lineup to be better than another. This work focused on the average number of runs scored per game, but we are aware that the variance also plays a key role here, as some lineups may be more all-or-none than others. For example, it is best to have a team scoring 8 runs on every game, than a lineup generating the following number of runs per match: 0, 0, 24, 0, 0, 24,... (which also amounts to an average of eight runs per game).

Adding those features to the current version of the simulator would yield even more realistic results. Such a tool can be extremely useful to general managers and managers in order to maximize their team's chances of performing well (by acquiring underrated players based on their NRGG statistic, by making the best decision at all times during a game and by using the best possible lineup on any given night). We also envision another crucial application to the NRGG statistic: drafting the right players. Indeed, the NRGG measure could be computed for all eligible players to the draft, accounting for the strength of opposition.

Acknowledgments: The author has been partially supported by a research grant from the Natural Sciences and Engineering Research Council of Canada. A special thanks to the Mathematics and Statistics Department at Laval for the use of its computing resources.

References

Ano, K. 2001. "Modified Offensive Earned-Run Average with Steal Effect for Baseball." *Applied Mathematics and Computation* 120(1–3): 279–288.

Baumer, B. S. 2009. "Using Simulation to Estimate the Impact of Baserunning Ability in Baseball." *Journal of Quantitative Analysis in Sports* 5(2): 1–16.

Beaudoin, D. and T. B. Swartz. 2010. "Strategies for Pulling the Goalie in Hockey." *The American Statistician* 64(3): 197–204.

Bennett, J. M. and J. A. Flueck. 1983. "An Evaluation of Major League Offensive Performance Models." *The American Statistician* 37: 76–82.

Bukiet, E. R., E. Harold, and J. L. Palacios. 1997. "A Markov Chain Approach to Baseball." *Operations Research* 45: 14–23.

Cover, T. M. and C. W. Keilers. 1977. "An Offensive Earned-Run Average for Baseball." *Operations Research* 25: 729–740.

D'Esopo, D. A. and B. Lefkowitz. 1977. "The Distribution of Runs in the Game of Baseball." pp. 55–62 in *Optimal strategies in sports*, edited by S.P. Ladany and R. E. Machal. New York: North Holland.

Hirotsu, N. and M. Wright. 2005. "Modelling a Baseball Game to Optimise Pitcher Substitution Strategies Incorporating Handedness of Players." *IMA Journal of Management Mathematics* 16: 179–194.

Hirotsu, N. and M. Wright. 2004. "Modelling a Baseball Game to Optimize Pitcher Substitution Strategies Using Dynamic Programming." pp. 131–161 in *Economics, Management, and Optimization in Sports*, edited by S. Butenko et al. Berlin: Springer.

James, B. 1981. *The Bill James Baseball Abstract*. New York: Ballantine Books.

James, B. 1987. *The Bill James Baseball Abstract*. New York: Villard Books.

Kinoshita, A. 1987. "Evaluation of Baseball Batters and Pitchers (in Japanese)." *Communications of the Operations Research Society of Japan* 32: 689–697.

Lackritz, J. 1990. "Salary Evaluation for Professional Baseball Players." *The American Statistician* 44: 4–8.

Lewis, M. 2003. *Moneyball: the art of winning an unfair game*. New York: W.W. Norton and Company.

Lindsey, G. R. 1977. "A Scientific Approach to Strategy in Baseball." *Optimal strategies in sports*. New York: Elsevier-North Holland.

McCracken, V. 2001. "Pitching and Defense: How Much Control Do Hurlers Have?l." http://www.baseballprospectus.com/article.php?articleid=878

Mills, E. and H. Mills. 1970. *Player win averages*. New Jersey: A.S. Barnes and Co., Cranbury.

Pankin, M. D. 1978. "Evaluating Offensive Performance in Baseball." *Operations Research* 26: 610–619.

Sueyoshi, T., K. Ohnishi, and Y. Kinase. 1999. "A Benchmark Approach for Baseball Evaluation." *European Journal of Operational Research* 115: 429–448.

Sugano, A. 2008. "A Player Based Approach to Baseball Simulation", University of California, Los Angeles (dissertation).

Tango, T. M., M. G. Lichtman, and A. E. Dolphin. 2006. *The book: playing the percentages in baseball*. Dulles, Virginia, USA: Potomac Books Inc.