Using Play-by-Play Baseball Data to Develop a Better Measure of Batting Performance

Jim Albert
Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, OH 43403

September 9, 2001

Abstract

Play-by-play data is used to measure the value of a baseball hitter. A state of an inning is defined by the number of outs and the runners on base. Every possible state has a run potential which is defined by the expected number of runs scored in the remainder of the inning. The value of a player's plate appearance is defined by the change in the expected runs between the before and after batting states plus the number of runs scored. A player can be evaluated by the total run value of his plate appearances or the run value per plate appearance. The proposed measure is shown to be better than other proposed batting measures in predicting the number of team runs. The measure is used to evaluate the best National League hitters in 1987.

1. Introduction

One issue in baseball research that has received much attention over the last 40 years is how to properly evaluate the performance of players. In this paper we focus on the evaluation of a batter. Given all of the hitting statistics of a player, such as hits, runs, and home runs, how can we evaluate the player's success as a batter?

The standard measure of batting effectiveness is the batting average (AVG), which divides the number of hits of a batter by the number of at-bats (AB). Although the batting crown is awarded each year to the player with the highest AVG, it is generally well-known among baseball researchers that the batting average is a relatively poor measure of the quality of a hitter. Bennett (1998, Chapter 2), James (1984), and Thorn and Palmer (1989) present alternative measures of hitting ability. Two essentials are needed to create runs: getting on base and advancing runners that are already on base. A good measure of the ability to get on base is the on-base percentage (OBP), which divides the number of times on-base by the number of plate appearances. A common measure of a batter's ability to advance runners is the slugging percentage (SLG), the result of dividing the total bases (the total base accumulation of all of the hits) by the number of at-bats. A number of proposed hitting measures combine the on-base and advancement abilities. The OPS statistic is found by simply adding the on-base and slugging percentages:

$$OPS = OBP + SLG.$$

Bill James created a measure called Runs Created (RC), that multiplies the number of times on-base (H + BB) by the number of total bases and then divides the result by the number of plate appearances (AB + BB):

$$RC = (H + BB)\ TB\ /\ (AB + BB)\ .$$

Another proposal (Thorn and Palmer, 1989), called Linear Weights (LW), considers a batting measure of the form

$$LW = c1\ (1B) + c2\ (2B) + c3\ (3B) + c4\ (HR) + c5\ (BB)$$

(1B, 2B, 3B, HR, BB are respectively the number of singles, doubles, triples, home runs, and walks) and finds values of the constants c1, c2, c3, c4, c5 using a computer simulation model for run production. Bennett and Flueck (1983) and Albert and Bennett (2001), Chapter 6, evaluate the goodness of these different hitting measures by seeing how well the measures predict the actual number of team runs scored. One conclusion of this study is that the batting average AVG is inferior to the measures OPS, RC, and LW, and the latter three measures are relatively similar in their predictive performance.

One problem with all of these measures is that they are based on the hitting statistics for a player, such as hits, walks, and doubles, and these statistics only indirectly measure the number of runs that a player creates for his team. It would seem desirable to create alternative measures of hitting that actually take into account the runs that are scored during an inning.

Lindsey (1963) was the first to actually use play-by-play data to better understand the run scoring ability of teams and assess different managerial strategies. Using play-by-play data hand collected for 373 games during the 1959 and 1960 baseball seasons, he constructed probability distributions for the number of runs scored in an inning from each bases and outs situation.

These probability distributions were used by Lindsey and Palmer to assess the value of each type of hit and to evaluate strategies such a sacrifice bunt.

Retrosheet and Project Scoresheet

Baseball historical play-by-play data has recently been publicly available due to the efforts of the Retrosheet organization (www.retrosheet.org) and Project Scoresheet. Retrosheet was founded in 1989 for the purpose of computerizing play-by-play accounts for as many pre-1984 major league games as possible. Other organizations, such as Project Scoresheet, have collected play-by-play data for games since 1984. One goal of these data collection efforts is to aid modern baseball statistical analyses that use performance measures that are dependent on this play-by-play information.

We focus on an analysis of the National League 1987 season, since this is one of the most recent seasons for which the play-by-play data is downloadable from the Retrosheet web site. This dataset contains extensive information on each game played in the 1987 season, including the names of all players, the ballpark attendance, the outside temperature, and the identity of each umpire. The central feature of this dataset is the chronological listing of all plays for every game. A "play" in baseball can consist of a "batting event" where the pitcher/batter matchup results in an out, a hit, or a non-hit (such as a walk or a hit by pitch) that results in the batter getting on-base. A play can also consist of a "non-batting event", such as a wild pitch or a steal or a balk, that does not change the batter, but changes the bases occupied, the runs scored, or the number of outs in the inning.

For each play in the game, a number of variables are recorded, including the

- the inning and the team at-bat (home or visitor)
- the current score (home team and visiting team)
- the name of the pitcher
- the name of the batter
- the number of outs
- the bases that are occupied by runners
- the event that occur during this play

At the beginning of the game, the visiting team bats with no runners on base and no outs and no runs scored. For the purpose of this paper, any play is recorded that changes the base/outs/runs situation. In the 1987 National League baseball, there were a total of 78,032 such plays. The goal of this analysis is to use this detailed play-by-play information to develop a good measure of a hitter's effectiveness. This measure, called the value added approach, was suggested by Lindsey (1963), proposed in detail by Skoog in James (1987), and has been used in a series of reports by Ruane (1999).

2. Batting States

As mentioned in Section 1, a hitter creates runs by getting on base and by advancing runners that are already on base. A hitter's opportunity to create runs depends on the number of outs in the inning and the runners on base. We define a "state" of an inning as the number of outs and the base situation. There are three possible number of outs (0, 1, and 2), and eight possible base situations (first, second, and third base can each be either occupied or not by a runner), and so there are 3 x 8 = 24 possible states. The inning ends with a 3-outs state, so there are a total of 24 + 1 = 25 states.

Each possible state of an inning has an associated run potential. For example, if there are two outs and no runners on base, then it is unlikely that the team will score any runs. In contrast, a team is likely to score one or more runs if the bases are loaded with no outs. Using this play-by-play dataset, it is possible to calculate the run potential of each possible state by summarizing the number of inning runs scored. For example, suppose that one is interested in the run potential of the (1 out, runner at second) state. In the dataset, we collect all of the plays where (1 out, runner at second) is the current state. From these plays, we collect the number of runs scored in the remainder of the inning starting at this (1 out, runner at second) state. We can summarize this collection of runs scored with a mean – this number estimates the expected number of inning runs scored starting from the (1 out, runner at second) state. Table 1 displays the expected number of runs scored from each possible state. (A similar table was prepared by Lindsay, 1963.)

Table 1

Expected runs scored in remainder of inning from each of 24 possible (runners, number of outs) states.

		Runners							
		none on	1^{st}	2^{nd}	3^{rd}	1^{st}, 2^{nd}	1^{st}, 3^{rd}	2^{nd}, 3^{rd}	bases loaded
Number of Outs	0	.49	.85	1.11	1.3	1.39	1.62	1.76	2.15
	1	.27	.51	.68	.94	.86	1.11	1.32	1.39
	2	.10	.23	.31	.38	.42	.48	.52	.65

Table 1 quantifies the run potentials of the 24 possible states. Generally note that the expected runs in an inning increase as the number of runners on base changes from 0 to 3, and likewise the run potentials decrease as the number of outs change from 0 to 2. As expected, the highest run potential occurs at (0 outs, bases loaded) – the expected runs scored in the remainder of the inning is 2.15. In contrast, the (2 outs, no runners on) state has the lowest run potential – only .10 runs are scored in an inning, on average, from this state. Which state is more preferable: (0 outs, runner at 2^{nd}) or (1 out, runners at 1^{st} and 3^{rd})? From the table, we see that both states have a run potential of 1.11 runs, so they can be viewed as equally valuable states using this expected runs criterion.

3. **Moving Between States**

A player comes to bat at a particular state in an inning with an associated run potential. For example, a player might come to bat with one out and runners at 1^{st} and 2^{nd}; from Table 1, we see that this "pre-batting" state has a potential of .86 runs. The player then has a "batting event" (a ball hit in the field or a strikeout or a walk or a hit-by-pitch) that results in a new "post-batting" state. In our example, suppose that the player hits a double – one run scores and runners are on 2^{nd} and 3^{rd} with one out. This new state has a potential of 1.32 runs.

When the player bats, a transition is made between the pre-batting state and the post-batting state. The run value of this transition can be measured by

$$\text{VALUE} = E_{POST}[R] - E_{PRE}[R] + \text{Runs\_scored} \,,$$

where $E_{POST}[R]$ denotes the expected runs in the post-batting state and $E_{PRE}[R]$ denotes the expected runs in the pre-batting state. (Skoog (1987) refers to this measure as RC1.)

The number of runs scored in this transition can be computed using a basic fact about baseball. The batter that completes his plate appearance represents an additional out, or an additional runner on base, or a run scored. Let NR_{PRE} and O_{PRE} denote the respective number of runners and outs at the pre-batting state and NR_{POST} and O_{POST} denote the number of runners and outs in the post-batting state. Then the runs scored in this transition is given by

$$\text{Runs Scored} = (NR_{POST} + O_{POST} + 1) - (NR_{PRE} + O_{PRE}) .$$

In our example, consider the value of the transition from (1 out, runners at 1st and 2nd) to (1 out, runners on 2nd and 3rd). From Table 1, we have $E_{PRE}[R] = 1.32$ and $E_{POST}[R] = .86$. In the pre-batting state, we had $NR_{PRE} = 2$ runners and $O_{PRE} = 1$ out, and $NR_{POST} = 2$ runners and $O_{POST} = 1$ out in the post-batting state. So the run value of this transition is

$$\text{VALUE} = 1.32 - .86 + (2 + 1 + 1) - (2 + 1) = 1.46$$

This is a good transition – by hitting a double in this situation, this batter has essentially created 1.46 runs for his team.

In contrast, let's look at a batting performance that is ineffective. Suppose that there are runners on 1st and 2nd with 0 outs – the run potential in this state is 1.39 runs. Suppose the batter hits a groundball to third – the third baseman steps on 3rd base and throws to first, completing a double play; the runner at 1st advances to 2nd. The new state is a runner on 2nd with 2 outs – the run potential is .31 runs. The value of the transition from (0 outs, runners on 1st and 2nd) to (2 outs, runner on 2nd) is

$$\text{VALUE} = .31 - 1.39 + (2 + 0 + 1) - (1 + 2) = -1.08.$$

Since the value of this transition is negative, the batter has essentially taken away one run from his team.

4. Transitions from Various States

From a given state, the batter's action results in a transition to one of a number of possible states. Some of these transitions will result in negative values and other transitions result in positive values. By examining the variation of these transition values, one can judge the importance of starting from a particular state. For example, consider the state (1 out, no runners at base). Table 2 shows the five possible transitions, the frequency of their occurrence, and their run values starting from this state. Here the two most likely transitions, to the (1 out, runner on 1st) and (2 outs, no runners on) states, have relatively small values (.24 and -.17). The most valuable transition to (1 out, no runners on) occurs when the batter hits a home run – this relatively infrequent transition has a value of 1 run. From this table, we see that the (1 out, no runners at base) state is a relatively unimportant state since there is small variation in the values of the likely transitions.

Table 2

Possible transitions, counts of these transitions, and values starting from the (1 out, no runners on) state.

Beginning state	Final state	Count	Proportion	Value
(1 out, no runners on)	(1 out, no runners on)	310	.024	1.00
	(1 out, runner on 3rd)	90	.007	.67
	(1 out, runner on 2nd)	584	.045	.41
	(1 out, runner on 1st)	3171	.246	.24
	(2 outs, no runners on)	8738	.678	-.17

In contrast, let's consider the (2 outs, bases loaded) state, which is possibly the most dramatic state in baseball. Here there are eight possible transitions listed in Table 3. Many of these transitions have high values. For example, a grand-slam home run results in a transition from (2 outs, bases loaded) to (2 outs, no runners) which has a run value of 3.45, and a double can give a transition to (2 outs, runner on 2nd) which has a value of 2.66 runs. The most likely transition, an out which moves the state to (3 outs), has a relatively large negative value of -.65. One reason why the (2 outs, bases loaded) situation is so dramatic is the sharp contrast in values between the good plays (such as a home run) and bad plays (like an out).

Table 3

Possible transitions, counts of these transitions, and values starting from the (2 out, bases loaded) state.

Beginning state	Final state	Count	Proportion	Value
(2 outs, bases loaded)	(2 outs, no runners)	19	.024	3.45
	(2 outs, runner on 3rd)	6	.008	2.73
	(2 outs, runner on 2nd)	16	.020	2.66
	(2 outs, runners on 2nd and 3rd)	24	.030	1.86
	(2 outs, runner on 1st, 3rd)	41	.051	1.82
	(2 outs, runner on 2nd , 3rd)	31	.039	1.77
	(2 outs, bases loaded)	81	.102	1
	(3 outs)	580	.727	-.65

Figure 1 displays the probability distribution of the run values of the transitions for each of the 24 starting states. The mean run value of the transition for each state is approximately zero and the 24 probability distributions differ mainly in their spreads. We can measure the spread of each probability distribution by a standard deviation; Table 4 shows the standard deviation of the run values of the transitions for all 24 states.

Figure 1

Probability distributions of run transitions starting from 24 possible states.

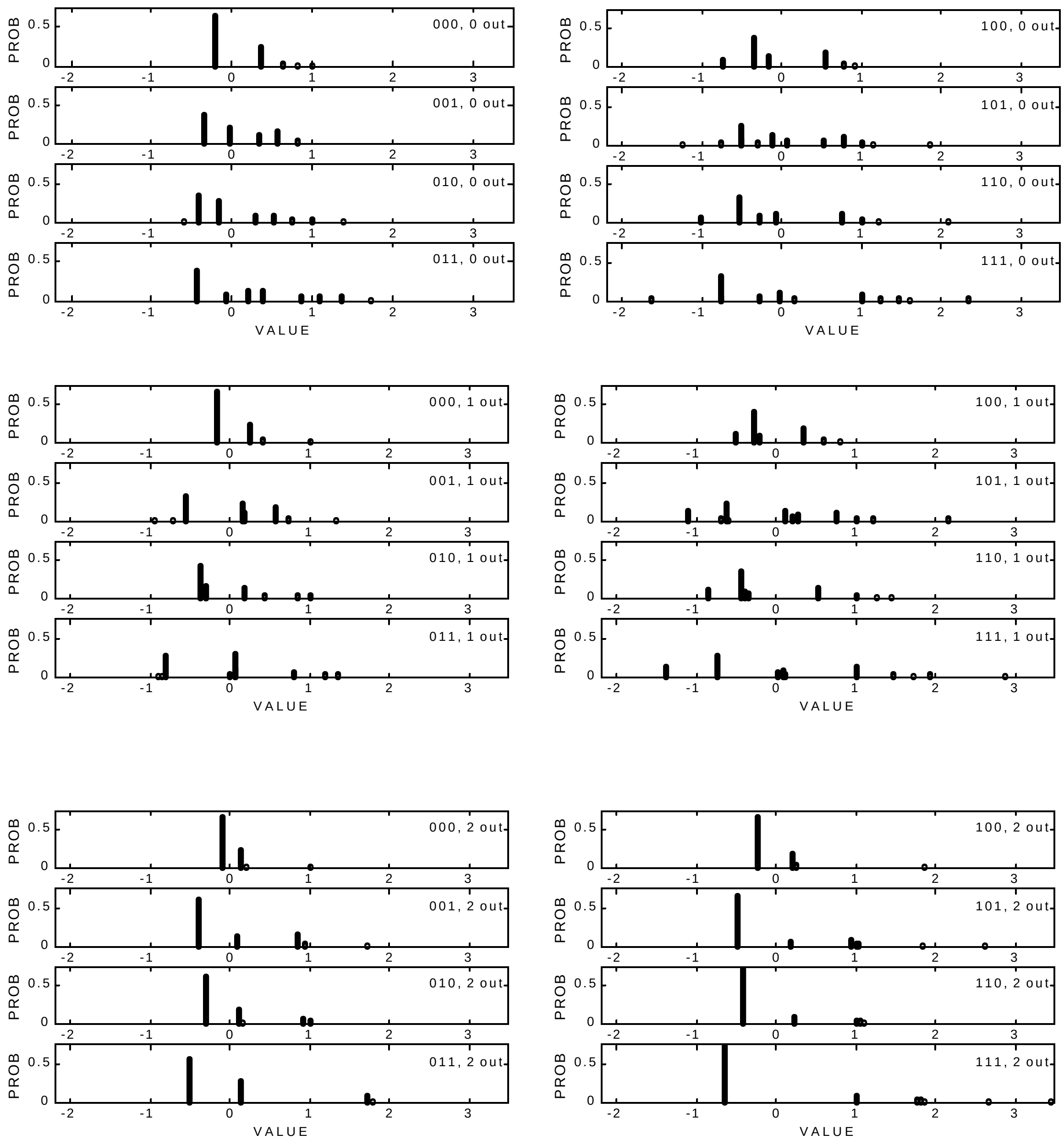

Table 4

Standard deviations of the run values of the transitions starting at each of the 24 possible states.

		Runners							
		none on	1^{st}	2^{nd}	3^{rd}	1^{st}, 2^{nd}	1^{st}, 3^{rd}	2^{nd}, 3^{rd}	bases loaded
Number of Outs	0	.34	.58	.50	.43	.80	.69	.66	1.03
	1	.26	.50	.51	.52	.79	.82	.70	1.11
	2	.19	.43	.50	.57	.73	.81	.80	1.13

Looking at the collection of standard deviations in Table 4, they appear to fall into three clusters depending on the number of runners on base: the standard deviations of the "no runner" transition values fall in the (.19, .34) range, the standard deviations of the 1-runner values fall in (.43, .58), the standard deviations of the 2-runner values fall in (.66, .82), and the standard deviations of the bases loaded values fall in (1.03, 1.13).

This analysis suggests that it is useful to categorize the 24 states by the number of runners on base. The importance of a plate appearance with no runners on-base is relatively low, since there is little difference between the run values of a good play (like a hit) and a bad play (like an out). In contrast, a plate appearance with 2 or 3 runners on base is very important, since there is a big difference in the run values of good and bad plays. In the evaluation of a player that we discuss next, his contribution will be categorized by the number of runners on base.

5. Evaluating a Hitter

The run value criteria developed in the previous sections can be applied in a straightforward manner to evaluate the hitting accomplishment of an individual player. The key element in this evaluation is the player's transition matrix that gives the number of moves between all of the possible states of an inning. Each move between two different states has an associated run value. One way to evaluate a player is to sum his run values over all his transitions. If we denote the player's transition matrix of counts by $\{n_{ij}\}$ and the corresponding matrix of run values by $\{v_{ij}\}$, then the total contribution of a player's hitting over the season, T_CONT, is given by

$$T\_CONT = \sum_{i,j} n_{ij} v_{ij}$$

Of course, one player may have a larger total contribution than another player for the simple reason that he had more opportunities to bat. One can standardize the total value statistic by dividing by the number of plate appearances $\sum n_{ij}$; we call this statistic a player's contribution per play CONT_PLAY:

$$CONT\_PLAY = \frac{\sum_{i,j} n_{ij} v_{ij}}{\sum_{i,j} n_{ij}}$$

Both statistics T_CONT and CONT_PLAY are useful in evaluating players. Players with large values of T_CONT have contributed significantly to the season total run production of their teams. Players with large values of CONT_PLAY are effective in that they produce a large run value for each plate appearance.

6. Evaluating the New Measure for Team Data

Since the goal of a hitter is to create runs, any hitting statistic, such as T_CONT, has to be evaluated by its relationship with runs scored. But we can't measure directly how many runs an individual batter creates. So we judge the goodness of a hitting statistic using team data, where we do have a direct measurement of the number of runs created or scored.

There were 455 players that batted in the National League in 1987. Each player has an associated transition matrix, and using this matrix and the transition values, one can compute his total contribution T_CONT. If we sum the contributions for players in each team, we obtain the team values of T_CONT shown in Table 5. Note the wide spread of values in this table, from Houston at –112.6 to New York at 78.0. Since the median team value of T_CONT is approximately 0, teams with negative values of T_CONT had below-average run production and positive values correspond to above-average run production.

Table 5

Values of T_CONT for all teams in the National League in 1987.

Team	T_CONT
Atlanta	15.5
Chicago	-15.2
Cincinnati	35.5
Houston	-112.6
Los Angeles	-100.2
Montreal	5.5
New York	78.0
Philadelphia	-34.7
Pittsburgh	-7.7
San Diego	-58.8
San Francisco	31.2
St. Louis	34.2

We use this team data to compare the T_CONT measure with the popular hitting measures OBP, OPS, Runs Created, and Linear Weights. (The linear weights measure in this case is the least-squares estimate of runs scored using the team hitting statistics.) Each of the five measures was used to predict, using a simple linear regression, the teams runs per game. Table 6 displays the residuals {Runs per game – Predicted Runs per Game} for each of the five measures. We can summarize the goodness of the five fits by the root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum (R_i - \hat{R}_i)^2}{12}}$$

where R_i is the observed runs of team i and $\hat{R}_i$ is the predicted runs from the linear regression. The RMSEs are shown at the bottom of Table 6. Generally it can be seen that the residuals using the fit using the T_CONT statistic are much smaller than the residuals from fits using any of the other four statistics. The RMSE for the T_CONT fit is .067 which is significantly smaller than the next best fit, Linear Weights, which has a RMSE value of .121. So the new statistic appears to be an excellent predictor of team run production.

Table 6

Residuals from simple linear fits using five different batting measures.

		Residual				
Team	Runs per Game	OBP	OPS	Runs Created	Linear Weights	T_CONT
Atlanta	4.64	-.19	.02	.09	-.09	-.04
Chicago	4.47	.03	-.32	-.34	-.13	-.02
Cincinnati	4.83	.27	.05	.05	.12	.02
Houston	4.00	-.20	-.08	-.09	-.05	.13
Los Angeles	3.92	-.01	-.05	-.05	.08	-.03
Montreal	4.57	.07	.09	.09	-.03	-.05
New York	5.08	.25	.13	.07	-.06	-.00
Philadelphia	4.33	-.14	-.24	-.20	.04	.03
Pittsburgh	4.46	-.10	-.07	-.08	-.14	-.07
San Diego	4.12	-.50	-.16	-.14	-.13	.09
San Francisco	4.83	.45	.08	.07	.18	.05
St. Louis	4.93	.06	.56	.52	.23	.12
RMSE		.242	.212	.202	.121	.067

7. Evaluating Players

We focus our player evaluation on the 105 hitters who had at least 300 plate appearances in 1987. By limiting our analysis to players with a large number of batting appearances, we are ignoring pitchers and other part-time players who were relatively ineffective batters. Figures 2 and 3 display dotplots of the values of T_CONT and CONT_PLAY for the players.

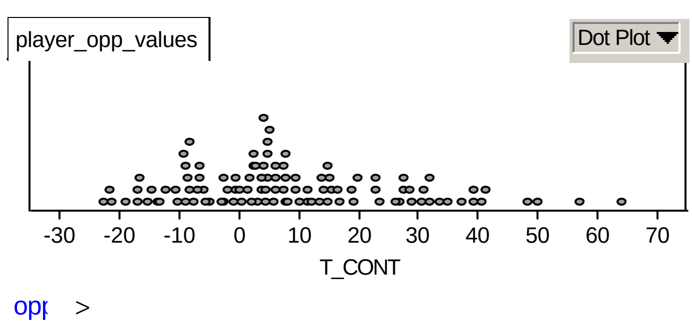

Figure 2

Dotplot of values of T_CONT for the 105 National League players in 1987 who had at least 300 plate appearances.

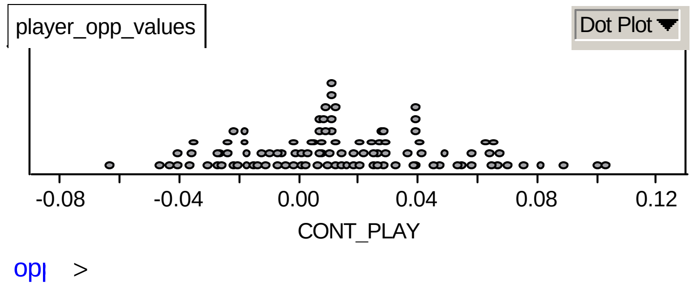

Figure 3

Dotplot of values of CONT_PLAY for the 105 National League players in 1987 who had at least 300 plate appearances.

Let's focus first on the values of T_CONT graphed in Figure 2. Note that a majority (actually 68%) of these regular players had positive total contributions and the median is equal to 5 runs. Four players (Darryl Strawberry, Jack Clark, Eric Davis, and Dale Murphy) appear to stand out with values of T_CONT 50 or higher. Also there were some players with negative run contributions around –20. Looking at Figure 3, the run contributions per play had a median of .01 with values ranging from -.06 to +.10. In Figure 3, two players (Jack Clark and Darryl Strawberry) appear to stand out on the high end. It is interesting that the players with the largest total run production were generally the players with the largest run production per play. This comment is reinforced by Figure 4, which shows a large positive association between the T_CONT and CONT_PLAY statistics.

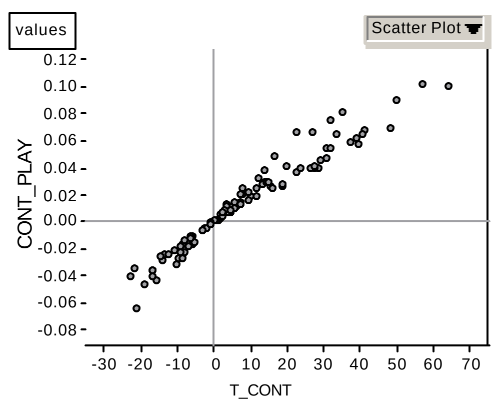

Figure 4

Scatterplot of the total contribution (T_CONT) and contribution per play (CONT_PLAY) statistics for the 1987 NL players with at least 300 plate appearances.

Table 7 displays run contribution statistics for four players – two that had successful 1987 batting seasons and two that had unsuccessful seasons. The first and second rows display respectively the number and percentage of plate appearances of the batter when there were 0, 1, 2, and 3 runners on base. The T_CONT and CONT_PLAY rows gives the total run contribution and contribution per plate appearance for each runner situation.

Darryl Strawberry had the largest value of T_CONT, 63.9, among the NL players. We see from Table 7 that most of this run contribution occurred when there were 1 or 2 runners on base. Likewise, Strawberry's per play contribution, CONT_PLAY, was greatest when there were 1 or 2 runners on base. Mike Schmidt's total run contribution of 41.2 was equally divided between the 0 runner, 1 runner, and 2 runners situations. His per play run contribution, however, was greatest with 2 runners on base.

Table 7

Values of T_CONT and CONT_PLAY for two successful and two unsuccessful players in 1987.

Darryl Strawberry – T_CONT = 63.9

	0 Runner	1 Runner	2 Runners	3 Runners
Count	340	194	95	11
%	53.1	30.3	14.8	1.7
T_CONT	11.0	30.4	22.4	.2
CONT_PLAY	.032	.157	.236	.019

Mike Schmidt – T_CONT = 41.2

TOTAL	0 Runner	1 Runner	2 Runners	3 Runners
Count	292	225	83	11
%	47.8	36.8	13.6	1.8
T_CONT	15.7	12.9	12.3	.3
CONT_PLAY	.054	.057	.148	.028

Glenn Wilson – T_CONT = -21.6

	0 Runner	1 Runner	2 Runners	3 Runners
Count	325	194	85	9
%	53	31.6	13.9	1.5
T_CONT	4.2	-10.9	-12.0	-2.9
CONT_PLAY	.013	-.056	-.141	-.319

Steve Jeltz – T_CONT = -21.3

TOTAL	0 Runner	1 Runner	2 Runners	3 Runners
Count	187	102	45	2
%	55.7	30.4	13.4	.6
T_CONT	2.0	-11.3	-11.4	-.6
CONT_PLAY	.010	-.111	-.253	-.286

It is interesting to contrast these players with Glenn Wilson and Steve Jeltz, who both had weak hitting seasons. Both players had positive total run contributions with no runners on base. But these positive contributions are overwhelmed in each case by the poor performance of both hitters with 1 and 2 runners on base. Neither runner did well with bases loaded, but the total negative run contribution is small since both hitters rarely had plate appearances with the bases loaded.

Figure 5 shows a scatterplot of the OPS statistic against the contribution per play CONT_PLAY for the 105 batters with at least 300 plate appearances. Although there is a positive association between the two measures, there is considerable scatter, which indicates that players have different rankings using the two measures. Five points are identified -- these correspond to the top players with respect to either the OPS or CONT_PLAY measures.. We see that although Jack Clark had the largest OPS and CONT_PLAY values, he was barely ahead of Darryl Strawberry with respect to the CONT_PLAY measure. Kal Daniels had a very large OPS statistic and relatively low T_CONT value.

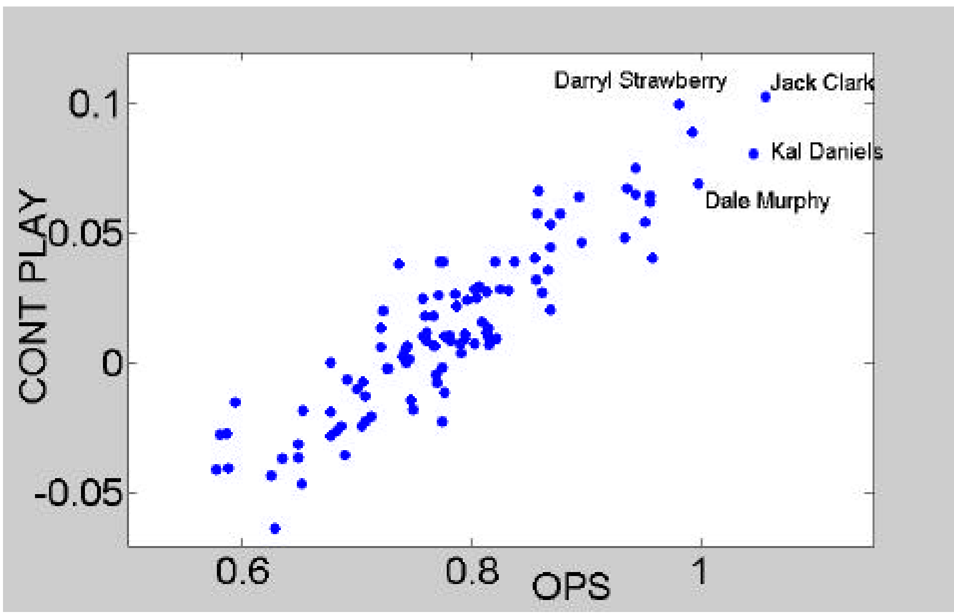

Figure 5

Scatterplot of OPS and CONT_PLAY values for 1987 National League regulars.

Tables 8, 9, 10 give the top ten hitters (at least 300 plate appearances) with respect to the OPS, T_VALUE, and CONT_PLAY statistics. To reinforce the comment made earlier, the top players with respect to total contribution generally were high with respect to the per play contribution. We see that Kal Daniels had a relatively low total contribution since he had only 428 plate appearances. Based on these tables, Darryl Strawberry seemed to have the best hitting season since he had a large lead over Jack Clark in total run contribution and nearly the same per-play contribution as Clark.

Table 8

1987 National League Players with the 10 largest values of OPS.

Player	PA	OPS	T_VALUE	CONT_PLAY
Jack Clark	556	1.056	57.0	0.1025
Kal Daniels	430	1.046	34.8	0.0810
Dale Murphy	692	0.997	48.2	0.0696
Eric Davis	561	0.992	50.0	0.0892
Darryl Strawberry	640	0.981	64.0	0.0999
Tony Gwynn	680	0.958	27.7	0.0407
Pedro Guerrero	628	0.955	40.8	0.0649
Tim Raines	626	0.955	39.1	0.0625
Will Clark	584	0.951	31.9	0.0546
Randy Ready	423	0.943	31.9	0.0753

Table 9

1987 National League Players with the 10 largest values of T_VALUE.

Player	PA	OPS	T_VALUE	CONT_PLAY
Darryl Strawberry	640	0.981	64.0	0.0999
Jack Clark	556	1.056	57.0	0.1025
Eric Davis	561	0.992	50.0	0.0892
Dale Murphy	692	0.997	48.2	0.0696
Mike Schmidt	611	0.936	41.2	0.0674
Pedro Guerrero	628	0.955	40.8	0.0649
Von Hayes	681	0.877	39.4	0.0579
Tim Raines	626	0.955	39.1	0.0625
Tim Wallach	643	0.857	37.2	0.0579
Kal Daniels	430	1.046	34.8	0.0814

Table 10

1987 National League Players with the 10 largest values of CONT_PLAY.

Player	PA	OPS	T_VALUE	CONT_PLAY
Jack Clark	555	1.056	57.0	0.1025
Darryl Strawberry	629	0.981	64.0	0.0999
Eric Davis	558	0.992	50.0	0.0892
Kal Daniels	428	1.046	34.8	0.0810
Randy Ready	417	0.943	31.9	0.0753
Dale Murphy	681	0.997	48.2	0.0696
Mike Schmidt	605	0.936	41.2	0.0674
Mike Aldrete	406	0.858	27.0	0.0665
Tim Teufel	350	0.943	22.9	0.0654
Pedro Guerrero	619	0.955	40.8	0.0649

8. Final Remarks

The main goal of this article is to introduce the new play-by-play data that is now available for baseball research and illustrate its usefulness in the evaluation of players. Although the use of the measures T_CONT and CONT_PLAY are generally not familiar among baseball fans, they seem to more directly measure a hitter's contribution to a team's run production.

We have only scratched the surface on the use of this play-by-play data in better understanding baseball. Here is a list of some topics for further study:

- **Pitcher evaluation.** The T_CONT measure, as defined in this paper, can be used in a straightforward way to measure the contributions of pitchers. This may be more useful, since it can be difficult to judge the effectiveness of pitching performances.
- **Evaluation of non-batting plays.** Likewise, this method makes it easy to evaluate the effect of any play in a game including a stolen base, a wild pitch, or a balk.
- **Value in terms of winning and losing games.** The value of a run scored depends on the game situation. For example, a home run hit in the ninth inning with the game tied is certainly more valuable that a home run hit when the offensive team is winning 10-0. Given the inning situation and the run differential, one can estimate the probability that a

team wins. The change in run values developed here can be translated to the changes of a team winning the game. (See Lindsey, 1963.)

- **Modeling the game by use of a Markov Chain**. The transition matrix defined in Section 4 can be used to develop a Markov Chain probability model for baseball. (See Trueman, 1977.)
- **Modeling transition probabilities.** The batting measure T_VALUE is an estimate of a player's ability and we are ignoring the distinction between a player's ability, as measured by the unknown transition probabilities, and the observed transition matrix counts. Due to the sparseness of the transition count matrix for individual players, the underlying probabilities are not well-estimated. So it is desirable to impose some statistical models that can be used to smooth the observed counts.

References

Albert, J. and Bennett, J. (2001). Curve Ball

Bennett, J. M. and Flueck, J. A. (1983). An evaluation of Major League Baseball offensive performance models. *The American Statistician*, 37 (1), 76-82.

Bennett, Jay (ed.) (1998), *Statistics in Sport*, London: Arnold.

James, B. (1984). *The Bill James Baseball Abstract 1984*. New York: Ballantine Books.

James, B. (1987). *The Bill James Baseball Abstract 1987*. New York: Ballantine Books.

Lindsey, G. R. (1963). An investigation of strategies in baseball. *Operations Research*, 11, 4, 477-501.

Ruane, T. (1999), Value added batting data 1980-1998, Baseball Think Factory.

Skoog, G. R. (1987). Measuring runs created: the value added approach. In James, B. *The Bill James Baseball Abstract 1987,* New York: Ballantine Books.

Thorn, J. and Palmer, P. (1985), *The Hidden Game of Baseball,* New York: Doubleday.

Trueman, R. E. (1977). Analysis of baseball as a Markov process. In Landany, S. P. and Machol, R. E. *Optimal Strategies in Sports*, North-Holland / American Elesevier.