By the Numbers

This is a special issue of By the Numbers, consisting of two articles. The first is "Finding Better Batting Orders," by Mark Pankin, in which he continues his explorations of batting order effects on team offense. Those of you who are not familiar with Mark's work will find it challenging, thought-provoking, and stimulating. He does an excellent job of clarifying the issues and formulating strategies to deal with the inevitable data problems. Furthermore, he is doing work from which we all can learn.

The second is a piece of mine titled "A Retention Model of Employment Discrimination: The Case of Major League Baseball," in which I consider whether the continued performance differential in favor of blacks in major league baseball is a result of differential retention standards applied by teams in deciding who to keep and who to let go.

We both welcome your comments on these pieces.

As usual, I am issuing a plea for additional material. Keep those articles and ideas coming.

Donald A. Coffin
Division of Business and Economics
Indiana University Northwest
3400 Broadway
Gary, IN 46408
219/980-6867

Finding Better Batting Orders

By Mark D. Pankin

Given the nine starting players, in what order should they bat? Traditional guidelines such as "the leadoff man should be a good base stealer", "number two should be a contact hitter who can hit behind the runner", "bat your best hitter third" abound. Due to computational complexities, there have been few studies that analyze the batting order question from a quantitative viewpoint. This article discusses what I believe is the most comprehensive mathematical and statistical approach to lineup determination. The models and the methods used to develop them are described, and some resulting principles of batting order construction are presented. Finally, the models are applied to the 1991 AL division winners and compared to the batting orders employed by the teams' managers.

The material presented here is an expanded version of the talk I gave at SABR XXI in New York during July, 1991. I have written several pieces on using Markov models applied to baseball; readers wanting more information may write to me [1018 N. Cleveland St., Arlington, VA 22201].

The study utilizes two mathematical/statistical models: 1) a Markov process model that calculates the long-term average (often called expected) runs per game that a given lineup will score, and 2) a statistically derived model that quantitatively evaluates the suitability of each of the nine players in each of the nine batting order positions. Data for the second model were generated by numerous runs of the Markov model. Hence, we see that the Markov model underlies the entire analysis.

The Markov Process Model. The Markov process model is based on the probabilities of moving from one runners and outs situation to another, possibly the same, situation. These probabilities, which depend on who is batting, are called transition probabilities. For example, one such transition is from no one on and no outs to a runner on first and no outs; and the transition probability is that of a single, walk, hit batsman, safe at first on an error, catcher interference, or striking out and reaching first on a wild pitch or passed ball. The Markov model employs matrix algebra to perform the complex calculations. However, once all the requisite probabilities have been determined, the matrix formulation enables the remaining calculations to be carried out without much difficulty.

It is important to note that assumptions made in determining the transition probabilities have an enormous influence on the the batting order results presented later. The goal is to choose a realistic set of assumptions, but, as always, some simplifying assumptions are quite helpful. Moreover, some of the assumptions are open to alternatives, the particular ones employed being a matter of judgement or study objectives. The key assumptions for the current analysis are:

1) Players bat the same in all situations. For this study, each player's 1990 full season data was used to determine how he would bat.
2) All base advancement, outs on the bases (including double plays), wild pitches, passed balls, balks, etc. occur according to major league average probabilities.
3) Stolen base attempts are permitted with a runner on first only.
4) Only pitchers attempt sacrifice bunts.
5) Overall 1990 pitcher batting is used for all pitchers.
6) Small adjustments to hit and walk frequencies are made in certain situations. In particular, there are more walks and fewer hits when there are runners on base and first base is not occupied.

Data for 2) and 6) are derived from combined AL and NL data for the 1986 season. I used this season because I had extracted the needed data from the Project Scoresheet database for a prior study. Since this is a time consuming operation, I decided not to repeat it using 1990 data. Comparable data for several seasons would be better, and I may do the computer work on the entire Project Scoresheet database covering 1984-91. However, I doubt that the essential results and lineup optimization models derived would be affected very much.

The first assumption is the most critical and most controversial. One of its consequences is that the differences in expected runs between batting orders tend to be relatively small. A previous, less extensive, study that incorporated situational performance assumptions (e.g. certain players hit better with runners on) showed much larger differences in expected scoring. I plan to explore various alternative assumptions about performance levels in future batting order studies.

Base advancement on hits certainly is not uniform since it depends on runner speed and where the particular batter tends to get his hits (e.g. the percentage of singles to left, center, or right). However, I did not have the data needed to

incorporate such effects. Data availability also prevented batter specific double play modeling.

The stolen base try restriction does not have a large effect because over 80% of steal attempts occur with a runner on first only. The restriction to this case greatly simplifies the computations and is not likely to affect comparisons between batting orders. Sacrifice bunt tries are not included for non-pitchers because they are game situation specific and reduce overall scoring, contrary to the study objective of finding the highest scoring lineups.

Data for the Statistical Models. The Markov model was used for two primary purposes. One purpose is to evaluate a specific batting order by calculating its expected runs per game. In this way, alternative lineups can be compared. The second purpose is the generation of data for use in the statistical models. For each of the 26 major league teams in 1990, 200 "batting rotations" were chosen at random. A batting rotation consists in specifying the order in which the players will bat by establishing who follows whom, but a rotation does not become a lineup or batting order until the leadoff hitter in the first inning is specified. Each batting rotation corresponds to nine lineups, one for each possible leadoff batter. The Markov calculations have the property that the computations needed for one lineup are also sufficient for the other eight lineups corresponding to the same batting rotation. There is nothing special about the choice of 200; it was a function of the computing power available to me and the amount of time I could spend on this phase of the study. More, as usually is the case for statistical analyses, would have been better.

Thus, the Markov model computed the expected runs per game for 1800 "semi-randomly" (a made up concept since only the batting rotations are chosen at random) generated batting orders incorporating the nine most frequent players, one for each position. One property of the 1800 lineups is that each of the nine players hits in each batting position exactly 200 times.

The next step was to select the best lineups for each team from the 1800 tested. I used two definitions of best. The first is obvious: select the ones with the highest expected runs per game. The second definition is more subtle. Each batting rotation will have one lineup that scores the best, and this lineup may or may not be one of the highest scoring lineups out of the 1800. Call the highest scoring lineup for each rotation, a *maximal* lineup. The reason a maximal lineup, which may not be a particularly high scoring lineup overall, is of interest is

that it can reveal advantages to batting certain players in certain positions although the overall scoring is held down by the batting positions of other players. Since there were 200 maximal lineups, one for each rotation, I decided to use them and the 200 highest scoring lineups as the basis for the statistical analysis. I did not determine how many of the maximal lineups were also in the 200 highest scoring.

Within each set of 200 best lineups, I computed how often each player hit in each batting position. For example, Wade Boggs leads off in 21% of Boston's highest scoring lineups. (This value, the highest on the team, means that Boggs is a good first hitter since the average is $100\%/9 = 11.1\%$) In this way, each player has a rating for his suitability for each batting order position.

For each player, I computed scores in 21 offensive measures relative to the group of nine starting players on his team. The offensive measures are batting average; on base average; slugging average; slugging average modified by counting walks as singles and SF as AB (which is the relationship of on base percentage to batting average); extra base average [=SA-BA, also called isolated power]; runs created per game; frequency per plate appearance of each type of hit, walks (including hit by pitch), and strikeouts; relative frequency of each type of hit (i.e. the percentage of players hits that are singles, doubles, etc.); percentage of plate appearances that are not walks or strikeouts (which measures putting the ball in play); secondary average [=(TB-H+BB+SB-CS)/AB, a Bill James idea]; run element ratio [=(BB+SB)/(TB-H), another Bill James idea]; steal attempt frequency [=(SB+CS)/(1B+BB)]; and stolen base success percentage [=SB/(SB+CS)]. No claim is made that the set of measures chosen is complete or perfect, just that it covers all the significant aspects of offensive performance.

I used two measures of player performance relative to the team: 1) percentage above or below the team mean in the category, and 2) the z-score, which is the number of standard deviations above or below the mean. By using z-scores, I am not claiming any of the these distributions is normal (given that there are only nine values for a team in each offensive category, the distributions are almost certainly not even approximately normal); I am just using z-scores as a measure of relative performance.

Regression Analysis. In the next phase, I applied regression analysis using the players' batting position ratings (e.g. Wade Boggs 21%

batting first) as the dependent variable and their relative scores for the various offensive measures as the candidate independent variables. For each batting position there are 236 data points,-one for each of the nine players on the 26 teams-used in the regression estimates. Because there were two measures for batting position ratings-one based on the highest scoring lineups and one based on the maximal lineups-and two measures of relative offensive performance-percentages above or below the team mean and z-scores, there are four possible categories of models that can be derived. I tested all four, as described below, decided on the one that seemed to yield the models with the best statistical properties, and focused on that one. The best combination from the first round of testing was highest scoring rather than maximal lineups as the basis of the dependent variable and z-scores for the independent variables.

To do the regressions, I used the stepwise regression procedure in the SHAZAM statistical package with a 10% significance level required for variables to enter or leave the equations. One equation is estimated for each batting order position, and the estimates are done independently. Since the nine batting position values for a given player must add to 100%, I experimented with some joint estimation techniques. However, they did not yield significantly different models from the independent estimates, so I used the independent estimates throughout this study. After performing stepwise regressions for each of the four categories of models described in the previous paragraph, I restricted further investigation to the highest scoring/z-scores category.

For this first set of regressions for highest scoring/z-scores models, the r^2 values ranges from a high of 0.914 (#9 position) to a low of 0.580 (#6). It is no surprise that the best fit is obtained for the #9 position because of the inclusion of NL teams with pitchers that bat. The number of independent variables in these equations range from a low of 4 (#2,#4) to 12 (#9). Overall, I judged this to be good and workable set of models. Three candidate variables-home runs per plate appearance, run element ratio, and stolen base success percentage (which is highly correlated with steal attempt frequency)-did not enter any of the nine model equations. The variables most frequently in the equations were runs created per game (in 7 equations, all but #4 and #5) and modified slugging average including walks (in 6, all but #2, #5, #7).

The offensive performance measures that are the basis of the independent variables are not truly

independent, and several measure similar player performance characteristics. Since the models usually included several such variables, often with opposite signs, I decided to see if a smaller set of independent variables could yield models with r^2 values almost as high, but which lend themselves to more sensible interpretations. After examining the equations and the correlation matrix of the candidate independent variables, I restricted the candidates to the following nine: on base average (OBA), slugging average (SA), extra base average (EBA), BB/PA, K/PA, 1B/H, HR/H, ball in play percentage (INPLAY), steal attempt frequency (SBTRY).

The resulting set of models had r^2 values from 0.885 (#9) down to 0.607 (#5) and 0.434 (#6). With the exception of #6, the decline in r^2 is not a major concern. In order to improve the model for the sixth position, I added RC/G to set of candidate independent variables for that equation only, which improved its r^2 to 0.557. The number of independent variables ranges from 3 (#3,#4,#7) to 7 (#9). Each candidate variable appeared in at least one of the model equations. The table that follows summarizes the models; a plus sign before the variable means high scores are best for the particular batting order position, and a minus sign indicates the opposite. There are numerical values, the model equation parameters, which are not shown, associated with each variable in the table. These values determine the relative importance of the variables.

I also did some regression analyses using each of the leagues separately because I wanted to see if the DH rule affected the models. In general, the statistical properties-goodness of fit and significance levels of the parameters-were poorer for the models based on the separate leagues. Also, I was not able to interpret the models in a way that could answer the DH question. I suspect that I need more and better data to do this analysis. More in that teams from seasons other than 1990 should be included, and better in that more than 200 batting rotations should be calculated to determine the player/batting position scores. Additional candidate independent variables should also be considered. Due to time constraints, I did not pursue these models further, but this is a topic worth further investigation if for no other reason than the feeling of some AL managers that the number nine hitter should considered as a second leadoff hitter.

4

Batting Order Pos.	Order of Importance 1	2	3	4	5	6
1	+OBA	+BB/PA	-INPLAY	-HR/H	-SBTRY	
2	+SLUG	+OBA	-EBA	+BB/PA	-INPLY	
3	+SLUG	+BB/PA	+INPLY			
4	+SLUG	+OBA	-HR/H			
5	+SLUG	-HR/H	+INPLY	+SBTRY		
6	-RC/G	+SLUG	+INPLY	+OBA	+K/PA	+SBTRY
7	-OBA	+INPLY	+SBTRY			
8	-SLUG	-OBA	-BB/PA	+HR/H	+INPLY	
9*	-INPLY	-K/PA	-SLUG	-OBA	-BB/PA	+1B/H

*-SBTRY is significant, and the 7th most important characteristic for ninth-place hitters.

Generating Lineups Based on the Batting Position Models. Once the batting position model equations are in hand, for a given team, we can compute a value in each of the nine batting order positions for each player. These values can be positive, meaning the player is better than average for the particular lineup position, or negative, which has the opposite meaning. These scores serve to rank the nine players for each lineup position and also to identify the best position for each player. The next step is using those values to find one or more high scoring lineups. Things would be easy if the best position for each player was the highest rating for that position on the entire team. This occurs, for example if Wade Boggs best spot is leadoff and the highest scoring leadoff man on the Red Sox is Boggs; Jody Reed's best spot is #2 and the Sox' best #2 is Reed; etc. However, such is rarely the case. Due to the nature of the models, it is common for the player with the best leadoff score to also have the best #2 score and a high #3 score. Also, the scores on the ends of the lineup (#1, #2, #8, #9) tend to be more extreme, both on the high and low sides, than the scores in the middle. This reflects the models' emphasis on the importance of having high on base average hitters at the top of the order, which is discussed later.

What we need is a method of assigning players to lineup positions so that total model scores from the assignments is high. This is a well known Operations Research topic known as an *assignment problem*. Fortunately, this type of problem can be solved used several methods, some of which are easy to implement on computers and run quickly. I chose an algorithm that not only finds the best possible assignment, but also finds the top n assignments, where n can be specified. For the purposes of this study, I set n equal to five. For each set of batting positions models-one based on the full set of independent variables and one based on the reduced set-I found the five highest assignments for a team, which were always quite close in total batting position values. These lineups were fed into the Markov model to find the expected runs per game. The lineup with the highest expected scoring was usually one of the top three solutions to the assignment problem, but the best solution did not seem to have an advantage over the next two. In some cases, a comparison of the expected scoring and the batting order differences among lineups led me to formulate a lineup with even better expected runs per game that was not in the five solutions to the assignment problem.

For each of the 1990 major league teams, I compared the expected runs of the best lineups found using the models described in the table with the best found using the models based on the full set of candidate independent variables. For 3 AL and 6 NL teams, the full variable models had a slight advantage (about 1-2 runs a season), and for 4 AL and 2 NL teams, the reduced variable set models had a similar advantage. For the rest of the teams, the two sets of models were virtually the same. Because the smaller variable set models are easier to comprehend, the discussion in the next section is based on those models.

Interpreting the Models. Due to the nature of the regression process, it can be misleading to draw conclusions about individual variables without considering the context provided by the entire set of variables. One example is the -SBTRY for the leadoff position. This is the fifth most important variable (its weight is about 10% that of OBA, which is by far and away the most important characteristic of a good leadoff hitter). Even so, does it mean that other things being equal, which they never are, it is better to have a leadoff hitter who doesn't try to steal? It might, but it also may just be the regression distinguishing certain slow effective leadoff hitters based on the Markov model, Wade Boggs for example. Additional statistical analysis, which I have not yet gotten to,

could determine if one or two specific players are the cause of the -SBTRY.

The less important explanatory variables often play a role of emphasizing or modifying the more important ones. For example, the -INPLAY in the #1 and #2 positions serves to emphasize BB/PA. If I wanted to try to find the best set of variables for each position, I would try to build these two models without INPLAY. To illustrate the idea of modification, the -EBA in #2 balances the +SLUG and +OBA. Often players with high OBA have high BA and above average SLUG since slugging average incorporates batting average. The negative EBA in effect puts more weight on the OBA and less weight on power. A more interesting instance is the -HR/H in the model for #4. Does this mean that the clean up hitter shouldn't hit homers? No, what it means is that among players with high slugging averages, it is better to have one who does not get his slugging average mainly from home runs-a Dave Kingman-but instead has a good batting average and hits a fair number of doubles-an Eddie Murray.

The model equations, which are not shown, can be interpreted to characterize the desirable abilities for each batting order position:

1) Getting on base is everything. To much lesser extent, home run hitters should not lead off. Stolen base ability is irrelevant.
2) Similar to the leadoff hitter, but not quite as crucial to get on base; some power is also desirable.
3) Should have fair power, be able to draw walks, and not strike out much.
4) Highest slugging average; also has a good on base percentage and is not necessarily the best home run hitter.
5) Good power; secondarily puts ball in play (i.e. does not walk or strike out a lot).
6) Hardest spot to characterize and probably least critical. Probably want to use player who doesn't fit well in other positions. Base stealing ability is a small plus.
7-9) Decreasing overall abilities as hitters as characterized by on base percentage and measures of power hitting.

One clear result from this and prior studies is the importance of having the right batters at the top of the order. This follows from the finding that most of the difference in expected runs between high and low scoring lineups using the same players occurs in the first inning. In particular, the leadoff batter must have a high on base percentage. Also, the second hitter must be good. The practice of leading off a fast runner who can steal bases, but doesn't get on base much, and putting a weak hitter "with good bat control who can bunt or hit behind the runner" second is a perfect prescription for a lower scoring batting order.

Applying the Models. To see them in action, consider what these models say about the 1991 ALCS teams, Toronto and Minnesota. Batter performance is based on full season 1991 data, and no righty-lefty splits are used. The lineups used by the teams were against right handed starting pitchers. Before Joe Carter was hurt in game three, Cito Gaston used the batting order:

1) D. White, 2) R. Alomar, 3) J. Carter, 4) J. Olerud, 5) K. Gruber, 6) C. Maldonado, 7) L. Mulliniks, 8) P. Borders, 9) M. Lee.

The Markov model expected runs per game for this lineup is 4.739. This value is about 0.5 higher than Toronto's 1991 actual of 4.222 runs per game. That the Markov values are higher than the actuals is to be expected for several reasons. The most important are: 1) the players listed are generally better than the substitutes who play for various reasons; 2) sacrifice bunt attempts, which decrease overall scoring, are not included in the Markov model; 3) relief pitchers brought in with men on base or to face particular hitters can reduce late inning scoring; and 4) a good team usually loses more innings in games won at home than it gains in extra inning games, but the Markov value is based on nine complete innings per game.

The highest scoring lineup found by the models is:

1) Mulliniks, 2) Olerud, 3) Maldonado, 4) White, 5) Alomar, 6) Carter, 7) Gruber, 8) Borders, 9) Lee

The Markov value for the above lineup is 4.795 runs per game, which is about 9 runs per 162 game season more than Gaston's, a difference that should be worth one extra win. (Keep in mind that differences in expected runs between lineups are small due to the assumption that each player's batting is the same in all situations.)

Mulliniks should lead off because he has an on-base average (OBA) of .364, the highest in this group, and little power. White, in contrast, has an OBA of .342 and the second best slugging average (.455, Carter's is .503), so he should not lead off despite his stolen base ability. The major surprise is that Carter bats sixth. The batting position equations score him as best on the team in the third, fourth, and fifth spots, but Maldonado, White, and Alomar rate so low as sixth, that Carter is put there instead. Tests using the Markov model showed its

makes virtually no difference if Carter bats fourth and White and Alomar fill the five and six slots in either order.

Minnesota's Tom Kelly employed the following order in the four games against right handed starters:

1) D. Gladden, 2) C. Knoblauch, 3) K. Puckett, 4) K. Hrbek, 5) C. Davis, 6) B. Harper, 7) S. Mack, 8) M. Pagliarulo, 9) G. Gagne.

The Markov process expected runs per game is 5.383 for this lineup, which is higher than the Twins 1991 average of 4.790 for the reasons given previously.

The best model generated lineup is:

1) Hrbek, 2) Davis, 3) Mack, 4) Puckett, 5) Harper, 6) Gagne, 7) Gladden, 8) Pagliarulo, 9) Knoblauch.

The Markov value of the model lineup is 5.431, about 8 runs higher than Kelly's, which might yield one more victory. Clearly, the model result flies in the face of "conventional wisdom", but one reason for building models is to gain new knowledge. Perhaps the best thing is getting Gladden out of the lead off spot because his 1991 OBA of .306 is by far the worst among the nine players. I am never ceased to be amazed by managers who are so fascinated by speed that they forget players can't steal first base! Davis and Hrbek have the two highest OBAs, and the model takes advantage of this by loading the top part of the order. One reason Davis with a slugging average of .507 can bat second is that Mack's at .529 is even better. Knoblauch is an interesting case because the model values him highest at either the top or bottom of the order. However, on this team, he is best suited to the bottom because his OBA is far from the best.

One important factor not considered is what assumptions, if any, the managers make about batting performance by their players. If I knew such, those levels could be put into the models, and then we could judge better how well the managers constructed their batting orders.

Those with computer baseball games that will automatically play hundreds or thousands of games may find it interesting to enter the 1991 data for these two teams and then compare the scoring of the lineups shown above for a large number of games. I would be interested in seeing how the results of the simulations compare with the Markov calculations.

As a test of how well the models work, I compared lineups found by the models with lineups used by the teams in 1990. For each team, I tabulated the number of times each player started a game in each batting order position. From this information, I constructed one or more typical lineups for each team. Some teams did not really have anything close to a set lineup, and others platooned certain fielding and batting order positions. In all cases, I developed batting orders that were typical of those used by the managers and that reflect their thinking. Using the Markov process expected runs calculations, I compared the best team lineup with the best lineup found by either of the two sets of models -one using all the candidate independent variables and one using the reduced variable set described above. The table shows the extent to which the model did better than the major league managers

Runs/Game	Approx. R/162G	AL	NL
Table 2: Advantage in Expected Runs of Model Over Managers			
.095 - .105	16.0	1*	0
.085 - .094			
.075 - .084	12.0		1**
.065 - .074	11.0	1	2
.055 - .064	9.5	2	1
.045 - .054	8.0	4	2
.035 - .044	6.5	1	2
.025 - .034	5.0	1	3
.015 - .024	3.25	2	
.005 - .014			
-.005 - .005	0	2***	1****

*Chicago White Sox
**Philadelphia
***Boston, Milwaukee
****San Francisco

A general rule of thumb is that an additional 10 runs a season leads to one more win. We see that the model lineups were better than the managers' in 23 of 26 cases with the other three being virtually equal. These comparisons are far from definitive because the models are based on the assumptions listed previously. Also, managers consider many factors when deciding on batting orders, some of which can't be modeled. For example, although Barry Bonds would be an outstanding leadoff hitter because he gets on base so much, according to an article in August 12, 1991 Sporting News he prefers to bat 5th where he can get more RBIs and hence more attention and presumably a higher salary.

Even if he has faith in my models, Jim Leyland might figure that a happy Bonds hitting fifth can help his team more than an unhappy Bonds leading off. Moreover, Bonds might not draw so many walks if he were batting first.

Conclusion. Although I believe this study is a major advance of our knowledge about batting orders, the models discussed are not intended to be the final word on this subject. In particular, incorporation of some situational batting effects should be considered. One, of particular interest, is how the strength or weakness of the next hitter(s) affects a player's batting performance. For example, is there really a tendency to "pitch around" a strong hitter if he is followed by a weak one. The primary problem is obtaining relevant data. Also, there is room for improvement in the statistical (regression) modeling process; additional candidate independent variables should be studied.

I hope that this article has convinced readers that mathematical and statistical techniques can be useful for tools for designing higher scoring batting orders. For those who are interested in actually using the models described, if all goes according to plans, they should be part of the 1992 edition of the APBA computer baseball game (contact the publisher, Miller Associates, 11 Burtis Ave., New Canaan, CT 06840 for details).

Mark Pankin is an operations research analyst. Comments about this piece can be directed to him at 1018 N. Cleveland St., Arlington, VA 22201.

****

A Retention Model of Employment Discrimination:
The Case of Major League Baseball

By Donald A. Coffin

Employment discrimination need not take the form of wage discrimination, discrimination in hiring, or occupational segregation. It may also take the form of discrimination in retention, in which members of a minority group must meet higher standards for retention or promotion.[1] This may represent the most difficult form of discrimination to identify or to combat, for at least two reasons. First, assuming firms use performance measures to make retention decisions, these performance measures may not be readily observable by persons outside the firm.

Second, retention decisions are generally made about employees whose performance is marginal. That is, firms will only rarely decide to terminate employees with above average performance. Because decisions about marginal performers are likely to be made on the basis of small differences in performance, some part of those decisions may be based on non-observed performance differences.

It therefore becomes important to examine those relatively limited instances in which performance data and retention decisions are observable. In this respect, professional sports provide us with excellent data on performance and clearly observable data on retention. Within professional sports, baseball may provide the best data set, in that all non-pitchers carry some offensive responsibility. In this study I examine differences in retention among white players, players born in Latin America, and blacks born in North America for the 1960 to 1989 period. Briefly, I find some evidence of retention discrimination against both blacks and Latins; this discrimination appears to occur only in the 1976-1989 period.

Retention Models and Employment Discrimination. The issue of differences in employee retention as a manifestation of employment discrimination has been addressed by a number of authors. Robert Hall (1982), in

---

1. See Spurr (1990) for an analysis of this issue as it applies to promotion to partner in law firms. Olson and Becker (1983) examined promotion opportunities mode broadly and found that women were less likely than men to be promoted, given performance levels. Hall (1982, p. 716), on the other hand, found that the "duration of employment among blacks is just as long as among whites."

exploring the extent of and importance of lifetime jobs in the U.S. economy, found that "(t)he duration of employment among blacks is just as long as among whites" (p. 719). If there are no performance differentials between blacks and whites, this suggests that there are no differences in retention. However, if performance differentials exist, equal employment duration may accompany discriminatory retention practices by employers.

Olson and Becker (1983) examined whether returns to promotion were similar for men and women, and whether men and women had equal promotion opportunities. Using data from the Quality of Employment Panel, they found that "the returns from promotion are largely the same for men and for women" (p. 641). However, they also found that, had the same promotion standards been applied to men and to women, roughly 32% of the women in the sample would have won promotions; in fact only about 19% of the women were promoted (p. 637). If some employers practice "up-or-out" retention systems, these findings suggest that women will be less likely to retain their jobs than will men.

Olson and Becker's findings are supported by Stephen Spurr's (1990) work on promotion in the legal profession, in which being promoted to partner is, in many firms, a requirement for retention. Spurr finds that men are significantly more likely to be promoted to partner than are women, although he also finds no significant differences in quality of training or in performance as lawyers (pp. 415-416). He concludes that "(a)t a minimum, the results...are strongly suggestive of discrimination, in Becker's sense of the term" (p. 416). This discrimination in promotion opportunities suggests that women must be more productive than men to be promoted and thus more productive than men to be retained.

In a study of major league baseball, Robert Jibou (1988) examines the "career mortality" (p. 525) of black, Hispanic, and white players. In the absence of performance differentials, he notes, length of playing career should be the same for all groups. Using data for the 1971-1985 period, he estimated a hazard function for retirement, using player age on entry into the major leagues, playing position, and performance (measured by career batting average, career slugging average, and career on-base average) (pp. 528-529). He finds that Hispanics are no more likely than whites to retire (given age, position, and performance), but that blacks apparently face some discrimination in retention. Blacks are less likely to be retained, at

any age and performance level than are whites (pp. 530-532).

Jibou's study, however, has several drawbacks. First, he combines into one data set information on player retention for a 15-year period. If changes occurred within this period in the relative performance levels of black, Hispanic, and white players, his results may be biased. Furthermore, his results will not necessarily capture changes that may have occurred in the decision-making of teams within this period (i.e., teams became either more likely or less likely to apply higher performance standards to blacks or to Hispanics). Second, he examines only career terminations as a measure of the retention hazard. If teams are more likely to demote blacks or Hispanics to the minor leagues, this form of retention discrimination will not be captured by Jibou's analysis.

Differences in Retention, Employment Experience, and Performance in Major League Baseball. If blacks and Latins experience discrimination in retention, we would expect to observe higher performance levels among blacks and Latins than among whites. Aaron Rosenblatt (1967) first explored the possibility of performance differentials in major league baseball, finding that career batting averages were significantly higher for black hitters than for white hitters and career earned run averages were significantly lower for black pitchers than for white pitchers during the 1953-1965 period. Rosenblatt's findings might, however, be expected during a period of integration, in which the best-performing blacks will gain employment more rapidly. Over time, and in the absence of discrimination, we might expect to observe that the performance gap between blacks and whites would close.

Using data for the period 1960 to 1989, we can track the differences in several measures of offensive performance for blacks and whites. I have chosen to present two such measures here. The first is a measure I refer to as IMP, which provides a measure of the amount of playing time allocated to black players and to white players.[2] For a full-time player, IMP will be approximately 100. If there are no performance differentials between blacks and whites, we would expect no significant differences in IMP between blacks and whites. However, in every year, the mean value of IMP for black players is significantly greater than the mean

2. It is calculated as IMP $= (1/3)*SQRT(GAMES*AB)$, where GAMES in the number of games played in a season and AB is the number of official at-bats in that season.

9

value of IMP for white players. Furthermore, there is no tendency for the difference between blacks and whites to narrow (see Figure 1)[3].

The second performance measure I use is slugging average (SA).[4] Throughout the 1960-1989 period, the mean SA for blacks was higher than the mean SA for whites, although the difference narrowed in the late 1970s and in the 1980s (see Figure 2). It is not clear why differences in IMP did not narrow along with narrowing differences in other performance measures.

Because black performance levels exceed those of whites, blacks can have higher retention rates and more experience than do whites, even in the presence of discrimination. I define player retention to occur if a player in on a major league roster both in year t and in year t+1; he need not be on the roster of the same team for two consecutive years. In general, black retention rates did exceed those of whites for most of the 1960-1989 period (see Figure 3), and blacks had, on average, more experience than did whites beginning in the mid-1960s (see Figure 4). There is no simple and apparent relationship between retention and player performance that emerges from an inspection of these data.

A Model of Retention. Major league baseball is an unusual industry in that the labor market, even following the abolition of the reserve clause in 1976, contains large elements of monopsony.[5] Players can still be transferred between teams without their consent (until they are 10-and-5 players-- 10 years in the major leagues with at least five years tenure with their current teams--or unless they have negotiated "no-trade" clauses in their contracts). There is substantial involuntary mobility. Nearly 25% (117/473) of non-pitchers who appeared on a major league roster in 1989 and in an earlier year (generally 1988) were on the roster of a different team in 1989. More than 80% of these moves were involuntary from the player's point of view.[6]

Because major league baseball players earn salaries that are (on average) substantially greater than their alternative earnings opportunities,[7] I assume that few players voluntarily surrender their positions. This allows me to assign decision-making responsibility for continuity of employment in major league baseball to the teams. I define a dichotomous variable RET, equal to 1 if a player in on a major league roster in years t and t+1 and zero if he is on a major league roster in year t, but not in year t+1.[8] Note that a player need not be on the same team in two successive years; he need only be on a major league roster in two successive years.

If team owners are profit motivated,[9] they will prefer players who are more productive to players who are less productive. This suggests retaining players for whom expected performance will be greater. As indicators of expected performance, I use immediate past performance (PERF) (in year t), player age (AGE),[10] and a measure of the quantitative aspects of a player's contribution to his team (IMP; see fn. 2 above). We would expect players with greater expected performance (higher past performance, higher IMP, younger) to be more likely to be retained. Performance data were obtained from Thorn and Palmer (1991).[11]

If there is discrimination in retention, then blacks and Latins will be less likely to be retained. Accordingly, I included in the analysis dummy variables for race (BLACK = 1 if the player was a North American-born black, = 0 otherwise) and ethnicity (LATIN = 1 if the player was born in Latin America, including Puerto Rico, = 0 otherwise). Data on place of birth were obtained from Thorn and Palmer (1991); data on race was obtained by a visual inspection of baseball card photographs in Topps (1991).[12]

3. All of the diagrams appear at the end of the article.
4. Data on batting average (BA) and on-base average (OBA) provide similar differences between blacks and whites. I use SA because it is the performance variable which is most frequently significant in the LOGIT models of retention.
5. Complete discussions of the structure of the labor market for player talent can be found in Scully (1989), Dworkin (1981), Jennings (1990), and Lowenstein (1991).
6. Counted from player records in Thorn and Palmer (1990), pp. 29-80.

7. The minimum salary in major league baseball in 1991 was over $100,000; the average salary was in excess of $800,000.
8. I have excluded players who died in between seasons (e.g., Lyman Bostock, Ken Hubbs, Thurman Munson), or players who were not on a major league roster in year t+1 because of serious injuries from which they recovered (e.g., Damaso Garcia).
9. See Scully (1989), chapter 6, for evidence on whether team owners are profit motivated.
10. There is substantial evidence (James [1982], pp. 191-206) that player performance declines predictably with age, rather than with experience.
11. I want to express my appreciation to Pete Palmer for making the computerized data base underlying Thorn and Palmer (1991) available to me.
12. I was unable to identify the race of only one player in my sample; he was, accordingly, eliminated from the sample.

The model then took the form of:

(1) RET = f(BLACK, LATIN, AGE, IMP, PERF)

There were too few pitchers to incorporate pitchers in the analysis;[13] I accordingly limited the analysis to non-pitchers. Because the dependent variable is dichotomous, I used a logit procedure to estimate the model each year for a thirty year period, beginning with players who were active in 1960 and either were (RET = 1) or were not active in 1961 (RET = 0) and ending with players who were active in 1989 and either were or were not active in 1990. In every year, the coefficients on AGE and IMP were significant and had the expected sign (negative and positive respectively). Performance variables were significantly and directly related to RET in 24 of the 30 years.[14] The sample sizes tended to grow over time (partly because there are more teams now than there were in 1960), from 316 to 465.

Because my primary concern is with the influence of race and ethnicity on retention, I report the coefficients on BLACK and LATIN in Table 1 (on p. 12). This table suggests little or no discrimination against blacks and Latins. The coefficient on BLACK is significant only four times in 30 years (once positive, three times negative); the coefficient on LATIN is significant only three times (twice positive, once negative).

However, on closer examination, the annual coefficients display an interesting pattern. Beginning in 1976, the coefficients on both BLACK and LATIN were much more likely to be negative than positive. In the first 16 years of the analysis, the coefficient on BLACK was negative five time; the coefficient on LATIN was negative nine times. Neither of these results was a surprise.

Assume the expected coefficient on BLACK and on LATIN is zero. We can then examine the estimated coefficients as a sample from a distribution

with an expected mean of zero, and with a probability of 0.5 that any observed coefficient will be negative. I calculate, using the binomial theorem, that there is a 90% probability that five or more of the coefficients on BLACK would be negative in the 1960-1975 period; there is a 39% probability that nine or more of the coefficients on LATIN would be negative.

However, in the 1976-1989 period, 11 of the coefficients on BLACK, and 12 of the coefficients on LATIN were negative (out of 14). Again based on the binomial theorem, I would expect 11 or more negative coefficients only 3% of the time, and 12 or more negative coefficients less than 1% of the time. This suggests that a change in the retention experience of blacks and Latins sometime in the mid-1970s.

To examine this possibility more closely, I computed retention rates for whites, blacks, and Latins for each year. Given the importance of AGE, IMP, and performance variables in the LOGIT models, I modeled annual retention rates as a function of AGE, IMP, and performance (on-base average--OBA--performed best in this model). Given the variations in retention rates, age, importance, and performance over time, I transformed all variables by dividing each year's value for each variable by the annual mean for that variable.[15] I also included a dummy variable, D76, equal to 1 for 1976 and subsequent years (and equal to 0 for 1960-1975). The best fit was a log-log model; the results are presented in Table 2. Note that the results are similar to those for the annual regressions--older players are less likely to be retained; players performing more (IMP) or better (OBA) were more likely to be retained. The coefficient on D76 was negative--retention rates were about 1% lower--and significant.

Disaggregation by race and ethnicity led to different results. First, the three disaggregated regressions differed significantly from the pooled regression. Second, the variables designed to measure expected performance were generally insignificantly related to retention. This is largely because the variations in these performance indicators varied relatively little within a group, but varied substantially across groups.

Most importantly, the coefficient on D76 was negative and significant for blacks and Latins, suggesting a 5% lower probability of retention for blacks and a 7% lower probability of retention for

---

13. By 1988, only about 10% of the pitchers in major league baseball were either Latin or black (Hadley and Gustafson, [forthcoming]).

14. The same performance variable does not appear in all regressions; batting average appears in five regressions--all in the 1960s; on-base average also appears five times (scattered throughout the period); slugging average appears 14 times, including 10 times between 1979 and 1989. Because teams may change their minds about which performance variable is most important, I decided not to constrain the regressions to include the same performance variable in every year. Complete results of the LOGIT estimation process are available on request.

---

15. The retention rate for blacks (etc.) in year t was divided by the retention rate for all players in year t, and so on.

Table 1: Effects of Race and Ethnicity on Retention				
	Black		Latin	
Year	Coeff.	t-Stat.	Coeff.	t-Stat.
1960	+1.44	+1.31	+1.70	+1.41
1961	-0.17	-0.27	-0.84	-1.21
1962	-0.63	-1.08	-1.01	-1.48
1963	-0.38	-0.83	+1.60*	+1.72*
1964	-0.32	-0.61	-1.25**	-2.37**
1965	+1.16*	+1.85*	+2.25**	+2.06**
1966	+0.11	+0.24	-0.01	-0.01
1967	+0.07	+0.16	-0.10	-0.20
1968	+0.24	+0.44	+1.75	+1.64
1969	+0.10	+0.22	-0.41	-0.83
1970	+0.24	+0.48	-0.05	-0.08
1971	-0.54	-1.27	-0.25	-0.50
1972	+0.36	+0.40	+0.59	+1.27
1973	+0.42	+0.87	+0.34	+0.63
1974	+0.33	+0.83	-0.01	-0.02
1975	+0.31	+0.72	+0.57	+0.93
1976	-0.85**	-2.02**	-0.59	-1.15
1977	-0.56	-1.54	-0.28	-0.64
1978	+0.30	+0.69	+0.55	+1.09
1979	-0.18	-0.42	-0.32	-0.68
1980	-0.80**	-2.24**	-0.50	-1.16
1981	+0.28	+0.66	-0.57	-1.34
1982	-1.09***	-3.06***	-0.06	-0.13
1983	-0.25	-0.61	-0.64	-1.60
1984	-0.57	-1.57	-0.45	-1.03
1985	+0.01	+0.02	-0.37	-0.91
1986	-0.32	-0.87	-0.32	-0.69
1987	-0.52	-1.44	+0.08	+0.18
1988	-0.48	-1.22	-0.25	-0.53
1989	-0.04	-0.08	-0.60	-1.23

* Significant at the 10% level.
** Significant at the 5% level.
***Significant at the 1% level.

Variable	Total	White	Black	Latin
Table 2: Explaining Variations in Annual Retention Rates				
Constant	+0.01 (+1.37)	-0.01 (-0.61)	-0.11 (-1.68)	+0.05** (+2.17)
LRAGE	-0.53*** (-3.31)	-0.16 (-0.29)	-2.08*** (-3.65)	-0.52** (-1.96)
LRIMP	+0.11** (+2.64)	+0.01 (+0.05)	+0.08 (+0.86)	-0.14 (-0.98)
LROBA	+0.45*** (+2.86)	+1.11 (+1.32)	-0.62 (-0.89)	+0.62 (+0.59)
D76	-0.01* (-1.83)	+0.004 (+0.46)	-0.05** (-2.60)	-0.07** (-2.62)
PCTBLACK			+0.71** (+2.43)	
Adj. R^2	+0.35	-0.02	+0.49	+0.26
F	+12.91	+0.89	+6.64	+3.61
N	90	30	30	30

*   Significant at the 10% level.
** Significant at the 5% level.
***Significant at the 1% level.

Latins beginning in 1976, compared to the earlier period. The coefficient on D76 was positive and insignificant for whites. Again, this suggests a change in the retention experience of blacks and Latins beginning in the mid-1970s.

How can we explain this change? There is substantial evidence that there was no salary discrimination against blacks and Latins in the late 1970s and 1980s.[1] Furthermore, the structure of the labor market changed in the mid-1970s, with the introduction of salary arbitration and free agency making the salary determination at once more competitive and more public. If team owners were seeking opportunities to continue discriminating against blacks and Latins in a climate that made continued salary discrimination less likely, then discrimination against marginal blacks and Latins in making retention decisions could have been a possibility.

Unfortunately, there is also no evidence that blacks and Latins experienced salary discrimination in the 1960s and early 1970s.[2] We cannot, therefore, accept this as an explanation.

A second possibility is the existence of a "quota" or effect. In this case, discrimination does not occur until minority representation in employment reaches a critical level. Once that level is reached, discrimination begins to manifest itself in one form or another. Note (in Figure 5) that the proportion of black and Latin non-pitchers in the major leagues rose fairly steadily until about 1975, reaching about 40%, at which point it plateaued. It is worth noting that this occurred along with continued performance differentials in favor of blacks. These performance differentials would ordinarily suggest continued displacement of whites by blacks until performance levels equalized. The failure of this to happen suggests that team owners may have altered the criteria for retention in the mid-1970s, so as to prevent minority representation among non-pitchers from continuing to rise.

Conclusions. The year-by-year analysis of employee retention in major league baseball does not, by itself, suggest that teams in major league baseball were engaging in discriminatory retention decisions. However, the fact that the coefficients on race and ethnicity were almost always negative beginning in 1976, and the significance of a dummy variable for the 1976-1989 period in regressions for annual relative retention rates for blacks and His-

panics, suggests a change in the decision-making of major league baseball teams in the mid-1970s.

One consequence of this change is that the proportion of non-pitchers who were either black or Latin stopped increasing in the mid-1970s, following 15 years of steady increases. While there is no direct evidence of a "quota" for blacks and Latins (in part because of substantial variations in minority representation between teams), the proportion of non-pitchers who are black or Latin has remained at about 40% since 1976. Had it continued to rise as it had before 1976, more than 50% of the non-pitchers in the major leagues would be blacks or Latins.

The increased difficulties in retention faced by black and Latin players is, of course, entirely among marginal players. Blacks and Latins continued, in the late 1970s and in the 1980s, to be more likely to be full-time players than did whites, and the differences between blacks and whites in my measure of playing time (IMP) did not narrow (although the performance differentials did narrow). It is not surprising that any retention discrimination will affect marginal players, since it is among marginal players that the performance differentials are smallest and the decisions perhaps more affected by subjective or non-observed factors.

Team managements often use terms such as "chemistry" is discussing the process of making choices among marginal players. If management perceives black or Latin players to be less congenial or less willing to accept marginal roles, then they will be perceived as being less likely to contribute to a desirable "chemistry" on the team. Among marginal players, where the performance loss will be less if a less able white is retained and a more able black is let go, this perception may have substantially greater force. The effect, however, is that blacks and Latins, with higher measured performance, are less likely to keep their jobs in major league baseball than are whites.

References

Christiano, Kevin J. 1986. "Salary Discrimination in Major League Baseball: The Effect of Race." Sociology of Sport Journal, Vol. 3, No. 2 (June), pp. 144-153.

----------. 1988. "Salaries and Race in Professional Baseball: Discrimination 10 Years Later." Sociology of Sport Journal, Vol. 5, No. 2 (June), pp. 136-149.

---

1. See Christiano (1986 and 1988), Cymrot (1985), Kahn (1989) and Raimondo (1983).
2. See Medoff (1975), Mogull (1975 and 1981), Pascal and Rapping (1972), and Scully (1974).

Cymrot, Donald J. 1985. "Does Competition Lessen Discrimination? Some Evidence." Journal of Human Resources, Vol. 20, No. 4 (Fall), pp. 605-617.

Dworkin, James B. 1981. Owners Versus Players: Baseball and Collective Bargaining. Boston, MA: Auburn House Publishing Co.

Hadley, Lawrence, and Elizabeth Gustafson. Forthcoming. "Arbitration and Salary Gaps in Major League Baseball." Quarterly Journal of Business and Economics.

Hall, Robert. 1982. "The Importance of Lifetime Jobs in the U.S. Economy." American Economic Review, Vol. 72, No. 4 (September), pp. 716-724.

James, Bill. 1982. The Bill James Baseball Abstract 1982. New York, NY: Ballantine Books.

Jennings, Kenneth M. 1990. Balls and Strikes: The Money Game in Professional Baseball. New York, NY: Praeger Publishers.

Kahn, Lawrence. 1989. "Managerial Quality, Team Success, and Individual Player Performance in Major League Baseball." Mimeo, University of Illinois.

Lowenfish, Lee. 1991. The Imperfect Diamond: A History of Baseball's Labor Wars. New York, NY: DaCapo Press,

Medoff, Marshall. 1975. "A Reappraisal of Racial Discrimination Against Blacks in Professional Baseball." Review of Black Political Economy, Vol. 5, No. 3 (Spring), pp. 259-268.

Mogull, Robert G. 1975. "Salary Discrimination in Major League Baseball." Review of Black Political Economy, Vol. 5, No. 3 (Spring), pp. 269-279.

----------. 1981. "Salary Discrimination in Professional Sports." Atlantic Economic Journal, Vol. 9, No. 3 (September), pp. 106-110.

Olson, Craig A., and Brian E. Becker. 1983. "Sex Discrimination in the Employment Process." Industrial and Labor Relations Review, Vol. 36, No. 4 (July), pp. 624-641.

Pascal, Anthony H., and Leonard A. Rapping. 1971. "The Economics of Racial Discrimination in Professional Baseball." In Anthony H. Pascal (ed). Racial Discrimination in Economic Life, Lexington, MA: D. C. Heath and Co., pp. 119-156.

Raimondo, Henry. 1983. "Free Agents' Impact on the Labor Market for Professional Baseball Players." Journal of Labor Research, Vol. 4, No. 2 (Spring), pp. 183-193.

Rosenblatt, Aaron. 1967. "Negroes in Baseball: The Failure of Success." Transaction, Vol. 4 (September), pp. 51-53.

Scully, Gerald M. 1974. "Discrimination: The Case of Baseball." In Roger Noll (ed.), Government and the Sports Business. Washington, DC: The Brookings Institution, pp. 221-273.

---------- 1989. The Business of Major League Baseball. Chicago, IL: University of Chicago Press.

Spurr, Stephen J. 1990. "Sex Discrimination in the Legal Profession: A Study of Promotion." Industrial and Labor Relations Review, Vol. 43, No. 4 (April), pp. 406-417.

Thorn, John and Pete Palmer (eds.). 1990. Total Baseball Update. New York, NY: Warner Books.

----------. 1991. Total Baseball, 2nd Ed. New York, NY: Warner Books.

Topps. 1991. Topps Baseball Cards: The Complete Picture Collection. New York, NY: Warner Books.

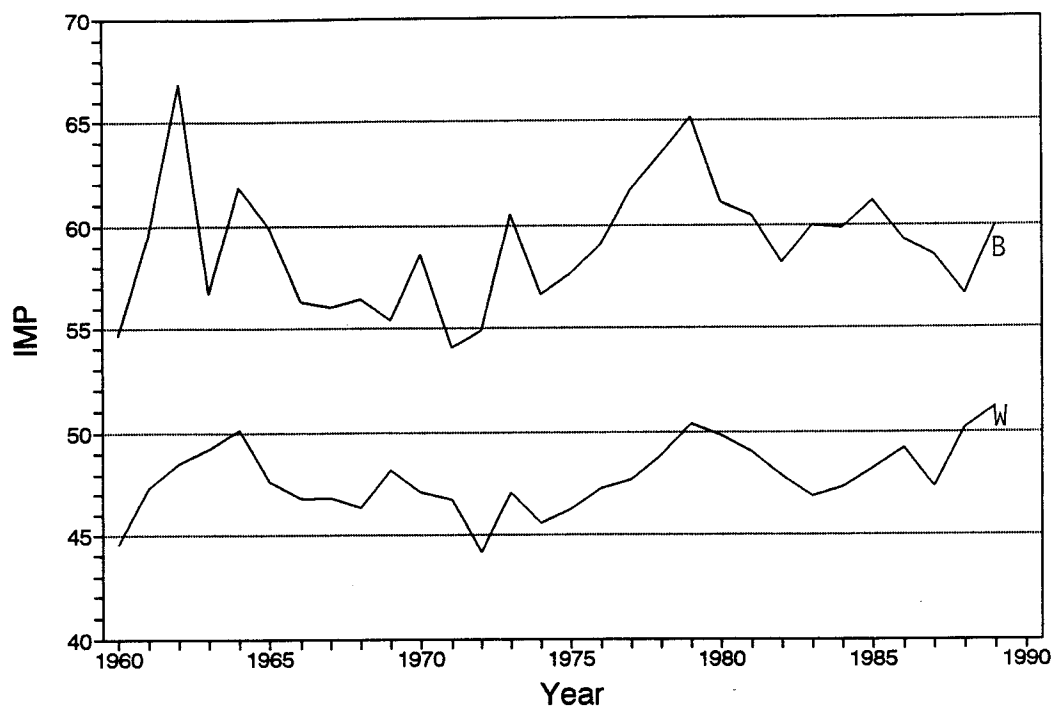## Figure 1: IMP for Blacks and Whites, 1960-1989

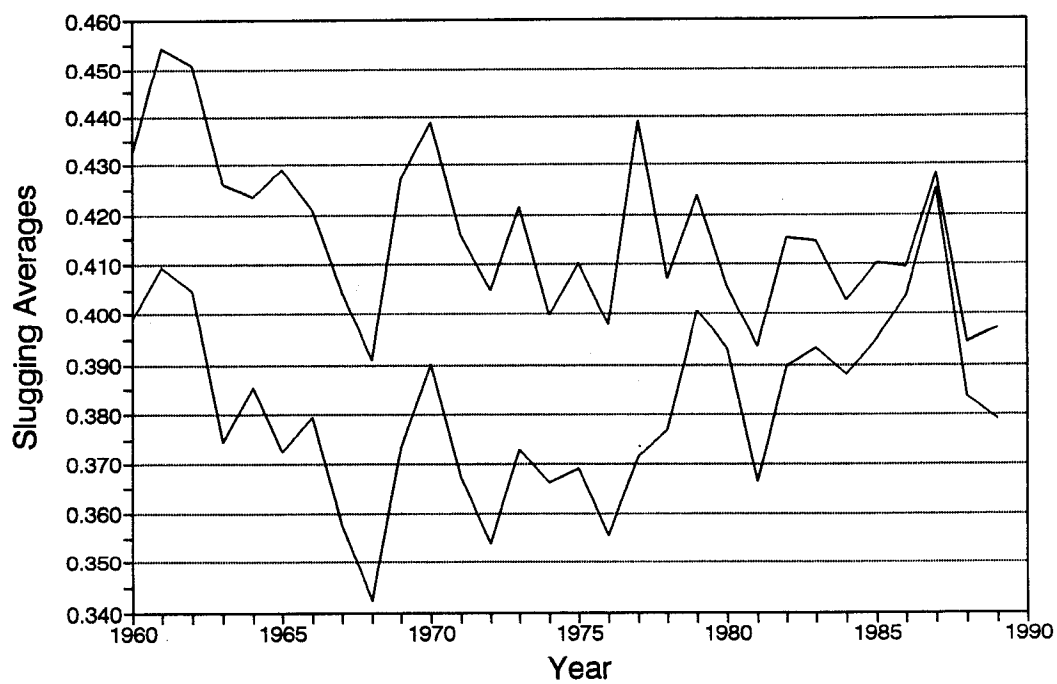## Figure 2: Slugging Averages for Blacks and Whites, 1960-1989
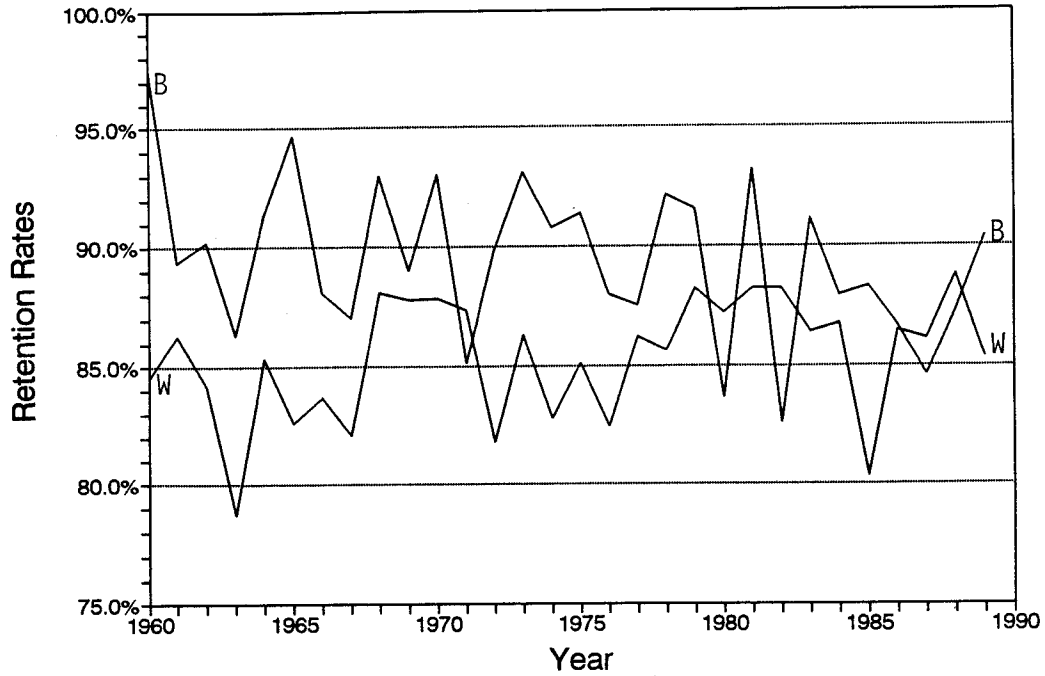
Figure 3: Retention Rates for
Blacks and Whites, 1960-1989

Figure 4: Years of Experience for
Blacks and Whites, 1960-1989