Wins Above Replacement and the MLB MVP Vote: A Natural Experiment

Shane Sanders*[1], Joel Potter†[2], Justin Ehrlich‡[1], and Justin Perline§[1]

[1]Syracuse University, Sport Analytics
[2]University of North Georgia, Economics

October 15, 2019

1 Introduction

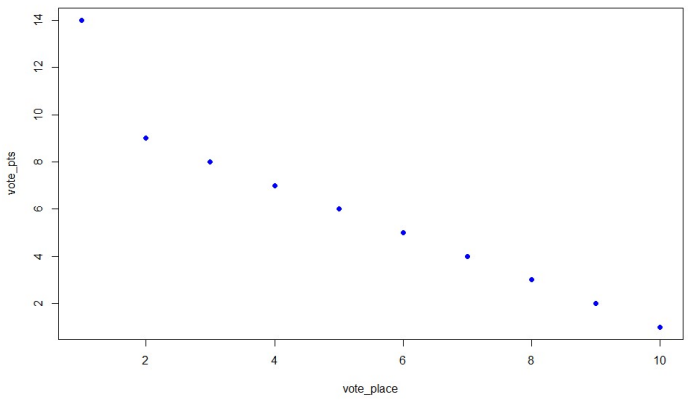
Major League Baseball was formed as a confederacy of two leagues—the National League (NL; 1876-) and the American League (AL; 1901-)—in 1903. Since 1911, the NL and AL have chosen separate Most Valuable Players (MVPs) following each regular season. From 1931, these Awards have been selected by the *Baseball Writers' Association of America* (*BBWAA*). The *Baseball Almanac* summarizes the early history of the Award:

There have been three different official most valuable player awards in Major League Baseball history, since 1911; the Chalmers Award (1911-1914), the League Award (1922-1929), and the Major League Baseball Most Valuable Player Award [1931-]. The MVP...is presented annually by the *BBWAA*. It is considered by MLB as the only official Most Valuable Player Award and symbolizes the pinnacle of a player's personal achievement during any single season of play.

In 1938, the *BBWAA* began electing MVPs via a vote of the *BBWAA* members. Initially, there were three NL (AL) Award voters for each NL (AL) team. That number was reduced to 2 in 1961. For several decades, then, there have been 60 MLB MVP voters, where 30 participate in the NL MVP Award Election and 30 participate in the AL MVP Election following each regular season. The voting rule employed is a weighted scoring rule that has been called a (corner-weighted) version of the Borda Rule. Each voter fills in a ten-place ballot with his or her first-place vote, second-place vote,...,and tenth-place vote. The voting system varies from a standard Borda Rule in two important respects: i) voters write in their choices (i.e., no candidates are specified on the ballot) and ii) first-place votes receive more weight than under a standard Borda Count. Namely, players receive 14 points for each first-place vote, 9 points for each second-place vote, 8 points for each third place vote,..., and 1 point for each tenth-place vote. Hence, a first-place vote has the value of a second and sixth place vote rather than that of a second and tenth place vote. Under this system, the player candidate with the highest total number of points for a given league-year is crowned MVP of that league-year. Below is a valuation plot for this weighted scoring rule.

Figure 1: MVP Vote Points for Each Vote Place

*sdsander@syr.edu
†joel.potter@ung.edu
‡jaehrlic@syr.edu
§jtperlin@syr.edu

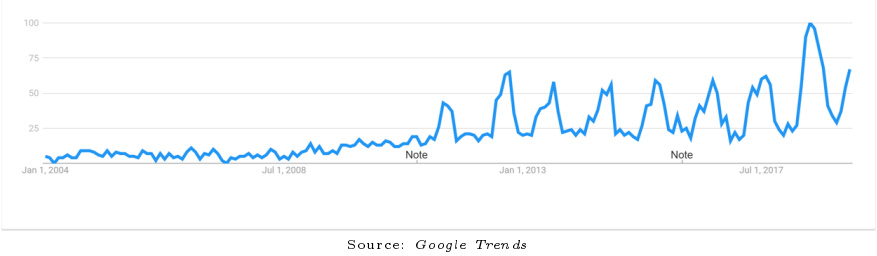


A standard Borda Count system would maintain a linear trend throughout.[1] Under this rule, there are $(14 + \Sigma_{i=1}^{9} i) \cdot 30 = 1,770$ total points for a given League race, where a player can score as many as $30 \cdot 14 = 420$ points.[2]

Baseball is a game of specialization such that cross-positional value comparisons have an apples-to-oranges quality, especially in the absence of a comprehensive, cross-positional measure of (win) value. Consider the comparison between an ace (starting) pitcher and a top shortstop. The shortstop accumulates win value through an everyday mix of offense and defense. The pitcher accumulates value almost exclusively through pitching and does so about once every 6 days. Without some exchange rate mechanism, it is difficult to compare these respective contributions. And yet, MVP voting demanded just such a comparison for decades. In 2004, *Wins Above Replacement* (*WAR*) was created by *Baseball Prospectus* writer Jay Jaffe (Jaffe, 2004) as a cross-positional measure of player win value (above league replacement level player at position). From 2004, *WAR* has been calculated for all MLB player-seasons (e.g., by *Baseball Prospectus* and other baseball publications). Importantly, it has also been calculated retrospectively for the universe of (preserved) professional baseball. *Baseball Reference* (baseball-reference.com) features retrospective calculations beginning with the first season of the first professional baseball league (*National Association of Professional Baseball Players*, 1871-75) and continuing through the respective histories of the *NL* (1876-) and *AL* (1901-). These histories were archived using the massive, crowd-sourced data collection project *retrosheet.org*, which was created by University of Delaware biology professor David Smith in 1989.

The concurrent collection of past and present *WAR* data since 2004 has created something of a natural experimental setting. Since 2004, *MLB MVP Award* voters have cast ballots largely with knowledge of player *WAR* values. Before 2004, voters were relatively uninformed. Given WAR's prominence (e.g., on baseballprospectus.com, fangraphs.com, and baseball-reference.com) and the institutional nature of the *BBWAA* (i.e., as a defined group with a well-archived website, as well as regular chapter and national meetings), *Baseball Writers* who vote for *MLB MVP Awards* each year represent informed voters relative to their pre-2004 counterparts (many of whom are an earlier version of the same person). Beyond frequent references to *WAR*, by *Baseball Writers* (see, e.g., Madden (2017, *NY Daily News*) or Slowinski (2010) for a summary of the measure's importance to *Baseball Writers*), we refer to a *Google Trends* time series graph of the term's search history prominence since 2004.

Figure 2: WAR Interest Over Time



Source: *Google Trends*

[1] Given its extra weight on first-place votes, MLB MVP voting may have a different level of susceptibility to violations of *Local Independence from Irrelevant Alternatives* (*LIIA*) when ranking the top two or three candidates, where Young (1995) states that the preservation of *LIIA* as it pertains to top candidates may be more important than the preservation of *IIA* (general form *Independence*) in a winner-take-all election.

[2] In our data, the NL and AL were imbalanced in terms of number of teams from 1998 (when the Brewers moved to the NL) through 2012 (after which the Astros moved to the AL). This affected the number of ballots designated to each League during those years. We account for this in the data by re-scaling vote points to 420 points possible for each player during those seasons.

From this graph, we observe that interest in *WAR* is cyclical. It builds throughout the season, typically peaking in August or September and falling off during the playoffs and into the offseason. The *MLB MVP* vote occurs just after the close of the *MLB* season in September, at or near the height of annual (search) interest. Though there was some lag until the measure "went public" so to speak, it is important to note that the measure originated from within the *Baseball Writers Association* community, an active and somewhat small organization with regular chapter-level and national-level meetings. After Jaffe published his seminal article on January 6, 2004, the measure quickly gained traction in leading baseball periodicals such as *Baseball Prospectus*. As such, the measure was visible within the *Baseball Writers* community soon after it was developed.

The present study uses the development and retrospective calculation of *WAR* as a natural experiment by which to ask a question that Banerjee et al. (2011) previously consider in a starkly different voting context. Namely, we ask whether informed voters make better choices vis-à-vis objective information about candidate quality. Whereas fans, participants, and voters may not want the *MLB MVP* race to boil down to a contest of highest *WAR*, neither do these parties likely wish for the race to be too far removed from considerations of objective player values. Banerjee et al. (2011) conducted field voting experiments in India and found evidence that non-partisan, third-party public disclosure of incumbent legislator report cards led to a higher voter share for high-performing incumbents. Within the present voting context, the *WAR* measure can be thought of as similar to a non-partisan, third-party evaluation of candidates. It evaluates candidates in a manner that imposes no *a priori* subjective criteria and is "computationally agnostic" to the MVP race. Thoth and Chytilek (2018) find that voters facing time pressure shift their information-gathering efforts from accuracy to efficiency when evaluating candidates. Specifically, they restrict their attention to a smaller set of policies in candidate evaluation. In the present study, *MVP* voters may face time pressure in that the performance differences of top *MLB* players are often slight and subtle and may require viewing hundreds of hours of game footage to perceive. In a given season, an MVP could play for any of the 30 MLB teams, each of which plays a 162 game schedule. There are 2,480 MLB games in a season, and the average game length is a little over 3 hours according to *Baseball Reference*. As such, there are approximately 7,300 hours (more than 304 24-hour periods) of MLB regular season games each year. Faced with this massive output of game performances, MVP voters may rely partly upon time-friendly measures such as basic counting statistics observable from box scores and video highlights to evaluate players throughout the season.

The remainder of the paper is structured as follows. Section II presents the data and methodology of the study. Using a set of fixed (team and season) effects negative binomial (vote) count regression models[3], Section III specifies and tests a model of informed voting. Namely, the model asks whether *MLB MVP* elections from 2004 relate more strongly to an objective, comprehensive, cross-positional measure of on-field value (i.e., *WAR*) than do prior elections dating from 1980. That is, are informed voters making choices that allow them to more closely identify the true "Most Valuable Player" (rank-ordering) for each league-season? In the absence of information on player *WAR*, we consider specific components of player performance and player characteristics that may have become more or less important in explaining the *MVP* race beginning in 2004. Section IV discusses the study's central results and concludes.

2 Data Description, Summary, and Visualization

We collected data on the top 50 seasonal *MLB WAR* leaders in each season from 1980 through 2017. This comprises 38 vote-years and 76 league-level *MVP* elections. We use the *Baseball Reference* version of *WAR* within the study rather than the *Fangraphs* version, the *Baseball Prospectus* version, the *openWAR* version, or another implementation. While all *WAR* measures generate positively correlated sets of *WAR* values of moderate to high strength (see, e.g., Baumer et al. 2015), there are slight methodological differences between the implementations that are beyond the scope of the present study.[4] We chose *Baseball Reference WAR* because it is popular, accessible, and was created with retrospective analysis in mind.[5] Allowing for possible ties in seasonal *WAR* value (at the fiftieth highest value) in each season, the sample includes 1,907 player-season observations rather than 1,900. Each *MLB MVP* voter is charged with identifying the ten most valuable players in a given league, where *Baseball Writers* typically have a high degree of consensus as to whom should be on a given ballot. For example, the union of all *AL MVP* ballots in the year 2000—approximately midway through our sample—featured 19 *AL* players and 22 *NL* players.

The present data set was scraped from *baseball-reference.com* (with permission) using *RVest* package in the statistical software program *R*. Variables at the player-season level include:

[3] Negative binomial models are chosen over Poisson models due to strong evidence of overdispersion in the data

[4] These measures use the same inputs but slightly different specifications.

[5] *Baseball Reference* Founder and CEO Sean Forman confirmed this point in a conversation with two of the present authors. Of course, retrospective analysis of the recent season is an input in *MVP* decision-making.

(MVP election) vote point count, *WAR value*, *age*, *whether in a big (top-5) market*, *whether traded during season*, *(primary) league-of-play*, *(primary) playing position*, *(primary) team-of-play*

Summary statistics for these variables are provided in the following table. Note that *votepts* is rescaled from 1998 through 2012 to account for league imbalance.

Table 1: Summary Statistis of Key Variables

Variable	**Obs**	***Mean***	***st dev (s)***	***Min***	***Max***
votepts	1,907	54.33	92.68	0	420
WAR	1,907	5.60	1.29	3	12.7
age	1,907	28.30	3.69	19	44
big market	1,907	0.30	0.46	0	1
multiple teams	1,907	0.02	0.15	0	1
AL	1,907	0.50	0.50	0	1
P	1,907	0.31	0.46	0	1
C	1,907	0.05	0.21	0	1
1B	1,907	0.10	0.30	0	1
2B	1,907	0.07	0.26	0	1
SS	1,907	0.07	0.26	0	1
3B	1,907	0.11	0.31	0	1
RF	1,907	0.09	0.29	0	1
CF	1,907	0.11	0.31	0	1
LF	1,907	0.08	0.27	0	1
DH	1,907	0.01	0.12	0	1

The summary statistics of Table 1 reveal some interesting features of the data. Player *age* in the sample ranges between 19 and 44, which represents most of the overall age range for *Major League Baseball*. Incredibly, the average *WAR* value in the sample is 5.60. This value suggests that *MLB* stars generate exceptional average win value. A team of replacement players is commonly estimated to win about 48 games in a season, *ceteris paribus* (*Baseball Reference*, 2012). Using this benchmark, we can consider the implied success of a (25-player) *MLB* team, where players average a 5.60 *WAR* value. In fact, such a high level of win production could not exist on one team (due to crowding out effects). A single team would have to win $48 + (5.60 \cdot 25) = 188$ games to support such a *WAR* average. However, there are only 162 games in an *MLB* regular season. The sample is quite balanced in terms of League representation, with roughly 50 percent of observations coming from the *AL* (*NL*). In the dataset, playing *position* is represented by a set of *position* dummy variables, *{P, C, 1B,...,DH}*. We observe substantial variation in sample representation by playing *position*. This stands to reason, as (players of) some positions are innately more valuable than (players of) others. We observe that more than 28 percent of sampled player-seasons were conducted in the service of a team in a city of top-5 market size (represented as the variable *big market*). As 8 of 30 (26.7 percent of) *MLB* teams were in a top-5 market throughout the sample, it appears that top talent does not gravitate disproportionately to big markets in the *MLB*. This appears to be the case despite the absence of a salary cap in *MLB*. We base this variable upon the U.S. metropolitan areas with the five largest populations according to the 1980, 1990, 2000, and 2010. Surprisingly, the five metropolitan areas making this list did not change from 1980 to 1990, 2000, or 2010. In the estimation section, we will consider the distinct question as to whether being in a "big market" helps a player in terms of *MVP vote points*. Lastly, we observe that there are 43 sampled player-seasons in which the player represented *multiple teams*.

The data contains player-level vote counts but not voter-level (ballot-level) data. Ballot-level data was not publicly disclosed by the *BBWAA* until 2018. As such, we treat aggregate *vote points* as our main left hand side variable. We take advantage of a long longitudinal data set with substantial variation in player characteristics. We also benefit from the natural experimental context of *WAR's* development (with subsequent retrospective calculation). From our 1,907 player-seasons, 1,125 received at least one vote point. With complete consensus across ballots (as to whom should receive a vote), 760 players would have received points. With no consensus (with respect to players in the data set), all 1,907 sampled player-seasons would have received votes. Thus, we conclude that *Baseball Writers* have a moderately strong degree of consensus as to whom should be on the ballot(s). The present study considers whether such a level of consensus is one among informed voters or one among uninformed voters (e.g., an echo chamber) over time. Figure 3 displays a kernel density plot of the variable *vote points* within our sample.

Figure 3: Kernel Density Plot of Vote Points

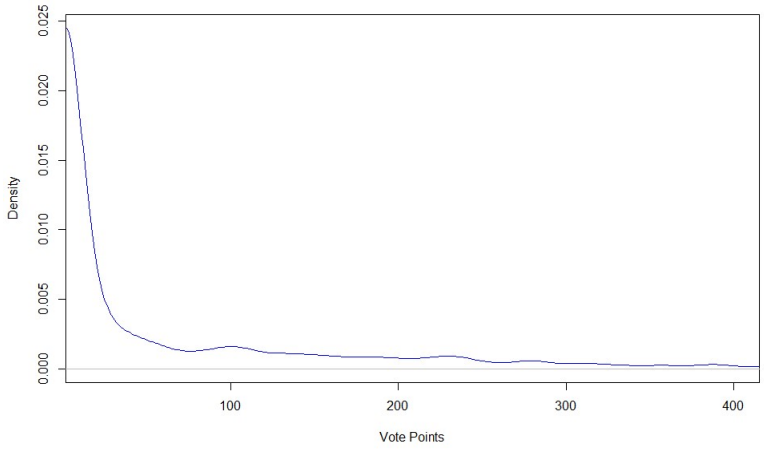
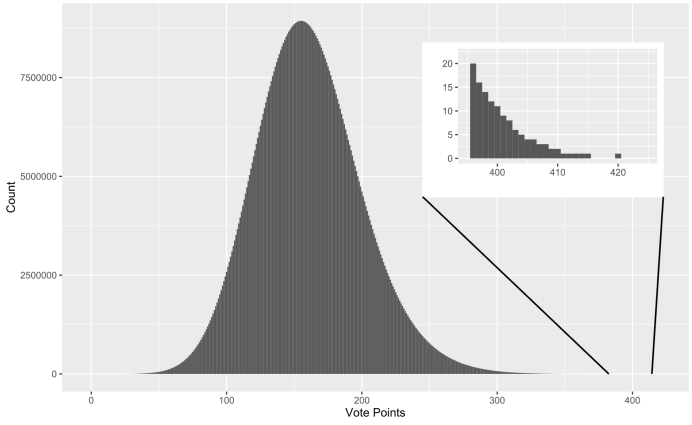


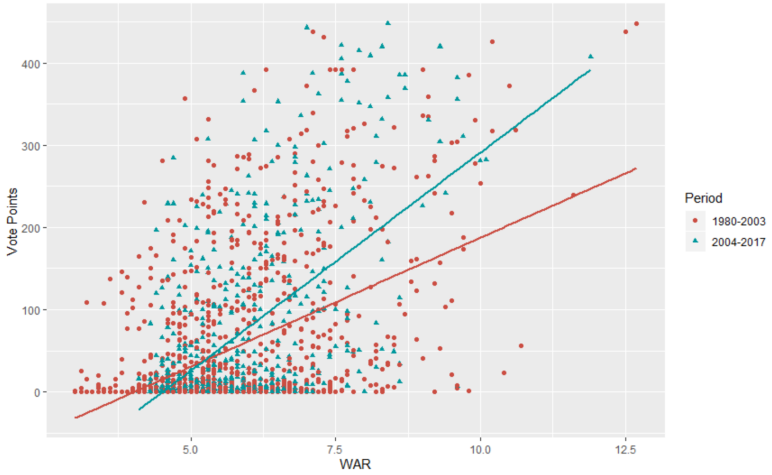
Figure 3 represents a smoothed representation of a discrete variable (*vote point* count). The figure suggests that, even among seasonal *WAR* leaders, most player-seasons result in a low number of vote points. The median *vote point* count in sample is 4, and the 75^{th} percentile is 64. Given our model selection (i.e., of a count data model), we account for the extreme right skewness of the dependent variable, its discrete distribution, and its left-truncation (at the value of zero). *Vote points* is a count data variable in that a) it is non-negative integer valued, b) vote points are in fact a count of how many points voters allocate to a given player (from their bundle of points), and c) the set of possible *vote points* for a player represents a set of consecutive, non-negative integers along with one possible non-consecutive value obtained by a player sweeping all first place votes. That is $VotePoints = \{v \in N : 0 \leq v \leq 415 \ \cup \ v = 420\}$. To demonstrate that *vote point count* possesses this co-domain, Figure 4 displays a histogram of the distribution of possible *vote point counts* for an individual (i.e., from the set of all possible $30x1$ voting vectors for a given individual), where the count represents the number of possible vectors yielding the corresponding number of *vote points*.

Figure 4: Distribution of Possible Player Vote Point Counts



Before formally modeling the relationship of interest, let us consider a scatter plot with player-season *vote point count* on the vertical axis and player-season *WAR* value on the horizontal axis. In Figure 5a to follow, we use color-coded data points and corresponding color-coded trend lines to depict the uncontrolled relationship between the two variables both before and after the creation of *WAR*, respectively.

Figure 5a: Plot of MVP Vote Points against WAR for 2 Time Periods (before and after information technological shift)

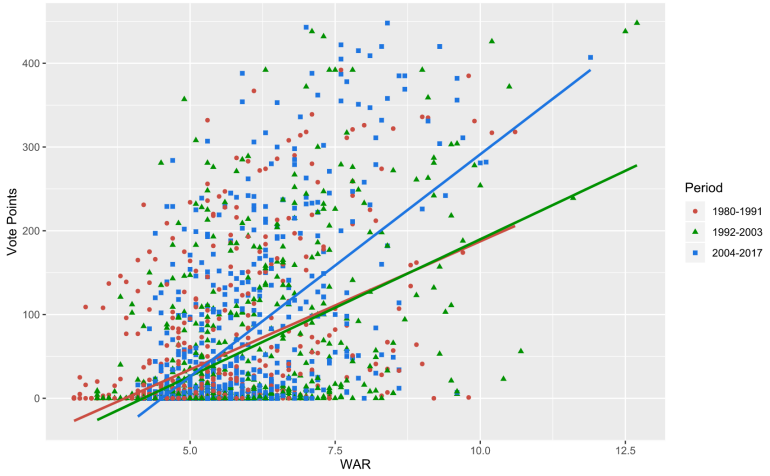


The uncontrolled relationship between *vote point count* and *WAR* was positive both from 1980 through 2003 and from 2004 through 2017. However, simple regression trend lines suggest that the responsiveness of *vote point count* to changes in *WAR* has been noticeably stronger in the latter time period. Multivariate regression analysis will inform us as to whether this apparent difference is significant and substantial conditional on a set of control variables. When considering Figure 4, one might wonder whether the demonstrated slope difference is indeed due to the creation of *WAR* or if it might suggest a more general (gradual) improvement in the knowledge and ability of voters over the course of the data set. We are dealing with a panel data herein with multiple observations per time period. Therefore, an endogenous structural break estimation is not supported. However, we do in fact know when the information technology break point of interest occurred (a few months after the 2003 *MVP* vote), such that an exogenous structural break estimation is appropriate. Moreover, we know that *WAR* is the first and only virally adopted cross-positional player value measure (i.e., the first such measure to be featured on leading baseball statistics sites *Baseball Prospectus*, *Baseball Reference*, *ESPN*, and *Fan Graphs*). As such, *WAR* was not established in a sea of comparable measures. Rather, it was a groundbreaking measure in the area of comprehensive sabermetric player analysis. As Baumer and Matthews (2014) state in an article subtitled *There is No Avoiding WAR*:

> While there have been many important contributions to the field, arguably the most prominent success story in recent years has been wins above replacement (WAR). WAR is an all-encompassing assessment of a baseball player's contribution to his team, measured in wins added relative to a hypothetical (and often vaguely defined) 'replacement player'...While predecessors of WAR (like Bill James's Win Shares) have been around for some time, the modern incarnation of WAR may now reach a broader audience than any sabermetric stat since OBP. Perhaps the most telling indication of WAR's permeation of the baseball landscape was the announcement that WAR will appear on the back of Topps baseball cards in 2013.

According to Baumer and Matthews, the only sabermetric measure comparable to *WAR* in impact is *On Base Percentage*, which measures a player's value in one aspect of the game. For the present setting, then, there appear to be no other technology break point candidates than 2004. Therefore, we will conduct a *Chow* (exogenous structural break) *Test*—through our empirical specification itself—in the estimation section to follow. The structural break testing will initially center upon the year 2004 before considering the possibility of an earlier break point. Despite not having a tractable endogenous structural break test at our disposal, we will use one means of testing for a more gradual change in voter behavior. In the visualization to follow, as well as in the estimation section, we divide our pre-2004 data period into two equal sub-periods: 1980-1991 and 1992-2003. If voters were simply gradually improving in their ability to measure player value, we might expect the trend line slope between *vote point count* and *WAR* to increase incrementally between both the first (1980-1991) and second (1992-2003) time period and between the second and third (2004-2017) time period. The following trinary color-coded plot helps us consider this relationship across time.

Figure 5b: Plot of MVP Vote Points against WAR for 3 Time Periods (before information technological shift in 2 periods and then after)



In terms of the (slope of) relationship between *vote points* and *WAR*, there appears to be no substantial difference between the first two time periods. As such, Figure 5 is quite similar to Figure 4 in terms of identifying a potential change in the relationship of interest beginning in 2004. In our estimation models, we will consider both binary and trinary (time period) partitions of the data set within a multivariate regression setting.

3 Model Specification

Given that the dependent variable, *vote point count* is a non-negative integer-valued variable, we specify a count model herein. A χ^2 test reveals over-dispersion in our dependent variable such that we use a *negative binomial* model rather than a *poisson* model. The baseline model features *team*, *season*, and playing *position* as fixed effect variables, where a Hausman specification test supports selection of fixed over random effects modeling for each variable. Given the institutional and empirical setting of the data, we forego the specification of *player* fixed effects. Many sampled players (283) appear in the data set in only one season. Several other players (70) appear in the data set in more than one season but do not earn a vote in any season. Moreover, *team* and *position* fixed effects are mutually compatible within a specification, whereas the specification of *player* fixed effects precludes the inclusion of both *position* and *team*, as a player tends to persist on the same team and in the same position. Thus, specification of *player* fixed effects would lead to a) loss of substantial data representing specific types of player-season (sub-sample selection bias) and b) preclusion of our other fixed effect variables. As such, we decide that this data does not have classic micro-panel characteristics and specify a panel structure at the levels of *team*, *position*, and *season*.

Our set of fixed effect, negative binomial regressions appear as follows. The baseline model tests for a general relationship between *vote point count* and *WAR* from 1980-2017. The baseline model is given as follows.

$$votepts_{i,t} = \beta_0 + \beta_1 \cdot WAR_{i,t} + \beta_2 \cdot age_{i,t} + \beta_3 \cdot age_{i,t}^2 + \beta_4 \cdot multiple\_teams_{i,t} + \mathbf{position_i} \cdot \beta_{\mathbf{5}} + \mathbf{team_i} \cdot \beta_{\mathbf{6}} + \mathbf{season_t} \cdot \beta_{\mathbf{7}} + \epsilon_{i,t} \quad (1)$$

$$votepts_{i,t} = \beta_0 + \beta_1 \cdot WAR_{i,t} + \beta_2 \cdot age_{i,t} + \beta_3 \cdot age_{i,t}^2 + \beta_4 \cdot multiple\_teams_{i,t} + \beta_5 \cdot big\_market_{i,t} + \mathbf{position_i} \cdot \beta_{\mathbf{6}} + \mathbf{season_t} \cdot \beta_{\mathbf{7}} + \epsilon_{i,t} \quad (2)$$

There is no *a priori* assurance that *vote point count* and *WAR* bear a significant relationship over the full data period. The purpose of the baseline model is to test whether any such relationship exists. As *WAR* was not created until 48 of the 76 sampled MVP votes had occurred, voters made decisions in the absence of *WAR* for almost two-thirds of the data period. Over this time, voters cast ballots presumably based on their observation of the game and on available (decentralized) player statistics. Further, there is no *a priori* assurance that voters adopted *WAR* in their voting behavior following its creation. While *Baseball Writers* have certainly featured *WAR* prominently in (public) baseball articles since 2004, it is an empirical question as to whether this adoption carried over to their (individually private or anonymous) voting behavior.

Following the baseline treatment, we use a *Chow Test* (for difference in slopes) type specification to test for presence of slope difference in the relationship of interest during the sub-period from 2004-2017. Like the scatter plots presented in the previous section, this test is conducted with 1980-2003 as the (single) comparison period and (in subsequent specifications) with 1980-1991 and 1992-2003 as the (dual) comparison periods. Note that a *Chow Test* for *intercept* difference is irrelevant herein. The intercept in our models represents the expected *vote point count* of a replacement player-season (i.e., a player with a *WAR* value of 0). Such player-seasons are well below average by definition and thus not in the data. Our two-period *Chow Test* specifications will appear as follows.

$$votepts_{i,t} = \beta_0 + \beta_1 \cdot WAR_{i,t} + \beta_2 \cdot \ (WAR_{i,t} \cdot after\_2003_{i,t}) + \beta_3 \cdot age_{i,t} + \beta_4 \cdot age_{i,t}^2 + \beta_5 \cdot multiple\_teams_{i,t} + \mathbf{position_i} \cdot \beta_{\mathbf{6}} + \mathbf{team_i} \cdot \beta_{\mathbf{7}} + \mathbf{season_t} \cdot \beta_{\mathbf{8}} + \epsilon_{i,t} \quad (3)$$

$$votepts_{i,t} = \beta_0 + \beta_1 \cdot WAR_{i,t} + \beta_2 \cdot \ (WAR_{i,t} \cdot after\_2003_{i,t}) + \beta_3 \cdot age_{i,t} + \beta_4 \cdot age_{i,t}^2 + \beta_5 \cdot multiple\_teams_{i,t} + \beta_6 \cdot big\_market_{i,t} + \mathbf{position_i} \cdot \beta_{\mathbf{7}} + \mathbf{season_t} \cdot \beta_{\mathbf{8}} + \epsilon_{i,t} \quad (4)$$

The variable $after\_2003$ is an indicator variable that equals 1 for vote years 2004-2017 and 0 otherwise. We interact this variable with *WAR* to compare the relationship between *vote point count* and *WAR* before and after the measure's creation. The variable $after\_2003$ does not enter the right hand side of the model alone for reasons discussed previously. Namely, there is no reason to expect a difference in intercept values across time periods. Moreover, the intercept is extrapolative within the model, as replacement players are not expected to receive *MVP* votes. We next consider a model that is consistent with a three-period *Chow Test* for change in slope.

$$votepts_{i,t} = \beta_0 + \beta_1 \cdot WAR_{i,t} + \beta_2 \cdot (WAR_{i,t} \ \cdot \ after\_2003_{i,t}) + \beta_3 \ \cdot (WAR_{i,t} \ \cdot \ between\_92\&03_{i,t}) + \beta_4 \cdot age_{i,t} + \beta_5 \cdot age_{i,t}^2 + \beta_6 \cdot multiple\_teams_{i,t} + \mathbf{position_i} \cdot \beta_{\mathbf{7}} + \mathbf{team_i} \cdot \beta_{\mathbf{8}} + \mathbf{season_t} \cdot \beta_{\mathbf{9}} + \epsilon_{i,t} \quad (5)$$

$$votepts_{i,t} = \beta_0 + \beta_1 \cdot WAR_{i,t} + \beta_2 \cdot (WAR_{i,t} \ \cdot \ after\_2003_{i,t}) + \beta_3 \ \cdot (WAR_{i,t} \ \cdot \ between\_92\&03_{i,t}) + \beta_4 \cdot age_{i,t} + \beta_5 \cdot age_{i,t}^2 + \beta_6 \cdot multiple\_teams_{i,t} + \beta_7 \cdot big\_market_{i,t} + \mathbf{position_i} \cdot \beta_{\mathbf{8}} + \mathbf{season_t} \cdot \beta_{\mathbf{9}} + \epsilon_{i,t} \quad (6)$$

The variable $between\_92\&03_{i,t}$ allows us to trisect the data so as to test for possible evidence that voting behavior actually began to change before 2004 (e.g., as some more gradual result of the sabermetric movement in general). In essence, models (5) and (6) allow us to conduct a *Chow Test* for changes in the relationship between *vote points* and *WAR* for the periods 1980-1991, 1992-2003, and 2004-2017. If the creation of *WAR* were integral to changing this relationship, we might expect the relationship to remain fairly stable before 2004 and to change significantly (and perhaps substantially) thereafter.

4 Estimation & Results

In this section, we report the estimation results of the 6 fixed effect, negative binomial models specified in the previous section. These results are presented in Table 2.

[See Appendix for Table 2]

Over the whole data period, Table 2 demonstrates that the relationship between *WAR* and *vote points* is positive and significant. Models 3 and 4 show that the estimated slope of this relationship increased significantly (became significantly more positive) following the creation and publication of *WAR*. Models 5 and 6 provide evidence that this increase did not arise as a gradual and vague response to the sabermetric era in general (e.g., not a process that began in years prior to 2004 and built from there). Rather, these models demonstrate that voters behaved in a statistically equivalent manner from 1992-2003 as they had from 1980-1991. It is only in the 2004-2017 data that we observe a change in voter behavior. As such, we have evidence that (informed) post-2003 MVP voters allocated *vote points* in a manner that is more consistent with actual player value as compared to their relatively uninformed counterparts of previous time periods. Further, we find evidence that this increased responsiveness to *WAR* leads

to a more explanatory model. For the purpose of assessing explanatory power before and after the development of *WAR*, we divide our data into two sub-periods (1980-2003 and 2004-2017) and run model specification (1) for each time period separately as an ordinary least squares, fixed effects model. We do this because fixed effect, negative binomial estimation does not yield an *R-squared* value or any other accessible measure of model explanatory power. We find that the overall *R-squared* rises from 0.366 in the former data period to 0.464 in the latter data period. That is, the percentage of variation in vote points explained by variation in player characteristics and performance rose by almost 10 percentage points in the latter period. It is not simply that voting was significantly and fairly substantially more responsive to objectively-measured player value beginning in 2004. From 2004, voting has been better explained by objective measures of player value (i.e., has been less subject to noise).

Negative binomial regression coefficient estimates do not indicate the marginal effect that a unit change in an explanatory variable has upon the dependent variable, as in a linear model. We therefore estimate model specification (5) as a fixed effect, OLS model to estimate marginal effects. Before 2004, we find that each additional unit of *WAR* increased expected *votepts* by 35.24 and that this marginal effect increased by 13.13 *votepts* after 2004. If pre-2004 voting had been as responsive to *WAR* as subsequent voting, *ceteris paribus*, several pre-2004 *MVP* races would have pivoted (e.g., the 1999 and 2001 *AL* races). Lastly, we replace *WAR* with *offensive WAR* and *defensive WAR* (for position players) and with *pitching WAR* (for pitchers) and estimate this modified version of model (3) so as to determine what specific aspects of performance were treated differently by voters from 2004. We run this s a negative binomial model for both position players (1) and for pitchers (2). Though we use the full specification from Model (3) of Table (2), we report only coefficient estimates for the variables of interest in Table 3 (for brevity).

Table 3: Additional Estimation Results

	(1) VotePts	(2) VotePts
VotePts		
oWAR	0.569****	
	(0.020)	
oWAR_after	0.070****	
	(0.015)	
dWAR	0.002	
	(0.036)	
dWAR_after	0.104**	
WAR		0.721****
		(0.058)
WAR_after03		0.066**
Observations	1324	583

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$

5 Discussion and Conclusion

Before the advent of *WAR*, there is no statistical evidence that *ceteris paribus* improvements in defensive performance (*defensive WAR* or *dWAR*) improved one's MVP *vote point* count by any amount! Rather, only *offensive WAR* (*oWAR*) and *pitching WAR* (*pWAR*) led to a significant increase in MVP *vote points* before 2004. From 2004, improvements in *dWAR* were rewarded significantly in the MVP race. As defense is the least salient source of value on the baseball field, its lack of (statistical) importance in MVP voting before 2004 stands to reason. With the publication of *dWAR* values, it appears there is evidence that defensive value became a productive input in the MVP race for the first time. We also find evidence that voting from 2004 rewarded marginal improvements in *oWAR* even more strongly. However, the reward for marginal improvements in *pWAR* did not change significantly in magnitude. Both before and after 2004, units of *pWAR* were rewarded in MVP voting but by much less than were units of *oWAR*. It is possible that (recently informed) voters purposely begrudge pitchers because the *Cy Young Awards* (specifically for pitchers) are announced just before the *MVP Awards*. Like some parents whose child has a December 24^{th} birthday, even informed voters may be reluctant to bestow consecutive sets of gifts to the same parties.

From these results, we conclude that *Baseball Writers* are not simply writing about advanced baseball statistics to add color to articles. Since 2004, we find evidence that the creation of *WAR* has changed decision-making behavior in a high stakes environment (MLB MVP Voting) both significantly and substantially.

6 References

Banerjee, Abhijit, Selvan Kumar, Rohini Pande, and Feliz Su. 2010. Do Informed Voters Make Better Choices? Experimental Evidence from Urban India. http://www.povertyactionlab.org/node/2764

Baumer, B. S., Jensen, S. T., & Matthews, G. J. (2015). openWAR: An open source system for evaluating overall player performance in major league baseball. Journal of Quantitative Analysis in Sports, 11(2), 69-84.

Baumer, B. & Matthews, G. (2014). A Statistician Reads the Sports Page: There is No Avoiding WAR. Chance, 27(3), 41-44.

Jaffe, Jay (6 January 2004). "The Class of 2004: Analyzing the Hitters." Baseball Prospectus. Retrieved 18 June 2019. Available online at: https://www.baseballprospectus.com/news/article/2502/the-class-of-2004-analyzing-the-hitters/

Madden, Bill. (8 December, 2017). "WAR shouldn't be so heavily emphasized when it comes to voting for the Baseball Hall of Fame." Retrieved 03 July 2019. Available online at: https://www.nydailynews.com/sports/baseball/war-shouldn-heavily-emphasized-baseball-hof-voting-article-1.3684618

Slowinski, S. (2010), "What is WAR?" http://www.fangraphs.com/library/index.php/misc/war/

Thoth, M., & Chytilek, R. (2018). Fast, frugal and correct? An experimental study on the influence of time scarcity and quantity of information on the voter decision making process. Public Choice, 177(1-2), 67-86.

Young, P. (1995). Optimal voting rules. Journal of Economic Perspectives, 9(1), 51-64.

7 Appendix

Table 2: Main Estimation Results

	(1) VotePts	(2) VotePts	(3) VotePts	(4) VotePts	(5) VotePts	(6) VotePts
VotePts						
WAR	0.545****	0.529****	0.528****	0.514****	0.544****	0.532****
Age	-0.007	0.052	0.024	0.113	0.032	0.115
Age^2	0.000	-0.001	-0.000	-0.002	-0.001	-0.002
MultipleTeams		-0.789***		-0.846***		-0.845***
bigmarket		0.039		0.044		0.052
WAR · after_2003			0.078****	0.077****	0.065****	0.062****
WAR · between_92&03					-0.024	-0.027*
1B	0.800****	0.825****	0.762****	0.788****	0.780****	0.816****
2B	-0.072	-0.075	-0.141	-0.172	-0.138	-0.159
3B	0.020	0.056	0.002	0.038	0.007	0.051
C	0.057	0.162	0.056	0.156	0.066	0.170
CF	-0.143	-0.100	-0.178	-0.155	-0.168	-0.135
DH	0.992****	1.017****	0.997****	1.025****	1.020****	1.056****
LF	0.085	0.057	0.072	0.034	0.078	0.060
P	-1.283****	-1.189****	-1.310****	-1.229****	-1.297****	-1.208****
RF	0.272**	0.330***	0.227*	0.281**	0.244*	0.308**
Constant	-4.336***	-5.055****	-4.909****	-5.992****	-5.073****	-6.097****
Observations	1907	1907	1907	1907	1907	1907

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, **** $p < 0.001$