1. Introduction

Major League Baseball (MLB) is now a multi-billion dollar operation with top players regularly receiving contracts in excess of 100 million dollars. It is necessary for owners to have accurate measures of player productivity to properly gauge the monetary value of players, who represent significant economic investments. Advanced statistical analysis is already the norm for MLB teams and informs management decisions regarding player compensation, trades, and other features of player personnel. The most widely used and accepted measure of player productivity in baseball is Wins Above Replacement (WAR), which estimates how many additional wins each player produces relative to their presumed replacement.

WAR is an attractive concept; it considers offense, defense, and pitching to give each player one number that represents their productivity measured by their direct effect on team success. WAR has entered even the casual fan’s lexicon; it is rare to see a day go by without WAR being mentioned on SportsCenter or Baseball Tonight, and ESPN.com’s MLB statistics have included WAR for several years already. The simple fact that WAR is now engrained in the minds of casual baseball fans shows its intuitive appeal as a measure of productivity.

To summarize WAR: the MLB labor market consists of a finite pool of players that exceeds the combined roster limitations its teams. Assuming all teams have win-maximizing strategies, it is only rational that teams would attempt to fill their rosters with the best players available. Players not considered valuable enough for a roster spot are then substitutes, or replacements, for players who are on rosters. The worst possible scenario for any MLB team is to have a roster consisting only of replacement-quality players. A team of replacements would be highly unsuccessful, but still expect to win a relatively small percentage of games. There is empirical evidence of this: the 2003 Detroit Tigers, the worst team modern baseball history, still won 26.5% of their games. A team consisting solely of replacement players should be considered the lowest possible level of performance for any MLB team. As such, WAR can be thought of as an estimate of the Marginal Productivity of each player relative to that replacement-level expectation, hence “Wins Above Replacement.”

Despite WAR’s conceptual simplicity, there is no standardized method of calculation. Several different sources calculate WAR, each doing so in a different way, all advertising as “Wins Above Replacement” (this paper focuses specifically on WARP, rWAR, and fWAR for comparisons). For example, 2013 Cy Young Award winner Clayton Kershaw was found to be worth 5.7, 7.1, or 7.8 additional wins that season depending on the source. A difference of 2 WAR corresponds to a variance of roughly $10 million in player compensation, a significant dollar amount by any
measure. Furthermore, the inner workings of existing WAR measures are largely inaccessible, as the full derivation of existing WAR measures is unavailable.

Further regarding the labor market, baseball players typically have very specialized roles requiring specialized skillsets. The most obvious division of labor is between pitchers and batters, but the majority of players have a primary position that they rarely deviate from. A win-maximizing team forced to select personnel from the available pool of replacement players cannot simply choose the best players overall, they are limited to the pool of available players at each distinct position. As such, the Marginal Productivity of individual players should be position-specific, rather than relative to broad league averages, because player value from the team perspective is relative to other players at the same position. Existing WAR measures fail to make these distinctions, and thus lack flexibility and specificity in their analyses.

This paper introduces a new measure of WAR, called zWins, which outperforms and overcomes the weaknesses of existing WAR measures by introducing major innovations and providing an open system of calculations with complete disclosure of all formulas and methods. zWins begins with the same underlying concept as existing WAR measures, that wins are a function of runs scored and allowed. Because wins are a function of runs, improving the accuracy of run estimation models should improve the accuracy of win models. In essence, zWins is calculated as such:

\[
\text{zWins} = f(\text{zOffense, zPitching, zDefense})
\]

\[
\text{zOffense} = f(\text{CRE, Outs})
\]

\[
\text{zPitching} = f(\text{CREID, Outs})
\]

\[
\text{zDefense} = f(\text{DRA, zRange, z Outs})
\]

Where Calculated Runs Expectancy (CRE) is an estimate of offensive run production, Calculated Runs Expectancy Independent of Defense (CREID) is an estimate of pitcher run allowance, Defensive Runs Allowed (DRA) is an estimate of run allowance attributable to fielding, and zRange and zOuts are estimates of runs saved due to defensive performance. zOffense, zPitching, and zDefense are Runs Above Replacement measures for the three primary aspects of baseball, specific to player position, league, and handedness of pitchers.

In this paper I derive zWins and show that it outperforms the existing WAR measures. This is not only for WAR, but also the subcomponents that make up offensive and defensive estimates of performance. In particular I show that CRE was consistently more accurate than both Weighted Runs Created and Base Runs, the production models used in various WAR measures. CREID and DRA are complementary statistics, calculated identically to CRE while using different sets of inputs, which when summed provide estimates of total run allowance comparable in accuracy to CRE. The sets of inputs were determined based on the division of labor between pitchers and fielders, but their accuracy when combined provides strong evidence of their ability to outperform comparable existing methods of pitching and defensive analysis. As a result of maintaining position-specific
analysis and improving run estimation models, zWins showed superior accuracy over competing WAR measures when tested against both actual team performance and Pythagorean Expectation.

## 2. Deriving zWins

As noted above, zWins is a function of zOffense, zPitching, and zDefense, all of which are measured in runs. Two assumptions and corresponding calculations (denoted W1 and W2) were used in converting runs to wins: 1) there exists some number of runs that mathematically equates to one win, specific to each season; 2) it is possible to derive expected replacement-level winning percentages specific to each team, every year. zWins is the mean of W1 and W2, with a multiplicative adjustment included to ensure that the sum of all zWins and all teams’ expected replacement-level wins is equal to the number of actual wins in the league (half of the total number of games played).

\[
zWins = \mu(W1, W2) \times optimization \ coefficient \tag{5}
\]

I denote a variable, Q, to represent the number of runs corresponding to roughly one marginal win in any given season. Existing WAR measures assume a constant exchange rate of about 10 runs, but Q itself is a function of league run-scoring averages, and has a positive relationship with runs with decreasing marginal returns.

\[
Q = \mu(RS_1, RS_2, RS_3, ... RS_n)^{0.49} - 15.1 \tag{6}
\]

W1 simply divides each player’s sum of zOffense, zPitching, and zDefense by Q. CRE, an exponential function, approaches infinity as outs approach zero, so W1 is multiplied by the percentage of games played (Pct. G) for all non-pitchers to combat highly efficient batters as outliers.

\[
W1 = \frac{\sum z\text{Offense, zPitching, zDefense}}{Q} \times Pct. G_{non-pitchers} \tag{7}
\]

W2 estimates the effect on team winning percentage attributable to each player relative to the expected replacement-level winning percentage. I derive expected replacement-level winning percentages (\(\epsilon_{rep}\)) using the Pythagorean Expectation (\(\epsilon\) formula, which estimates winning percentage as a function of Runs Scored (RS) and Allowed (RA), and team-specific sums of zOffense, zPitching, and zDefense. I used the same version as baseball-reference.com, which has 1.83 as exponents to minimize root-mean-square error.

\[
\epsilon = \frac{RS^{1.83}}{RS^{1.83} + RA^{1.83}} \tag{8}
\]
\[ \varepsilon_{rep} = \frac{(RS - \sum z_{Offense})^{1.83}}{(RS - \sum z_{Offense})^{1.83} + (RA + \sum z_{Pitching} + \sum z_{Defense})^{1.83}} \] (9)

W2 is the measured change in expected winning percentage as a function of an individual’s zOffense, zPitching, and zDefense, multiplied by the number of games played by the team, G. zPitching is multiplied by a constant 1.33 in this equation because the average marginal returns to wins on RS consistently exceed the marginal returns on RA by approximately that rate.

\[ \Delta \varepsilon_{rep} = \varepsilon_{rep} - \frac{(RS + z_{Offense})^{1.83}}{(RS + \sum z_{Offense})^{1.83} + (RA - z_{Pitching} \times 1.33 - z_{Defense})^{1.83}} \] (10)

\[ W2 = \Delta \varepsilon_{rep} \times G \] (11)

**Figure 1: zWins vs. Actual Wins, 2009-2014**

\[
y = 0.9769x + 48.677 \\
R^2 = 0.83484
\]

*WARP data could not be collected as it is not freely available for team statistics*
The intercepts of Figure 1 (\(\beta_0\)) approximate the average expected number of wins by a replacement-level team. For comparison, WAR’s general replacement-level expectation (0.294 winning percentage) predicts slightly under 48 victories over a 162-game season. All else equal, the slopes of these models (\(\beta_1\)) report the average change in team wins when 1 WAR is added. On average, fWAR was the closest to maintaining a perfect linear relationship, but it is clear from the \(R^2\) values that zWins better explained actual team success than existing measures.

I included \(\epsilon\) in Figure 2 to illustrate the inefficiency of existing WAR estimates. Assuming that Pythagorean Expectation provides the most accurate approximations of winning percentage, it seems an obvious shortcoming that existing WAR measures fail to achieve that level of accuracy. This regression also illustrates the effect of dynamic replacement-level expectations on the accuracy of WAR calculations. The \(R^2\) values for both fWAR and rWAR remain the same in both regressions because both measures regard replacement-level expectations as constant, rather than as a function of team personnel. zWins derived...
team-specific replacement-level expectations to produce a tighter correlation, which rivals Pythagorean Expectation in accuracy.

\[ y = 1.0423x - 3.3973 \]
\[ R^2 = 0.98012 \]

Figure 3 explains that zWins rivals ∈ in accuracy because the two have a near-linear relationship. With such a strong correlation to Pythagorean wins, zWins is limited in accuracy only by the accuracy of the win estimation model being used. Likewise, error observed in the second regression can be largely explained by error existing within Pythagorean Expectation.

**2.1. Deriving CRE, CREID, and DRA**

The primary metric for zWins is a run estimation model called Calculated Runs Expectancy (CRE). CRE combines traditional offensive statistics as inputs to produce runs as an output. CRE is not a linear Econometric model because I believe run production in baseball is poorly explained by linear coefficients. Weighted Runs Created (wRC) uses weighted coefficients of selected outcomes to model run production (i.e. home runs, on average, were
worth 2.065 runs), which I refer to as an outcome-driven approach. It assumes that the most productive batters are those who hit many home runs and draw many walks because those plays frequently produce beneficial outcomes. While that is true, the rules of the game dictate that the process of run production resets each inning after 3 outs, forcing teams to take a strategic, objective-driven approach to scoring runs. A manager would never send a batter to the plate and say “get an extra base hit because, on average, that increases our scoring expectation by 1.7 runs.” Instead the manager might say “move the runner over to third” or “just put the ball in play here.” The team wants to score runs, and thus will look for smaller objectives that increase the likelihood of scoring.

I concluded that the existing measures Total Average (TA) and Runs Created (RC) most closely matched my empirical logic about run production. Total Average is the ratio of bases acquired to outs made, which implies that run production is nothing but a series of acquiring bases (the offensive objective) and making outs (the defensive objective). It assumes that the most productive batters are those who acquire many bases while making few outs. Runs Created splits run production into two objectives: reaching base safely, and advancing among the bases. Runs Created attempts to quantify those objectives, controlled for opportunities, with the resulting value being an estimate of runs scored. It assumes that the most productive batters are those who frequently reach base and effectively advance the scoring process. Both TA and RC have been criticized as being simplistic and outdated; to counter that, I have expanded both equations to encompass all statistical offensive outcomes as inputs.

\[
TA = \frac{\text{Bases}}{\text{Outs}} \quad (12)
\]

\[
RC = \frac{\text{On} - \text{Base Factor} \times \text{Advancement Factor}}{\text{Opportunity Factor}} \quad (13)
\]

Where Bases is the sum of all bases acquired through hitting and baserunning. Outs are taken directly from baseball-reference.com, On-Base Factor is equal to the number of times batters safely reached base less the number of outs made on the basepaths, Advancement Factor quantifies baserunner advancement by hits, sacrifices or baserunning, and Opportunity Factor is the sum of plate appearances and the number of baserunning plays.

\[
\alpha = \left(\sqrt{\frac{\text{Bases} \times \text{RC}}{\text{Outs}^2}} \times \text{Opportunity Factor}\right) + \text{Home Runs + hr} \quad (14)
\]
Where hit-and-run plays (hr) are instances in which a baserunner advanced a greater number of bases than Total Bases produced by the batter during the play; they were taken from baseball-reference.com, and are denoted 1s3/4, 2s4, and 1d4 there.

\[
CRE = \frac{\alpha}{\gamma}
\]  

(15)

The denominator of equation 15, \(\gamma\), represents the league-specific mean of all divisors required to produce actual runs as an output for each team in that league.

\[
\gamma = \mu(\frac{\alpha_1}{RS_1}, \frac{\alpha_2}{RS_2}, \frac{\alpha_3}{RS_3}, \ldots, \frac{\alpha_n}{RS_n})
\]  

(16)

For the years 1975-2014, I observed mean \(\gamma\) values of 3.7631 (MLB), 3.7292 (AL), and 3.7978 (NL), indicating that National League teams, on average, scored fewer runs than American League teams. Over that same period, standard deviations were just 0.0268 (MLB), 0.0377 (AL), and 0.0366 (NL), suggesting a somewhat constant relationship between \(\alpha\) and runs scored. To test the model, I ran regressions on RS for all 1,114 teams for the seasons 1975-2014, evaluating CRE against both wRC and Base Runs (BsR), two run estimation models used in existing WAR measures.
Although wRC showed a $\beta_1$ value that more closely estimated a perfect linear relationship, CRE held a clear advantage in both $R^2$ and error, indicating a consistent and reliable improvement in accuracy. This accuracy was maintained when altering the arrangement of inputs to reflect run allowance by pitchers and fielders, as the sums of Calculated Runs Expectancy Independent of Defense (CREID) and Defensive Runs Allowed (DRA) were found to have a mean error of 14.82 runs when regressed on team runs allowed using the same dataset.

Rather than use traditional pitching or fielding statistics, I took an unconventional approach by using opponent offensive statistics to estimate runs allowed. Baseball statistics abide by the fundamental accounting principle that every debit is matched by a credit, meaning that every offensive statistic debited to a batter has a corresponding statistic credited to either a pitcher or fielder. If $\alpha$ is calculated using the full set of offensive inputs, imagine $\rho$ represents that same calculation using only opposing offensive inputs credited to pitchers, while $\delta$ includes only outcomes attributable to fielders. CREID calculations use the same league-specific $\gamma$ as CRE calculations because no true defense-independent run value exists, although CREID more closely estimates earned runs than total runs allowed.

\[
CREID = \frac{\rho}{\gamma} \tag{17}
\]

Defensive Runs Allowed is the marginal effect on team CREID observed when including all inputs credited to a individual fielder. For example, assume a team’s pitchers combined to allow a CREID of 500. If, after including an individual fielder’s defensive statistics, that value becomes 505, that fielder is then credited with 5 Defensive Runs Allowed. DRA is not comparable across positions because $\delta$ varies by position.

\[
DRA = \sum \frac{\rho}{\gamma} - \sum \frac{\rho + \delta}{\gamma} \tag{18}
\]

2.2. Deriving zOffense and zPitching

CRE, CREID, and DRA measure raw run production or run allowance without regard to cost or efficiency. Two pitchers can allow an equal number of runs, but it’s impossible to determine who was more valuable without context. As noted earlier in the paper, zWins is the marginal productivity relative to a position-specific replacement-level expectation. I
chose outs made per CRE produced as the units of expected production for batters, and outs made per CREID allowed for pitchers, both denoted $\epsilon$.

$$\epsilon_{\text{batters}} = \frac{\sum \text{Outs}_i}{\sum \text{CRE}_i} \quad (19)$$

$$\epsilon_{\text{pitchers}} = \frac{\sum \text{Outs}_i}{\sum \text{CREID}_i} \quad (20)$$

Replacement-level production expectations were observed by classifying players as "starters" or "replacements" based upon the amount of playing time received. Each position thus has three expectations, $\epsilon_{\text{avg}}$, $\epsilon_{\text{starters}}$, and $\epsilon_{\text{reps}}$, denoting the mean production rates among all players at that position, as well as among those classified as either starters or replacements. The criteria for being labeled a "starter" were a minimum of 300 plate appearances for all non-pitchers, 100 innings pitched for all starting pitchers, 60 innings pitched for all right-handed relief pitchers, and 40 innings pitched for all left-handed relief pitchers. I defined “starting pitcher” as any pitcher who started in at least half of their appearances, or any pitcher who started at least 15 games. As expected, observed production expectations were greater for replacement batters and smaller for replacement pitchers than the rates of starters, who, on average, are more skilled players.

The ratio between starter and replacement expectations (denoted $\theta\epsilon$) estimates the percent of production maintained when replacing a starter. A ratio of 1 would suggest no loss of production when making a replacement, while a ratio of 0.5 would suggest that an average replacement would only be expected to produce half (as a batter), or allow double (as a pitcher), the output of an average starter at the same position. Expectations change every year based on the performance of starters and replacements, and over time ratios can reveal scarcity or depth of positions and can provide valuable information regarding contracts and personnel decisions.

$$\theta\epsilon_{\text{batters}} = \frac{\epsilon_{\text{starter}}}{\epsilon_{\text{reps}}} \quad (21)$$

$$\theta\epsilon_{\text{pitchers}} = \frac{\epsilon_{\text{reps}}}{\epsilon_{\text{starter}}} \quad (22)$$

The counterfactual replacement-level production expectations used in zWins are a function of $\epsilon_{\text{avg}}$ and $\theta\epsilon$, specific to league and position, and are denoted Replacement $\epsilon$. 
\[ \text{Replacement } \varepsilon_{\text{batters}} = \frac{\varepsilon_{\text{avg}}}{\theta \varepsilon_{\text{batters}}} \]  

\[ \text{Replacement } \varepsilon_{\text{pitchers}} = \varepsilon_{\text{avg}} \cdot \theta \varepsilon_{\text{pitchers}} \]  

With the assumption that Replacement \( \varepsilon \) is the true replacement-level, \( z\text{Offense} \) and \( z\text{Pitching} \) estimate the difference, in runs, between each player’s demonstrated production and the expected production of a replacement at the same position given the same number of outs.

\[ z\text{Offense} = \text{CRE} - \frac{\text{Outs}}{\text{Replacement } \varepsilon_{\text{batters}}} \]  

\[ z\text{Pitching} = \frac{\text{Outs}}{\text{Replacement } \varepsilon_{\text{pitchers}}} - \text{CRE1D} \]  

3. Deriving \( z\text{Defense} \)

It is important to recognize that defense remains the most abstract and difficult facet of baseball to quantify. No matter how advanced the data, any defensive calculations should be met with healthy skepticism and an understanding that consensus is uncommon. Despite that, each of the unique defensive metrics used in \( \text{rWAR} \), \( \text{WARP} \), and \( \text{fWAR} \) have been generally accepted as valid, albeit imperfect, measures of defensive performance. I ran a regression matrix of each combination of \( z\text{Defense} \) (used in \( z\text{Wins} \)), Defensive Runs Saved (\( \text{DRS} \), used in \( \text{rWAR} \)), Fielding Runs Above Average (\( \text{FRAA} \), used in \( \text{WARP} \)), and Ultimate Zone Rating (\( \text{UZR} \), used in \( \text{fWAR} \)), observing that \( z\text{Defense} \) showed the highest harmonic mean of \( R^2 \) values. While this does not prove that \( z\text{Defense} \) is a superior measure of defense, it does indicate that \( z\text{Defense} \) produces valuations of defensive performance that generally align with those produced by existing metrics. As such, \( z\text{Defense} \), while imperfect, appears to be a valid estimate of defensive performance.

**Figure 5: Defensive Metric Regression Matrix**

<table>
<thead>
<tr>
<th></th>
<th>( z\text{Defense} )</th>
<th>DRS</th>
<th>FRAA</th>
<th>UZR</th>
</tr>
</thead>
<tbody>
<tr>
<td>( z\text{Defense} )</td>
<td>0.4882</td>
<td>0.5098</td>
<td>0.4010</td>
<td>0.5916</td>
</tr>
<tr>
<td>DRS</td>
<td>0.5098</td>
<td>0.4248</td>
<td>0.2741</td>
<td>0.6888</td>
</tr>
<tr>
<td>FRAA</td>
<td>0.4010</td>
<td>0.2741</td>
<td>0.2884</td>
<td>0.2348</td>
</tr>
<tr>
<td>UZR</td>
<td>0.5916</td>
<td>0.6888</td>
<td>0.2348</td>
<td>0.4053</td>
</tr>
</tbody>
</table>

*Values in italics represent the harmonic mean of each metric’s three \( R^2 \) values.*
zDefense is a sum of three subcomponents which attempt to quantify different aspects of fielding ability. Each subcomponent is measured relative to positional averages, rather than to derived replacement-level expectations.

\[ z_{Defense} = z_{DRA} + z_{Range} + z_{Outs} \]  

(27)

Where \( z_{DRA} \) is the relative number of Defensive Runs Allowed, \( z_{Range} \) estimates the number of runs saved by using positional zone data to quantify range, and \( z_{Outs} \) estimates runs saved by a player’s ability to produce outs in the field. All three subcomponents are position-specific and can be either positive or negative.

3.1. Deriving \( z_{DRA} \)

\( z_{DRA} \) measures the difference between observed DRA and expected DRA, derived from positional averages. Pitchers and catchers (denoted “battery”) use innings per DRA as units of production, while all other positions (“non-battery”) use DRA per out produced.

\[ \epsilon_{battery} = \frac{\sum Innings_i}{\sum DRA_i} \]  

(28)

\[ \epsilon_{non-battery} = \frac{\sum DRA_i}{\sum Outs_i} \]  

(39)

\[ z_{DRA_{battery}} = \frac{Innings}{\epsilon_{battery}} - DRA \]  

(30)

\[ z_{DRA_{non-battery}} = \frac{\epsilon_{non-battery}}{Outs} - DRA \]  

(31)

3.2. Deriving \( z_{Range} \)

\( z_{Range} \), which attempts to quantify the number of runs saved attributable to range, uses John Dewan’s Revised Zone Rating (RZR) as a range input. RZR divides a baseball field into positional “zones” such that each zone represents the area of the field in which an average fielder at that position is able to convert at least half of his chances into outs. RZR refers to the percentage of Balls in Zone (BIZ) that a fielder successfully converted into outs (“Plays” refers to outs made only on BIZ). All outs made outside of the positional zone (OOZ) are counted as well. Because RZR data is not collected for pitchers or catchers, those positions have no \( z_{Range} \) calculation.

\( z_{Range} \) makes several assumptions, the first being that players who exhibit relatively high RZR and frequently make plays outside of their positional zones possess exceptional range.
This assumption can be problematic, though, because zones remain static regardless of where a fielder is positioned at the start of a play, meaning defensive shifts may cause biased estimates. zRange also assumes that every ball hit to the same positional zone is of equal run value. Without context-specific data available, I cannot overcome this assumption.

\[ zRange = \varphi_{outs} \ast \mu_{outval} \]  

(32)

zRange estimates runs saved or lost by multiplying the number of outs made relative to the positional expectation, \( \varphi_{outs} \), by the average run value of an out hit to each positional zone, \( \mu_{outval} \).

\[ \varphi_{outs} = \left( \text{Plays} - \frac{\sum \text{Plays}_i \ast BIZ_i}{\sum BIZ_i} \right) + \left( \text{OOZ} - \frac{\sum \text{OOZ}_i \ast Innings_i}{\sum Innings_i} \right) \]  

(33)

To derive average positional out values, I first assumed that CRE per offensive out was the true average value (\( v_{avg} \)) of all outs because CRE had unrestricted inputs.

\[ v_{avg} = \frac{\sum \text{CRE}}{\sum \text{Outs}} \]  

(34)

I also assumed that positional DRA per out (\( \omega_i \)) were restricted estimates of out values specific to each position. By finding the weighted (by number of players at each position, \( n_i \)) average of all \( \omega_i \) estimates, I derived a restricted estimate of the average value of all outs (\( v_{res} \)).

\[ v_{res} = \frac{\sum (\omega_i n_i)}{\sum n_i} \]  

(35)

I then assumed that the ratio between position-specific out values and the restricted estimate, denoted \( \theta v_i \), was indicative of the run value of outs hit to each position, relative to all other positions. When multiplied by the unrestricted out value estimate, the result is the position-specific run value of an out hit to any zone on the field.

\[ \theta v_i = \frac{\omega_i}{v_{res}} \]  

(36)

\[ \mu_{outval} = v_{avg} \ast \theta v_i \]  

(37)
3.3. Deriving $zOuts$
While $zDRA$ and $zRange$ have clear units, it was not intuitively obvious that $zOuts$ estimated runs saved. I theorized that $z$-scores can estimate production, in units of runs/points above average, when weighted by measures of volume (i.e. playing time, opportunities, attempts, etc.). Supporting evidence for this is provided in Section 3.4 of this paper.

$$zOuts = z_{OP} \times \gamma \times \log BIZ \times Pct. Innings \times \text{positional adjustment}$$  \hspace{1cm} (38)

Where $z_{OP}$ is the $z$-score of a derived statistic known as Out Product (OP), an altered calculation of RZR that includes double plays and the percentage of balls fielded successfully converted into outs (F2O%) to estimate efficiency at producing outs. Using $z$-scores is beneficial in this type of analysis because $z$-scores indicate how values differ from the underlying distribution, in units of standard deviations from the population mean. Assuming OP has a normal distribution, $z_{OP}$ will provide a standardized range of continuous values indicative of each value's relationship with the league average.

$$z_{OP} = \frac{OP - \mu_{OP_i}}{\sigma_i \text{ (min.100 innings)}}$$  \hspace{1cm} (39)

Out Product assumes that double plays occur at constant rates regardless of outside factors. This assumption can cause biased estimates because double play opportunities can vary greatly by pitcher, batter, and situation. Ideally I would include some measure of effectiveness at producing double plays controlled for the number of chances, but without context-specific data available I cannot overcome this assumption. I took F2O% data directly from baseball-reference.com.

$$OP = \frac{(Plays \times F2O\%) + \text{Double Plays}}{BIZ}$$  \hspace{1cm} (40)

The positional adjustment used in $zOuts$ attempts to control for the fact that balls are hit to different parts of the field at varying rates. Similar to $\theta v$ used in $zRange$, these adjustments estimate the relative frequency with which each position fields balls in play. I grouped by each distinct position as well as by “infield” and “outfield” categories, using BIZ as my measure of fielding volume.

$$\text{positional adjustment} = \frac{(\sum BIZ_i)^2}{\mu(\sum BIZ_{IF}) \times \mu(\sum BIZ_{OF})}$$  \hspace{1cm} (41)
3.4. Using Z-Scores in Football

The assumption necessary to derive zOuts was that z-scores describe production in sports, in units of runs/points above average, when multiplied by measures of volume. To test this assumption, I developed a model that estimates mean margin of victory ($\mu_{MOV}$) using football statistics. I collected data from all 1,802 Division I FBS (formerly I-A) college football teams for the seasons 2000-2014.

$$\mu_{MOV} = MOV_{OFF} + MOV_{DEF}$$

(42)

I evaluated offensive performance ($MOV_{OFF}$) using team offensive statistics and defensive performance ($MOV_{DEF}$) using opponent offensive statistics, both calculated identically. In particular, I used the z-scores of completion percentage ($z_{CMP}$), yards per pass attempt ($z_{Yds}$), yards per rush attempt ($z_{RYds}$), turnovers ($z_{TO}$), offensive plays ($z_{Plays}$), and penalty yards ($z_{penalty}$). Means and standard deviations used included all teams active in each particular season of play. I collected data using per game statistics taken from sports-reference.com. Measures of volume included completed passes ($Cmp$), attempted passes ($Passes$), and rushing attempts ($Rushes$).

$$MOV_i = \frac{z_{CMP} \cdot Cmp + z_{Yds} \cdot Passes + z_{RYds} \cdot Rushes}{10} + z_{TO} + z_{Plays} + z_{penalty}$$

(43)

Figure 6: $\mu_{MOV}$ vs. Actual Point Mean Differential, NCAA Division I FBS, 2000-2014

$$y = 0.8781x + 1.027$$

$R^2 = 0.90206$
Applying these results back to baseball, this regression presents strong evidence that zOuts estimates runs saved defensively, provided that certain assumptions are true. For this to be the case, it must be true that Out Product estimates the ability to produce outs defensively, and that \( z_{OP} \) estimates the ability to do so relative to positional averages. If this is the case, Out Product must be negatively correlated with runs allowed because run production does not cease until 3 outs are recorded each inning. Finally, it must be the case that the measures of volume used in calculating zOuts are valid indicators of playing time and defensive chances. I believe all of these to be true assumptions based on the regression matrix earlier in this paper, which indicated that zDefense estimates compared favorably to those produced by existing measures of defensive performance.

4. Conclusion and Discussion

The concept of Wins Above Replacement as an estimate of the marginal productivity of labor has strong support in economic theory. Furthermore, the fact that WAR has been so widely accepted by baseball analysts and media shows the support WAR receives from within the industry. However, existing measures of WAR have major shortcomings in methodology, transparency, and reproducibility. In this paper, I have presented an alternative calculation of WAR that not only outperforms existing measures, but also discloses all calculations, and is reproducible without the use of proprietary data.

There remain several limitations of zWins that provide additional research opportunities. First, it may be beneficial to derive estimate intervals of zWins, rather than to report single values. Despite showing improved accuracy over existing measures, there remains error and uncertainty in the subcomponents, and thus it is unlikely that zWins estimates are universally accurate.

As derived in the paper, zDefense makes numerous assumptions that affect the accuracy of estimates. The use of RZR as a measure of defensive range presents problems because RZR does not control for defensive positioning, complicating the definition of zRange as a measure. Additionally, zOuts is derived largely upon theoretical concepts, and thus warrants further examination as a measure of fielding.

Currently zWins does not incorporate single-game win expectations or context-specific data, instead using only season totals for all calculations. As a result, there is no differentiation in the value of production dependent upon situation. Runs saved, allowed, or produced vary in their effect on team success on the basis of score, inning, and league standing. It would be very interesting to develop a framework that controls for situational production, in terms of both single-game win expectations and league standings.

I suspect that zWins estimates for most players would be relatively unchanged, as contextual performance likely would not significantly vary from total seasonal production across robust data. However, there may be significant effects observed in high-leverage
role players such as pinch hitters, set-up relievers, or closers, who are used disproportionately during critical game situations. When using league standings as a basis for differentiation there may be significant bias in favor of players on successful teams, as last-place teams are often eliminated from playoff contention with many games remaining.

zWins is uniquely innovative in that it implements a uniform approach to offensive, pitching, and defensive analysis while rejecting generalizations about the concept of replacement-level production. In addition, zWins is entirely replicable using publicly available datasets and the processes outlined in this paper. I demonstrated in this paper that the subcomponents used in zWins – namely CRE, CREID, and zDefense – are comparable or superior in accuracy than those used by existing measures. Additionally, my methodology in calculating both runs and zWins consistently produced more explanatory models when regressed on actual team wins than existing measures. As such, zWins should be considered a viable alternative to fWAR, rWAR, and WARP, perhaps even a replacement.
Appendix

CRE was derived using expanded versions of both Total Average and Runs Created, where I defined each measure's inputs as such:

1. Bases = Total Bases + Walks + Hit by Pitch + Reach on Error + Catcher Interference + Sacrifice Hits + Sacrifice Flies + Stolen Bases + Bases Taken

2. On-Base Factor = Hits + Walks + Hit by Pitch + Reach on Error + Catcher Interference - Caught Stealing - Double Plays Grounded Into - Outs on Base

3. Advancement Factor = Total Bases + Stolen Bases + Sacrifice Hits + Sacrifice Flies + Bases Taken

4. Opportunity Factor = Plate Appearances + Stolen Bases + Caught Stealing + Bases Taken + Outs on Base

In deriving CREID and DRA, I categorized each offensive outcome as attributable to the pitcher (ρ) or to the defense (δ), as defined in the table below.

<table>
<thead>
<tr>
<th>Attributable to Pitcher (ρ)</th>
<th>Attributable to Defense (δ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Bases (all Hits)</td>
<td>Reach on Error</td>
</tr>
<tr>
<td>Walks</td>
<td>Catcher Interference</td>
</tr>
<tr>
<td>Hit by Pitch</td>
<td>Stolen Bases*</td>
</tr>
<tr>
<td>Sacrifice Hits</td>
<td>Caught Stealing*</td>
</tr>
<tr>
<td>Sacrifice Flies</td>
<td>Bases Taken</td>
</tr>
<tr>
<td>Double Plays Induced</td>
<td>Outs on Base</td>
</tr>
<tr>
<td>Plate Appearances</td>
<td>Hit-and-Run Plays</td>
</tr>
</tbody>
</table>

*Stolen Bases and Caught Stealing are credited to both catcher and pitcher DRA