Ronald Yurko*, Samuel Ventura and Maksim Horowitz

nflWAR: a reproducible method for offensive player evaluation in football

https://doi.org/10.1515/jqas-2018-0010

Abstract: Existing methods for player evaluation in American football rely heavily on proprietary data, are often not reproducible, lag behind those of other major sports, and are not interpretable in terms of game outcomes. We present four contributions to the study of football statistics to address these issues. First, we develop the R package nflscrapR to provide easy access to publicly available play-by-play data from the National Football League (NFL). Second, we introduce a novel multinomial logistic regression approach for estimating the expected points for each play. Third, we use the expected points as input into a generalized additive model for estimating the win probability for each play. Fourth, we introduce our nflWAR framework, using multilevel models to isolate the contributions of individual offensive skill players in terms of their wins above replacement (WAR). We assess the uncertainty in WAR through a resampling approach specifically designed for football, and we present results for the 2017 NFL season. We discuss how our reproducible WAR framework can be extended to estimate WAR for players at any position if researchers have data specifying the players on the field during each play. Finally, we discuss the potential implications of this work for NFL teams.

Keywords: generalized additive models; multilevel models; multinomial logistic regression; R; reproducibility.

1 Introduction

Despite the sport's popularity in the United States, public statistical analysis of American football ("football") has lagged behind that of other major sports. While new statistical research involving player and team evaluation is regularly published in baseball (Albert 2006; Jensen, Shirley, and Wyner 2009; Piette and Jensen 2012; Baumer, Jensen, and Matthews 2015), basketball (Kubatko et al. 2007; Deshpande and Jensen 2016), and hockey (Macdonald 2011; Gramacy, Taddy, and Jensen 2013; Thomas et al. 2013), there is limited new research that addresses on-field or player personnel decisions for National Football League (NFL) teams. Recent work in football addresses topics such as fantasy football (Becker and Sun 2016), predicting game outcomes (Balreira, Miceli, and Tegtmeyer 2014), NFL TV ratings (Grimshaw and Burwell 2014), the effect of "fan passion" and league sponsorship on brand recognition (Wakefield and Rivers 2012), and realignment in college football (Jensen and Turner 2014). Additionally, with the notable exception of Lock and Nettleton (2014), recent research relating to on-field or player personnel decisions in football is narrowly focused. For example, Mulholland and Jensen (2014) analyze the success of tight ends in the NFL draft, Clark, Johnson, and Stimpson (2013) and Pasteur and Cunningham-Rhoads (2014) both provide improved metrics for kicker evaluation, Martin, Timmons, and Powell (2017) examine the NFL's change in overtime rules, and Snyder and Lopez (2015) focus on discretionary penalties from referees. Moreover, statistical analysis of football that does tackle on-field or player personnel decisions frequently relies on proprietary and costly data sources, where data quality often depends on potentially biased and publicly unverified human judgment. This leads to a lack of reproducibility that is well-documented in sports research (Baumer et al. 2015).

In this paper, we posit that (1) objective on-field and player personnel decisions rely on two fundamental categories of statistical analysis in football: play evaluation and player evaluation, and (2) in order to maintain a standard of objectivity and reproducibility for these two fundamental areas of analysis, researchers must agree on a dataset standard.

1.1 Previous work: evaluating plays

Traditionally, yards gained/lost have been used to evaluate the success of a play. However, this point of view strips away the importance of context in football (Carter and Machol 1971; Carroll et al. 1988). For instance, three yards gained on 3rd and 2 are more valuable than three

***Corresponding author: Ronald Yurko,** Carnegie Mellon University, Statistics and Data Science, Pittsburgh, PA 15213, USA, e-mail: ryurko@andrew.cmu.edu
Samuel Ventura: Carnegie Mellon University, Statistics and Data Science, Pittsburgh, PA 15213, USA; and Pittsburgh Penguins, Pittsburgh, PA 15213, USA, e-mail: sventura@stat.cmu.edu
Maksim Horowitz: Carnegie Mellon University, Statistics and Data Science, Pittsburgh, PA 15213, USA, e-mail: bklynmaks@gmail.com

yards gained on 3rd and 7. This point, that not all yards are created equal, has been the foundation for the development of two approaches for evaluating plays: expected points and win probability. The expected points framework uses historical data to find the number of points eventually scored by teams in similar situations, while the win probability framework uses historical data to find how often teams in similar situations win the game. One can obtain the difference between pre-snap and post-snap values of a play to estimate the value of the play itself – expected points added (*EPA*) or win probability added (*WPA*).

These approaches have been recently popularized by Brian Burke's work at www.advancedfootballanalytics.com and ESPN (Burke 2009; Katz and Burke 2017). Burke (2009) and ESPN provide an intuitive explanation for what expected points means, however they do not go into the details of the calculations. "Nearest neighbors" algorithms have been used to identify similar situations based on down, yards to go, and the yard line to then average over the next points scored (Carter and Machol 1971; Dasarathy 1991). Goldner (2017) describes a Markov model and uses the absorption probabilities for different scoring events (touchdown, field goal, and safety) to arrive at the expected points for a play. Causey (2015) takes an exact-neighbors approach, finding all plays with a set of identical characteristics, taking the average outcome, and conducting post-hoc smoothing to calculate expected points.

Compared to expected points models, there is considerably more literature on different methodologies for estimating the win probability of a play in football. Goldner (2017) uses a Markov model, similar to the approach taken by Tango, Lichtman, and Dolphin (2007) in baseball, by including the score differential, time remaining, and timeouts to extend the expected points model. Burke's approach is primarily empirical estimation by binning plays with adjustments and smoothing. In some published win probability analyses, random forests have been shown to generate well-calibrated win probability estimates (Causey 2013; Lock and Nettleton 2014). The approach taken by Lock and Nettleton (2014) also considers the respective strengths of the offensive (possession) and defensive (non-possession) teams.

There are many areas of research that build off of these approaches for valuing plays. For example, analyses of fourth down attempts and play-calling are very popular (Romer 2006; Alamar 2010; Goldner 2012; Quealy, Causey, and Burke 2017). This paper focuses on using play evaluation to subsequently evaluate players, and we discuss prior attempts at player evaluation below.

1.2 Previous work: evaluating players

Most analysis for comparing players at offensive skill positions such as quarterback (QB), running back (RB), wide receiver (WR), and tight end (TE) relies on the usage of basic box score statistics such as yards gained, number of attempts, etc. Linear combinations of these box score statistics, such as passer rating (Smith, Siwoff, and Weiss 1973), are often used to compare players at the same position while taking into account more than just a single box score measure. Similarly, Pro Football Reference's adjusted net yards per attempt ("ANY/A") expands upon passer rating in that it accounts for sacks and uses a different linear weighting scheme (Pro-Football-Reference 2018). However these measures are not based on *EPA* or *WPA*, so they are not interpretable in terms of points or wins.

There have been multiple irreproducible attempts at player evaluation with frameworks considering *EPA*-like values. Schatz (2003) provides a metric called "defense-adjusted value over average" which evaluates plays in manner similar to expected points while also accounting for the strength of the opposing defense. However, specifics on the modeling techniques are not disclosed. ESPN's total quarterback rating ("QBR") accounts for the situational contexts a QB faces throughout a game (Katz and Burke 2017; Oliver 2011). ESPN uses the following approach when computing QBR. First, they determine the degree of success or failure for each play. Second, they divide credit for each play amongst all players involved. Third, additional adjustments are made for plays of very little consequence to the game outcome. This approach has several important advantages. In the first step, the EPA is used to assign an objective value to each play. Another advantage is that some attempt is made to divide credit for a play's success or failure amongst the players involved. We loosely follow these steps in our proposed approach for NFL player evaluation. However ESPN's QBR is not directly reproducible since it relies on human judgment and is limited only to QBs.

For positions other than QB, RB, WR, and TE, data are limited, since the NFL does not publicly provide information about which players are on the field for a particular play, the offensive and defensive formations (other than the "shotgun" formation on offense), or the pre- and post-snap locations of players on the field. With these positions, it is difficult to obtain adequate within-positional comparisons of player value, let alone across-position comparisons. Pro Football Focus assigns grades to every player in each play, but this approach is solely based on human judgment and proprietary to PFF (Eager, Chahrouri, and

Palazzolo 2017). The only public approach for evaluating players at all positions according to common scale is Pro Football Reference's "approximate value" (AV) statistic (Drinen 2013). Using a combination of objective and subjective analysis, AV attempts to assign a single numerical value to a player's performance in any season since 1950, regardless of the player's position. AV has some subjective components, such as whether or not a lineman was named to the NFL's "all-pro" team, and is not interpretable in terms of game outcomes. There are no publicly known attempts for developing a *Wins Above Replacement* (*WAR*) measure for every individual NFL player, as made popular in baseball (Schoenfield 2012) and other sports (Thomas and Ventura 2015).

1.3 Our framework for evaluating NFL plays and players

In order to properly evaluate players, we need to allocate a portion of a play's value to each player involved (Katz and Burke 2017). Baumer and Badian-Pessot (2017) details the history of division of credit modeling as a primary driver of research in sports analytics, with origins in evaluating run contributions in baseball. However, in comparison to baseball, every football play is more complex and interdependent, with the 22 players on the field contributing in many ways and to varying degrees. A running play depends not only on the running back but the blocking by the linemen, the quarterback's handoff, the defensive matchup, the play call, etc. A natural approach is to use a regression-based method, with indicators for each player on the field for a play, providing an estimate of their marginal effect. This type of modeling has become common in basketball and hockey, because it accounts for factors such as quality of teammates and competition (Rosenbaum 2004; Kubatko et al. 2007; Macdonald 2011; Gramacy et al. 2013; Thomas et al. 2013).

We present four contributions to the study of football statistics in order to address the issues pertaining to play evaluation and player evaluation outlined above:

1. The `R` package `nflscrapR` to provide easy access to publicly available NFL play-by-play data (Section 1.4).
2. A novel approach for estimating expected points using a multinomial logistic regression model, which more appropriately models the "next score" response variable (Section 2.1).
3. A generalized additive model for estimating the win probability using the expected points as input (Section 2.2).
4. Our *nflWAR* framework, using multilevel models to isolate offensive skill player contribution and estimate their *WAR* (Section 3).

We use a resampling procedure similar to Baumer et al. (2015) to estimate uncertainty in each player's seasonal *WAR*. Due to the limitations of publicly available data, the primary focus of this paper is on offensive skill position players: QB, RB, WR, and TE. However, we present a novel metric that serves as a proxy for measuring a team's offensive line performance on rushing plays. Furthermore, the reproducible framework we introduce in this paper can also be easily extended to estimate *WAR* for all positions given the appropriate data. Researchers with data detailing which players are on the field for every play can use the framework provided in Section 5.2 to estimate *WAR* for players at all positions.

Our *WAR* framework has several key advantages. First, it is fully reproducible: it is built using only public data, with all code provided and all data accessible to the public. Second, our expected points and win probability models are well-calibrated. Third, player evaluation with *WAR* is easily interpretable in terms of game outcomes, unlike prior approaches to player evaluation in the NFL discussed above. The replacement level baseline informs us how many wins a player adds over a readily available player. This is more desirable than comparing to average from the viewpoint of an NFL front office, as league average performance is still valuable in context (Baumer et al. 2015). Fourth, the multilevel model framework accounts for quality of teammates and competition. Fifth, although this paper presents *WAR* using our expected points and win probability models for play evaluation, researchers can freely substitute their own approaches for play evaluation without any changes to the framework for estimating player *WAR*. Finally, we recognize the limitations of point estimates for player evaluation and provide estimates of the uncertainty in a player's *WAR*.

1.4 Play-by-play data with `nflscrapR`

Facing the challenge of limited publicly available NFL data and the need for a dataset standard, we develop an `R` package (R Core Team 2017) called `nflscrapR` (Horowitz, Yurko, and Ventura 2017). This package was inspired largely by other `R` packages facilitating the access of sports data such as `nhlscrapR` (Thomas and Ventura 2017), `pitchRx` (Sievert 2015), `Lahman` (Lahman 1996 – 2017), `openWAR` (Baumer et al. 2015), and `ballR` (Elmore and DeWitt 2017). The results in this paper

are generated entirely from play-by-play data accessed with the `nflscrapR` package. A detailed description of its functionality is provided in our supplementary materials.

2 Evaluating plays with expected points and win probability

As described in Section 1.1, expected points and win probability are two common approaches for evaluating plays. These approaches have several key advantages: They can be calculated using only the NFL's publicly available data, they provide estimates of a play's value in terms of real game outcomes (i.e. points and wins), and, as a result, they are easy to understand for both experts and non-experts.

Below, we introduce our own novel approaches for estimating expected points (*EP*) and win probability (*WP*) using publicly available data via `nflscrapR`.

2.1 Expected points

While most authors take the average "next score" outcome of similar plays in order to arrive at an estimate of *EP*, we recognize that certain scoring events become more or less likely in different situations. As such, we propose modeling the probability for each of the scoring events directly, as this more appropriately accounts for the differing relationships between the different categories of the "next score" response and the covariates. Once we have the probabilities of each scoring event, we can trivially estimate expected points.

2.1.1 Multinomial logistic regression

To estimate the probabilities of each possible scoring event conditional on the current game situation, we use multinomial logistic regression. For each play, we find the next scoring event within the same half (with respect to the possession team) as one of the seven possible events: touchdown (7 points), field goal (3 points), safety (2 points), no score (0 points), opponent safety (−2 points), opponent field goal (−3 points), and opponent touchdown (−7 points). Here, we ignore point after touchdown (PAT) attempts, and we treat PATs separately in Section 2.3.

Figure 1 displays the distribution of the different type of scoring events using data from NFL regular season games between 2009 and 2016, with each event located on the y-axis based on their associated point value y. This data consists of 304,896 non-PAT plays, excluding QB kneels (which are solely used to run out the clock and are thus assigned an *EP* value of zero). The gaps along the y-axis between the different scoring events reinforce our decision to treat this as a classification problem rather than modeling the point values with linear regression. While we use seven points for a touchdown for simplicity here, our multinomial logistic regression model generates the probabilities for the events regardless of the point value. This is beneficial, since it allows us to flexibly handle PATs and two-point attempts separately. We can easily adjust the point values associated with touchdowns to reflect changes in the league's scoring environment.

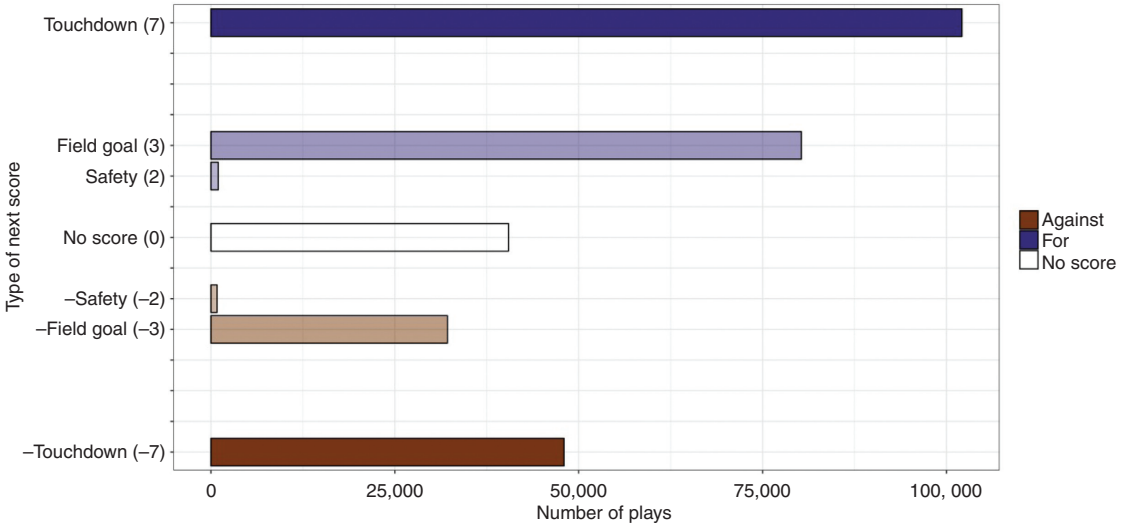

Figure 1: Distribution of next scoring events for all plays from 2009 to 2016, with respect to the possession team.

We denote the covariates describing the game situation for each play as $\mathbf{X}$, which are presented in our supplementary materials, and the response variable

$$\begin{aligned} Y \in \{&\text{Touchdown (7)}, \text{ Field Goal (3)}, \text{ Safety (2)}, \\ &\text{No Score (0)}, -\text{Touchdown (-7)}, -\text{Field Goal (-3)}, \\ &-\text{ Safety (-2)}\}. \end{aligned} \tag{1}$$

The model is specified with six logit transformations relative to the "No Score" event with the form

$$\begin{aligned} \log\left(\frac{P(Y = Touchdown|\mathbf{X})}{P(Y = No\ Score|\mathbf{X})}\right) &= \mathbf{X} \cdot \boldsymbol{\beta}_{Touchdown}, \\ \log\left(\frac{P(Y = Field\ Goal|\mathbf{X})}{P(Y = No\ Score|\mathbf{X})}\right) &= \mathbf{X} \cdot \boldsymbol{\beta}_{Field\ Goal}, \\ &\vdots \\ \log\left(\frac{P(Y = -Touchdown|\mathbf{X})}{P(Y = No\ Score|\mathbf{X})}\right) &= \mathbf{X} \cdot \boldsymbol{\beta}_{-Touchdown}, \end{aligned} \tag{2}$$

where $\boldsymbol{\beta}_y$ is the corresponding coefficient vector for the type of next scoring event. Using the generated probabilities for each of the possible scoring events, $P(Y = y|\mathbf{X})$, we simply calculate the expected points (*EP*) for a play by multiplying each event's predicted probability with its associated point value y:

$$EP = E[Y|\mathbf{X}] = \sum_y y \cdot P(Y = y|\mathbf{X}). \tag{3}$$

2.1.2 Model selection with Calibration

Since our expected points model uses the probabilities for each scoring event from multinomial logistic regression, the variables and interactions selected for the model are determined via calibration testing, similar to the criteria for evaluating the win probability model in Lock and Nettleton (2014). The estimated probability for each of the seven scoring events is binned in five percent increments (20 total possible bins), with the observed proportion of the event found in each bin. If the actual proportion of the event is similar to the bin's estimated probability then the model is well-calibrated. Because we are generating probabilities for seven events, we want a model that is well-calibrated across all seven events. To objectively compare different models, we first calculate for scoring event y in bin $b \in \{1, \ldots, B\}$ its associated error

$$e_{y,b} = |\hat{P}_b(Y = y) - P_b(Y = y)|, \tag{4}$$

where $\hat{P}_b(Y = y)$ and $P_b(Y = y)$ are the predicted and observed probabilities, respectively, in bin b. Then, the overall calibration error e_y for scoring event y is found by averaging $e_{y,b}$ over all bins, weighted by the number of plays in each bin, $n_{y,b}$; that is

$$e_y = \frac{1}{n_y} \sum_b n_{y,b} e_{y,b}, \tag{5}$$

where $n_y = \sum_b n_{y,b}$. This leads to the model's calibration error e as the average of the seven e_y values, weighted by the number of plays with scoring event y, n_y:

$$e = \frac{1}{n} \sum_y n_y e_y, \tag{6}$$

where $n = \sum_y n_y$, the number of total plays. This provides us with a single statistic with which to evaluate models, in addition to the calibration charts.

We calculate the model calibration error using leave-one-season-out cross-validation (LOSO CV) to reflect how the `nflscrapR` package will generate the probabilities for plays in a season it has not yet observed. The selected variables and interactions for the model yielding the best LOSO CV calibration results are available in our supplementary materials. Figure 2 displays the selected model's LOSO CV calibration results for each of the seven scoring events, resulting in $e \approx 0.013$. The dashed lines along the diagonal represent a perfect fit, i.e. the closer to the diagonal points are the more calibrated the model. Although time remaining is typically reserved for win probability models (Goldner 2017), including the seconds remaining in the half, as well as the indicator for under two minutes, improved the model's calibration, particularly with regards to the "No Score" event. We also explored the use of an ordinal logistic regression model which assumes equivalent effects as the scoring value increases, but found the LOSO CV calibration results to be noticeably worse with $e \approx 0.022$.

Our supplementary materials provides further details on our novel weighting scheme to handle potential issues regarding score differential (Burke 2014) and the difference in drives from the next score without excluding any observations. These materials also cover our use of a separate model for field goals and PAT attempts.

2.1.3 Expected points by down and yard line

For reference, Figure 3 displays the relationship between the field position and the *EP* for our multinomial logistic regression model available via `nflscrapR` compared to the previous relationships found by Carter and Machol (1971) and Carroll et al. (1988). We separate the `nflscrapR`

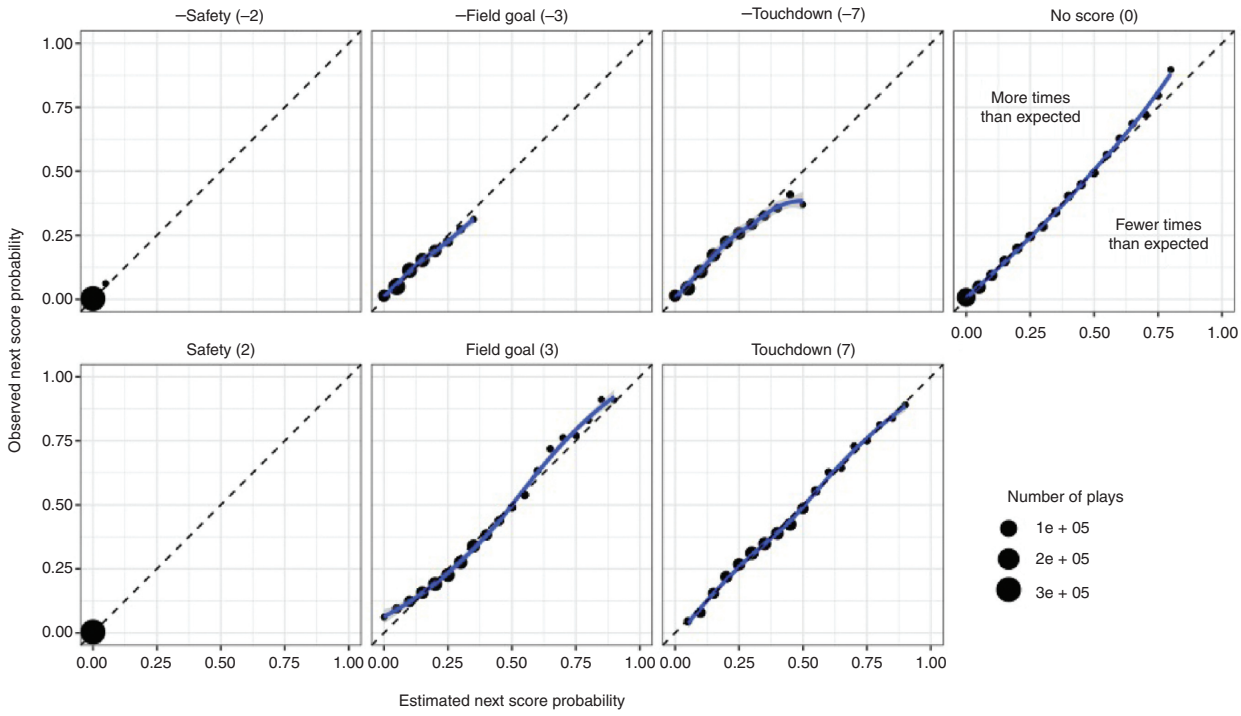

Figure 2: Expected points model LOSO CV calibration results by scoring event.

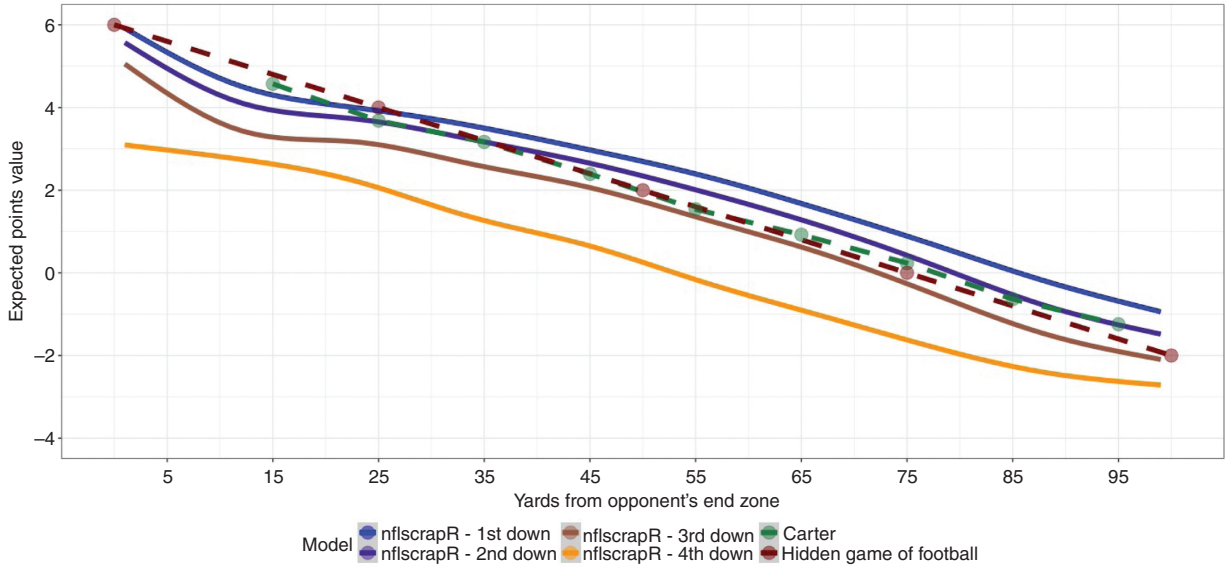

Figure 3: Comparison of historical models and `nflscrapR` expected points value, based on distance from opponent's end zone by down.

model by down to show its importance, and in particular the noticeable drop for fourth down plays and how they exhibit a different relationship near the opponent's end zone as compared to other downs. To provide context for what is driving the difference, Figure 4 displays the relationship between each of the next score probabilities and field position by down. Clearly on fourth down, the probability of a field goal attempt overwhelms the other possible events once within 50 yards of the opponent's end zone.

2.2 Win probability

Because our primary focus in this paper is in player evaluation, we model win probability without taking into

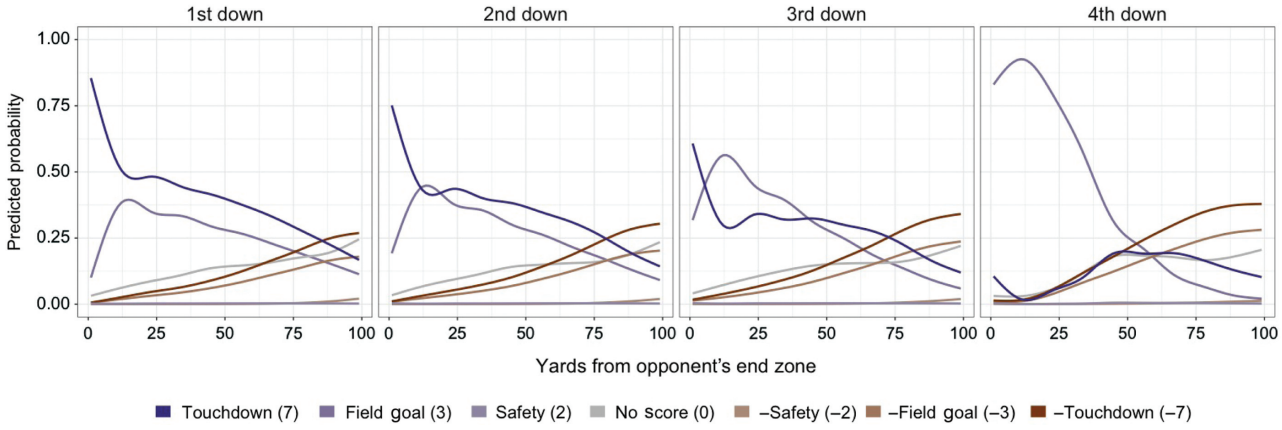

Figure 4: Relationship between next score event probabilities and field position by down.

Table 1: Description of selected variables for the win probability model.

Variable	Variable description
$E[S]$	Expected score differential = $EP + S$
s_g	Number of seconds remaining in game
$E[\frac{S}{s_g+1}]$	Expected score time ratio
h	Current half of the game (1st, 2nd, or overtime)
s_h	Number of seconds remaining in half
u	Indicator for whether or not time remaining in half is under two minutes
t_{off}	Time outs remaining for offensive (possession) team
t_{def}	Time outs remaining for defensive team

S is the score differential at the current play.

account the teams playing (i.e. we do not include indicators for team strength in the win probability model). As a result, every game starts with each team having a 50% chance of winning. Including indicators for a team's overall, offensive, and/or defensive strengths would artificially inflate (deflate) the contributions made by players on bad (good) teams in the models described in Section 3, since their team's win probability would start lower (higher).

Our approach for estimating *WP* also differs from the others mentioned in Section 1.1 in that we incorporate the estimated *EP* directly into the model by calculating the expected score differential for a play. Our expected points model already produces estimates for the value of the field position, yards to go, etc. without considering which half of the game or score. When including the variables presented in Table 1, we arrive at a well-calibrated *WP* model.

2.2.1 Generalized additive model

We use a generalized additive model (GAM) to estimate the possession team's probability of winning the game conditional on the current game situation. GAMs have several key benefits that make them ideal for modeling win probability. They allow the relationship between the explanatory and response variables to vary according to smooth, non-linear functions. They also allow for linear relationships and can estimate (both ordered and unordered) factor levels. We find that this flexible, semi-parametric approach allows us to capture nonlinear relationships while maintaining the many advantages of using linear models. Using a logit link function, our *WP* model takes the form

$$\log\left(\frac{P(\text{Win})}{P(\text{Loss})}\right) = s(E[S]) + s(s_h)\cdot h + s\left(E\left[\frac{S}{s_g+1}\right]\right) + h\cdot u\cdot t_{off} + h\cdot u\cdot t_{def}, \quad (7)$$

where s is a smooth function while h, u, t_{off}, and t_{def} are linear parametric terms defined in Table 1. By taking the inverse of the logit we arrive at a play's *WP*.

2.2.2 Win probability Calibration

Similar to the evaluation of the *EP* model, we again use LOSO CV to select the above model yielding the best calibration results. Figure 5 shows the calibration plots by quarter, mimicking the approach of Lopez (2017) and Yam and Lopez (2018), who evaluate both our *WP* model and that of Lock and Nettleton (2014). The observed proportion of wins closely matches the expected proportion of wins within each bin for each quarter, indicating that the model is well-calibrated across all quarters of play and across the spectrum of possible win probabilities. These findings match those of Yam and Lopez (2018), who find "no obvious systematic patterns that would signal a flaw in either model." An example of a single game's win probability chart is available in our supplementary materials.

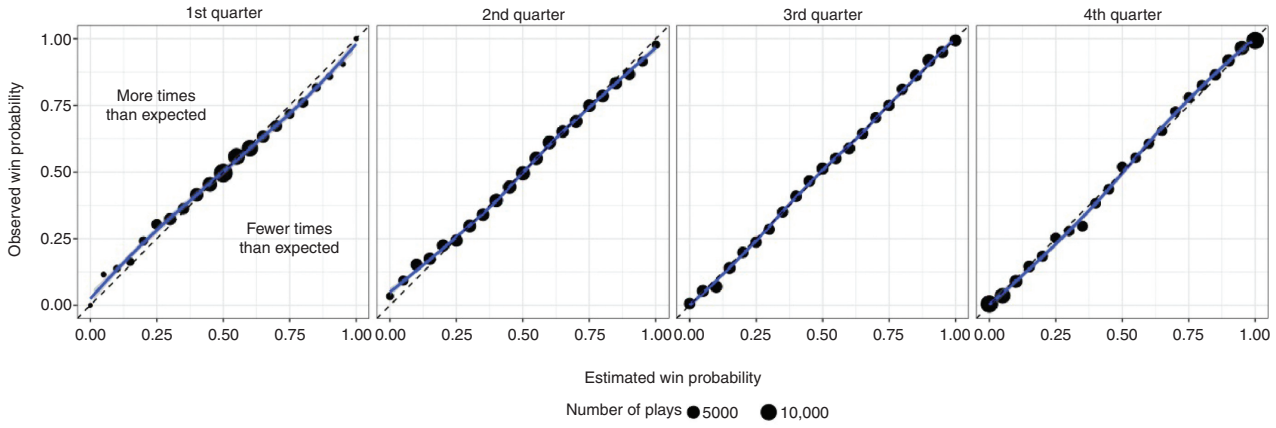

Figure 5: Win probability model LOSO CV calibration results by quarter.

2.3 Expected points added and win probability added

In order to arrive at a comprehensive measure of player performance, each play in a football game must be assigned an appropriate value $\delta_{f,i}$ that can be represented as the change from state i to state f. We define

$$\delta_{f,i} = \boldsymbol{V}_f - \boldsymbol{V}_i, \tag{8}$$

where $\boldsymbol{V}_f$ and $\boldsymbol{V}_i$ are the associated values for the ending and starting states, respectively. We represent these values by either play i's expected points (EP_i) or win probability (WP_i).

Plugging our *EP* and *WP* estimates for the start of play i and the start of the following play f into Equation 8's values for $\boldsymbol{V}_i$ and $\boldsymbol{V}_f$, respectively provides us with the two types of play valuations $\delta_{f,i}$: (1) the change in point value as expected points added (*EPA*), and (2) the change in win probability as win probability added (*WPA*). For scoring plays, we use the associated scoring event's value y as $\boldsymbol{V}_f$ in place of the following play's *EP* to reflect that the play's value is just connected to the difference between the scoring event and the initial state of the play. As an example, during Super Bowl LII the Philadelphia Eagles' Nick Foles received a touchdown when facing fourth down on their opponent's one yard line with thirty-eight seconds remaining in the half. At the start of the play the Eagles' expected points was $\boldsymbol{V}_i \approx 2.78$, thus resulting in $EPA \approx 7 - 2.78 = 4.22$. In an analogous calculation, this famous play known as the "Philly special" resulted in $WPA \approx 0.1266$ as the Eagles' increased their lead before the end of the half.

For passing plays, we can additionally take advantage of *air yards* (perpendicular distance in yards from the line of scrimmage to the yard line at which the receiver was targeted or caught the ball) and *yards after catch* (perpendicular distance in yards from the yard line at which the receiver caught the ball to the yard line at which the play ended), for every passing play available with `nflscrapR`. Using these two pieces, we can determine the hypothetical field position and whether a turnover on downs occurs to separate the value of a play from the air yards versus the yards after catch. For each completed passing play, we break the estimation of *EP* and *WP* into two plays – one comprising everything leading up to the catch, and one for the yards after the catch. Because the models rely on the seconds remaining in the game, we make an adjustment to the time remaining by subtracting the average length of time for incomplete passing plays, 5.7 seconds[1]. We then use the *EP* or *WP* through the air as $\boldsymbol{V}_f$ in Equation 8 to estimate $EPA_{i,air}$ or $WPA_{i,air}$, denoting these as $\delta_{f,i,air}$. We estimate the value of yards after catch, $\delta_{f,i,yac}$, by taking the difference between the value of the following play, $\boldsymbol{V}_f$, and the value of the air yards, $\delta_{f,i,air}$. We use this to calculate both $EPA_{i,yac}$ and $WPA_{i,yac}$.

3 Evaluating players with nflWAR

We use the play values calculated in Section 2 as the basis for a statistical estimate of wins above replacement (*WAR*) for each player in the NFL. To do this, we take the following approach:

- estimate the value of each play (Section 2),
- estimate the effect of each player on play value added (Section 3.1),
- evaluate relative to replacement level (Section 3.2),

1 This estimate could be improved in future work if information about the time between the snap and the pass becomes available.

– convert to a wins scale (Section 3.3), and
– and estimate the uncertainty in *WAR* (Section 3.4).

This framework can be applied to any individual season, and we present results for the 2017 season in Section 4. Due to data restrictions, we currently are only able to produce *WAR* estimates for offensive skill position players. However, a benefit of our framework is the ability to separate a player's total value into the three components of WAR_{air}, WAR_{yac}, and WAR_{rush}. Additionally, we provide the first statistical estimates for a team's rush blocking based on play-by-play data.

3.1 Division of credit

In order to properly evaluate players, we need to allocate the portion of a play's value $\delta_{f,i}$ to each player on the field. Unfortunately, the NFL does not publicly specify which players are on the field for every play, preventing us from directly applying approaches similar to those used in basketball and hockey discussed in Section 1.2, where the presence of each player on the playing surface is treated as an indicator covariate in a linear model that estimates the marginal effect of that player on some game outcome (Kubatko et al. 2007; Macdonald 2011; Thomas et al. 2013). Instead, the data available publicly from the NFL and obtained via `nflscrapR` are limited to only those players directly involved in the play, plus contextual information about the play itself.

For rushing plays, this includes at the player level the rusher and/or tackler(s), with contextual information for the run gap (end, tackle, guard, middle) and direction (left, middle, right). Further details regarding the nomenclature of run gaps are in our supplementary materials. For passing plays, player level information includes the passer, targeted receiver, tackler(s), and/or interceptor, along with contextual information about the air yards, yards after catch, pass location (left, middle, right), and if the passer was hit on the play.

3.1.1 Multilevel modeling

All players in the NFL belong to positional groups that dictate how they are used in the context of the game. For example, for passing plays we have the QB and the targeted receiver. However, over the course of an NFL season, the average QB will have more pass attempts than the average receiver will have targets, because there are far fewer QBs (more than 60 with pass attempts in the 2017 NFL season) compared to receivers (more than 400 targeted receivers in the 2017 season).

Because of these systematic differences across positions, there are differing levels of variation in each position's performance. Additionally, since every play involving the same player is a repeated measure of performance, the plays themselves are not independent. To account for these structural features of football, we use multilevel models (also referred to as hierarchical, random-effects, or mixed-effects models), which embrace this positional group structure and account for the observation dependence. Multilevel models have recently gained popularity in baseball statistics due to the development of catcher and pitcher metrics (Brooks, Pavlidis, and Judge 2015; Turkenkopf, Pavlidis, and Judge 2015), but have been used in sports dating back at least to 2012 (Cafarelli, Rigdon, and Rigdon 2012). Here, we extend their use for assessing offensive player contributions in football, using the play values $\delta_{f,i}$ from Section 2 as the response.

In order to arrive at individual player effects we use varying-intercepts for the groups involved in a play. A simple example of modeling $\delta_{f,i}$ with varying-intercepts for two groups, QBs as Q and receivers as C, with covariates X_i and coefficients β is

$$\delta_{f,i} \sim N\big(Q_{q[i]} + C_{c[i]} + X_i \cdot \beta,\ \sigma_\delta^2\big),\ \text{for } i = 1, \ldots, n \text{ plays}, \tag{9}$$

where the key feature distinguishing multilevel regression from classical regression is that the group coefficients vary according to

$$\begin{aligned} Q_q &\sim N\big(\mu_Q,\ \sigma_Q^2\big),\ \text{for } q = 1, \ldots, \#\text{ of QBs}, \\ C_c &\sim N\big(\mu_C,\ \sigma_C^2\big),\ \text{for } c = 1, \ldots, \#\text{ of receivers}. \end{aligned} \tag{10}$$

By assigning a probability distribution (such as the normal distribution) to the group intercepts, Q_q and C_c, with parameters estimated from the data (such as μ_Q and σ_Q for passers), each estimate is pulled toward their, respective group mean levels μ_Q and μ_C. In this example, QBs and receivers involved in fewer plays will be pulled closer to their overall group averages as compared to those involved in more plays and thus carrying more information, resulting in partially pooled estimates (Gelman and Hill 2007). This approach provides us with average individual effects on play value added while also providing the necessary shrinkage towards the group averages. All models we use for division of credit are of this varying-intercept form, and are fit using penalized likelihood via the `lme4` package in `R` (Bates et al. 2015). While these models are not explicitly Bayesian, as Gelman and Hill (2007) write, "[a]ll multilevel models are Bayesian in the sense of

Table 2: Description of variables in the models assessing player and team effects.

Variable name	Variable description
Home	Indicator for if the possession team was home
Shotgun	Indicator for if the play was in shotgun formation
NoHuddle	Indicator for if the play was in no huddle
QBHit	Indicator for if the QB was hit on a pass attempt
PassLocation	Set of indicators for if the pass location was either middle or right (reference group is left)
AirYards	Orthogonal distance in yards from the line of scrimmage to where the receiver was targeted or caught the ball
RecPosition	Set of indicator variables for whether the receiver's position was either TE, FB, or RB (reference group is WR)
RushPosition	Set of indicator variables for whether the rusher's position was either FB, WR, or TE (reference group is RB)
PassStrength	EPA per pass attempt over the course of the season for the possession team
RushStrength	EPA per rush attempt over the course of the season for the possession team

Table 3: Description of groups in the models assessing player and team effects.

Group	Individual	Description
Q	q	QB attempting a pass or rush/scramble/sack
C	c	Targeted receiver on a pass attempt
H	ι	Rusher on a rush attempt
T	τ	Team-side-gap on a rush attempt, combination of the possession team, rush gap and direction
F	v	Opposing defense of the pass

assigning probability distributions to the varying regression coefficients", meaning that we are taking into consideration all members of the group when estimating the varying intercepts rather than just an individual effect.

Our assumption of normality for $\delta_{f,i}$ follows from our focus on *EPA* and *WPA* values, which can be both positive and negative, exhibiting roughly symmetric distributions[2]. We refer to an intercept estimating a player's average effect as their *individual points/probability added* (*iPA*), with points for modeling *EPA* and probability for modeling *WPA*. Similarly, an intercept estimating a team's average effect is their *team points/probability added* (*tPA*). Tables 2 and 3 provide the notation and descriptions for the variables and group terms in the models apportioning credit to players and teams on plays. The variables in Table 2 would be represented by X, and their effects by β in Equation 9.

3.1.2 Passing models

Rather than modeling the $\delta_{f,i}$ (*EPA* or *WPA*) for a passing play, we take advantage of the availability of air yards and develop two separate models for $\delta_{f,i,air}$ and $\delta_{f,i,yac}$. We do not credit the passer solely for the value gained through the air, nor the receiver solely for the value gained from after the catch. Instead, we propose that the passer, receiver, and opposing defense should each have credit divided amongst them for both areas. We let Δ_{air} and Δ_{yac} be the response variables for the air yards and yards after catch models, respectively. Both models consider all passing attempts, but the response variable depends on the model

$$\begin{aligned}\Delta_{air} &= \delta_{f,i,air} \cdot \mathbf{1}(\text{completion}) + \delta_{f,i} \cdot \mathbf{1}(\text{incompletion}),\\ \Delta_{yac} &= \delta_{f,i,yac} \cdot \mathbf{1}(\text{completion}) + \delta_{f,i} \cdot \mathbf{1}(\text{incompletion}),\end{aligned} \tag{11}$$

where $\mathbf{1}$(completion) and $\mathbf{1}$(incompletion) are indicator functions for whether or not the pass was completed. This serves to assign all completions the $\delta_{f,i,air}$ and $\delta_{f,i,yac}$ as the response for their respective models, while incomplete passes are assigned the observed $\delta_{f,i}$ for both models. In using this approach, we emphasize the importance of completions, crediting accurate passers for allowing their receiver to gain value after the catch.

The passing model for Δ_{air} is

$$\begin{aligned}\Delta_{air} &\sim N\big(Q_{air,q[i]} + C_{air,c[i]} + F_{air,v[i]} + \boldsymbol{A}_i \cdot \boldsymbol{\alpha},\ \sigma_{\Delta_{air}}\big) \quad \text{for } i = 1, \ldots, n \text{ plays},\\ Q_{air,q} &\sim N\big(\mu_{Q_{air}}, \sigma^2_{Q_{air}}\big), \text{ for } q = 1, \ldots, \#\text{ of QBs},\\ C_{air,c} &\sim N\big(\mu_{C_{air}}, \sigma^2_{C_{air}}\big), \text{ for } c = 1, \ldots, \#\text{ of receivers},\\ F_{air,v} &\sim N\big(\mu_{F_{air}}, \sigma^2_{F_{air}}\big), \text{ for } v = 1, \ldots, \#\text{ of defenses},\end{aligned} \tag{12}$$

where the covariate vector $\boldsymbol{A}_i$ contains a set of indicator variables for Home, Shotgun, NoHuddle, QBHit, Location, RecPosition, as well as the RushStrength value while $\boldsymbol{\alpha}$ is the corresponding coefficient vector. The passing model for Δ_{yac} is:

$$\Delta_{yac} \sim N\big(Q_{yac,q[i]} + C_{yac,c[i]} + F_{yac,v[i]} + \boldsymbol{B}_i \cdot \boldsymbol{\beta},\ \sigma_{\Delta_{yac}}\big) \quad \text{for } i = 1, \ldots, n \text{ plays},$$

2 Residual plots for models from the 2017 season are available in our supplementary materials.

$$
\begin{aligned}
Q_{yac,q} &\sim N\left(\mu_{Q_{yac}},\, \sigma^2_{Q_{yac}}\right),\ \text{for } q = 1, \ldots, \#\text{ of QBs},\\
C_{yac,c} &\sim N\left(\mu_{C_{yac}},\, \sigma^2_{C_{yac}}\right),\ \text{for } c = 1, \ldots, \#\text{ of receivers},\\
F_{yac,v} &\sim N\left(\mu_{F_{yac}},\, \sigma^2_{F_{yac}}\right),\ \text{for } v = 1, \ldots, \#\text{ of defenses},
\end{aligned} \tag{13}
$$

where the covariate vector $\boldsymbol{B}_i$ contains the same set of indicator variables in $\boldsymbol{A}_i$ but also includes the AirYards and interaction terms between AirYards and the various RecPosition indicators, with $\boldsymbol{\beta}$ as its respective coefficient vector. We include the RushStrength in the passing models as a group-level predictor to control for the possession team's rushing strength and the possible relationship between the two types of offense. For QBs, their estimated $Q_{air,q}$ and $Q_{yac,q}$ intercepts represent their iPA_{air} and iPA_{yac} values, respectively (same logic applies to receivers). Likewise, the opposing defense values of $F_{air,v}$ and $F_{yac,v}$ are their tPA_{air} and tPA_{yac} values.

3.1.3 Rushing models

For rushing plays, we again model the play values $\delta_{f,i}$. However, we build two separate models, with one rushing model for QBs and another for all non-QB rushes. This is because we cannot consistently separate (in the publicly available data) designed QB rushes from scrambles on broken plays, the characteristics of which result in substantially different distributions of play value added. It is safe to assume all non-QB rushes are designed rushes. Our rushing model for QBs consists of all scrambles, designed runs, and sacks (to account for skilled rushing QBs minimizing the loss on sacks). The QB rushing model is

$$
\begin{aligned}
\delta_{f,i} &\sim N\left(Q_{rush,q[i]} + F_{rush_Q,\, v[i]} + \boldsymbol{\Gamma}_i \cdot \boldsymbol{\gamma},\ \sigma_{\delta_{rush_Q}}\right)\\
&\qquad \text{for } i = 1, \ldots, n \text{ plays},\\
Q_{rush,q} &\sim N\left(\mu_{Q_{rush}},\, \sigma^2_{Q_{rush}}\right),\ \text{for } q = 1, \ldots, \#\text{ of QBs},\\
F_{rush_Q,v} &\sim N\left(\mu_{F_{rush_Q}},\, \sigma^2_{F_{rush_Q}}\right),\\
&\qquad \text{for } v = 1, \ldots, \#\text{ of defenses},
\end{aligned} \tag{14}
$$

where the covariate vector $\boldsymbol{\Gamma}_i$ contains a set of indicator variables for Home, Shotgun, NoHuddle, as well as the PassStrength variable where $\boldsymbol{\gamma}$ is the corresponding coefficient vector.

For the designed rushing plays of non-QBs, we include an additional group variable T. As detailed in Table 3, T serves as a proxy for the offensive linemen or blockers involved in the rushing attempt. Each team has seven possible T levels of the form team-side-gap. For example, the Pittsburgh Steelers (PIT) have the following levels: PIT-left-end, PIT-left-tackle, PIT-left-guard, PIT-middle-center, PIT-right-guard, PIT-right-tackle, PIT-right-end. The non-QB rushing model is

$$
\begin{aligned}
\delta_{f,i} &\sim N\left(H_{\iota[i]} + T_{\tau[i]} + F_{rush,v[i]} + \boldsymbol{P}_i \cdot \boldsymbol{\rho},\ \sigma_{\delta_{rush}}\right)\\
&\qquad \text{for } i = 1, \ldots, n \text{ plays},\\
H_\iota &\sim N\left(\mu_H,\, \sigma^2_H\right),\ \text{for } \iota = 1, \ldots, \#\text{ of rushers},\\
T_\tau &\sim N\left(\mu_T,\, \sigma^2_T\right),\ \text{for } \tau = 1, \ldots, \#\text{ of team-side-gaps},\\
F_{rush,v} &\sim N\left(\mu_{F_{rush}},\, \sigma^2_{F_{rush}}\right),\ \text{for } v = 1, \ldots, \#\text{ of defenses},
\end{aligned} \tag{15}
$$

where the covariate vector $\boldsymbol{P}_i$ contains a set of indicator variables for Home, Shotgun, NoHuddle, RushPosition, and PassStrength, and where $\boldsymbol{\rho}$ is the corresponding coefficient vector. The resulting $Q_{rush,q}$ and $H_{rush,\iota}$ estimates are the iPA_{rush} values for the QB and non-QB rushers, respectively. Additionally, the T_τ estimate is the $tPA_{rush,side-gap}$ for one of the seven possible side-gaps for the possession team, while $F_{rush,v}$ and $F_{rush_Q,v}$ are the tPA_{rush} and tPA_{rush_Q} values for the opposing defense for non-QB and QB rushes.

3.1.4 Individual points/probability added

Let $\kappa_{p,pass}$ and $\kappa_{p,rush}$ refer to the number of attempts for player p on passing and rushing plays, respectively. Using an estimated type of *iPA* value for a player p and multiplying by the player's associated number of attempts provides us with an *individual points/probability above average* ($iPAA_p$) value. The three different types of $iPAA_p$ values for each position are

$$
\begin{aligned}
iPAA_{p,air} &= \kappa_{p,pass} \cdot iPA_{p,air},\\
iPAA_{p,yac} &= \kappa_{p,pass} \cdot iPA_{p,yac},\\
iPAA_{p,rush} &= \kappa_{p,rush} \cdot iPA_{p,air},
\end{aligned} \tag{16}
$$

where the values for $\kappa_{p,pass}$ and $\kappa_{p,rush}$ depend on the player's position. For QBs, $\kappa_{p,pass}$ equals their number of pass attempts, while $\kappa_{p,rush}$ is the sum of their rush attempts, scrambles, and sacks. For non-QBs $\kappa_{p,pass}$ equals their number of targets and $\kappa_{p,rush}$ is their number of rush attempts. Summing all three components provides us with player p's total individual points/probability above average, $iPAA_p$.

3.2 Comparing to replacement Level

As described in Section 1.2, it is desirable to calculate a player's value relative to a "replacement level" player's

performance. There are many ways to define replacement level. Thomas and Ventura (2015) define a concept called "poor man's replacement", where players with limited playing time are pooled, and a single effect is estimated in a linear model, which is considered replacement level. Others provide more abstract definitions of replacement level, as the skill level at which a player can be acquired freely or cheaply on the open market (Tango et al. 2007).

We take a similar approach to the *openWAR* method, defining replacement level by using a roster-based approach (Baumer et al. 2015), and estimating the replacement level effects in a manner similar to that of Thomas and Ventura (2015). Baumer et al. (2015) argue that "replacement level" should represent a readily available player that can replace someone currently on a team's active roster. Due to differences in the number of active players across positions in football, we define replacement level separately for each position. Additionally, because of usage for the different positions in the NFL, we find separate replacement level players for receiving as compared to rushing. In doing so, we appropriately handle cases where certain players have different roles. For example, a RB that has a substantial number of targets but very few rushing attempts can be considered a replacement level rushing RB, but not a replacement level receiving RB.

Accounting for the 32 NFL teams and the typical construction of a roster (Lillibridge 2013), we consider the following players to be "NFL level" for each the non-QB positions:

- rushing RBs = $32 \cdot 3 = 96$ RBs sorted by rushing attempts,
- rushing WR/TEs = $32 \cdot 1 = 32$ WR/TEs sorted by rushing attempts,
- receiving RBs = $32 \cdot 3 = 96$ RBs sorted by targets,
- receiving WRs = $32 \cdot 4 = 128$ WRs sorted by targets,
- receiving TEs = $32 \cdot 2 = 64$ TEs sorted by targets.

Using this definition, all players with fewer rushing attempts or targets than the NFL level considered players are deemed replacement level. This approach is consistent with the one taken by Football Outsiders (Schatz 2003). We combine the rushing replacement level for WRs and TEs because there are very few WRs and TEs with rushing attempts.

Due to the nature of QB usage in the NFL, we proceed in a different manner to find replacement level QBs. Figure 6 displays the distribution of the percentage of a team's plays in which a player is directly involved (passer, receiver, or rusher) by position using data from 2009 to 2017. This does not represent the percentage of team snaps by a player, but rather for a given position that is directly

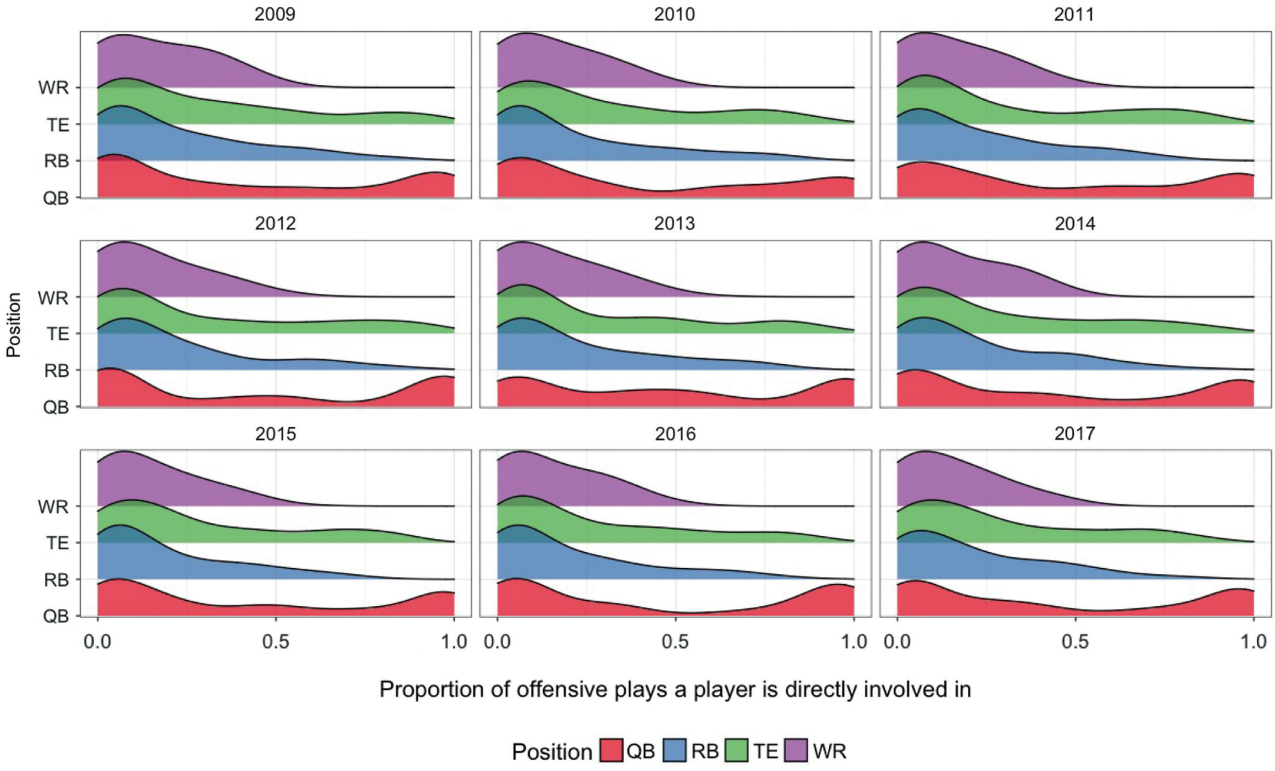

Figure 6: Distribution of the proportion of offensive plays a player is directly involved in by position (2009–2017).

involved in a play, it shows the distribution of team play percentages for every player of that position (e.g. New Orleans Saints' RB Alvin Kamara was involved in 38.39% of all Saints plays that directly involved a RB). While the distributions for RB, WR, and TE are unimodal and clearly skewed right, the distributions for QBs are bimodal for each season. This is an unsurprising result, since most NFL teams rely on a single QB for an entire season, resulting in them being involved in more than 80% of the team's plays at QB.

Observing this clear difference in the distribution for QBs, we define a replacement level QB as any QB with less than ten percent involvement in their team's plays that directly involve QBs. This approach essentially asserts that backup QBs with limited playing time should represent replacement level for QBs, rather than assuming all NFL teams have at least a certain number of NFL level QBs on their roster. We recognize this definition is far from perfect, we cover an additional "one QB" approach in our supplementary materials, but we hope this provides a starting point for defining replacement level which researchers can improve upon.

Prior to fitting the models discussed in Section 3.1, every player who is identified as replacement level is replaced in their corresponding play-by-play data with their replacement label (e.g. Replacement QB, Replacement RB-rushing, Replacement RB-receiving, etc). By doing so, all replacement level players for a particular position and type (receiving versus rushing) have the same iPA^{repl} estimate. We then calculate a player's value above replacement, *individual points/probability above replacement* ($iPAR_p$) in the same manner as Baumer et al. (2015) and Thomas and Ventura (2015), by calculating a replacement level "shadow" for a particular player. For a player p, this is done by first calculating their replacement "shadow" value, $iPAA_p^{repl}$, by using their respective number of attempts:

$$\begin{aligned} iPAA_{p,air}^{repl} &= \kappa_{p,pass} \cdot iPA_{air}^{repl}, \\ iPAA_{p,yac}^{repl} &= \kappa_{p,pass} \cdot iPA_{yac}^{repl}, \\ iPAA_{p,rush}^{repl} &= \kappa_{p,rush} \cdot iPA_{rush}^{repl}, \end{aligned} \tag{17}$$

which leads to natural calculations for the three *iPAR* values:

$$\begin{aligned} iPAR_{p,air} &= iPAA_{p,air} - iPAA_{p,air}^{repl}, \\ iPAR_{p,yac} &= iPAA_{p,yac} - iPAA_{p,yac}^{repl}, \\ iPAR_{p,rush} &= iPAA_{p,rush} - iPAA_{p,rush}^{repl}. \end{aligned} \tag{18}$$

Taking the sum of the three, we arrive at a player's total $iPAR_p$.

3.3 Conversion to wins

If the play's value used for modeling purposes was *WPA* based, then the final *iPAR* values are an individual's win probability added above replacement, which is equivalent to their *wins above replacement* (*WAR*). However, for the *EPA*-based play value response, the *iPAR* values represent the individual expected points added above replacement, and thus require a conversion from points to wins. While other, more complicated approaches exist for addressing this (see Baumer et al. (2015) for a review), we take a simple approach, similar to that of Zhou and Ventura (2017) and others: we use linear regression to estimate the relationship between team t's regular season win total and their score differential (S) during the season,

$$Wins_t = \beta_0 + \beta_S S_t + \epsilon_t, \text{ where } \epsilon_t \overset{iid}{\sim} N(0, \sigma^2). \tag{19}$$

The resulting coefficient estimate $\hat{\beta}_S$ (typically about 0.03, depending on the season) represents the increase in the number of wins for each one point increase in score differential. Since one additional point of score differential raises the expected number of wins by 0.03, then roughly 33 additional points will raise the expected number of wins by 1. We can now estimate *WAR* for the *EPA* based approach by taking the *iPAR* values and dividing by 33, the estimated points per win. This is equivalent to multiplying *iPAR* by $\hat{\beta}_S$.

3.4 Uncertainty

Similar to the approach taken by Baumer et al. (2015) for estimating the variability in their *openWAR* metric, we use a resampling strategy to generate distributions for each individual player's *WAR* values. Rather than resampling plays in which a particular player is involved to arrive at estimates for their performance variability, we resample entire team drives. We do this to account for the fact that player usage is dependent on team decision making, meaning that the random variation in individual events is dependent upon the random variation in team events. Thus, we must resample at the team level to account for the variability in a player's involvement. The decision to resample whole drives instead of plays is to represent sampling that is more realistic of game flows due to the possibility of dependencies within a drive with regards to team play-calling. In Section 4, all uncertainty estimation uses this drive-resampling approach, with 1000 simulated seasons.

4 Results

Given the definitions in Section 3.2, we found the following replacement level designations for the 2017 NFL season:

- rushing: 52 of the 148 RBs are replacement level,
- rushing: 278 of the 310 WR/TEs are replacement level,
- receiving: 52 of the 148 RBs are replacement level,
- receiving: 73 of the 201 WRs are replacement level,
- receiving: 45 of the 109 TEs are replacement level,
- passing: 25 of the 71 QBs are replacement level.

We compare the distributions for both types of *WAR* estimates, *EPA*-based and *WPA*-based, by position in Figure 7. For all positions, the *EPA*-based *WAR* values tend be higher than the *WPA*-based values. This could be indicative of a player performing well in meaningless situations due to the score differential, particularly for QBs. It is clear that QBs have larger *WAR* values than the other positions, reflecting their involvement in every passing play and potentially providing value by rushing. Although this coincides with conventional wisdom regarding the importance of the QB position, we note that we have not controlled for all possible contributing factors, such as the specific offensive linemen, the team's offensive schemes, or the team's coaching ability due to data limitations. Researchers with access to this information could easily incorporate their proprietary data into this framework to reach a better assessment of QB value.

Following Major League Baseball's 2017 MVP race, *WAR* has received heavy criticism for its unclear relationship with wins (James 2017; Tango 2017). For this reason, we focus on the *WPA*-based version of *WAR*, with its direct relationship to winning games. Figure 8 displays the top five players based on total *WAR* for each position in the 2017 season. Each chart is arranged in descending order by each player's estimated *WAR*, and displays the three separate *WAR* values of WAR_{air}, WAR_{yac}, and WAR_{rush}. By doing this separation, we can see how certain types of players vary in their performances. Tom Brady for instance is the only QB in the top five with negative WAR_{rush}. Alvin Kamara appears to be providing roughly equal value from rushing and receiving, while the other top RB performances are primarily driven by rushing success.

Elaborating on this separation of types of players, we can use the random intercepts from the multilevel models, the *iPA* values, to see the underlying structure of players in terms of their efficiency. Figures 9 and 10 reveal the separation of types of QBs and RBs, respectively. The origin points for both charts represent league averages. For QBs, we plot their estimates for iPA_{air} against iPA_{yac}, providing an overview of the types of passers in the NFL. The two components represent different skills of being able to provide value by throwing deep passes through the air, such as Jameis Winston, as compared to short but accurate passers such as Case Keenum. We can also see where the replacement level QB estimates place for context. For RBs, we add together their iPA_{air} and iPA_{yac} estimates to summarize their individual receiving effect and plot this against their iPA_{rush} estimates. This provides a separation

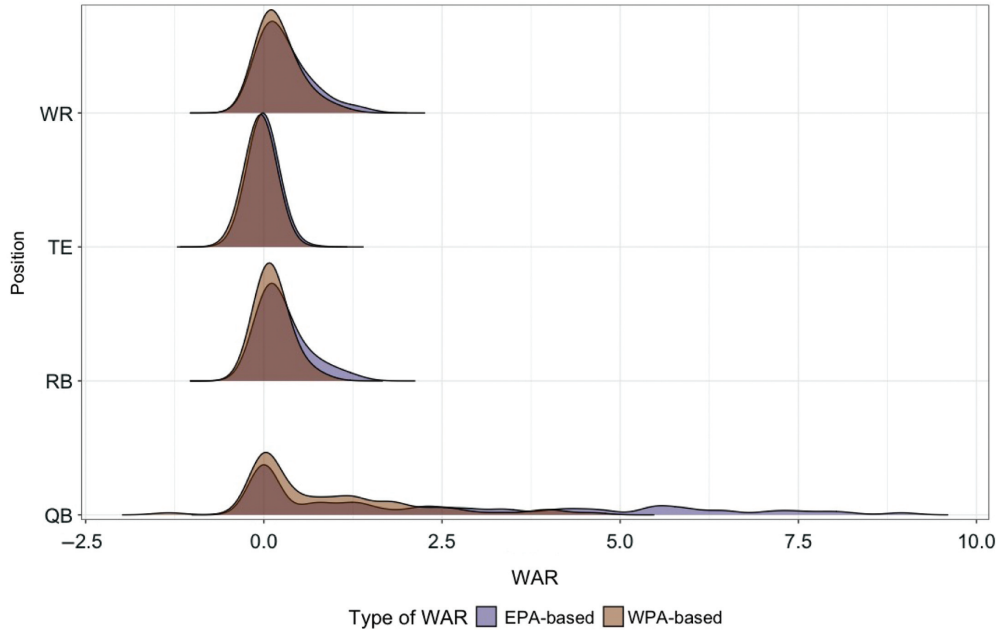

Figure 7: Distribution of *WAR* in 2017 season by type and position.

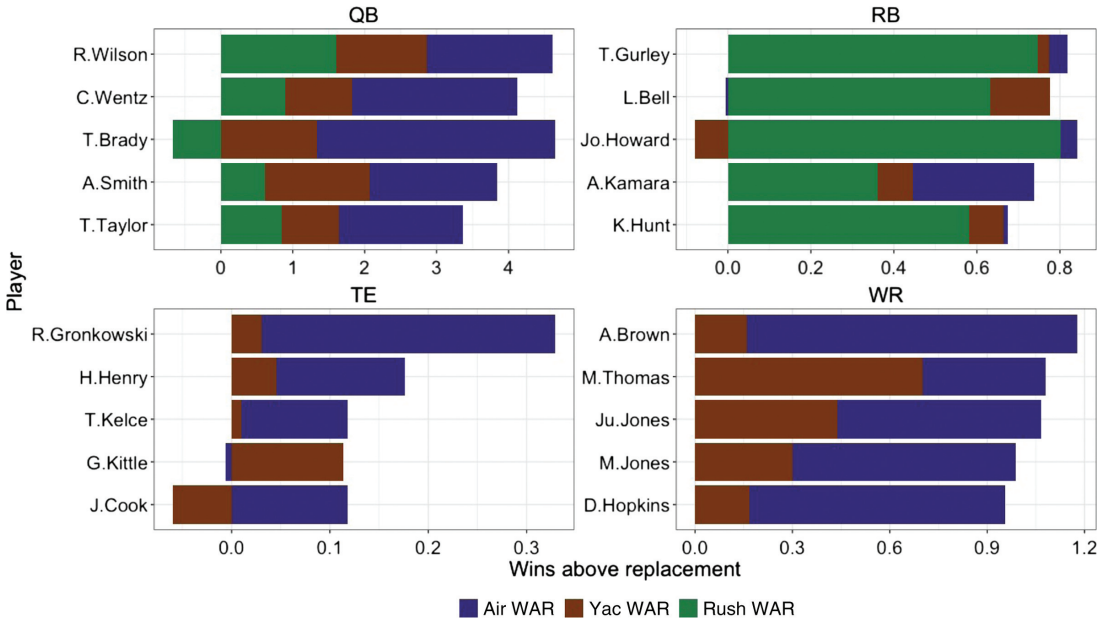

Figure 8: Top five players in *WAR* by position for the 2017 season.

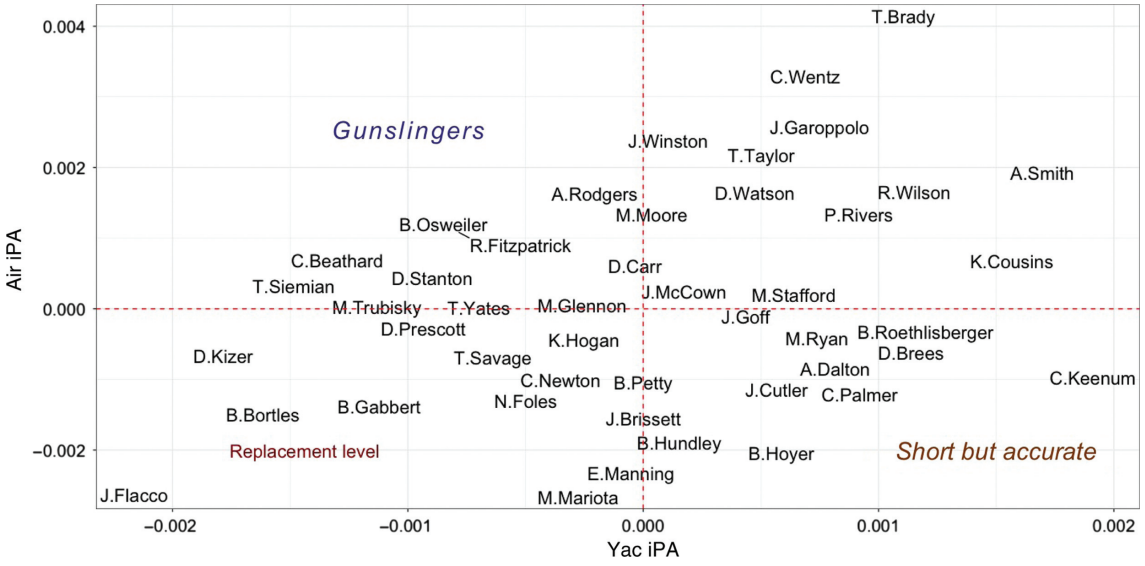

Figure 9: Estimates for 2017 QB efficiency from iPA_{air} against iPA_{yac}.

between RBs that provide value as receivers versus those who provide positive value primarily from rushing, such as Ezekiel Elliott. New Orleans Saints RB Alvin Kamara stands out from the rest of the league's RBs, providing elite value in both areas.

Using the drive resampling approach outlined in Section 3.4, we can compare the variability in player performance based on 1000 simulated seasons. Figure 11 compares the simulation distributions of the three types of *WAR* values (WAR_{air}, WAR_{yac}, WAR_{rush}) for selected QBs in the 2017 NFL season, with a reference line at 0 for replacement level. We can clearly see that the variability associated with player performance is not constant, which is not suprising given the construction of the resampling at the drive level. However, we can see some interesting features of QB performances, such as how Seattle Seahawks QB Russell Wilson's three types of *WAR* distributions are overlapping significantly, emphasizing his versatility. Additionally, New England Patriots QB Tom Brady displays large positive WAR_{air} and WAR_{yac} values, but a clearly negative WAR_{rush} value. Finally, Joe Flacco's 2017 performance was at or below replacement level in the vast majority of simulations across all three types of *WAR*, indicating that he is not elite.

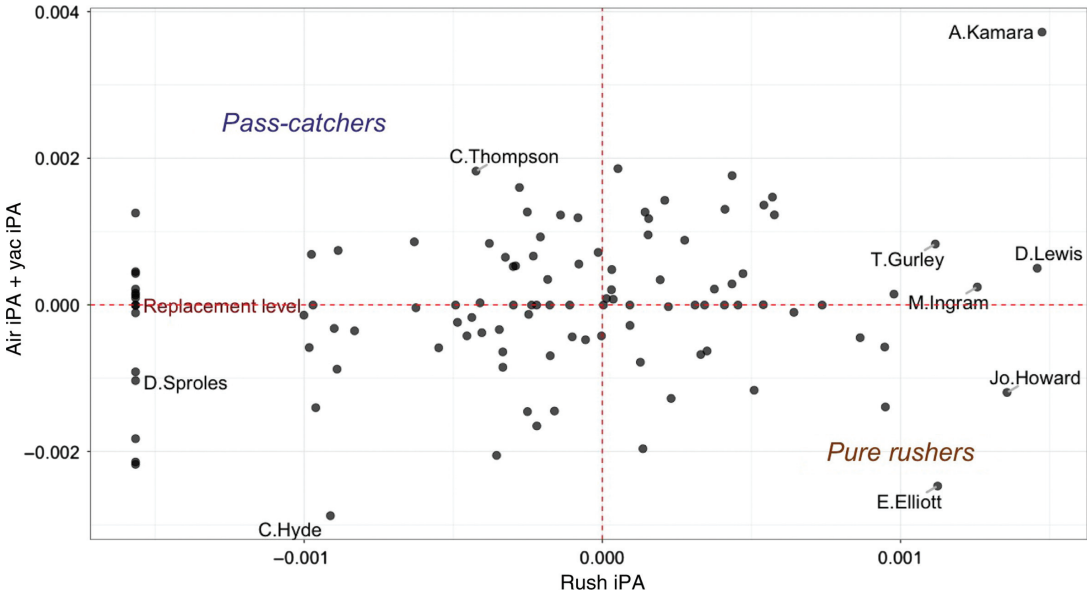

Figure 10: Estimates for RB efficiency from receiving ($iPA_{air} + iPA_{yac}$) against rushing (iPA_{rush}) for the 2017 season.

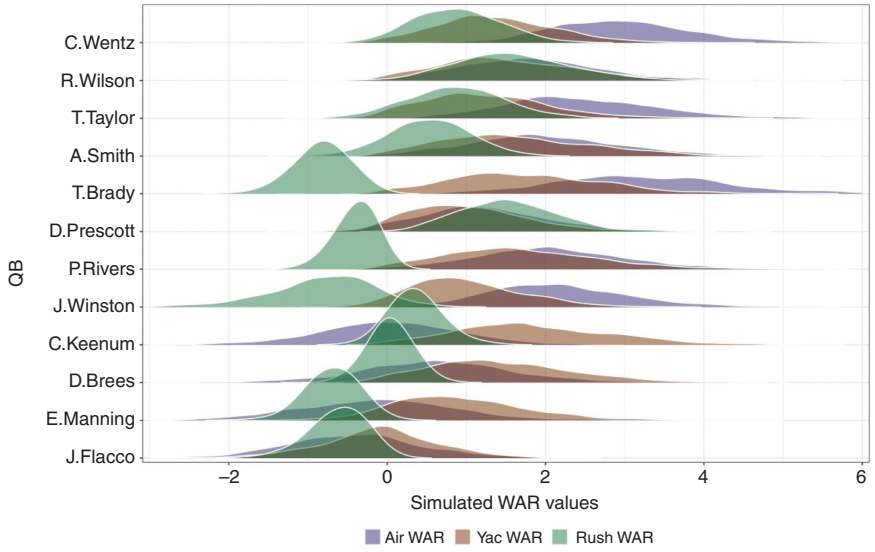

Figure 11: Simulation distributions of 2017 *WAR* by type for a selection of 12 QBs.

Figure 12 displays the simulation distributions for the top ten RBs during the 2017 NFL season, as ranked by their average total *WAR* across all simulations. Relative to the *WAR* values for QBs in Figure 12, the best RBs in the league are providing limited value to their teams. This is in agreement with the recent trend of NFL teams, who have been paying QBs increasing salaries but compensating RBs less (Morris, 2017). Two of the top RBs in the 2017 were rookies Alvin Kamara and Kareem Hunt, leading to discussion of which player deserved to be the NFL's rookie of the year. Similar to Baumer et al. (2015) we address this question using our simulation approach and display the joint distribution of the two player's 2017 performances in Figure 13. In nearly 71% of the simulated seasons, Kamara leads Hunt in *WAR* providing us with reasonable certainty in Kamara providing more value to his team than Hunt in his rookie season. It should not come as a surprise that there is correlation between the player performances as each simulation consists of fitting the various multilevel models resulting in new estimates for the group averages, individual player intercepts as well as the replacement level performance.

Additionally, we examine the consistency of the *WPA*-based *WAR* from season-to-season based on the correlation within players between their 2016 and 2017 seasons (excluding replacement level) and compare this

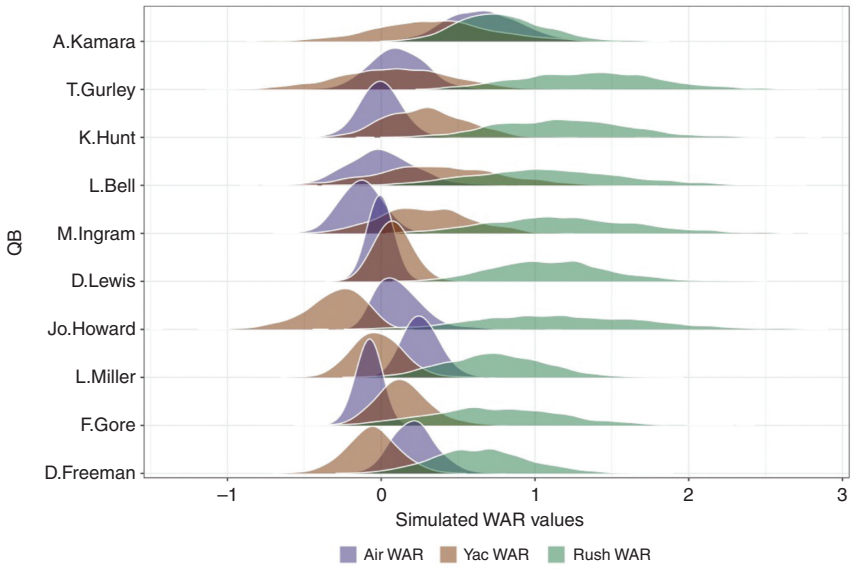

Figure 12: Simulation distributions of 2017 *WAR* value by type for top 10 RBs.

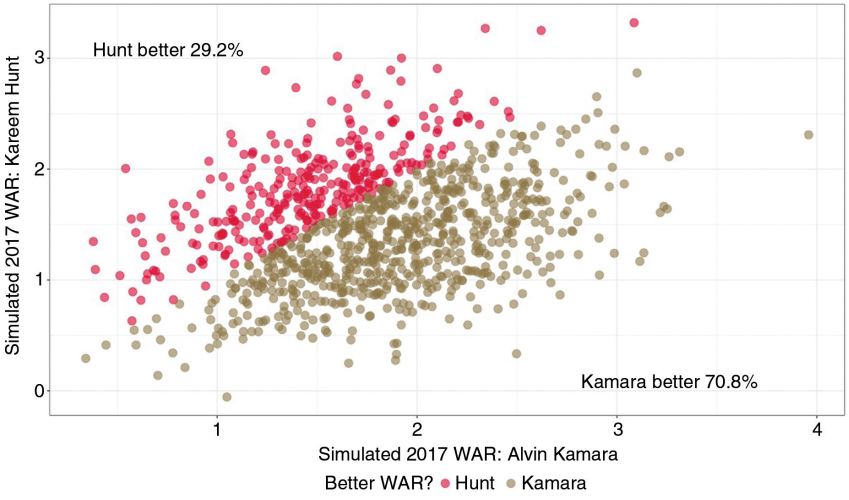

Figure 13: Joint distribution of *WAR* for Alvin Kamara and Kareem Hunt in 2017.

to other commonly used statistics for QBs and RBs. Seen in Table 4, our estimates for QB *WAR* displayed higher correlations than both the commonly used passer rating statistic as well as ANY/A. We also see in Table 5 higher correlations for RB *WAR* as compared to Brian Burke's success rate (percentage of rush attempts with *EPA* greater than zero) and rushing yards per attempt. Future work should consider a proper review and assessment of football statistics accounting for the number of attempts needed for determing the reliability of a statistic as well as accounting for when a player changes teams (Yurko, Ventura, and Horowitz 2017), and also apply the framework laid out by Franks et al. (2017).

Although it does not provide a measure for individual players' contributions, we can sum together the seven possible $tPA_{rush,side-gap}$ estimates for a team providing a proxy for their offensive line's overall efficiency in contributing to rushing plays. We can also look at

Table 4: Correlation of QB statistics between 2016 and 2017 seasons.

	WAR	Passer Rating	ANY/A
Correlation	0.598	0.478	0.295

Table 5: Correlation of RB statistics between 2016 and 2017 seasons.

	WAR	Success Rate	Yards per Attempt
Correlation	0.431	0.314	0.337

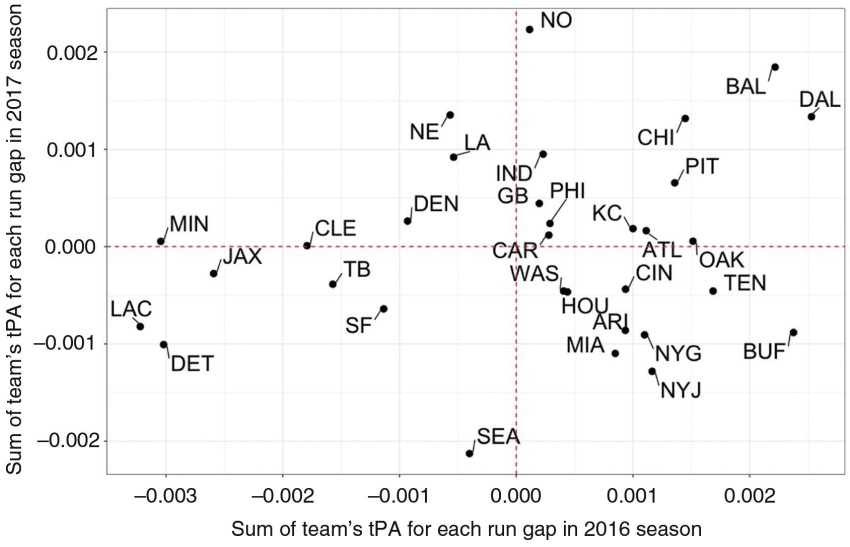

Figure 14: Team offensive line measures.

individual side-gaps for specific teams to assess their offensive line's performance in particular areas. Figure 14 displays the $tPA_{rush,side-gap}$ sum in 2017 against 2016 for each NFL team. The red lines provide indication to average performances in each year, so teams in the upper right quadrant performed above average overall in both years such as the Dallas Cowboys (DAL) which are known to have one of the best offensive lines in football.

5 Discussion and extensions

In this work, we have provided four major contributions to the statistical analysis of NFL football, in areas that can impact both on-field and player personnel decisions. We discuss these contributions and some extensions below.

5.1 Reproducible play and player evaluation

Data and software development: We provide an `R` package, `nflscrapR`, to provide easy access to publicly available NFL play-by-play data for researchers to use in their own analyses of the NFL. This package has already been used by researchers to further research into NFL decision-making (Yam and Lopez 2018).

Play evaluation: We make two contributions in the area of play evaluation. First, we introduce a novel approach for estimating expected points using a multinomial logistic regression model that is fully reproducible. Second, we use a generalized additive model for estimating in-game win probability, incorporating the results of the expected points model as input. With these two play evaluation models, we can calculate measures such as expected points added and win probability added, which are commonly used to evaluate both plays and players. All of these measures are included directly into the play-by-play data provided by `nflscrapR`. Moreover, all code used to build these expected points and win probability models is provided in `nflscrapR` and available on GitHub https://github.com/ryurko/nflscrapR-models. By taking these important steps, we ensure that all of our methods are fully reproducible, and we make it as easy as possible for researchers to use, explore, and improve upon our work.

Player evaluation: With our *nflWAR* framework, we use multilevel models to isolate offensive skill player contribution and estimate their individual wins above replacement.

Our estimates of *WAR* are given for several different areas of the game and are available for all offensive skill position players, unlike previous approaches. By compartmentalizing our estimates of player *WAR*, we are able to better characterize players and how they achieved success. For example, New Orleans Saints RB Alvin Kamara was unique in his success as both a rusher and a receiver in the 2017 NFL season. Similarly, Seattle Seahawks QB Russell Wilson was unique in his success as a rusher as well as from passing through the air and for yards after the catch, with about equal *WAR* contributions in all three areas in the 2017 NFL season. While these findings may not surprise knowledgeable football fans, our framework also reveals the value of potentially overlooked skills, such as the rushing ability of Tyrod Taylor and Dak Prescott, as seen in Figure 11. The value they provide by their rushing

WAR reflects not only their ability to scramble, but their ability to limit the loss from sacks as compared to other QBs. This is a skill that cannot be captured with basic statistics such as sack rate (which treats all sacks as equal).

Our multilevel modeling approach provides the first statistical estimate of an offensive line's value towards run-blocking that also controls for factors such as RB ability, opposing defense, etc. We recognize, however, that this is not a perfect measure of offensive line performance since it does not necessarily capture individual linemen, and does not account for potential selection bias that could influence specific side-gap estimates (since RBs are likely to run towards holes and away from defenders). We lack information about which specific offensive linemen are on the field or even involved in plays, preventing us from fitting player-specific terms in our multilevel model that would provide *WAR* estimates for individual offensive linemen. Researchers with access to this data can build this into our modeling framework with minimal issues, as we discuss in Section 5.2.

Adopting a resampling procedure similar to that of Baumer et al. (2015), we provide estimates of uncertainty on all *WAR* estimates. Our approach resamples at the drive-level, rather than individual plays, to preserve the effects of any within-drive factors, such as play sequencing or situational play-calling tendencies.

Finally, our *WAR* models are fully reproducible, with all data coming directly from `nflscrapR`, and with all code provided on GitHub https://github.com/ryurko/nflWAR.

5.2 Extensions relevant to NFL teams

The road to WAR for players at all positions: One key benefit to our approach is that it can easily be augmented with the inclusion of additional data sources, e.g. player-tracking data or proprietary data collected by NFL teams. Given data about which players are present on the field for each play, we can update our multilevel models from Section 3.1 by including additional positional groups. For example, for the non-QB rushing model, we can update the model as follows:

$$\begin{aligned}
\delta_{f,i} &\sim N\left(\sum_k O^k_{rush,v_k[i]} + \sum_g D^g_{rush,v_g[i]} + \boldsymbol{P}_i \cdot \boldsymbol{\rho},\ \sigma_{\delta_{rush}}\right) \\
&\text{for } i = 1, \ldots, n \text{ plays}, \\
O^k_{rush,v_k} &\sim N\left(\mu_{O^k_{rush}},\ \sigma^2_{O^k_{rush}}\right), \\
D^g_{rush,v_g} &\sim N\left(\mu_{D^g_{rush}},\ \sigma^2_{D^g_{rush}}\right),
\end{aligned}$$

where O^k_{rush,v_k} are the intercepts for offensive positions, indexed by k, D^g_{rush,v_g} are the intercepts for defensive positions, indexed by g, with both types of positions varying according to their own model, where $\boldsymbol{P}_i$ and $\boldsymbol{\rho}$ are described as above. Similar updates can be made to the models representing QB rushing, passing through the air, and passing for yards after catch. After doing so we can calculate the *iPAA* for any player at any position and, with adequate definitions for replacement level at each position, will have statistical *WAR* estimates for players of any position, offensive and defensive.

Roster construction: If NFL teams have player participation data (detailing which players are on the field for each play) and are able to implement the framework above, they would then have *WAR* estimates for players at all positions dating back as far back as they have complete player participation data. Teams that are able to do this could potentially gain substantial advantages in important areas of roster construction.

First, teams could more appropriately assess the contract values of players in free agency, similar to what is commonly done in baseball (Paine 2015). Second and perhaps most importantly, teams would be able to substantially improve their analysis for the NFL draft. Using an approach similar to that of Citrone and Ventura (2017), teams could substitute an objective measure of *WAR* in place of the more subjective measure of AV, in order to project the future career value in terms of *WAR* for all players available in the NFL draft. Additionally, teams employing this approach could create updated, *WAR*-based versions of the "draft pick value chart", first attributed to Jimmy Johnson and later improved by Meers (2011) and Citrone and Ventura (2017). In doing so, teams could more accurately assess the value of draft picks and potentially exploit their counterparts in trades involving draft picks.

References

Alamar, B. 2010. "Measuring Risk in NFL Playcalling." *Journal of Quantitative Analysis in Sports* 6(2). https://doi.org/10.2202/1559-0410.1235.

Albert, J. 2006. "Pitching Statistics, Talent and Luck, and the Best Strikeout Seasons of All-Time." *Journal of Quantitative Analysis in Sports* 2(1). https://doi.org/10.2202/1559-0410.1014.

Balreira, E. C., B. K. Miceli, and T. Tegtmeyer. 2014. "An Oracle Method to Predict NFL Games." *Journal of Quantitative Analysis in Sports* 10:183–196.

Bates, D., M. Machler, B. Bolker, and S. Walker. 2015: "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67:1–48.

Baumer, B., and P. Badian-Pessot. 2017. "Evaluation of Batters and Base Runners". Pp. 1–37 in *Handbook of Statistical Methods*

and Analyses in Sports, edited by J. Albert, M. E. Glickman, T. B. Swartz and R. H. Koning, Boca Raton, Florida: CRC Press.

Baumer, B., S. Jensen, and G. Matthews. 2015. "Openwar: An Open Source System for Evaluating Overall Player Performance in Major League Baseball". *Journal of Quantitative Analysis in Sports* 11:69–84.

Becker, A. and X. A. Sun. 2016. "An Analytical Approach for Fantasy Football Draft and Lineup Management". *Journal of Quantitative Analysis in Sports* 12:17–30.

Brooks, D., H. Pavlidis, and J. Judge. 2015. "Moving Beyond Wowy: A Mixed Approach to Measuring Catcher Framing." URL https://www.baseballprospectus.com/news/article/25514/moving-beyond-wowy-a-mixed-approach-to-measuring-catcher-framing/.

Burke, B. 2009. "Expected Point Values". URL http://archive.advancedfootballanalytics.com/2009/12/expected-point-values.html.

Burke, B. 2014. "Expected Points and Expected Points Added Explained". URL http://www.advancedfootballanalytics.com/index.php/home/stats/stats-explained/expected-points-and-epa-explained.

Cafarelli, R., C. J. Rigdon, and S. E. Rigdon. 2012. "Models for Third Down Conversion in the National Football League." *Journal of Quantitative Analysis in Sports* 8(3). https://doi.org/10.1515/1559-0410.1383.

Carroll, B., P. Palmer, J. Thorn, and D. Pietrusza. 1988. *The Hidden Game of Football*. New York, New York: Total Sports, Inc.

Carter, V. and R. Machol. 1971. "Operations Research on Football". *Operations Research* 19:541–544.

Causey, T. 2013. "Building a Win Probability Model Part 1". URL http://thespread.us/building-a-win-probability-model-part-1.html.

Causey, T. 2015. "Expected Points Part 1: Building a Model and Estimating Uncertainty". URL http://thespread.us/expected-points.html.

Citrone, N. and S. L. Ventura. 2017. "A Statistical Analysis of the NFL Draft: Valuing Draft Picks and Predicting Future Player Success". Presented at the Joint Statistical Meetings.

Clark, T. K., A. W. Johnson, and A. J. Stimpson. 2013. "Going for Three: Predicting the Likelihood of Field Goal Success with Logistic Regression". *MIT Sloan Sports Analytics Conference*.

Dasarathy, B. V. 1991. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*.

Deshpande, S. K. and S. T. Jensen. 2016. "Estimating an NBA Player's Impact on his Team's Chances of Winning". *Journal of Quantitative Analysis in Sports* 12:51–72.

Drinen, D. 2013. "Approximate Value: Methodolgy." URL https://www.sports-reference.com/blog/approximate-value-methodology/.

Eager, E. A., G. Chahrouri, and S. Palazzolo. 2017. "Using PFF Grades to Cluster Quarterback Performance". *Pro Football Focus Research and Development Journal* 1:4–14.

Elmore, R. and P. DeWitt. 2017. *Ballr: Access to Current and Historical Basketball Data*. URL https://CRAN.R-project.org/package=ballr, r package version 0.1.1.

Franks, A. M., A. D'Amour, D. Cervone, and L. Bornn. 2017. "Meta-Analytics: Tools for Understanding the Statistical Properties of Sports Metrics". *Journal of Quantitative Analysis in Sports* 12:151–165.

Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, United Kingdom: Cambridge University Press.

Goldner, K. 2012. "A Markov Model of Football: Using Stochastic Processes to Model a Football Drive". *Journal of Quantitative Analysis in Sports* 8(1). https://doi.org/10.1515/1559-0410.1400.

Goldner, K. 2017. "Situational Success: Evaluating Decision-Making in Football". Pp. 183–198 in *Handbook of Statistical Methods and Analyses in Sports*, edited by J. Albert, M. E. Glickman, T. B. Swartz, and R. H. Koning. Boca Raton, Florida: CRC Press.

Gramacy, R. B., M. A. Taddy, and S. T. Jensen. 2013. "Estimating Player Contribution in Hockey with Regularized Logistic Regression". *Journal of Quantitative Analysis in Sports* 9:97–111.

Grimshaw, S. D. and S. J. Burwell. 2014. "Choosing the most Popular NFL Games in a Local tv Market". *Journal of Quantitative Analysis in Sports* 10:329–343.

Horowitz, M., R. Yurko, and S. L. Ventura. 2017. *nflscrapR: Compiling the NFL Play-by-Play API for Easy use in R*. URL https://github.com/maksimhorowitz/nflscrapR, r package version 1.4.0.

James, B. 2017. "Judge and Altuve." URL https://www.billjamesonline.com/judge_and_altuve/.

Jensen, J. A. and B. A. Turner. 2014. "What if Statisticians Ran College Football? A Re-Conceptualization of the Football Bowl Subdivision". *Journal of Quantitative Analysis in Sports* 10:37–48.

Jensen, S., K. E. Shirley, and A. Wyner. 2009. "Bayesball: A Bayesian Hierarchical Model for Evaluating Fielding in Major League Baseball". *The Annals of Applied Statistics* 3:491–520.

Katz, S. and B. Burke. 2017. "How is Total QBR Calculated? We Explain our Quarterback Rating". URL http://www.espn.com/blog/statsinfo/post/_/id/123701/how-is-total-qbr-calculated-we-explain-our-quarterback-rating.

Kubatko, J., D. Oliver, K. Pelton, and D. T. Rosenbaum. 2007. "A Starting Point for Analyzing Basketball Statistics". *Journal of Quantitative Analysis in Sports* 3(3). https://doi.org/10.2202/1559-0410.1070.

Lahman, S. 1996–2017. *Lahman's Baseball Database*. URL http://www.seanlahman.com/baseball-archive/statistics/.

Lillibridge, M. 2013. "The Anatomy of a 53-Man Roster in the NFL". URL http://bleacherreport.com/articles/1640782-the-anatomy-of-a-53-man-roster-in-the-nfl.

Lock, D. and D. Nettleton. 2014. "Using Random Forests to Estimate Win Probability before Each Play of an NFL Game". *Journal of Quantitative Analysis in Sports* 10:1–9.

Lopez, M. 2017. "All Win Probability Models are Wrong Some are Useful." URL https://statsbylopez.com/2017/03/08/all-win-probability-models-are-wrong-some-are-useful/.

Macdonald, B. 2011. "A Regression-Based Adjusted Plus-Minus Statistic for NHL Players". *Journal of Quantitative Analysis in Sports* 7(3). https://doi.org/10.2202/1559-0410.1284.

Martin, R., L. Timmons, and M. Powell. 2017. "A Markov Chain Analysis of NFL Overtime Rules". *Journal of Sports Analytics* 4:95–105.

Meers, K. 2011. "How to Value NFL Draft Picks". URL https://harvardsportsanalysis.wordpress.com/2011/11/30/how-to-value-nfl-draft-picks/.

Morris, B. 2017. "Running Backs are Finally Getting Paid What Theyre Worth." URL https://fivethirtyeight.com/features/running-backsare-finally-getting-paid-what-theyre-worth/.

Mulholland, J. and S. T. Jensen. 2014. "Predicting the Draft and Career Success of Tight Ends in the National Football League". *Journal of Quantitative Analysis in Sports* 10:381–396.

Oliver, D. 2011. "Guide to the Total Quarterback Rating." URL http://www.espn.com/nfl/story/_/id/6833215/explaining-statistics-total-quarterback-rating.

Paine, N. 2015. "Bryce Harper should have Made $73 Million More". URL https://fivethirtyeight.com/features/bryce-harper-nl-mvp-mlb/.

Pasteur, D. and K. Cunningham-Rhoads. 2014. "An Expectation-Based Metric for NFL Field Goal Kickers". *Journal of Quantitative Analysis in Sports* 10:49–66.

Piette, J. and S. Jensen. 2012. "Estimating Fielding Ability in Baseball Players Over Time". *Journal of Quantitative Analysis in Sports* 8(3). https://doi.org/10.1515/1559-0410.1463.

Pro-Football-Reference. 2018. "Football Glossary and Football Statistics Glossary". URL https://www.pro-football-reference.com/about/glossary.htm.

Quealy, K., T. Causey, and B. Burke. 2017. "4th Down Bot: Live Analysis of Every n.f.l. 4th Down". URL http://nyt4thdownbot.com/.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Romer, D. 2006. "Do Firms Maximize? Evidence from Professional Football". *Journal of Political Economy* 114:340–365.

Rosenbaum, D. T. 2004. "Measuring how NBA Players Help their Teams Win". URL http://www.82games.com/comm30.htm.

Schatz, A. 2003. "Methods to Our Madness". URL http://www.footballoutsiders.com/info/methods.

Schoenfield, D. 2012. "What we Talk about when we Talk about War". URL http://espn.go.com/blog/sweetspot/post/_/id/27050/what-we-talk-about-when-we-talk-about-war.

Sievert, C. 2015. *pitchRx: Tools for Harnessing 'MLBAM' 'Gameday' Data and Visualizing 'pitchfx'*. URL https://CRAN.R-project.org/package=pitchRx, r package version 1.8.2.

Smith, D., S. Siwoff, and D. Weiss. 1973. "Nfl's Passer Rating." URL http://www.profootballhof.com/news/nfl-s-passer-rating.

Snyder, K. and M. Lopez. 2015. "Consistency, Accuracy, and Fairness: A Study of Discretionary Penalties in the NFL". *Journal of Quantitative Analysis in Sports* 11:219–230.

Tango, T. 2017. "War Podcast". URL http://tangotiger.com/index.php/site/comments/war-podcast.

Tango, T., M. Lichtman, and A. Dolphin. 2007. *The Book: Playing the Percentages in Baseball*. Washington, D.C: Potomac Book, Inc.

Thomas, A. C. and S. L. Ventura. 2015. "The Road to War". URL http://blog.war-on-ice.com/index.html%3Fp=429.html.

Thomas, A. and S. L. Ventura. 2017. *nhlscrapr: Compiling the NHL Real Time Scoring System Database for easy use in R*. URL https://CRAN.R-project.org/package=nhlscrapr, r package version 1.8.1.

Thomas, A. C., S. L. Ventura, S. T. Jensen, and S. Ma. 2013. "Competing Process Hazard Function Models for Player Ratings in Ice Hockey". *The Annals of Applied Statistics* 7:1497–1524.

Turkenkopf, D., H. Pavlidis, and J. Judge. 2015. "Prospectus Feature: Introducing Deserved Run Average (DRA) and all its Friends". URL https://www.baseballprospectus.com/news/article/26195/prospectus-feature-introducing-deserved-run-average-draand-all-its-friends/.

Wakefield, K. and A. Rivers. 2012. "The Effect of Fan Passion and Official League Sponsorship on Brand Metrics: A Longitudinal Study of Official NFL Sponsors and Roo". *MIT Sloan Sports Analytics Conference*.

Yam, D. R. and M. J. Lopez. 2018. "Quantifying the Causal Effects of Conservative Fourth down Decision Making in the National Football League." URL https://statsbylopez.files.wordpress.com/2018/01/quantifying-causal-effects.pdf, under Review.

Yurko, R., S. Ventura, and M. Horowitz. 2017. "Nfl Player Evaluation Using Expected Points Added with Nflscrapr". Presented at the Great Lakes Sports Analytics Conference.

Zhou, E. and S. Ventura. 2017. "Wins and Point Differential in the NFL". URL https://www.cmusportsanalytics.com/wins-point-differential-nfl/.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/jqas-2018-0010).