

# Evaluating NFL plays: Expected points adjusted for schedule

Konstantinos Pelechrinis<sup>1</sup>, Wayne Winston<sup>2</sup>, Jeff Sagarin<sup>3</sup>, and Vic Cabot<sup>2</sup> \*

<sup>1</sup> University of Pittsburgh

kpele@pitt.edu,

<sup>2</sup> Indiana University

{winston, NA}@indiana.edu,

<sup>3</sup> USA Today

jsagarin@attglobal.net

**Abstract.** “Not all yards are created equal”. A 3rd and 15, where the running back gains 12 yards is clearly less valuable than a 3rd and 3 where the running back gains 4 yards, even though it will not necessarily show up in the yardage statistics. While this problem has been addressed to some extent with the introduction of expected point models, there is still another inequality omission in the creation of yards and this is the opposing defense. Gaining 6 yards on a 3rd and 5 against the top defense is not the same as gaining 6 yards on a 3rd and 5 against the worst defense. Adjusting these expected points model for opponent strength is thus crucial. In this paper, we develop an optimization framework that allows us to compute offensive and defensive ratings for each NFL team and consequently adjust the expected point values accounting for the opposition faced. Our framework allows for assigning different point values to the offensive and defensive units of the same play, which is the rational thing to do especially in a league with an uneven schedule such as the NFL. The average absolute difference between the raw and adjusted points is 0.07 points/play (p-value < 0.001), while the median discrepancy is 0.06 (p-value < 0.001). This might seem negligible, but with an average of 130 plays per game this translates to approximately 125 points/season discredited. The opponent strength adjustment that we introduce in this work is crucial for obtaining a better evaluation of personnel based on the actual competition they faced. Furthermore, our work allows to evaluate special teams’ performance, a unit that has been identified as crucial for winning but has never been properly evaluated. We firmly believe that our developed framework can make significant strides towards computing an accurate estimate of the true monetary worth of a given player.

## 1 Introduction

Suppose it is a 3<sup>rd</sup> down and 15 yards for the Cowboys on their own 40-yard line and Ezekiel Elliott gains 12 yards. From the standpoint of Fantasy Football and

---

\* 1965-1994

yards-per-rushing attempt this is a very good play. In terms of the Cowboys' expected margin of victory in the game, this play is bad because the Cowboys will probably punt to the opposition. What is needed to properly evaluate each NFL play is a measure of the "expected point value" generated by each play. Cincinnati Bengal's QB Virgil Carter and his thesis advisor Robert Machol [5] were the first to realize the importance of assigning a point value to each play. Carter and Machol estimated the value of first and 10 on every 10-yard grouping of yard lines by averaging the points scored on the next scoring event. In *The Hidden Game of Football* [4] Carroll, Palmer and Thorn attempted to determine a linear function that gave the value of a down and yards to go situation. In this work we make use of the expected points model developed by Winston, Cabot and Sagarin [8, 9] (WCS for short) that models the game of American football as an infinite horizon zero-sum stochastic game. This model defines the value of the game in its current *state* to be the expected number of points by which the offensive team will beat the defensive team if the current score is tied. The state is defined by the triplet:  $\langle \text{Down}, \text{Yards} - \text{to} - \text{Go}, \text{Field Position} \rangle$ . This state definition results in 19,799 possible states. Since there are roughly 40,000 plays in an NFL season most states will occur rarely (if at all!). Examining the play-by-play data for the 2014-2016 seasons we identify that there are only 4,972, 4,989 and 4,966 states appearing in each of the season respectively. Furthermore, among these there is a prominent state (that is, the touchback state) in each season with almost 5x more appearances than the second most frequent state. This makes it challenging to use play-by-play data to accurately evaluate the value of each state, since many of these states do not appear in the data (or they appear very rarely). WCS details are beyond the scope of this work but in brief, WCS computes for each of the 19,799 states the value of the game and the corresponding mixed optimal strategies by using a linear programming approach [7] (Appendix A). For example, WCS provides a value of -2.575 for the state  $\langle 4, 4, 7 \rangle$ , which means that for a team that has a fourth down, at their own 7-yard line and need 4 more yards for a first down, they will be outscored by approximately 2.6 points from that point on. Using the values of the starting and ending state of a play we can obtain the corresponding play value. For instance, consider a team with the ball 1st and 10 on their own 23, which corresponds to a state value of  $v_1 = -0.014$ . Their first down play gains 4 yards and hence, the state now is 2nd and 6 on their own 27, which has a value of  $v_2 = 0.109$ . The value of this play is  $v_2 - v_1 = 0.123$ . More recently, other expected points models have appeared in the literature. For example, Yurko *et al.* [10] used a multinomial logistic regression to estimate the state values from play-by-play data, while Brian Burke [1] grouped *similar* states together and used the next score (i.e., which team scores next - offense or defense) to obtain an estimate for the current state's value.

Each one of these approaches clearly have its own merits and drawbacks. However, a crucial drawback that all of them exhibit is that none adjusts the value of a play for the strength of opponent. For example, a running play with a value of .10 points against a team with an above average rushing defense should

be more valuable than the same play against a team with a below average running defense. Adjusting for the strength of the opponent can provide us with a more holistic view of a play's value. Therefore, the major contribution of our work is determining the points per play for each team's running, passing, and special teams units, **adjusted for schedule strength**<sup>4</sup>.

In the following section we present our adjustment optimization framework, while Section 3 presents specific case studies that make use of this framework. Finally, Section 4 concludes our work, describing also possible future directions.

## 2 Adjusting for Opponent Strength

Schedule strength plays a crucial role in the number of games an NFL team wins. During 2015-2016 the easiest schedule was 4 points per game easier than the toughest schedule. This difference of 64 points is worth approximately 2 wins per season. Therefore, it is important to adjust the raw points per play that we showcased in the previous section, based on the strength of the schedule a team faces. For example, let us consider that the average running play gained 0.06 points (and the defense gave up 0.06 points). Suppose Team A averages 0.10 points per run but in each game played against running defenses that gave up 0 points per run. Then clearly Team A's running offense was better than 0.1 points per run, since it faced better than average defense.

We illustrate our adjustment procedure by demonstrating how to compute the adjusted points on running plays - for the rest of the play types (passes, punts, etc.) the method is the same. This method involves first calculating offensive and defensive ratings for the rushing offense and defense of each team. These ratings essentially capture how much better or worse per play a rushing offense/defense is compared to an average rushing offense/defense. We consider the unadjusted points  $v_i$  on each running play  $i$  to be the score of a single-play game between the offensive team's running offense and the defending team's running defense. Then we can predict for play  $i$  the point margin gained for the offense as:

$$\hat{p}_i = m_{rush} + r_{o_i} + r_{d_i} + \epsilon \quad (1)$$

where:

- $\hat{p}_i$  is the point margin gained for the offense for rushing play  $i$
- $m_{rush}$  is a constant representing the league average points per rush
- $r_{o_i}$  ( $r_{d_i}$ ) is the offensive (defensive) ability of the offensive (defensive) team of play  $i$
- $\epsilon$  is the error term of the prediction

Note here that a positive value for  $r_{d_i}$  indicates a below average rushing defense while a positive  $r_{o_i}$  indicates an above average rushing offense. To find the rushing (offensive and defensive) ratings for each team we need to solve the following constraint optimization problem:

---

<sup>4</sup> While similar metrics exist, their method is not public [2], while from the limited description they seem to still suffer from schedule strength biases (see Appendix B).

$$\begin{aligned}
& \underset{m_{rush}, \mathbf{r}_o, \mathbf{r}_d}{\text{minimize}} && \sum_{i=1}^N (v_i - (m_{rush} + r_{o_i} + r_{d_i}))^2 \\
& \text{subject to} && \sum_{j=1}^{32} n_{o_j} \cdot r_{o_j} = 0, \\
& && \sum_{j=1}^{32} n_{d_j} \cdot r_{d_j} = 0.
\end{aligned} \tag{2}$$

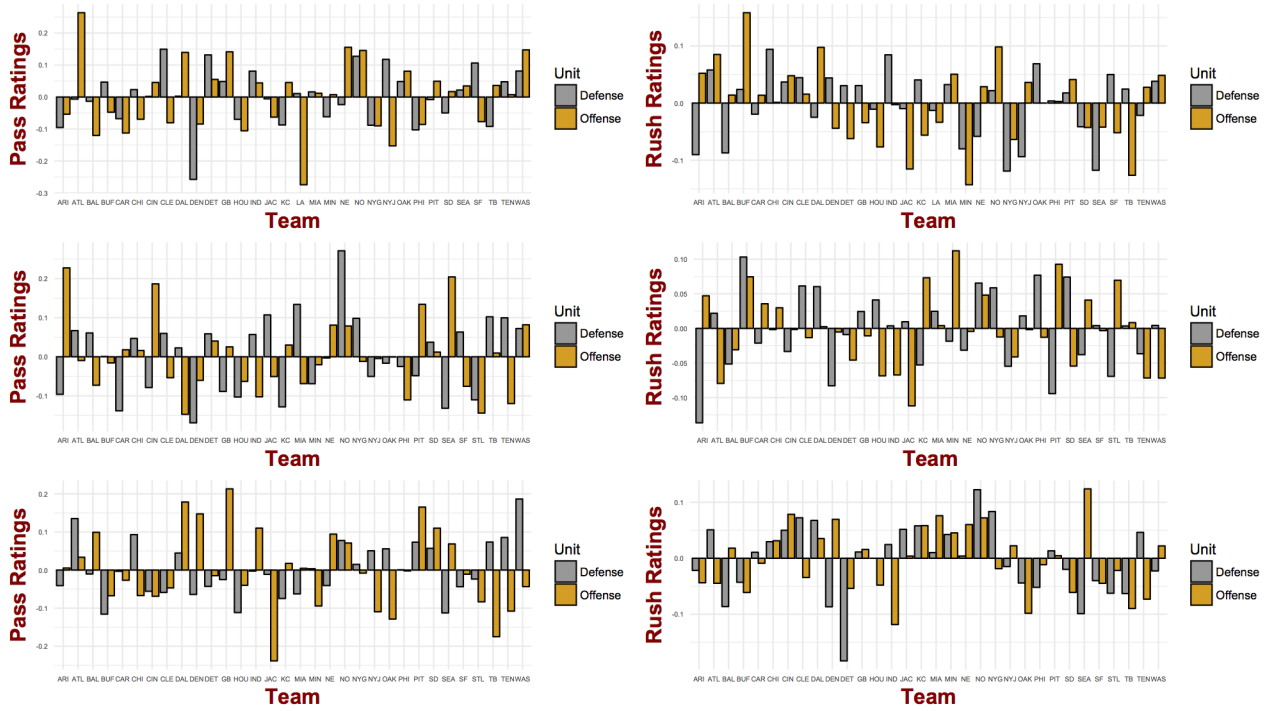
where  $n_{o_j}$  and  $n_{d_j}$  are the number of offensive and defensive rushing plays for team  $j$ ,  $v_i$  is the raw point value for play  $i$  and  $N$  is the total number of rushing plays. The two constraints normalize the weighted (by the number of plays) average of the teams rushing offensive and defensive abilities to both be equal to 0.

After solving optimization problem 2 we can use the team's rushing ratings to obtain the adjusted point value for each rushing play. The adjustment process will provide two adjusted point values, one from the standpoint of the offense and one from the standpoint of the defense. As it will become evident from our description below this is rational. Let us assume that a rushing play of team A against team B provided an unadjusted value of  $v$ . If the rushing defense ability of team B is  $r_{dB}$ , then the adjusted value for the offense is  $v - r_{dB}$ , i.e., we simply adjust the value of the play by subtracting the defensive team's rushing ability. Similarly, if the team A's offense rushing ability is  $r_{oA}$ , the adjusted value for B's defense is  $v - r_{oA}$ . For example, if the Cowboys are playing the Patriots and Ezekiel Elliott has a rushing play with an unadjusted value of 0.40 points and the Patriots' defensive rush ability is -0.05 points per play (which is better than average). Then the adjusted value of the play for the Cowboys offense would be  $0.40 - (-0.05) = 0.45$  points. Conversely, if the Cowboys' offensive rushing ability is +0.10 points per play then from the standpoint of evaluating the Patriots defense this play should have an adjusted value of  $0.40 - 0.10 = 0.3$  points.

## 2.1 What is the impact of the value adjustment?

To better understand the importance of our adjustment mechanism, we calculate the absolute difference between the raw/unadjusted values and the adjusted values for each play and for both the offense and the defense for the season 2014-2016. Overall, the average absolute discrepancy is 0.06 points per offensive play (p-value < 0.001) and 0.07 points per defensive play (p-value < 0.001). Offense includes rushing and passing offense as well as special teams plays, i.e., kicking, punting and field goal attempts. Similarly, defense includes rushing and passing defense as well as special teams plays, i.e., kick and punt returns as well as field goal defense.

This discrepancy might appear small in a first glance, however, its compounding effect on evaluating plays and players can be significant. In a typical game, there are approximately 130 plays (both offensive and defensive). A discrepancy



**Fig. 1.** The rushing and passing ability (both offensive and defensive) as provided by solving the optimization problem 2 for 2014 (bottom) to 2016 (top) NFL seasons

of 0.06 then translates to  $130 \cdot 16 \cdot 0.06 = 124.8$  points. Given that 33 points correspond approximately to 1 win (Appendix C), this translates to approximately 3.8 wins for every team not being credited correctly!

Figure 1 presents the solutions to the optimization problem 2 that provides us with the ratings for the teams with respect to the rushing and passing game (both offensively and defensively). The top two figures correspond to the 2016 season, where as we can see Broncos had the best passing defense, while the Rams had the worse passing offense, with the Falcons having the best passing offense with Matt Ryan having his MVP season. With respect to rushing, Bills and Cowboys were the two top rushing teams (LeSean McCoy's and Ezekiel Elliott's performances during that season had a lot to do with this).

Focusing on the results from the 2016 season it is interesting to notice that the Jets are top-3 rushing defense when we adjust the expected values. If one were to use traditional statistics that are typically used in broadcasting, such as yards/game, points/game etc. the Jets would be not make the top-10 of the list. The reality though is that the Jets had a very good rushing defense in 2016. Points against is also fatally flawed as an indicator of defense ability, since a poor offense often results in many points scored against the team as well. Why

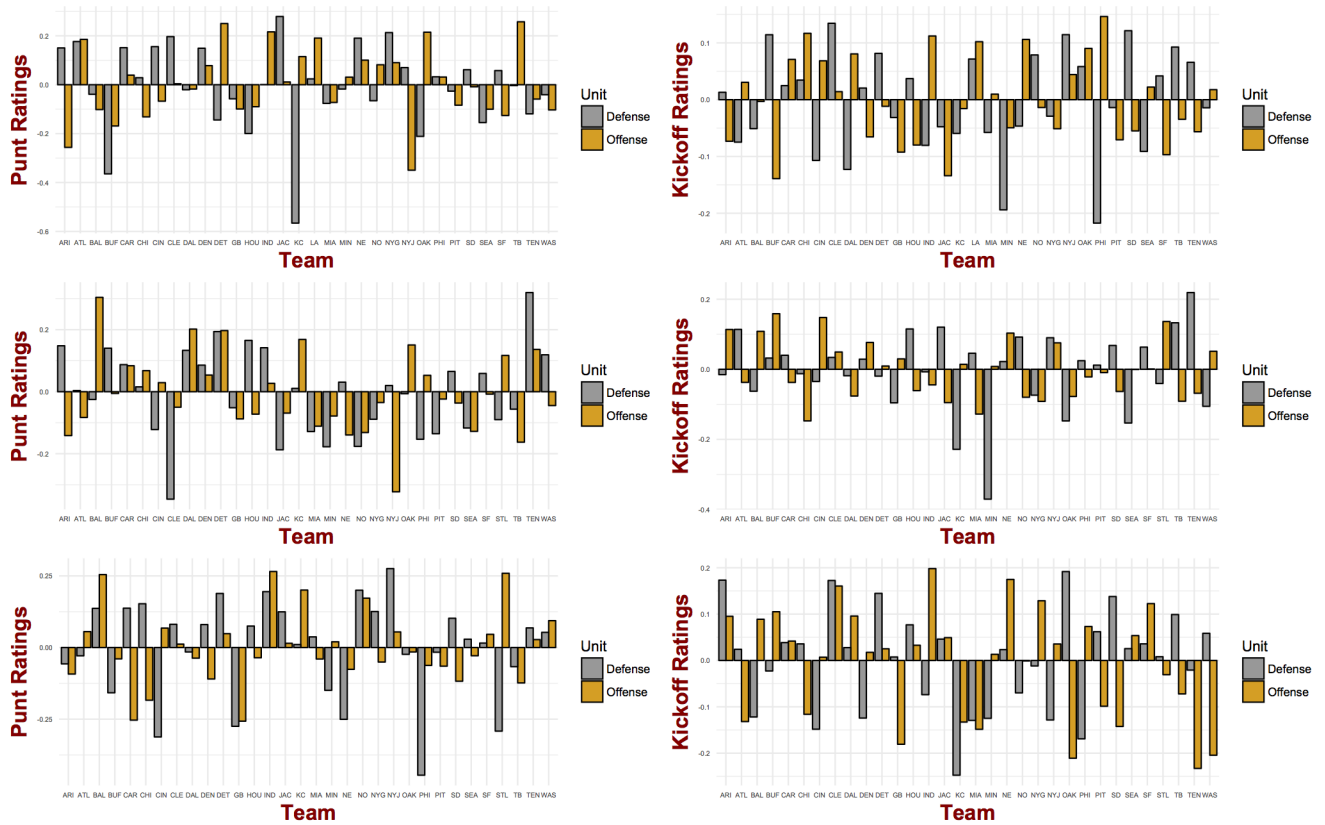
should a pick-6 count against the defense? Similarly, why a fumble that gives the opposing offense a short field should assign all the blame to the defense in case of an ensuing score? With expected points, the offense will be held accountable for these situations with negative value plays. Also yards given up don't reflect defenses that "bend but don't break". We hope that in the future more evaluation will happen with improved metrics that utilize the notion of point value per play.

This problem is even more pronounced when trying to evaluate the plays from the special teams. While a kick-off or punt return for a touchdown is undeniably a valuable play, it only captures a specific event (which also is rare). Metrics such as average net yards per punt can be extremely misleading. A 50-yards net punt from your own 45 will pin the opponent deep in their territory, while the same punt from your end zone will give great field position to the opposing offense. However, again in the latter case it is not the special team's fault that the offense couldn't move the ball down the field. So, these subtleties are hard (if not impossible) to be evaluated and quantified with the traditional statistics used by mainstream media. Figure 2 presents the special teams ratings for the 2014-2016 NFL seasons for kickoffs and punts respectively. The offense here refers to the team kicking/punting and ultimately covering, and a positive value indicates a good play. On the contrary the defense refers to the returning team, and a negative value indicates a better than average play.

To reiterate, one of the novelties of using expected point values is that we can evaluate the special teams play. Focusing again in the 2016 season we can see that the Cardinals were worse than average in almost all special teams facets of the game (they were slightly above average in kickoff returns). After accounting for the number of plays, we found that Arizona's terrible teams performed 40 points below average and cost Arizona nearly 1.2 wins in 2016! On the contrary, Chiefs special teams had the best punt return unit in the league. Just the punt return game for the Chiefs performed 40 points above average!

### 3 Case Studies

**Passing-vs-Rushing:** As it should be evident from the results presented earlier, on average passing plays are much better than running plays: in 2016 running plays averaged 0.069 points per play and pass attempts averaged 0.26 points per play. In prior years, the average points per run and pass attempt are very similar. So why not pass all the time? Isn't this what the analysis suggests? No, and this is where context needs to be integrated in the analytics process to interpret the results. First, the data that were used in the analysis are observational and were obtained based on the teams' specific game plan. If a team calls the same play every time we should expect its efficiency to drop [6] and thus, relying heavily on one type of play might lead to different observations. If the opposing defense is sure the offense will pass, they will be more likely to play a pass oriented defense that reduces passing effectiveness. In fact, we analyzed the expected points gained in 3rd downs from passing and rushing plays when the offense uses the shotgun formation, which is typically indicative of a passing play. For 3rd and

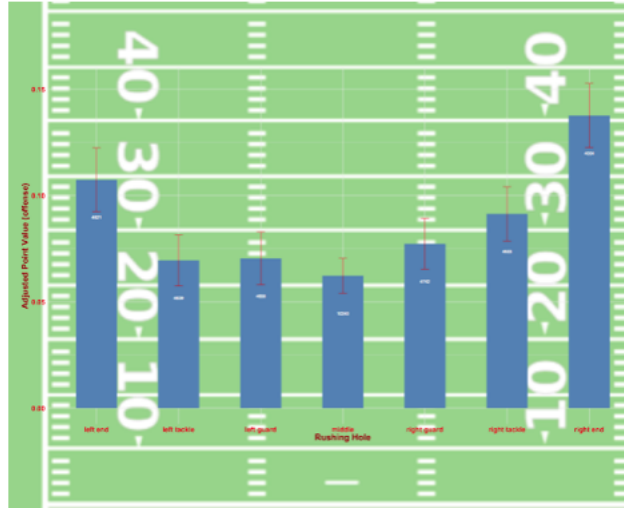


**Fig. 2.** Adjustment of point values per play allows us to evaluate special teams, an aspect of the game that while analysts agree on its importance it has been ignored in its quantification.

less than 5 yards to go, rushing plays gain on average 0.2 points, while passing plays gain on average 0.16 points. This is probably because the shotgun formation signals a probable pass and the defense is *surprised* by a running play (which again might be the reason for the increased efficiency of a rushing play from the shotgun).

Also passing is riskier than running. The standard deviation for the points per rushing play was 0.89 in 2016, while the same number for the passing places was 1.48. Riskier plays may make a successful sustained drive less likely, so the increased riskiness of passes can reduce their overall attractiveness and increase the need for a good rushing offense.

**Rushing Gaps:** Point values can also be used to evaluate rushing plays based on the rushing hole. During the 2014-2016 seasons, approximately 25% of all the rushing attempts were up the middle, while each of the rest rushing gaps accounted for approximately for 12% of the rushing attempts. Figure 3



**Fig. 3.** Rushing attempts closer to the middle provide less expected points per run as compared to attempts towards the end rushing gaps.

presents the average adjusted points per rush for the offense based on the location of the rushing gap. As we can see rush attempts towards the outside provide more value as compared to rushes up the middle! Of course, this can either be by design or because things up the middle break down and the running back is able to find space towards the outside.

One can also solve optimization problem 2 to obtain offensive and defensive ratings for each team and for each type of run. This can give teams several insights with regards to either their personnel and/or their opponent’s personnel. For instance, we have obtained the (offensive and defensive) rushing ratings per gap for the 2016 season (results are omitted due to space limitations), which show that the Cardinals were the best team in defending runs through the right guard, but they were not as successful through the left guard.

## 4 Discussion and Conclusions

In this study, we have introduced a framework for adjusting point for opponent strength the expected point values per play in NFL. Assigning a point value for a play is not a new concept and in fact there are several approaches that have been proposed for identifying these values, which we discussed earlier. However, none of them adjusts the values for opponent strength. Allowing 0.5 points on a rushing play from LeVeon Bell is certainly not as *bad* as allowing 0.5 points on a rushing play from Paul Perkins. The main contribution of our work is the design of the adjustment process that accounts for opposition strength. In particular, we formulate a constrained optimization problem, whose solution provides us with team ratings for each unit (i.e., rush offense, rush defense, punt return etc.)



that can then be used to adjust the raw point values for each play. We further show, how these values can be used to evaluate special teams' units. This has been identified as a crucial aspect for the success of a team but there has been little if any effort in evaluating and quantifying special teams' performance. One can easily utilize our results to develop a customized database which allows coaches to determine the effectiveness of their team's (and their opponents) various offensive and defensive plays in any game situation. For example, the 2016 Packers were a poor running team outside of the red zone. However, inside the red zone their rushing game was great, probably due to defenses being worried about the great Aaron Rodgers and his phenomenal passing game. We have created such an interactive database where users can explore some of the capabilities: <https://athlytics.shinyapps.io/nfl-pv/>. Furthermore, one does not have to use the WCS for obtaining the raw values. Any of the other (raw) point value models that we discussed in the introduction can be adjusted using the optimization framework we introduced in this work.

We have just scratched the surface in showing the uses of the Adjusted Points per Play (APP). A number of possibilities for further research are open. For example, if the coaching staff of a team has ratings for players on their ability on running and passing plays, then APP can be used to determine a football version of WAR (Wins Above Replacement). Let's assume that players are rated on a scale 1-100. For a team's offense, we would run two regression models (one for passing and one for rushing plays). Let us consider the regression model where the dependent variable  $Y$  is the team's Added Points per 100 Pass Attempts, while the independent variables are the average team rating for each of 11 offensive positions on passing plays. Suppose for our pass attempt regression the regression coefficient for left tackle was 0.10 and the 20th percentile of left tackle's passing rating is 30 (we consider the 20th percentile to be the replacement level [3, 9]). Then a left tackle with a passing rating of 90 who played on 500 pass attempt snaps would have added  $(90 - 30) \cdot 5 \cdot (0.10) = 30$  points above replacement on passing plays (note that the dependent variable is defined per 100 pass attempts). By a similar logic, using the rushing regression, assume the left tackle added 5 points on rushing plays. Then this left tackle added 35 points above replacement, or approximately 1 WAR. This type of analysis would allow NFL teams to move closer to identifying the real value of a player. Of course, building these models requires accurate player ratings which teams might have (either through third parties such as Pro Football Focus or through their own efforts) but unfortunately it is not available to us.

Finally, the adjustment framework we introduced in this work can be useful in evaluating college prospects. This is rather important since college football exhibits even higher degree of uneven strength schedule. In particular, if we use college football play-by-play data and determine an Adjusted Points Per Play Value for each NCAA team's rushing and passing defense, then we could adjust a college player's contributions based on the strength of the defenses they faced. This can further enhance the efforts of NFL teams to better evaluate draft picks based on their college performance.

## References

1. Expected points (ep) and expected points added (epa) explained. <http://archive.advancedfootballanalytics.com/2010/01/expected-points-ep-and-expected-points.html>, accessed: 2018-06-19
2. Football outsiders: Methods to our madness. <https://www.footballoutsiders.com/info/methods>, accessed: 2018-06-19
3. Prospectus toolbox: Value over replacement player. <https://www.baseballprospectus.com/news/article/6231/prospectus-toolbox-value-over-replacement-player/>, accessed: 2018-06-19
4. Carroll, B., Palmer, P., Thorn, J.: The hidden game of football. Warner books (1989)
5. Carter, V., Machol, R.E.: Operations research on football. *Operations Research* **19**(2), 541–544 (1971)
6. Pelechrinis, K.: The passing skill curve in the nfl. The Cascadia Symposium on Statistics in Sports (2018)
7. Shapley, L.S.: Stochastic games. *Proceedings of the national academy of sciences* **39**(10), 1095–1100 (1953)
8. Winston, W., Cabot, A., Sagarin, J.: Football as an infinite horizon zero-sum stochastic game. Technical Report (1983)
9. Winston, W.L.: *Mathletics: How gamblers, managers, and sports enthusiasts use mathematics in baseball, basketball, and football*. Princeton University Press (2012)
10. Yurko, R., Ventura, S., Horowitz, M.: nflwar: A reproducible method for offensive player evaluation in football. arXiv:1802.00998 (2018)

## Appendix A Intuition Behind the WCS Expected Points Model

In this appendix, we are going to present the main idea behind the WCS model for the value of each play. Let us consider for simplicity a 7-yard field as shown in the figure below. The rules of this 7-yard field football are simple and similar to the original game. We have one play to make a first down. It takes only 1 yard to get a first down. We have a 50% chance of gaining 1 yard and 50% chance of gaining 0 yards on any play. When we score, we get 7 points and the other team gets the ball 1 yard away from their goal line.

If we assume  $V(i)$  is the expected point margin by which we should win an infinite game if we have the ball on yard line  $i$ , then we can use the following systems of equations to solve for every  $V(i)$ :

$$V(1) = 0.5 \cdot V(2) - 0.5 \cdot V(5) \tag{3}$$

$$V(2) = 0.5 \cdot V(3) - 0.5 \cdot V(4) \tag{4}$$

$$V(3) = 0.5 \cdot V(4) - 0.5 \cdot V(3) \tag{5}$$

$$V(4) = 0.5 \cdot V(5) - 0.5 \cdot V(2) \tag{6}$$

$$V(5) = 0.5 \cdot (7 - V(1)) - 0.5 \cdot V(1) \tag{7}$$



**Fig. 4.** A 7-yard football field for illustrating the WCS point value model.

To derive the above equations we condition on whether we gain a yard or not. For example, for equation (3), suppose we have the ball on the 1-yard line. Then with probability 0.5 we gain a yard (and the situation is now worth  $V(2)$ ), while with probability 0.5 we do not gain this yard and the other team gets the ball one yard away from our goal line. Hence, their worth is  $V(5)$ , which for our team means that this situation is  $-V(5)$ . Equations (4)-(7) are obtained in a similar manner. For equation (7) we have to note that if we gain a yard (which happens with probability 0.5), we gain 7 points and the opponent will get the ball on its own 1-yard line (a situation worth  $V(1)$  to the opponent). Therefore, with probability 0.5 we will gain  $7 - V(1)$  points. By solving these equations we get:  $V(1) = -5.25$ ,  $V(2) = -1.75$ ,  $V(3) = 1.75$ ,  $V(4) = 5.25$  and  $V(5) = 8.75$ . Thus, in this simplified football game each yard line closer to our goal line is worth 3.5 points (or half a touchdown).

Adapting this method to the actual football game requires identifying the “transition probabilities” that indicate the chance of going from say, first and 10 on your own 25-yard line to second and 4 on your own 31-yard line. These transition probabilities are difficult to estimate simulations through the *Pro Quarterback* game<sup>5</sup> that allows the offensive team to choose one of 12 plays and the defensive team to choose one of 9 plays were used; it allowed us to generate a distribution of yards gained, which can then be used to obtain the corresponding recursive equations (similar to (3)-(7)). Today one could replicate this model using data from play-by-play logs, or (to avoid the sparsity problem of the situations) video game simulations in much of a similar manner that Pro-Quarterback was used by Winston, Cabot and Sagarin.

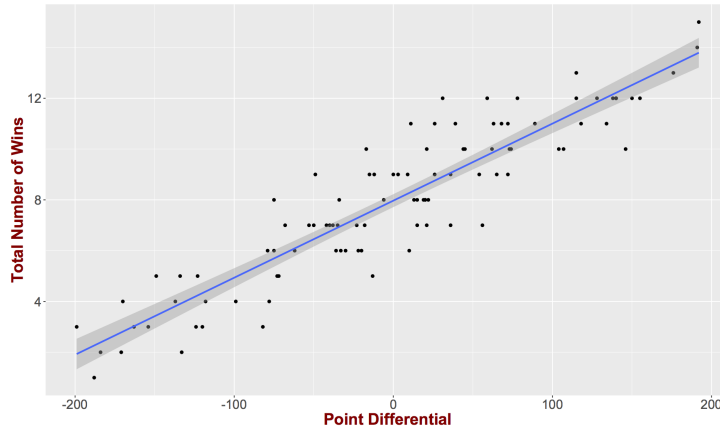
## Appendix B Other Adjustment Approaches

While there is not any academic literature (to the best of our knowledge) tackling the adjustment for schedule strength, there are sports websites (e.g., Football

<sup>5</sup> <https://boardgamegeek.com/boardgame/9821/pro-quarterback>

Outsiders) that have their own defense-adjusted metrics. The exact description and methodology behind the calculations of these metrics is not public (possibly available through subscription), and only a short description is available: “*passing plays are also adjusted based on how the defense performs against passes to running backs, tight ends, or wide receivers. Defenses are adjusted based on the average success of the offenses they are facing.*” [2]. However, it is not clear how this average success (both offensive and defensive) is calculated. If it is a simple average, then this is still biased from the competition faced by the team. Simply put, one can imagine that our optimization approach resembles a very large number of iterations where the team ratings are re-calculated, while using a simple average is one single iteration of a similar process.

## Appendix C 33 Points Equal 1 Win in NFL



**Fig. 5.** There is a linear relationship between the point differential of a team and its total number of wins.

Where did the equivalence of 33 points to 1 win come from? An easy way to understand this is by examining the relationship between the point differential for a team (i.e., points for minus points against) for a whole season with the total number of wins for the team. Using the data for the season 2014-2016 we obtain the following relationship. Fitting a linear regression model on the data we obtain an intercept of 7.9 and a slope of 0.03. Simply put, a team with 0-point differential is expected to have an 8-8 record, while every increase in the point differential by 1 point, corresponds to 0.03. Equivalently, 1 win corresponds to  $\frac{1}{0.03} \approx 33$  points.