STOP SHY OF THE FIRST DOWN

On the first down, an NFL player should stop shy of the first down marker unless they can make several extra yards on the play. The decision is clarified when we introduce the notion of implied yards per possession.

1. The First Down. Even a casual observer of the NFL will notice that commentators, fans, and coaches alike encourage their players to lunge, stretch, hurdle, or bulldoze their way across that last yard to complete the first down. Getting the first down is a motherhood issue. It's applauded. For example, how often do we see a wide receiver break the imaginary yellow plane in an acrobatic fashion, often one arm outstretched, as they careen out of bounds?

But using data from the 2009-2013 NFL seasons compiled by Ben Dilday [Dilday, 2016], I question this "decision" (or perhaps we should call it a non-decision) to complete the first down on the first attempt. We'll see that teams are better off stopping shy unless a few extra yards are thrown into the bargain.

2. The Choice. Let me be clear about this scenario. First, we are only talking about the first of four downs.[1] Second, I assume a player with the ball is certain he can't progress terribly far beyond the first down marker. We're talking about that last effort to break the plane whereas obviously, a player in full flight who is likely to achieve the first down and also many extra yards would be silly to stop one yard short.

Under my stated conditions, we can assess whether choosing to complete the first down helps or hinders. Lest the simplicity of the argument be lost, I shall assume that the ball carrier is determining which of these two outcomes will ensue:

1. First down and ten yards.
2. Second down and one yard, one yard further back.

The first possibility seems to be the strongly preferred option, given two advantages:

- The field position is advanced one yard.
- The team will have one extra down.

However, there is one big disadvantage. The number of yards required for the next down reset will be ten, rather than one. I ask the reader to focus on this downside when considering the completion of the first down.

Conversely, a near first down miss probably isn't a failure. You're going to get the first down in most cases, and further down the field. Indeed in comparing the two options, it's easier to compare a completed first down (option one - we break the plane) with a *hypothetical* first down completion *on the next play or the subsequent one* (option two - we stop short). Both game states differ only by translation down the field.

Due to this symmetry, it's clear that the trade-off comes down to how much we value yards gained as compared to the small chance of losing possession before this hypothetical eventuates. To formalize this and provide what I hope is a simple perspective, this article is broken down as follows.

[1] The "first down" refers to the first of four attempts to advance ten yards, and also refers to the act of achieving the 10-yard advance, thereby resetting the down count back to the first down. Due to penalties, the first attempt may sometimes require 5 or 15 yard gain in order for a completion of a first down to occur, but usually completion of a first down is synonymous with a ten yard gain or more.

1. A comment on ways to estimate the value of possession, measured in yards.
2. An estimate of *implied yards per possession* when a player chooses to complete the first down, based on second and third down completion statistics.

We shall find that the second number exceeds the first by as much as a factor of two. Players, and by implication coaching staff, are implicitly favoring possession over field position in a way that does not help their chances of winning the game.

Before continuing, let me remark that this finding applies to defensive players as well as offensive. Indeed defensive coordinators should not be cursing players who allow a ten yard gain when a nine yard stop seemed possible.[2]

3. The value of possession, in yards. The value of possession versus yards is an old topic. I'm not going to get into endgame analysis or special situations, nor the value of the clock when one team is ahead - though I think it should be clear how the rationale extends to these situations.

Rather than present a definitive calculus for possession value, I'll mention several ways to come at this that are all more than adequate. I would prefer the reader make the choice between accuracy and simplicity of calculation, because evaluation of the decision to stop shy of the first down (my only objective) is not predicated on a precise estimate of points per yard.

I shall be content with somewhat typical field and game position, which, if you prefer, can be assumed to occur in the first three-quarters of a relatively even game. I assume that the game is not so lopsided that one team has significantly diverged from a strategy that maximizes mean points scored.

The value of field position in points was considered by Virgil Carter and Robert E. Machol back in 1970 [Carter and Machnol, 1970]. Their table of field position values is presented in Figure 1. While some improvements are possible, we can read off the difference between 15 and 85 yards. This provides a very loose estimate of how expected points varies - namely one point per 12 yards. Combined with some reasoning about punts, this early analysis may suffice.

Coming at things another way, the value of a yard is arguably easier to estimate on fourth downs rather than first - at least if the team is in field goal range and will with certainty kick. Then, the value of a yard can be inferred from field goal kicking probability - for instance 15 yards might be worth a point if it drops from 0.75 to 0.42, as we move from a 45 yard attempt to a 60 yard shot.[3]

Benjamin Morris points out that the value of "a" fourth down yard (or at least some yards on the field, within reasonable field goal range) changed between 2004 and 2014 [Morris, 2015]. As kickers got better, the value of a yard on fourth down decreased.

The value of a yard on first down might be viewed as an averaging of the fourth down yardage values. The intervening plays smooth over some the differences, and extend the validity of the approximation to a greater range of typical field positions.

An alternative approach suggested by Brian Burke computes the the expected value of the next score by either team as a function of both field position and down count.[4] For example, Figure 2 shows the expected points as a function of field position and we read an approximate value of possession from this plot or the accompanying

[2]Some appreciate this already, based on their initial reaction to this idea.

[3]The possibility of a runback increases this slightly.

[4]A small flaw in the use of "expected points" is the value of possession conceded after a score is made - though as can be inferred on the plot this is small near the region where the defence are likely to restart play. Thus, the notion of expected next score is quite similar to a value function approach, as is common in control theory. The discrepancy does not warrant an extended discussion here since

TABLE I

THE EXPECTED POINT VALUES OF POSSESSION OF THE FOOTBALL WITH FIRST DOWN AND TEN YARDS TO GO FOR VARIOUS TEN-YARD STRIPS

Center of the ten-yard strip (yards from the target goal line): X	Expected point value: $E(X)$
95	−1.245
85	−0.637
75	+0.236
65	0.923
55	1.538
45	2.392
35	3.167
25	3.681
15	4.572
5	6.041

FIG. 1. *Table of field position values from Carter and Machol [Carter and Machnol, 1970].*

data, as follows.

Notice that 66 yards (i.e., our own 34-yard line) corresponding to +1 points, assuming we have possession. Now imagine we punt it to our opposition. They will start their next play on their own 34-yard line if we achieve only a mildly disappointing net punt of 32 yards. That would leave them at +1 points. Net, we have lost two points.

To emphasize the role of the punt, notice that the intercept with the x-axis occurs near the 85 yard line. A Herculean punter who pins an opposition to the 15-yard line when kicking from his own 15-yard line has saved his team any loss at all. Should punters ever become better than this, football might degenerate into punting back and forth repeatedly.

Nearer the other end of the field, failure to secure a first down forces us to settle for a field goal. The loss may be greater than two points, on average, though only if we are very close to scoring. It is roughly equal to the probability of a touchdown, multiplied by four points, assuming that the field goal is a near-certainty.

We also note in Figure 2 that the value of possession on first down varies a little less than it does for fourth down, as we expect. Granted, there is still some variation here that has not been eliminated completely. Look closely and you will see the field goal effect is still there - intermingled with touchdown possibilities.

We can read the value of a yard too. You can see that between the red zones (i.e. not within 20 yards of the end zone) it takes about 60 yards to go from four points to zero, or 15 yards per point. Switching from professional to college football doesn't change this very much, as noted by Saiem Gilani [Gilani, 2020].

The takeaway: from the value of possession in points (not much more than 2) and the number of yards equating to a point (around 15) we reason that, as a rule of thumb, a possession is *unlikely to be worth substantially more than 30 yards.*

we require only a rough estimate of possession value to make the primary point. See [Cotton, 2021] for more estimates.

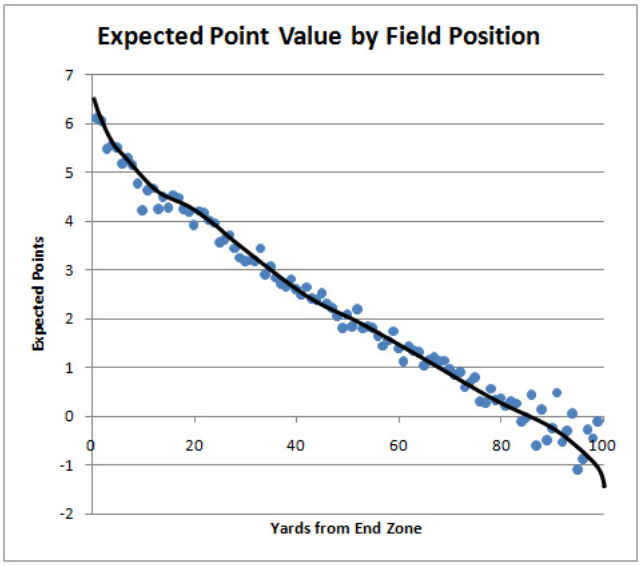

FIG. 2. *Expected value by field position, from an article by Brian Burke [Burke, 2008]. The slope is indicative of the value of a yard on the first down.*

4. Third and one. Armed with the basic rule that possession is worth 30 yards, give or take, we return to our key strategy question and reconsider "opting out" of a completion of the first down on the first attempt. This choice, which I have asserted is the superior one, might lead to a first down completion on the next play. Of course, we might also find ourselves in a third and one - a situation we must therefore analyze.

Strictly speaking these aren't the only possibilities - and the reader will note there is a small chance of a loss on the second down play, or yards lost, or even a catastrophic turnover. However one can, for the sake of argument, assume a rushing play, where the probability of the worst case scenarios is greatly diminished.[5]

Whether running or passing is chosen, the upside of "option two" is the number of yards gained on a successful second and one play (or if needed, a successful third and one play).

For avoidance of any doubt, I refer to the conditional averages, not the average including unsuccessful plays that don't advance the line of scrimmage. The data reveals that when rushing, this conditional average gain for a third down play is about five yards. When passing, it is 12.75 yards.

This is a large discrepancy between passing and running and as an aside, it suggests that teams might consider passing on third and one more often than they do. Passing plays are only successful 61% of the time, versus 72% when rushing. However, an 11% chance of lost possession corresponds to only a few yards, according to our analysis above. And this gets swamped by the massive 7.75 yard differential when passing.

The possible existence of sub-optimal third and one strategy is certainly interesting, and may even occlude the benefits of stopping shy. However for now the only thing we need is that a successful third and one results in an advance of field posi-

[5]Turnovers are increasingly rare in NFL games and averaged 2.6 per game. The majority are from interceptions, not fumbles on running plays.

tion of almost five yards, even if you choose to run the ball. Perhaps that's worth rephrasing for those who fear the third down. *Third and one is probably first and six, on average.*

(That's especially true if play restarts with less than one yard to the first down marker. I do not consider second and inches or third and inches plays in this note because in the data used, distances were rounded.[6] However, it is clear that these possibilities only strengthen the case for stopping shy of the first down.)

5. Second and one. With that in mind, let's roll back one play. How should we feel about finding ourselves in second and one? When we look at second and one plays, rather than third and one, we find that the rushing play percentage goes even higher - up to 80% - though the average number of yards drops slightly (down to 4.73). That isn't surprising, since the defensive team has less incentive to prevent the first down completion.

Another aside: the differential between passing and rushing yards gained decreases, as compared with the same differential for third and one, with passing leading to only 6.5 yards of gain on average compared with 7.75 for passing on third down. The natural urge to pass on second and run on third might be working against the better interest of teams.

That's interesting - but not crucial to the case I make for stopping shy of the first down. What's important, and now evident by multiplication (assuming conditional independence of third down play outcomes) is that if you tell your team to run on second and one, and then again from third and one (should that be necessary) then you will have a 94.5% chance of getting the first down. In the process you will advance an average of 4.75 valuable yards.

Perhaps it is apparent why you don't want those chains moved on the first down.

6. Implied value of possession. Can we imagine teams stopping shy on the first down deliberately?

Our star receiver takes a catch a yard short of the first down. Flat-footed, he turns to see a defensive player bearing down at great speed. Risking a season-ending injury he can, most certainly, dive forward with outstretched hands and make the hero play - securing the first down. Alternatively, he can casually step out of bounds, leaving his team at second and one.

Or perhaps a tight end has broken one tackle and staggers toward the first down marker dragging a defender who has grasped his leg. Should he break the plane or voluntarily stop his progress, if he knows he won't get any extra yards beyond the first down?

For a potentially controversial topic, the calculus is alarmingly easy and I frame it in terms of implied yards per possession. Using the values above, the wide receiver's lunge suggests that a 5.5% chance of losing possession in this series of downs is more important than 3.75 yards of field position.

He is wrong! Since 5.5% is roughly 1 in 18, this means that possession must be 17 times more important than 3.75 yards. The receiver implies a value of possession of 17*3.75=63.75 yards! But there is no way on God's green football field that a possession is worth over sixty yards. It is closer to half that number, as we have seen.

[6] As another aside, yardages are a little "sticky" because officials have a tendency to place the ball close to a yardage line. Indeed on first down there is a tendency to cheat one team out of half a yard so as to make measuring the next ten yards visually easy (thus avoiding excessive use of the chains which can slow play).

Extra yards on first down	Implied value of a possession	Assessment
0	64 yards	Incorrect
1	48 yards	Incorrect
2	30 yards	Break-even
3	13 yards	Correct

TABLE 1

The value of a possession, measured in yards, implied by a player's "decision" to achieve the first down on the first down. This assumes that a player might instead stop one yard shy of the first down marker, leaving his team at second and one. It also assumes a "typical" game and (middle) field position, and that a punt will be taken on fourth down. The implied yards per possession is conservative for several reasons. It assumes running plays will be chosen on second and third downs - though our discussion suggests this is sub-optimal. It doesn't take into account gains of less than a yard on second down. It also ignores the possible option value of running or passing on the fourth down, and it applies a conservative estimate of a gain of 4.75 *yards conditional on success (on either down). This represents a conditional mean gain of* 3.75 *yards - averaged over* 17 *of* 18 *occasions when possession is maintained. See Table 2 for a breakdown by play strategy.*

7. Extra yards on first down. It becomes difficult to justify achieving the first down even if some gain is made. Some values are tabulated in Table 1. If possession is valued at 30 yards, it is clear that players should "decline" the first down (voluntarily stop progress of the ball) quite often - and not just to reduce the risk of injury.

The calculus for the 11-yard gain reads $2.75 * 17 = 46.75$, which is still way too high a value (in yards) to put on possession. However more detailed variations on this calculation are presented in [Cotton, 2021] where a small turnover probability is introduced - though this does not materially alter the findings.

So, contrary to commonly accepted wisdom, a celebration is generally only warranted if a player can advance two or three yards past the first down marker. A marginal first down completion could be setting the team up for failure.

Conversely, defensive players need not suffer one extra concussion to bring about an abrupt deceleration of the ball carrier. Let them get the ten yards, or eleven if necessary. And the implication extends beyond individual player decisions. Offensive and defensive teams should design plays to make the +13 and +9 yard gains more likely, and the middle ground less so.

A more fine-grained analysis assumes a choice of play style on second and, if necessary, third down. One can also introduce empirical probabilities of yardage losses on the second down. The obvious extension of our simple argument leans on all empirical completion and conditional yardages. The findings are reported in Table 2. This lists potentially viable alternatives to obtaining the first down, even when at least one extra yard is possible.

Compared to Table 1 these numbers are slightly more conservative, from the perspective of a team looking to adopt a radical stopping-short strategy. Yet they ignore the small possibility of a touch-down on second or third down, which is yet another reason to stop shy. Even without this consideration, and with the possibility of an interception added to the mix, a team might still consider curtailing a 12-yard first-down play.

It may be argued that stopping precisely one yard shy is difficult - though one is inclined to believe that top tier players capable of precise routes are up to the task. Table 2 reveals that there is margin for error. Stopping two yards shy can also be beneficial.

Good passing teams should pay particular attention. They might go so far as to

Extra yards	Yards to go	On second	On third	Implied
1	1	rush	rush	43
1	1	rush	pass	43
1	1	pass	rush	44
1	1	pass	pass	43
1	2	pass	rush	39
1	2	pass	pass	37
1	3	pass	pass	32
2	1	rush	pass	31
2	1	pass	rush	36
2	1	pass	pass	37
2	2	pass	rush	32
2	2	pass	pass	31
3	1	pass	pass	31

TABLE 2

A team given the opportunity to gain a first down and extra yards in the process might nonetheless choose to stop shy with 1*,* 2*, or even* 3 *yards to go. The decision is expressed in terms of an implied belief in the value of a possession, measured in yards. For example if possession is believed to be worth less than* 30 *yards, then a team might "decline" a* 12*-yard first down completion by stopping two yards shy instead, assuming they plan to pass on second down (though very often, we see teams rush on second down - perhaps based on the assumption that obtaining the first down is the primary mission.)*

eschew a 13-yard first down completion! They can instead adopt the seemingly brave strategy of passing on second and third down with one yard to go. Though this seems counter-intuitive, it speaks to the relatively high frequency of passing gains of less than nine yards (as when receivers run slant routes, for example).

8. Evidence of poor strategy. Now that we know what is optimal, let's take a look at what happens on the field. Figure 3 show a histogram of first down yards gained rushing, where for simplicity we are restricting attention to cases where it is first and ten. One would think that with a rushing play, the offensive ball carrier would have good ability to aim for nine yards or 12, but not accidentally end up in between.

The data, which shows *some* mass moved from 10-yard gains back to nine yard gains, seems to indicate one of two things:

1. Some offensive players are aware that second and one is better than first and ten, and they are deliberately acting so as to achieve more nine yard gains than they otherwise would (though not to the extent they should).
2. Defensive players are trying too hard to prevent the first down, mistakenly believing that a stop at nine yards is better than conceding ten.

I'm not sure how we can disentangle these two effects using only this data - but fortunately we don't have to. It is abundantly clearly from watching live games that the latter is more plausible. There is plenty of room for improvement in strategy by both offense and defense.

From the perspective of the offensive team, the fact that 12-yard advances are less common than 11-yard advances is also a clear sign of poor strategy. That mass on the 11-yard gain should be moved back to 9 yards. Injury risking heroics used to get to 10 yards should, in fact, be reserved for going from 11 to 12, or 12 to 13, when the opportunity arises.

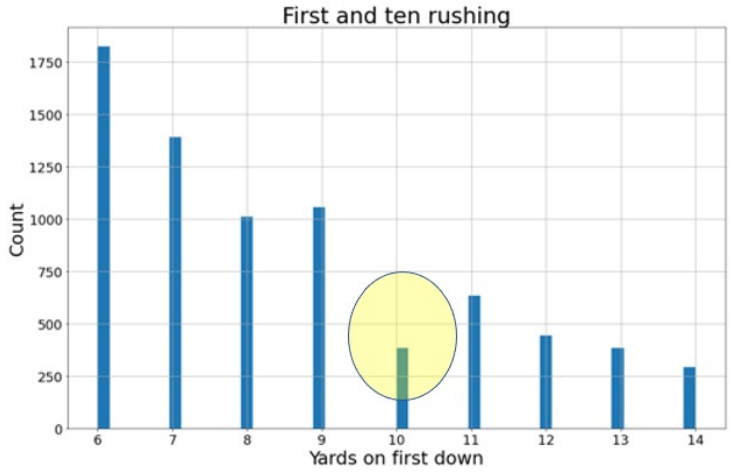

FIG. 3. *The empirical distribution of yards gained on the first down for rushing plays. There is potentially some evidence here that some, but not all, offenses appreciate the nine-yard gain versus ten. On the other hand, this might also suggest poor defensive strategy, with players trying very hard to prevent the first down. (Anecdotally, the latter explanation is more in line with a small survey of video replays, where the intent can usually be determined.)*

Defensive teams can no doubt benefit from coverage patterns that deliberately allow the first down if a player is likely to make nine yards. These changes, both to plays and execution, are likely to bring other benefits, such as reducing the chance of being caught flat footed by a long pass, or completely missing a tackle when trying to stop a ball carrier at nine yards rather than ten.

A more scathing indictment of strategy, both offensive and defensive, is delivered by Figure 4 which shows first down yards gained on the rarer occasions when, due to a defensive penalty, we are at first and five yards to go. Here our analysis suggests that a four yard gain is to be preferred to five, but there is an extremely pronounced preference for the former when passing plays are chosen.

What are they thinking?! My data suggest that teams' chances of making four yards when they want it is significantly higher than their chance of making five yards (when they think they want it) - about 10% higher. For this, we have to use second down data due to a dearth of first and four situations. But it is clear that teams are running riskier plays to achieve a worse outcome.

9. Conclusion. The lunge for the first down is part of football culture, and it seems almost distasteful to point out that this act fails to assist a team's chances of winning - at least on the first down. Instead, players should slide, run out of bounds, or otherwise stop the advance at the nine-yard line.

Admittedly this makes for quite the break in tradition. It is perhaps unfortunate that under optimal strategy, the great game of football is not as clean as fans might like. Team advantage is not mononotic in yards gained on the first down.

Setting aesthetics aside, the optimality of stopping shy might also be obscured by mental "chunking": the breaking down of the offense's task into one first down completion after the next. (This mindset is seemingly prevalent, but ignores the fact that if a team stops at nine yards, they'll probably make more yards per first down completion - and thus need fewer of them per touchdown.)

I have framed the decision on whether to stop shy as a trade-off between yards and

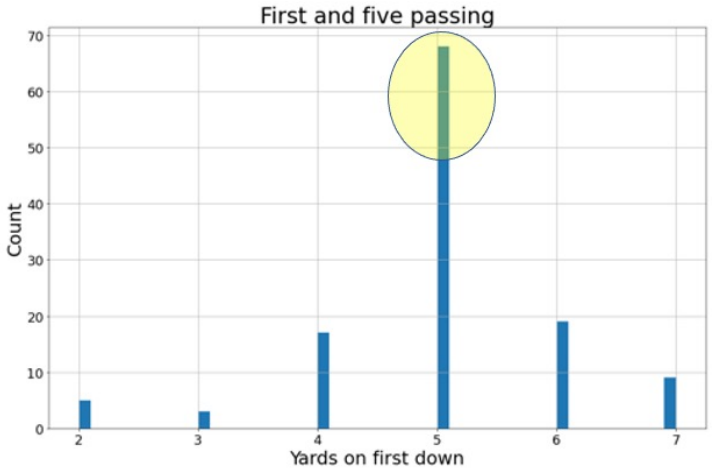

FIG. 4. *Yards gained on first down when there is five yards required. This scenario results from a penalty applied to the defense. Though the sample size is much smaller, this provides an even stronger indication that teams value a first down rather than stopping shy with one yard to go, since on a passing play the offensive team can plan the route. However, as with Figure 3 there is ambiguity. Video replays provide a more compelling condemnation of strategy.*

possession. Very often the likely yardage gain outweighs the small risk of being forced to punt or settle for a field goal. A player who stretches out his arm to break the plane of the first down marker is only justified in doing so if a possession is believed to be worth fifty or sixty yards. This strains credulity, given that the value of possession is typically half that number, however computed.

But there are other informal ways to convince skeptical players and coaches. Indeed, stopping at nine yards can be seen as a way of maintaining possession - whereas completion of the first down sets up the stiffer challenge: the necessity to make ten yards on three plays. Even coaches and players who believe in "possession at all cost" should come around to this way of thinking.

Perhaps offensive coordinators could ask their running backs if they would prefer five downs to make eleven yards (with 95% probability), or three to make ten.

I hope that this provokes a more rigorous treatment of offensive and defensive strategy, possibly starting with the reproduction and critique of these findings [Cotton, 2021]. Playbooks need to be rewritten, with a view to the importance of the nine yard gains. And though I have not tried to cover all situations, this should not present an excuse.

Many cases deserves careful consideration once we stray outside "typical" field and game situations, or introduce clock management. When touchdowns have a high probability, remaining downs are more valuable - but so is the marginal value of a yard. A future analysis might consider more granular yardages - and even the possibility of stopping inches shy of the first down marker on the second down, not only the first.

REFERENCES

[Burke, 2008] Burke, B. (2008). Expected Points. https://web.archive.org/web/20210302074507/http://archive.advancedfootballanalytics.com/2008/08/expected-points.html.

[Carter and Machnol, 1970] Carter, V. and Machnol, R. E. (1970). Operations research on football. Technical report, Northwestern University.

[Cotton, 2021] Cotton, P. (2021). First Down Repository. https://github.com/microprediction/firstdown.

[Dilday, 2016] Dilday, B. (2016). NFL Markov. https://raw.githubusercontent.com/microprediction/nflMarkov/master/inputData/pbp_nfldb_2009_2013.csv.

[Gilani, 2020] Gilani, S. (2020). College Football Expected Points Model Fundamentals. https://www.tomahawknation.com/florida-state-football-fsu-noles/2020/6/16/21212025/series-on-sports-analytics-college-football-expected-points-model-fundamentals-part-3.

[Morris, 2015] Morris, B. (2015). Kickers are Forever.