Tackling Endogeneity: Estimating Optimal Pass Rate in the NFL Using Instrumental Variables[*]

Daniel S. Putman[†] Tor N. Tolhurst[‡]

This Draft: February 27, 2025
Click Here For Most Recent Version

Abstract

What is the optimal rate of passing in professional football? We derive a simple model of playcalling efficiency in the NFL. Efficiency is highest when the play comes as a surprise and falls when a more expected play is called. That is, playcalling efficiency is determined by a usage curve. We use an instrumental variables strategy to identify these usage curves on early downs—estimating the effect of passing, expected passing, and their interaction on per-play efficiency. We propose two new instruments, the first of which is based on fumbles lost. More specifically, we argue that conditional on total cumulative fumbles by each team, cumulative fumbles lost serves as a valid instrument in each of these cases. The second exploits documented negative autocorrelation in playcalling: the play after a pass tends to be a rush, and *vice versa*. Using 15 years of play-by-play data, we use this strategy to estimate linear usage curves, from which we recover the optimal early-down passing rate. On first and second down, coaches have their quarterbacks drop back to pass about 51.7% of the time, whereas the optimal rate is around 59.1%. These results imply coaches deviate from the optimal pass rate, overestimating the value of surprise.

Keywords: Football Analytics, Decision-Making under Risk and Uncertainty, Causal Inference, Usage Curves, Mixed Strategies

JEL Codes: L83, Z20, L19, D81, C72

[*] This paper has benefited from a conversation with Peter Hull and excellent research assistance from Samantha Fenton. All errors herein are our own.

[†] University of Pennsylvania

[‡] Michigan State University

1 Introduction

The optimal pass-rush ratio in professional football has long been contentiously debated. The conventional wisdom instructs coaches to "establish the run," as it is safer, and sets up the pass. More recent analytical work suggests a more pass heavy approach (Schatz, 2003). Alamar (2006) re-frames this debate as the passing-premium puzzle: National Football League (NFL) teams rush an approximately equal number of passing and rushing plays, despite higher returns to passing. Over the years, many approaches have been proposed to rationalize the apparent difference in optimal pass rates and the observed rate of passing. These explanations include risk aversion and defensive adjustments (Rockerbie, 2008; Jordan et al., 2009; McGarrity and Linnen, 2010). The availability of play-by-play data to the public has increased dramatically since the majority of work was done. Open-source packages now allow anyone to pull data directly from league sources. This increased access has allowed for the development of a robust public analytics discussion.

Since 2006, passing has increased dramatically in the NFL, around 4.5 percentage points (Figure 1). Absent increases in passing efficiency (relative to rushing), this might suggest that the passing-premium puzzle has been solved. That is, coaches have become wise to the work of analysts and have responded by passing more. However, given the efficiency of passing has also improved relative to rushing, a clear alternative explanation exists: coaches have selected into passing as the returns have increased. Given this large increase in both pass rate and passing efficiency, does the passing-premium puzzle persist?

To answer this question, we first need to return to a fundamental question in football analytics: what is the optimal pass rate? Following work positing defensive adjustments as the primary reason for persistent differences, we use a simple model of per-play efficiency which allows the efficiency of passing and rushing to vary with their respective usage. Estimating the effect of the probability of passing on passing and rushing efficiency, we can infer the optimal pass rate.

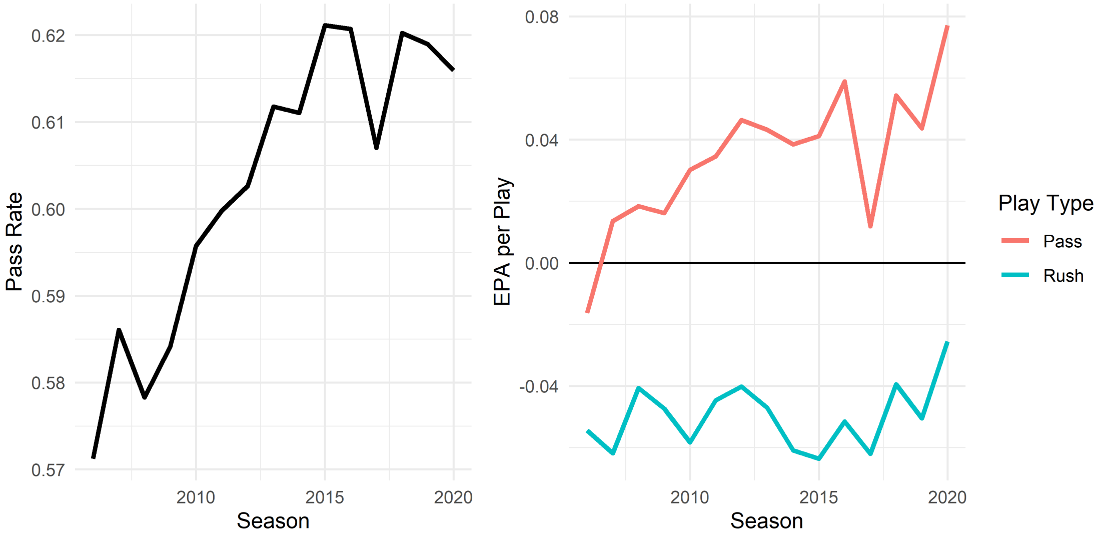

Figure 1: Evolution of NFL Pass Rate (left) and Per-Play Efficiency (right), 2006-2020

However, just as selection into passing over the past 15 years may be driven by improvements in passing offense, estimating usage curves means contending with how teams select into passing. For example, coaches with strong passing offenses may pass more often than their counterparts, which would bias the estimation of usage curves.

To remove selection bias in estimating the parameters of this model, we introduce two new instrumental variables for passing and expected pass rate. First, fumbles lost. In particular, we argue that conditional on the number of fumbles by a team, the fumbles that team lose is an important—and essentially random—determinant of game state, measured through win probability or net score. This change in game state forces teams to become more aggressive than they otherwise would be, passing more often in order to score the points necessary to win the game. Importantly, both the fumbling team and the opposing team realize this necessity, meaning our instrument also forces changes in passing expectations for the defense. In this way, our approach mimics the defensive adjustments of a long-term increase in passing probability. Second, we draw on an negative serial correlation in playcalling documented by Emara et al. (2017). That is, playcallers excessively switch between rushing and passing plays. To construct this instrument, we use the previous play's pass rate over expected (actual pass decision minus expected pass rate). This strips out game state—which we treat as endogenous—leaving only the "shock" of the previous pass or rush decision.

We draw on play-by-play data for 2006–2020 from the NFL to investigate these questions, as well as predictive models of expected passing and expected points (Yurko et al., 2019; Baldwin, 2021). Using several instruments generated from fumbles lost (conditional on the number of total fumbles, lost or otherwise), we use two-stage least squares (2SLS) to estimate: (1) the casual effects of passing, (2) expected pass rate, and (3) the interaction of the two variables on per-play efficiency. We argue our instruments and specification fulfill the conditions outlined in Angrist and Krueger (1999), so our estimates can be interpreted as local average treatment effects (LATE). In the case of instrumental variables, LATE weights individual treatment effects by how strongly they respond to the chosen instruments.

We find coaches deviate substantially from the optimal usage of passing, assuming usage curves are linear. Using the 2SLS estimates on early downs, when the opponent expects a sure rush, passing increases EPA by 1.35 points/play relative to a rush. However, when a sure pass is expected, passing increases EPA by only 0.23 points/play. Embedding our instrumental variables results with the model of usage curves, we find an optimal pass rate of 59.1%—compared to an actual pass rate of 51.7%—a considerable deviation. However, this deviation is smaller than would be recovered by naively estimating OLS. We attribute this to 2SLS not suffering the same issues of selection into passing as OLS. This result suggests that coaches may overestimate the value of surprise in football strategy, opting to rush too often on early downs.

We contribute to a literature that studies decision-making in professional football by introducing causal inference techniques to estimate usage curves. Alamar (2006) proposed the passing-premium puzzle, setting the agenda for this line of research with the expected value of yards and play success as an outcome. Rockerbie (2008) notes the importance of risk preferences in decisions made under uncertainty. Studying the passing-premium puzzle as a portfolio problem, they find that teams pass too often. Neither of these papers are able to trace usage curves and account for the value of surprise in playcalling. In contrast Jordan et al. (2009) and McGarrity and Linnen (2010) study playcalling in the context of game theory. McGarrity and Linnen (2010) considers the decision to pass or rush as a mixed-strategy game, formally considering surprise as an important element.[1] If surprise matters, passing should not drop off considerably as passing offense declines due to quarterback injuries, something they document empirically. We build on this premise, which is implicit in our modeling approach. However, their empirical set-ups do not allow these studies to manage omitted-variable bias in the decision to pass or rush.[2] Emara et al. (2017) studies a related topic, documenting negative serial correlation in playcalling. They explain this negative serial correlation as a behavioral bias—that people have difficulty recognizing and producing random sequences. In addition to drawing on this bias for an instrument, our results complement this result. Specifically, by using instrumental variables to estimate usage curves, we document that playcallers' decision-making departs from expected points maximization in their selection of rush and pass plays.

2 A Simple Model of Playcalling Efficiency

The intuition of a usage curve is straightforward: the relative efficiency of a given play is a function of how often it is used. More specifically, the more a play is used, the more opponents will prepare for that play, which lowers its efficiency. A stylized model of playcalling can formalize this intuition. In this model, a playcaller chooses a rush or a pass. The playcaller's goal is to maximize efficiency, which is the decision-weighted sum of passing and rushing efficiency. While passing

[1] Jordan et al. (2009) is similar, studying the optimization of playcalling over a wider variety of margins.

[2] McGarrity and Linnen (2010) may be the strongest in this regard. However, selection bias could also drive this null result (omitted factors could include game plans, game state, or weather conditions).

efficiency tends to be higher for NFL teams, we allow the efficiency of both passing and rushing to vary with the decision weights.[3] In this way, we take the perspective that the more expected a play call is, the easier it is to defend. In doing so, we draw implicitly on game theoretical models of decision-making (Jordan et al., 2009; McGarrity and Linnen, 2010).

2.1 Optimization Problem

We write the playcaller's optimization problem,

$$\max_{p} \mathrm{E}(p) = p \times \mathrm{PE}(p) + (1-p) \times \mathrm{RE}(p) \quad \text{subject to} \quad 0 \leq p \leq 1, \tag{1}$$

where E is overall efficiency, RE is rushing efficiency and PE is passing efficiency, p is the pass rate, and $1-p$ is the rush rate. We allow passing efficiency and rushing efficiency to rely on rush rate, and call these efficiencies usage curves.

2.2 General Case

The first-order condition for an interior solution to (1) is:

$$p^* \times \frac{\partial \mathrm{PE}(p^*)}{\partial p} + \mathrm{PE}(p^*) + (1-p^*) \times \frac{\partial \mathrm{RE}(p^*)}{\partial p} - \mathrm{RE}(p^*) = 0. \tag{2}$$

A closed form, explicit solution for p^* requires further assumptions, laid out in Appendix A.1. Moreover, to use the implicit function theorem, we would need to introduce some parameters that should influence the optimal rate of p. Eschewing this, we work within a linear case. While introducing functional form is a strong assumption, linearity is a natural first approach due to its

[3] We recognize this simple model abstracts from important aspects of the pre-play situation, which impact the optimality of the play called. One can think of this as a model of playcalling with all else held equal. For example, this model might apply to a specific situation, e.g., 1st & 10, or to a number of similar situations controlling for pre-play observables (specifically, down, distance, and field position). We consider these factors further in the empirical work.

simplicity.[4]

2.3 Linear Case

For simplicity of exposition (and later estimation), we assume affine usage curves, i.e., $\frac{\partial \text{PE}(p)}{\partial p}$ is constant. Written another way: $\text{RE}(p) = a + bp$, $\text{PE}(p) = c + dp$ where $\frac{\partial \text{PE}(p)}{\partial p} = d$ and $\frac{\partial \text{RE}(p)}{\partial p} = b$. Solving this optimization problem gives the optimal pass rate,

$$p^* = \frac{1}{2}\left(\frac{c-a}{b-d} + \frac{b}{b-d}\right). \tag{3}$$

For details on solving the linear case, see Appendix A.2. How do we interpret this optimal pass rate? First, note the denominator $b - d$. We call this the total relative change in efficiency, the rate at which rushing gains efficiency minus the rate passing loses efficiency as p increases. We can break out the optimal pass rate into two effects:

1. The Efficiency Effect: $\frac{1}{2}\left(\frac{c-a}{b-d}\right)$. The numerator suggests that the better the pass game and worse the rush game are overall (scaled by the total relative change in efficiency), the more one should pass.

2. The Responsiveness Effect: $\frac{1}{2}\left(\frac{b}{d-b}\right)$. This captures how responsive rushing is to passing as a fraction of the total relative change in efficiency. The greater the rushing response, the less one should pass.

Finally, are there ever times we should only pass or only rush? Yes. We can characterize these corner solutions based on this optimal rush rate:

1. Only Pass: if $c + 2d \geq a + b$, then $p^* = 1$.

[4] There are added benefits when we consider estimation of usage curves, it also limits the potential for overfitting. This is particularly important when we consider low and high density regions of support. A moderately flexible specification may sacrifice fit in low density regions for better fit high density regions.

2. Only Rush: if $a - b \geq c$, then $p^* = 0$.

Given the existing passing premium it is hard to imagine case 2. We expect $b \geq 0$, which means this restriction bounds $c - a \leq 0$ This would require the average rush when opponents are certainly expecting a rush to be more efficient than the average pass, a very counterintuitive condition. Case 1 is potentially plausible. First, given $c \geq a$, this becomes an empirical question of the slope of the usage curves. While the marginal value of passing is falling ($d \leq 0$), it may remain high enough to outpace marginal value of rushing. Nevertheless, it would be surprising if this were the case generally.

2.4 Translating Theory to Estimation

In the empirical work below, we will estimate regression models of the following form:

$$E(p, \text{Pass}) = \alpha + \beta \times p + \gamma \text{Pass} + \delta \times \text{Pass} \times p \tag{4}$$

where p is once again the probability of a pass on that play and Pass is an indicator equal to 1 if the called play was a pass, and 0 if it was a rush. This specification re-parameterizes our theoretical model: $\alpha = a$, $\beta = b$, $\gamma = c - a$, and $\delta = d - b$. When a rush is called, we are estimating the usage curve for rushing,

$$RE(p) = E(p, \text{Pass} = 0) = \alpha + \beta \times p, \tag{5}$$

and when a pass is called, we are estimating the usage curve for passing,

$$PE(p) = E(p, \text{Pass} = 1) = (\alpha + \gamma) + (\beta + \delta) \times p. \tag{6}$$

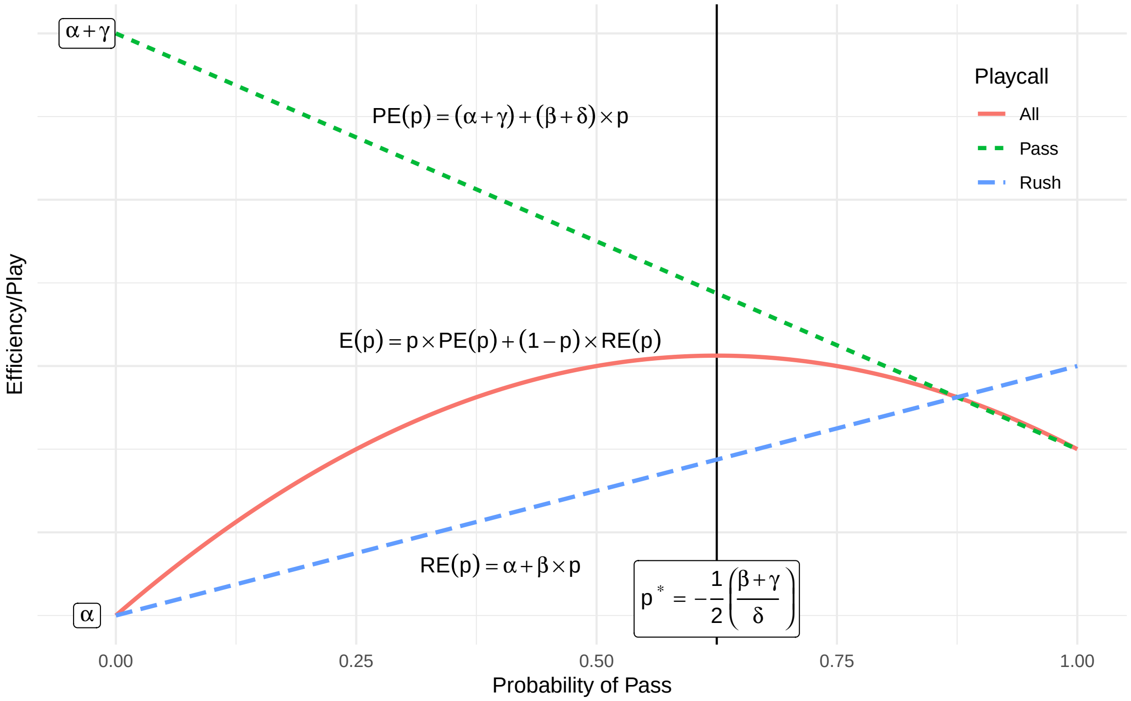

Figure 2: Our model of playcalling efficiency. Plotted are the usage curves for passing (PE) and rushing (RE). Additionally, we plot the efficiency curve, which is the play call weighted sum of these usage curves. For this example, we choose parameters so that we have an interior solution ($\gamma + \beta > 0$ and $\beta + \gamma + 2\delta < 0$).

Within this parametrization, optimal pass rate can be written as $p^* = -\frac{1}{2}\left(\frac{\gamma+\beta}{\delta}\right)$. Likewise, the efficiency effect could be restated as $-\gamma/2\delta$ and the responsiveness effect as $-\beta/2\delta$. The usage curves and the efficiency curve are plotted in Figure 2 for an interior solution.[5] Given credible estimates of α, β, γ, and δ, we can gain a good idea of what the optimal rush rate would be, holding situation constant.

[5] Figure A1 plots a corner solution where one should always pass and Figure A2 plots a corner solution where one should always rush.

3 Play-by-Play Data

We use play-by-play data from NFL games during 2006 to 2020. These data include comprehensive contextual and outcome variables, which document the state-of-the-game before the play and describe what happened on each play. Contextual variables include down, distance to a first down, time remaining in the half and game, yard line, current score, and the game's time and location. Additionally, outcomes of plays include whether a play was called as a pass or a rush, the yardage gained, if the pass was intercepted, if the ball was fumbled and who recovered, whether a sack was recorded, or if the team scored. This data was obtained using the nflfastR package in R (Carl and Baldwin, 2021).[6]

Included in the data are outputs from several public analytics models. These include expected points and expected pass probability. We will describe these models briefly, but urge readers to consult more comprehensive sources for deeper understanding if they are not familiar. In particular, for details on how each of these variables were computed, see Baldwin (2021), which draws on the framework of Yurko et al. (2019).

First, we need a measure of efficiency, for which we choose expected-points added (EPA). Expected points have long been used to as a measure of football efficiency, beginning with Carter and Machol (1971). Using XGBoost, the expected-points model predicts the probabilities of a set of scoring plays (touchdown, field goal, opponent touchdown, etc.) using variables related to time in game, yard line, home, field type, down and distance, era, and timeouts remaining (Baldwin, 2021). Then the expected-points after a play is the probability weighted sum of these scoring plays in that game state. EPA is the difference in expected points before and after a play. This will serve as our main measure of efficiency.

Second, to estimate usage curves, we need the probability of a pass. The Expected Pass (XPass) model estimates this using variables related to yard line, home, field type, time in game, down

[6] Details about the package and downloading the data can be found at https://www.nflfastr.com/.

and distance, score differential, times, win probability, and era (Baldwin, 2021). These values are provided for each play in the play-by-play dataset.

4 Empirical Strategy

4.1 Identifying Pass and Rush Usage Curves

4.1.1 Ideal Variation: A Playcalling Experiment

Suppose we could run an experiment, in which teams were induced to rush at a higher rate on average (say, some proportion above what might be expected in a given situation). This experiment would run throughout a season. In early games a team chosen to rush more might reduce their efficiency through this more conservative approach. However, when preparing for future games, opposing coaches watch and chart game film from previous matchups to understand their "tendencies," or the rate and situations in which teams call certain plays. After running this experiment for the first three weeks of the season, tendencies would be set.[7] If teams ran more, game plans would be adjusted to be "stout against the run," perhaps by changing defensive personnel or alignments to "put more men in the box" (Football Outsiders).[8] In this way, they would recoup some of this lost efficiency through greater success in their less frequently used passing game. Given the exogenous variation generated by the experiment, we could estimate the causal impact of usage on passing and rushing efficiency in future games.

Using the linear model specified above, we would estimate an efficiency regression,

$$\text{Efficiency}_{igp} = \alpha + \beta \bar{p}_{igp} + \gamma Pass_i + \delta Pass_i \times \bar{p}_{igp} + u_{igp} \tag{7}$$

[7] Anecdotally, teams often watch the last three games an opponent has played or the last time a team played that team, coach, or key player. These tendencies are then used in setting future game-plans.

[8] From Football Outsiders Glossary (cited in text), the box is "the defensive area between the offensive tackles extending approximately seven yards deep in the defensive backfield. The defense will put more players "in the box" the more intent they are on stopping a running play."

where i indexes team, g indexes game, and $\bar{p}_{igp}$ is the experimental pass rate for play p. This research design would estimate the effect of passing, and the effect of passing rate on rushing and passing efficiency.

In the real world, any such experiments in playcalling are few and far between. If and when they do occur, they would be hard to detect without inside information (i.e., $\bar{p}_{igp}$ would remain unobservable). Moreover, in the absence of such experiments, there is real and serious selection bias. Teams who are better at passing will pass more. Teams that have fallen behind will also pass more. Therefore, the endogenous nature of passing will bias observational estimates.

4.1.2 Quasi-Experimental Variation: Fumbles Lost as an Instrument

The playcalling experiment described above is just one possible cause for which we could estimate an effect. Just as coaches might respond to a persistent change in rush or pass rate, there are many other causes for coaches to adjust their defensive decision-making to counter the pass or the rush. Coaches do not just choose a gameplan and then rigidly execute it (in fact, this may be the hallmark of a bad coach). Instead, they make adjustments on the fly, often in response to the game state, which might encompass the net score and the probability a given team will win the football game. It is well documented an increase in a team's win probability decreases the propensity to pass. However, using the natural variation in game state (within a game) faces a similar issue with reverse causality: teams that are behind likely already have worse efficiency on offense, which is what caused them to be behind in the first place.

We propose to circumvent this challenge using an instrumental variables approach. Specifically, we use a set of instruments based on cumulative fumbles lost to the opposing team to remove the selection bias in expected pass rate and actual passing decisions. Football analysts have identified fumble recoveries as a random yet crucial component of winning football games. For example, Massey-Peabody Analytics, which prepares predictive team ratings based on historical data, down-weights statistics related to recovered fumbles in their team ratings: "recovered fumbles, which greatly influence the outcome of games, [...] are completely random" (Massey-Peabody Analytics, 2012). Likewise, ESPN Analyst and Writer Bill Connelly identifies fumbles and fumble recoveries as a one of the five main factors influencing the outcome of college football games (Connelly, 2014). More specifically, he identifies it as a *random* factor in winning football games. As he so adeptly puts, if you want to win football games, "you want that damned, pointy ball to bounce in a favorable way."

Therefore, cumulative fumbles lost is an attractive candidate for an instrument for pass rate and expected pass rate. Specifically, we propose a set of instruments that includes fumbles lost by the possessing team to the opposing team so far in the game and *vice versa*. These instruments are conditional on cumulative fumbles by each team, which need not be random. The properties of cumulative fumbles lost relate closely to the three conditions needed for a valid instrument.[9] First, if the number of past fumbles lost is random (conditional on total cumulative fumbles by that team), the instrument is as good as randomly assigned and the independence assumption is satisfied. Second, if fumbles are important in driving swings in game state, and game state is important determinant of pass rate, the instrument will be strong in the sense of Stock and Yogo (2005). This condition, referred to as instrumental relevance, can of course be checked in the data. Third, if fumbles themselves are a poor predictor of future performance (which they tend to be), the instrument is related to the outcome only through the endogenous variable(s), and the exclusion restriction is satisfied.

Given these three conditions are met, our approach will identify the LATE under one assumption: monotonicity. Monotonicity assumes everyone responds weakly in the same direction. In our context, this means if the majority of playcallers respond to a fumble lost by increasing their expected passing rate, monotonicity requires the remaining coaches do not change their expected

[9] See, for example, Theorem 4.5.1 in (Angrist and Pischke, 2009).

passing rate. That is, monotonicity is violated if and only if some playcallers responded to fumbles lost by rushing more.[10] The LATE differs from the Average Treatment Effect (ATE) in terms of weighting when treatment effects are heterogeneous—individuals with stronger responses to the instrument (in terms of passing) are weighted higher.

4.1.3 Lagged Pass Rate Over Expected

In addition to fumbles lost, we utilize a second instrument, which utilizes the fact that the decision to pass exhibits negative serial correlation (Emara et al., 2017). That is, coaches are more likely to call a pass after a rush (and vice-versa). This is an example of a more general behavioral phenomenon: that people have difficulty producing random sequences of actions. This also points to an instrument for passing. As we show, this negative serial correlation results in instrumental relevance. It is likely that this instrument is also monotonic in that if coaches avoid negative serial correlation, they do not consistently build strategies around clustering plays.

To satisfy independence and the exclusion restriction, however, we must adjust the instrument. Game state dictates that expected pass-rate is correlated across plays, in spite of this negative serial correlation. Therefore, we propose to instrument pass rate with lagged pass-rate over expected. We net expected pass-rate out of the decision to pass. Pass-Rate Over Expected is defined: $\text{PassOE}_{igt} = \text{Pass}_{igt} - \text{XPass}_{igt}$. In particular, expected pass-rate is the estimated probability of passing conditional on yard line, home team, field type, time in game, down and distance, score differential, time, win probability, and era (Baldwin, 2021). This measure captures the degree of surprise at the decision to pass on the previous play and, thus, should pass independence and the exclusion restriction. That is, the surprise of passing on the previous play should not impact EPA on the next play, except through play choice.

[10] There are circumstances when a coach might respond to a fumble lost by rushing more (e.g., in "garbage time" when the game is all but decided) but these situations are rare and controlled for in the estimates through measures of game state.

4.1.4 Fixed Effects

For additional robustness, we also propose the use of panel fixed-effects to control for offensive quality. With panel data, fixed effects are widely used to control for time-invariant unobservable characteristics. In this case we want to control for team-level propensity to pass (a function of underlying talent), so we use team-year fixed effects to control for the propensity of a given team to pass or rush.[11]

4.2 Main Specifications

4.2.1 Instrumental Variables Specification

We aim to replicate an experimental setup as closely as possible. Our intuition is reflected in a simple approach: we define games in which a fumble was lost as the treatment group and games in which the offense fumbled but did not lose possession as the control group. However, such a set-up would neglect the richness of the data. Therefore, we construct indicator variables for each value of cumulative fumbles lost and each value of cumulative fumbles for both the possessing and opposing teams.

Rather than a single treatment, this experiment has multiple treatments. For example, among teams that fumble three times, we have three possible treatments corresponding to whether one, two, or all three of these fumbles were lost. Finally, we interact our instrumental variables with whether a pass was called on a given play. We use the resulting set of instruments in our two-stage

[11] While one might want to utilize a great deal of fixed effects to control for various situations, many of these will overlap. In particular, it seems reasonable that a quarterbacks' talent, a coaches' offensive system, an offenses' supporting cast would affect the probability of calling a rush or calling a pass. However, many coach-quarterback combinations are themselves invariant for long periods of time, so estimating coach-quarterback combinations might be a bit fruitless. Moreover, while contracts might last five years in the NFL, talent surrounding the quarterback fluctuates at a faster rate than the quarterback does. Using quarterback-year fixed effects controls for injuries to starting quarterbacks, which tend to adjust the propensity to pass and using the year scale tends to proxy for the turnover of the rest of team by coinciding with free agency. In addition, this coincides with installations of new playbooks, a key point where offensive philosophy might change. Of course, we could cut these finer, for example, estimating QB-coach-year fixed effects that differ from QB-year fixed effects only when there is a mid-season firing, injury, or quarterback change.

least squares estimator, with the number of cumulative fumbles included as controls.

Before presenting the estimating equations we use to estimate usage curves, we build intuition by estimating the causal effect of calling a passing play on efficiency.

$$\begin{aligned}\text{Pass}_{igt} &= \alpha_1 \\ &+ \sum_{k=1}^{K} \eta_{1k}\text{Post-}k\text{ Team Fumbles}_{igt} + \sum_{l=1}^{L} \theta_{1l}\text{Post-}l\text{ Opp. Fumbles}_{igt} \\ &+ \sum_{n=1}^{N} \lambda_{1n}\text{Post-}n\text{ Team Fumbles Lost}_{igt} + \sum_{m=1}^{M} \pi_{1m}\text{Post-}m\text{ Opp. Fumbles Lost}_{igt} \\ &+ \kappa_1\text{PassOE}_{i,g,t-1} + \varepsilon_{1igt},\end{aligned} \tag{8}$$

where i indexes team t play, and g game. Post-k Team Fumbles$_{igt}$ are indicator variables for number of cumulative fumbles by the team to that point, and Post-l Opp. Fumbles$_{igt}$ plays the same role for fumbles by the opponent. Likewise, Post-n Team Fumbles Lost$_{igt}$ are indicator variables for the number of fumbles lost by the team to that point and Post-m Opp. Fumbles Lost$_{igt}$ plays the same role for their opponent. PassOE$_{i,g,t-1}$ is passing over expected on the previous play by the same team. Finally, Pass$_{1igt}$ is an indicator equal to one if team i passed on play t of game g. The fitted value of this variable appears in the second stage, denoted with a hat:

$$\begin{aligned}y_{2igp} &= \alpha_2 + \beta_2\widehat{\text{Pass}}_{igt} \\ &+ \sum_{k=1}^{K} \eta_{2k}\text{Post-}k\text{ Team Fumbles}_{igt} + \sum_{l=1}^{L} \theta_{2l}\text{Post-}l\text{ Opp. Fumbles}_{igp} + \varepsilon_{2igt},\end{aligned} \tag{9}$$

where y_{2igp} is an efficiency variable of choice. While we will use EPA, it could be win probability added (WPA), play success (defined as passing some yards or expected points threshold). Based on the assumptions presented in Section 4.1.2, β_2 is identified as the LATE of passing on efficiency.

4.2.2 Usage Curves Specification with Instrumental Variables

Our main results feature a more complex estimation. Since we have two endogenous variables as well as their interaction, we instrument using both fumbles lost, lagged passing over expected, and their interactions. We specify the first stage regressions:

$$
\begin{aligned}
y_{1igt} &= \alpha_1 \\
&+ \sum_{k=1}^{K} \eta_{1k}\text{Post-}k\text{ Team Fumbles}_{igt} + \sum_{l=1}^{L} \theta_{1l}\text{Post-}l\text{ Opp. Fumbles}_{igt} \\
&+ \sum_{n=1}^{N} \left(\lambda_{1n}\text{Post-}n\text{ Team Fumbles Lost}_{igt} + \mu_{1n}\text{Post-}n\text{ Team Fumbles Lost} \times \text{PassOE}_{i,g,t-1}\right) \\
&+ \sum_{m=1}^{M} \Big(\pi_{1m}\text{Post-}m\text{ Opp. Fumbles Lost}_{igt} + \phi_{1m}\text{Post-}m\text{ Opp. Fumbles Lost}_{igt} \times \text{PassOE}_{i,g,t-1}\Big) \\
&+ \kappa_1\text{PassOE}_{i,g,t-1} + \epsilon_{1igp},
\end{aligned}
\tag{10}
$$

where y_{1igp} stands in for the two endogenous variables and their interaction. Again, the fitted values of these endogenous variables appear in the second stage, denoted with a hat:

$$
\begin{aligned}
y_{2igt} &= \alpha_2 + \beta_2\widehat{\text{Pass}}_{igt} + \gamma_2\widehat{\text{XPass}}_{igt} + \delta_2\text{Pass}\widehat{_{igt} \times \text{X}}\text{Pass}_{igt} + \\
&+ \sum_{k=1}^{K} \eta_{2k}\text{Post-}k\text{ Team Fumbles}_{igt} + \sum_{l=1}^{L} \theta_{2l}\text{Post-}l\text{ Opp. Fumbles}_{igt} + \epsilon_{2igt},
\end{aligned}
\tag{11}
$$

where y_{2igp} is an efficiency variable of choice. We identify β_2, γ_2, and δ_2 as their respective LATEs, which gives us credible estimates to fill our our model of optimal playcalling. This will serve as our preferred specification for estimating parameters of the usage curves model.

4.2.3 Naive and Fixed-Effects Specifications

For the uninstrumented naive specification, we estimate,

$$y_{2igp} = \alpha_{2i} + \beta_2 \text{Pass}_{igp} + \gamma_2 \text{XPass}_{igp} + \delta_2 \text{Pass}_{igp} \times \text{XPass}_{igp} + \upsilon_{2igp}. \tag{12}$$

We use fixed-effects estimation to control for unobservable differences in teams, such as quarterback skill or offensive philosophy. In particular, we use a team-game fixed effect. Of course, making a similar adjustment to specification (9) and (11) allows us to estimate the 2SLS specification with fixed effects. We do so as a robustness check.

4.3 Optimal Pass Rate

Plugging the coefficient estimates into the parameters of the model of playcalling efficiency, we can draw usage curves and find a point estimate for the optimal pass rate,

$$\hat{p}^* = -\frac{1}{2}\left(\frac{\hat{\gamma} + \hat{\beta}}{\hat{\delta}}\right). \tag{13}$$

We do so using the coefficient estimates from our usage curves specification, and infer the uncertainty around this estimate using a cluster-robust jackknife. Hansen (2024, 2025) shows the cluster-robust jackknife has a number of attractive properties in the context of a potentially heteroskedastic and cluster-dependent data-generating process, including that the variance estimator is never downwards biased. As with the regressions, we cluster the jackknife at the team-game level.

Table 1: The Effects of Called QB Dropbacks on Efficiency. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	OLS	FE	2SLS		FE 2SLS	
			1st Stage	2nd Stage	1st Stage	2nd Stage
	EPA	EPA	Pass	EPA	Pass	EPA
	(1)	(2)	(3.1)	(3.2)	(4.1)	(4.2)
1 Pos. Fum. Lost			0.021***		0.023***	
			(0.005)		(0.004)	
2 Pos. Fum. Lost			0.044***		0.046***	
			(0.009)		(0.009)	
3 Pos. Fum. Lost			0.099***		0.100***	
			(0.019)		(0.019)	
1 Opp. Fum. Lost			−0.024***		−0.023***	
			(0.005)		(0.004)	
2 Opp. Fum. Lost			−0.055***		−0.059***	
			(0.009)		(0.009)	
3 Opp. Fum. Lost			−0.130***		−0.142***	
			(0.021)		(0.020)	
Lag PassOE			−0.001***		−0.001***	
			(0.00002)		(0.00002)	
QB Dropback	0.162***	0.162***		0.379***		0.400***
	(0.004)	(0.004)		(0.040)		(0.036)
Observations	352,984	352,984	344,148	344,146	344,148	344,146
R^2	0.005	0.009	0.014		0.027	
Note:				*p<0.1; **p<0.05; ***p<0.01		

5 Results

5.1 The Causal Effect of Passing on Early Downs

Estimating the naive regressions (Table 1 columns 1 and 2) first, we find a positive relationship between QB Dropbacks (i.e., called passes) and offensive efficiency on early down plays. Qualitatively, this is the effect we expect—passing has consistently outperformed rushing in this efficiency metric. However, when we estimate the 2SLS regressions with cumulative fumbles lost and lagged pass over expected as instruments, we find a much larger effect. In particular, while the "naive" effect was an increase of 0.16 EPA per called pass, 2SLS estimates a 0.38 EPA increase in EPA (Table 1 column 3.2). We interpret this downward bias in the naive coefficient as a function of biases that might cause better passing teams to pass less. For example, if teams with better passing offenses score earlier, they might move away from passing as they extend their lead.

Second, the results show the behavioral relevance of the instruments for passing. Cumulative fumbles lost drive pass rate as we would expect them to. A fumble lost puts a coach into a worse game state, which increases their probability of dropping back to pass—when coaches lose a fumble, they pass more. On the other hand, an opponent losing a fumble (i.e., fumbles gained) decreases the coach's probability of dropping back—when coaches get lucky with fumbles, they pass less. Moreover, the instruments tell a coherent story. The effect size of fumbles lost and gained is relatively symmetric (opponent fumbles have a slightly stronger effect) and the effects increase as fumbles mount up (two fumbles lost makes us pass more than one, and three more than two). For lagged pass over expected, we also find that our estimates reflect the negative serial correlation described in Emara et al. (2017).

Formally testing for instrumental relevance, we estimate a first stage Cragg-Donald statistic of 299.45 from the non-fixed effects 2SLS regression, reducing concerns about issues related to weak instruments. This is significant at least at the 5% level (bias of 5% relative to OLS, one endogenous variable, 13 instruments: 5% critical value of 21.1; Stock and Yogo, 2005, does not present critical values for smaller levels).[12] We also estimate single instrument variables models and find similar first and second stage results (see Table B1 for results and Table B2 for first stage results). Lag pass over expected tends to result in a slightly larger effect (0.41) as compared to fumbles lost instruments (0.26). We interpret this difference in terms of weighting, with the fumbles lost instrument leading to higher weighting of individual treatment effects in games where ball handling is difficult. Further results and discussion are presented in Appendix B.

5.2 Estimating Usage Curves and Optimal Pass Rate on Early Downs

5.2.1 Naive Estimates

The main effects on early downs (all first and second down plays) are presented in Table 2. While the value of passing falls as the expectation of passing increases, rushing did not become more valuable as it becomes more surprising. That rushing efficiency falls as passes are expected betrays the endogeneity present in this regression: this should tell us that less efficient offenses find themselves needing to pass later in games, and are poor at rushing and passing, so both decline for teams in such a game state. Placing these estimates into our simple model of optimal pass rates, the naive estimates suggest an optimal pass rate of approximately 86.7% on early downs. This is considerably higher than the average pass rate on early downs, 51.7%. However, we cannot take this number at face value: the high optimal pass rate is in part due to the endogenous reduction in rushing efficiency associated with greater expected passing rate.

5.2.2 Two-Stage Least Squares Estimates

Focusing on the pooled 2SLS regression in Table 2 column 3, estimates differ considerably from the naive estimates. Usage curves bear out: we estimate rushing efficiency increases by 0.16 EPA

[12] The Cragg-Donald Statistic is estimated using ivregress in Stata.

Table 2: Estimates of Usage Curves in Professional Football. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	EPA			
	OLS	FE	2SLS	FE 2SLS
	(1)	(2)	(3)	(4)
XPass ($\hat{\beta}$)	-0.010	0.032^{**}	0.164	0.129
	(0.014)	(0.014)	(0.783)	(0.809)
Pass ($\hat{\gamma}$)	0.244^{***}	0.239^{***}	1.346	1.058
	(0.013)	(0.013)	(1.162)	(1.126)
XPass × Pass ($\hat{\delta}$)	-0.135^{***}	-0.137^{***}	-1.278	-0.905
	(0.024)	(0.024)	(1.803)	(1.810)
Constant ($\hat{\alpha}$)	-0.077^{***}		-0.394	
	(0.007)		(0.455)	
Optimal Pass Rate ($\hat{p}^{*}$)	86.7%	98.9%	59.1%	65.6%
Fixed Effects	No	Yes	No	Yes
Observations	352,032	352,032	343,201	343,201
R^2	0.005	0.009		
Note:		*p<0.1; **p<0.05; ***p<0.01		

when a pass is sure versus when a rush is sure. Likewise, passing efficiency decreases by 1.28 EPA when a pass is sure versus when a rush is sure. Figure C1 visualizes the estimated usage curves from this specification. Usage curves are much steeper in the 2SLS estimates as compared to OLS. When the probability of passing approaches zero, 2SLS estimates a premium of 1.35 EPA/play. However, when passing probability is close to the average pass rate (51.7%), the passing premium is estimated at around 0.77 EPA. Finally, in situations where the probability of passing approaches one, the premium estimated falls to 0.23 EPA. In Appendix C.1 we discuss how LATE-weighting could impact these estimates.

One limitation to these results is that coefficient estimates on the endogenous variables tend to be noisy. This is due to three facts. First, unique to our set-up is that the endogenous variables are related, and thus have an unusually high degree of multicollinearity. Second, instrumental variables strategies by their nature chisel away at variation which is not quasi-random, which increases multicollinearity as a side effect (Rhoads, 1991). Third, we are using robust standard errors clustered at the team-game level. These factors can all lead to a loss of precision in estimating effects. Indeed, none of our coefficients of interest are significant in this specification.

Our ultimate goal is not the usage curves themselves, however, but the optimal pass rate. To better understand the optimal pass rate implied by our estimates and its variation, we compare the observed pass rate to the IV-implied optimal pass rate using a cluster-robust jackknife (Hansen, 2024, 2025). For this process, we use our preferred regression (column 3 of Table 2). We find that the optimal pass rate is statistically different from the observed empirical pass rate (significant at any conventional level). The results of this process are plotted in Figure 3. The upper panel depicts the efficiency curves estimated in each iteration of the jackknife. These result in a tight bound around the efficiency curve and a tight clustering of optimal pass rates, plotted in the panel below. These results suggest that offenses should pass about 59.1% of early down plays as opposed to the actual pass rate of 51.7% on early downs, implying the average coach overestimates the element of surprise (at least on early downs) and rushes too often as a result. Despite this, according to the 2SLS model, the benefits to moving to the optimal pass rate are not massive, about 0.007 EPA/play. This translates to an extra point every 143 plays.

5.3 Addressing Threats to Validity

5.3.1 Relevance with Multiple Endogenous Variables

While we demonstrated relevance in single equation models, we also provide evidence here around the relevance of the instruments when we estimate regressions with multiple endogenous

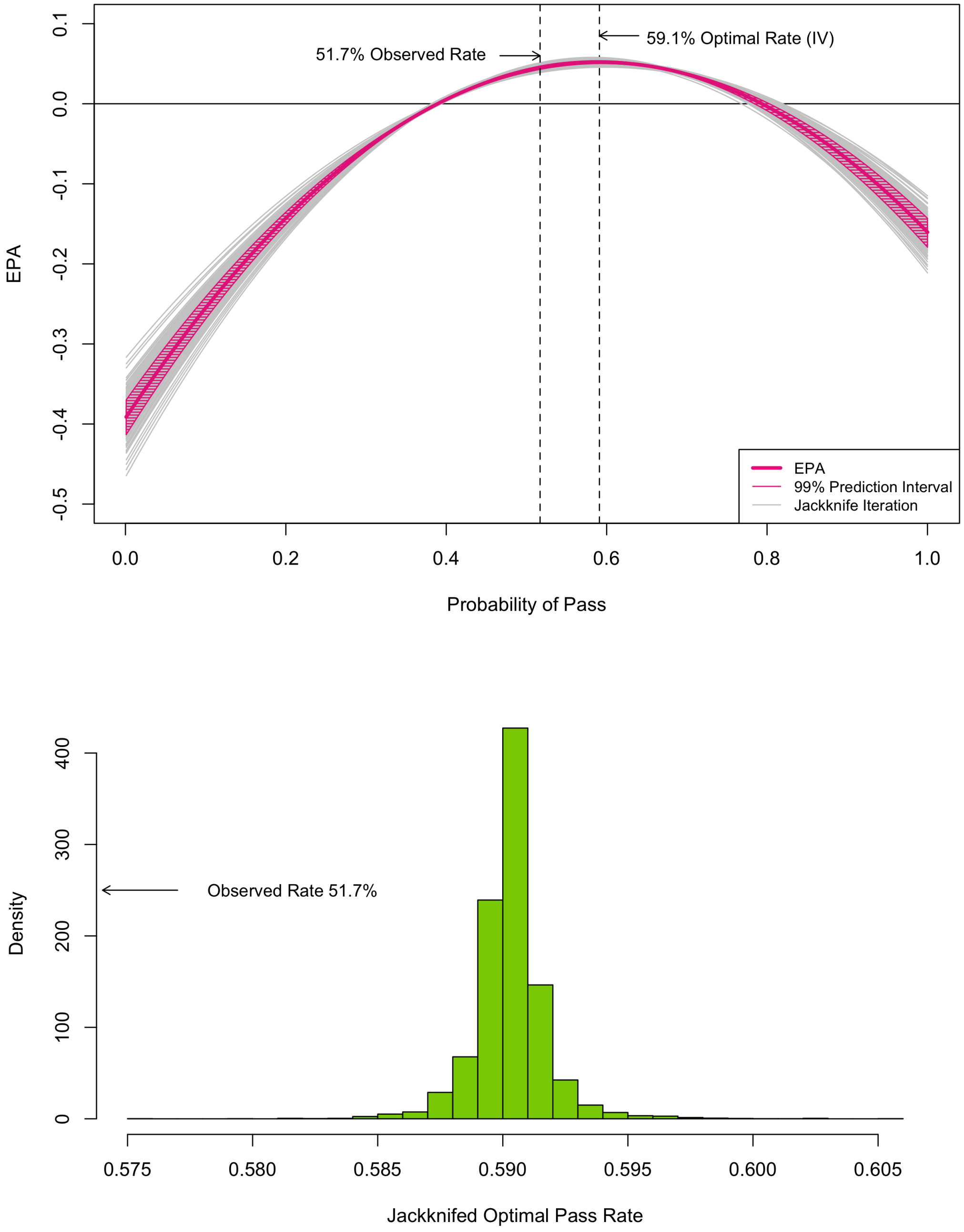

Figure 3: Jackknife IV Estimates of The Efficiency Curve (above) and Optimal Pass Rate (below).

variables. Results from the first stage of the Pooled 2SLS estimation are presented in Table 3. The coefficients tell a coherent and intuitive story: losing fumbles increases pass rate, while gaining fumbles decreases pass rate. Likewise, lagged pass over expected retains its negative serial correlation. Finally, there is some slight attenuation in the negative correlation in passing when fumbles are lost by either team. This may be because this takes teams out of neutral game states into pass heavy or rush heavy game plans. In each first strage regression instruments are strongly associated with the relevant endogenous variable (as can be seen in Section 5.1). We find similar results for the fixed effects specification (see Table C1). Despite these clear indications of relevance, we obtain a Cragg-Donald Statistic of 3.38. This is comparable to a 5% critical value of 4.41 (for 30% bias relative to OLS, three endogenous regressors, and 13 instruments, see Stock and Yogo, 2005). We cannot reject the null. This is a surprising result until considering that Cragg-Donald statistic tests for near under-identification (Stock and Yogo, 2005). While such near under-identification is almost always due to relevance, here it seems to be due to the high degree of multicollinearity.

5.3.2 Multicollinearity and Optimal Pass Rate

As discussed above, we find high degrees of multicollinearity in our variables of interest. While we have a sufficiently large sample to handle this issue when estimating OLS, IV estimates have higher variance inflation factors and are affected (see Table D1). As is the case with multicollinearity in simultaneous equations, this leads to noisier individual coefficients (e.g., fluvial systems, as in Rhoads, 1991). Such multicollinearity affects inference on individual coefficients; however, the optimal pass rate, $\hat{p}^*$, depends on a nonlinear combination of these correlated coefficients which, when put together, are statistically different from the observed pass rate. In fact, the $\hat{p}^*_{IV}$ is estimated quite precisely, with the underlying correlation structure of the coefficients stabilizing what otherwise might appear as noisy estimates. Intuitively, changes in the numerator are offset by (correlated) changes in the denominator, and vice versa (see Table D2).

Table 3: First Stage Results from 2SLS Usage Curves Estimation. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	Dependent variable:		
	Pass	XPass	XPass x Pass
	(1)	(2)	(3)
1 Pos. Fum. Lost	0.021***	0.018***	0.021***
	(0.005)	(0.003)	(0.004)
2 Pos. Fum. Lost	0.044***	0.039***	0.045***
	(0.009)	(0.007)	(0.008)
3+ Pos. Fum. Lost	0.098***	0.079***	0.097***
	(0.019)	(0.015)	(0.019)
1 Opp. Fum. Lost	-0.024^{***}	-0.023^{***}	-0.025^{***}
	(0.005)	(0.003)	(0.004)
2 Opp. Fum. Lost	-0.055^{***}	-0.047^{***}	-0.053^{***}
	(0.009)	(0.006)	(0.008)
3+ Opp. Fum. Lost	-0.129^{***}	-0.101^{***}	-0.113^{***}
	(0.021)	(0.014)	(0.016)
1 Pos. Fum. Lost x Lag PassOE	0.0003***	0.0001***	0.0002***
	(0.0001)	(0.00002)	(0.00004)
2 Pos. Fum. Lost x Lag PassOE	0.001***	0.0001***	0.0004***
	(0.0001)	(0.0001)	(0.0001)
3 Pos. Fum. Lost x Lag PassOE	0.0003	0.0001	0.0002
	(0.0003)	(0.0001)	(0.0002)
1 Opp. Fum. Lost x Lag PassOE	0.0002***	0.0001***	0.0002***
	(0.0001)	(0.00002)	(0.00004)
2 Opp. Fum. Lost x Lag PassOE	0.001***	0.0001**	0.0003***
	(0.0001)	(0.00005)	(0.0001)
3 Opp. Fum. Lost x Lag PassOE	0.001*	-0.00001	0.0003
	(0.0003)	(0.0001)	(0.0002)
Lag PassOE	-0.001^{***}	-0.0005^{***}	-0.001^{***}
	(0.00003)	(0.00001)	(0.00002)
Observations	344,148	343,203	343,203
R^2	0.014	0.038	0.029
Note:			*p<0.1; **p<0.05; ***p<0.01

5.3.3 Instrumental Validity and Aggression on Passing Plays

Other margins of aggression in decision-making by coaches may threaten the exclusion restriction if they are correlated with efficiency. While it is never possible to rule out every causal pathway by which an instrument might violate the exclusion restriction, it is useful to provide suggestive evidence about likely pathways. One pertinent pathway is passing aggression, by which we mean the aggressiveness of the design of passing plays, holding the decision to pass constant. If fumbles lost increases passing aggression, we might worry that our strategy underestimates the optimal passing rate holding this factor constant. We investigate this in Appendix E and find that our instruments do not systematically impact passing aggression.

6 Conclusion

Throughout professional and college football, passing has become the predominant offensive strategy. While passing rates have increased, so too has its efficiency. This naturally leads to questions about the optimal passing rate, and whether or not enough (or too many) passes are being called. Empirically, this is a difficult question to answer, because defenses can and do change their strategies. In this paper, we argue for an instrumental-variables approach to tackle the endogeneity of offensive and defensive decision-making. We use fumbles lost (conditional on cumulative fumbles) as the basis for a series of instruments, which we argue meet the conditions for interpreting the results as local average treatment-effects. We use lagged pass over expected as a second instrument, and construct a full set of instruments using their interactions.

We find NFL playcallers do not pass enough. Our results indicate, all else equal, the optimal passing rate on early downs is 59.1%. This is considerably lower than the 86.7% rate from a naive regression, illustrating the significant bias created by ignoring the endogeneity of decision-making. However, it is higher than the league-average pass rate (51.7%), suggesting more passing would increase offensive efficiency, even accounting for defensive responses. However, coaches do find themselves close to peak EPA/play, less than one hundredth of a point per play. This optimum rate is based on assumptions, including the linearity of usage curves and that defensive responses to increased passing in the future would be similar to defensive responses to higher passing probabilities in the past. Regardless, our findings suggest that coaches have moved toward optimal pass rates over the period in our sample. Further work is needed to settle this issue, including on the myriad contextual dimensions of playcalling, such as heterogeneity in game scenarios and team composition, as well as possible behavioral reasons for deviating from optimality (e.g. Romer, 2006; Slade and Tolhurst, 2019).

References

B. C. Alamar. The Passing Premium Puzzle. *Journal of Quantitative Analysis in Sports*, 2(4), 2006. ISSN 2194-6388. doi: 10.2202/1559-0410.1051.

J. D. Angrist and A. B. Krueger. Empirical strategies in labor economics. *Handbook of Labor Economics*, 3(23), 1999.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* 2009.

B. Baldwin. Open Source Football: nflfastR EP, WP, CP xYAC, and xPass models. https://www.opensourcefootball.com/posts/2020-09-28-nflfastr-ep-wp-and-cp-models/, 2021.

S. Carl and B. Baldwin. nflfastR: Functions to Efficiently Access NFL Play by Play Data, 2021.

V. Carter and R. E. Machol. Technical Note—Operations Research on Football. *Operations Research*, 19(2):541–544, Apr. 1971. ISSN 0030-364X, 1526-5463. doi: 10.1287/opre.19.2.541.

B. Connelly. The five factors: College football's most important stats, 2014.

N. Emara, D. Owens, J. Smith, and L. Wilmer. Serial correlation in National Football League play calling and its effects on outcomes. *Journal of Behavioral and Experimental Economics*, 69: 125–132, 2017. ISSN 22148051. doi: 10.1016/j.socec.2017.01.007.

Football Outsiders. Glossary: General Football Terms. https://www.footballoutsiders.com/info/glossary_general.

B. E. Hansen. Jackknife standard errors for clustered regression. *Working paper, University of Wisconsin*, 2024.

B. E. Hansen. Standard errors for difference-in-difference regression. *Journal of Applied Econometrics*, 2025.

J. D. Jordan, S. H. Melouk, and M. B. Perry. Optimizing Football Game Play Calling. *Journal of Quantitative Analysis in Sports*, 5(2), 2009. ISSN 2194-6388. doi: 10.2202/1559-0410.1176.

Massey-Peabody Analytics. Methodology, Aug. 2012.

J. P. McGarrity and B. Linnen. Pass or Run: An Empirical Test of the Matching Pennies Game

Using Data from the National Football League. *Southern Economic Journal*, 76(3):791–810, 2010. ISSN 0038-4038. doi: 10.4284/sej.2010.76.3.791.

B. L. Rhoads. Multicollinearity and Parameter Estimation in Simultaneous-Equation Models of Fluvial Systems. *Geographical Analysis*, 23(4):346–361, Oct. 1991. ISSN 0016-7363, 1538-4632. doi: 10.1111/j.1538-4632.1991.tb00244.x.

D. W. Rockerbie. The Passing Premium Puzzle Revisited. *Journal of Quantitative Analysis in Sports*, 4(2), 2008. ISSN 2194-6388. doi: 10.2202/1559-0410.1093.

D. Romer. Do firms maximize? Evidence from professional football. *Journal of Political Economy*, 114(2):340–365, 2006. ISSN 00223808. doi: 10.1086/501171.

A. Schatz. The Establishment Clause, 2003.

P. Slade and T. Tolhurst. Job Security and Risk-Taking: Theory and Evidence From Professional Football. *Southern Economic Journal*, 85(3):899–918, 2019. ISSN 00384038. doi: 10.1002/soej.12313.

J. H. Stock and M. Yogo. Testing for weak instruments in linear IV regression. (February):80–108, 2005.

R. Yurko, S. Ventura, and M. Horowitz. nflWAR: A reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports*, 15(3):163–183, 2019. ISSN 15590410. doi: 10.1515/jqas-2018-0010.

A Optimization Problem

A.1 Interior Solutions: General Case

We outline the necessary and sufficient conditions for an interior solution to optimal pass rate. The following conditions will guarantee an interior solution:

1. $\mathrm{E}(p)$ is continuous and differentiable
2. $\frac{\partial \mathrm{E}(p)}{\partial p}|_{p=0} > 0$
3. $\frac{\partial \mathrm{E}(p)}{\partial p}|_{p=1} < 0$
4. $\mathrm{E}(p)$ is concave down in p, i.e., $\frac{\partial^2 \mathrm{E}(p)}{\partial p^2} < 0$

If these hold, FOC (2) will deliver an optimal pass rate.

A.2 Interior Solutions: Linear Case

We outline the necessary and sufficient conditions for an interior solution for the optimal pass rate in the linear case. First, FOC (2) must hold. Second, note that $\mathrm{E}(p)$ can be re-written as a parabola:

$$\begin{aligned}\mathrm{E}(p) &= p \times \mathrm{PE}(p) + (1-p) \times \mathrm{RE}(p) = p \times (c + d \times p) + (1-p) \times (a + b \times p) \\ &= p \times c + d \times p^2 + a + b \times p - a \times p - b \times p^2 = (d-b) \times p^2 + (c-a) \times p + a\end{aligned}$$

Therefore, in the case that $d - b < 0$, $\mathrm{E}(p)$, points downwards, i.e., is concave down. Third, as shown in Section 2.3, $c + 2d > a + b$ and $a - b > c$.

To obtain the optimal pass rate, we simply plug in the linear usage curves to FOC (2):

$$
\begin{aligned}
(1-p^*) \times \frac{\partial RE(p^*)}{\partial p} - RE(p^*) + p * \times \frac{\partial PE(p^*)}{\partial p} + PE(p^*) = 0 \\
(1-p^*) \times b - (a + b \times p^*) + p^* \times d + (c + d \times p^*) =& 0 \\
2p^* \times (d-b) + c + b - a =& 0 \\
p^* =& \frac{1}{2}\left(\frac{a-c}{d-b} - \frac{b}{d-b}\right).
\end{aligned}
$$

A.3 Visualizing Corner Solutions

Figure A1 presents a case where one should always pass. As can be seen in the figure the loss of efficiency in passing and the gains in rushing as passing becomes expected are not enough to outweigh the premium from passing.

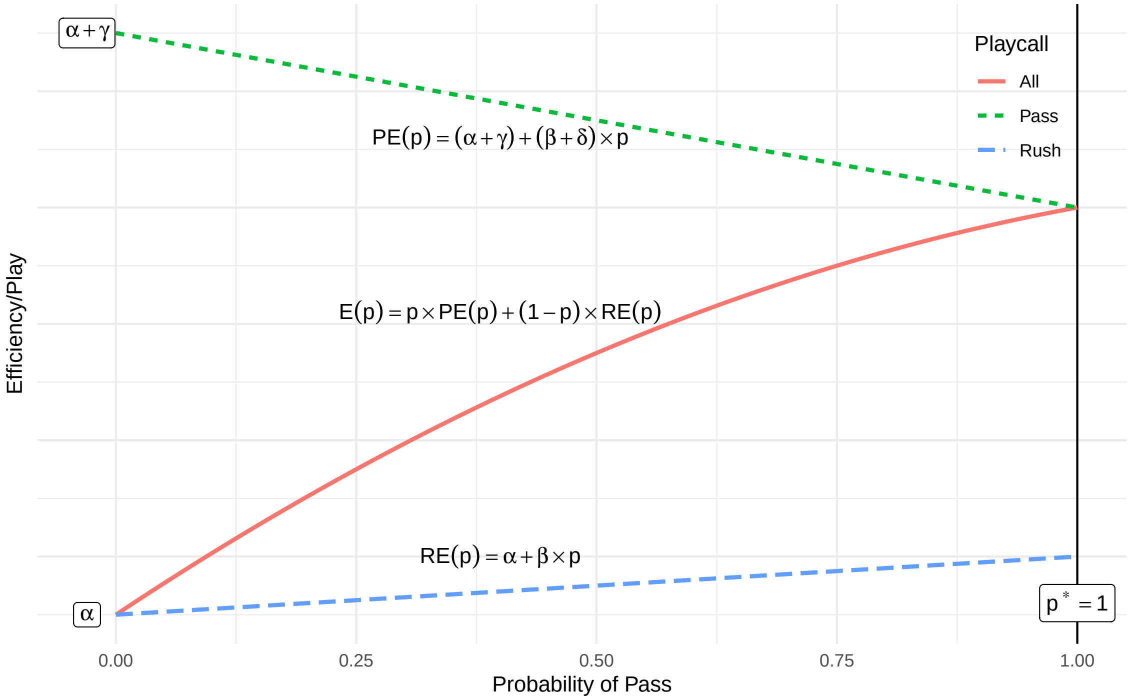

Figure A1: Our model of playcalling efficiency. Plotted are the usage curves for passing (PE) and rushing (RE). Additionally, we plot the efficiency curve, which is the play call weighted sum of these usage curves. For this example, we choose parameters so that the playcaller should always pass ($\beta + \gamma + 2\delta \geq 0$).

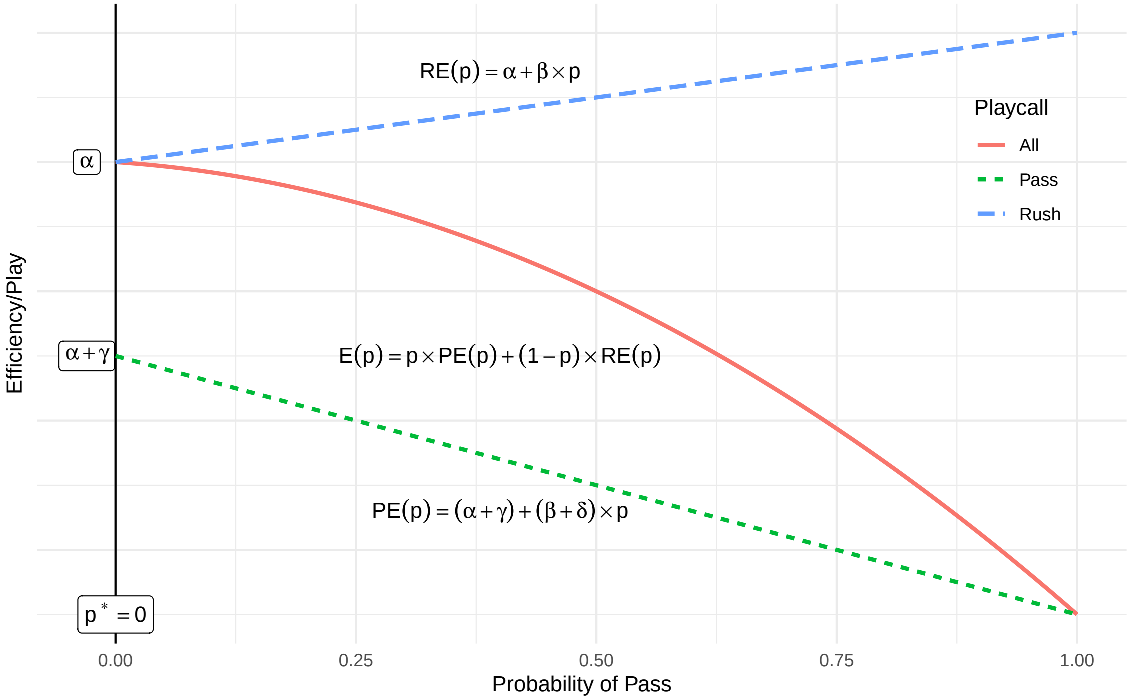

Figure A2: Our model of playcalling efficiency. Plotted are the usage curves for passing (PE) and rushing (RE). Additionally, we plot the efficiency curve, which is the play call weighted sum of these usage curves. For this example, we choose parameters so that the playcaller should always rush ($\gamma + \beta \leq 0$).

B The LATE of Passing on Early Downs: Single Instruments

Table B1 presents the treatment effects from estimating single instrument models. These models tend to reveal similar results as utilizing both instruments, with and without fixed effects. In all specifications passing leads to greater per play efficiency, with the premium ranging from 0.26 EPA to as much as 0.41 EPA.

Table B2 presents the first stage results for these models. The coefficients on the instruments are all significant at the 0.1% level and point in the theoretically expected direction: Team fumbles lost increase passing rate, opposing fumbles lost decrease passing rate, and passing exhibits negative serial correlation.

Why does lagged pass over expected tend to result in higher impacts than fumbles lost? We hypothesize this is the case because we are estimating LATEs. Specifically, if there are more fumbles in outdoor games with poor ball handling conditions (due to wind, rain, or temperature), this would suggest higher weighting of these situations in the LATE. However, as these conditions would also reduce passing efficiency relative to rushing efficiency, this would mean that this LATE might lead to a lower estimate when compared to lag pass over expectation. As negative serial correlation has a behavioral root (i.e., coaches attempting to mimic randomness), it should apply just as well to fair weather and games played inside domes.

Table B1: Causal Effects of Called QB Dropbacks on Efficiency. Single instrument estimates. First and second down plays from 2006-2020. Controls omitted from table.

	EPA			
	Fumbles Lost		Lagged Pass OE	
	(1)	(2)	(3)	(4)
QB Dropback Fit	0.262*	0.308**	0.386***	0.405***
	(0.143)	(0.133)	(0.041)	(0.037)
Fixed Effects	No	Yes	No	Yes
Observations	352,984	352,984	344,146	344,146
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table B2: Causal Effects of Called QB Dropbacks on Efficiency. First stage. First and second down plays from 2006-2020. Controls omitted from table.

	Pass			
	Fumbles Lost		Lagged Pass OE	
	(1)	(2)	(3)	(4)
1 Pos. Fum. Lost	0.021^{***}	0.023^{***}		
	(0.005)	(0.004)		
2 Pos. Fum. Lost	0.045^{***}	0.046^{***}		
	(0.009)	(0.009)		
3 Pos. Fum. Lost	0.098^{***}	0.099^{***}		
	(0.018)	(0.018)		
1 Opp. Fum. Lost	-0.024^{***}	-0.024^{***}		
	(0.004)	(0.004)		
2 Opp. Fum. Lost	-0.055^{***}	-0.059^{***}		
	(0.009)	(0.009)		
3 Opp. Fum. Lost	-0.129^{***}	-0.140^{***}		
	(0.020)	(0.020)		
Lag PassOE			-0.001^{***}	-0.001^{***}
			(0.00002)	(0.00002)
Observations	352,986	352,986	344,148	344,148
R^2	0.004	0.015	0.010	0.023
Note:		*p<0.1; **p<0.05; ***p<0.01		

C Estimating Usage Curves using Instrumental Variables

C.1 LATEs and Optimal Pass Rates

Instrumental variables estimates are LATEs, meaning that the individual effects are weighted by response to the instruments. Therefore, in understanding While we expected lagged passing over expected to affect response relatively symmetrically across teams and games, we expect that our fumbles lost instrument tends to weight outdoor games with poor ball handling relatively more as there are more fumbles in these games—situations that correspond with reduced passing efficiency relative to rushing. To the degree that fumbles lost drives weighting, we would expect that our optimal passing rate is is local to more adverse conditions than are faced on average in the NFL. However, we are not particularly worried about this possibility. First, if anything, this would imply a higher optimal pass rate than we report meaning our results are conservative. Second, as two stage least squares will balance weighting across the two instruments, this should also balance out the weighting of the LATEs. This interpretation is consistent with our results on the causal effect of early down passing (see Appendix B).

C.2 Visualizing Usage Curves Using OLS and IV Results

The comparison between OLS and IV results are illustrated as usage curves and optimal pass rates in Figure C1. The usage curves for IV are much steeper than those for OLS.

C.3 Relevance with Multiple Endogenous Variables: Fixed Effects

Table C1 presents evidence of instrumental relevance from the fixed effects specifications. Once again, these results tell a coherent story: fumbles lost increase the probability of passing, the decision to pass, and the interaction of these two variables. Likewise, fumbles gained reduce these three outcomes. Lagged pass over expected exhibits negative serial correlation, though this

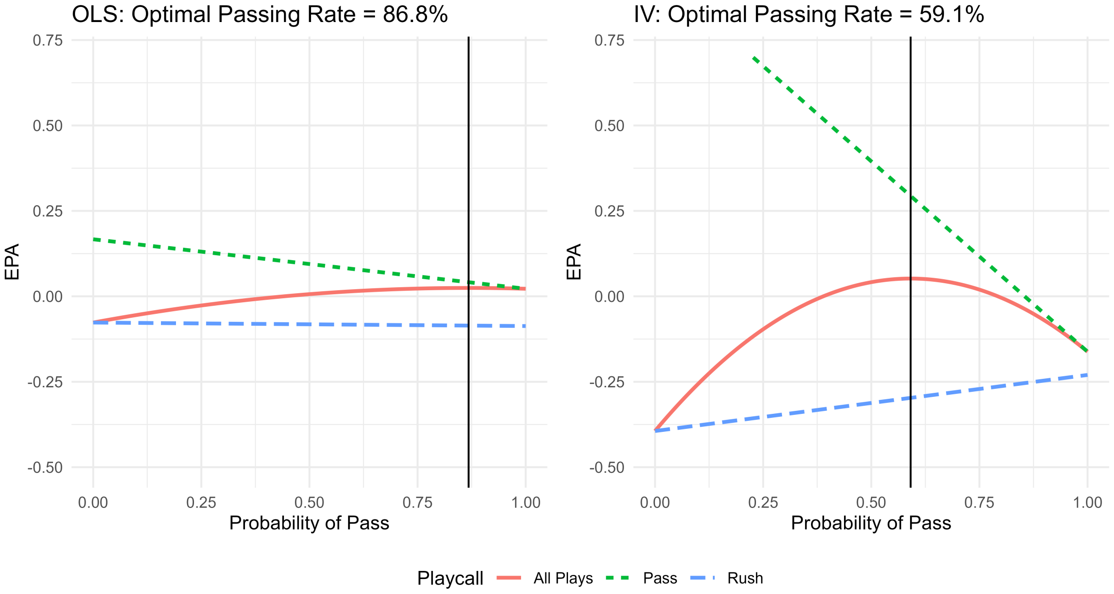

Figure C1: OLS and IV Estimates of Optimal Pass rates with Linear Usage Curves

is attenuated when there have been many lost fumbles.

Table C1: First Stage Results from FE 2SLS Usage Curves Estimation. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	Dependent variable:		
	Pass	XPass	XPass x Pass
	(1)	(2)	(3)
1 Pos. Fum. Lost	0.023^{***}	0.020^{***}	0.024^{***}
	(0.004)	(0.003)	(0.004)
2 Pos. Fum. Lost	0.046^{***}	0.040^{***}	0.046^{***}
	(0.009)	(0.006)	(0.008)
3+ Pos. Fum. Lost	0.099^{***}	0.081^{***}	0.097^{***}
	(0.019)	(0.014)	(0.018)
1 Opp. Fum. Lost	-0.023^{***}	-0.023^{***}	-0.024^{***}
	(0.004)	(0.003)	(0.004)
2 Opp. Fum. Lost	-0.059^{***}	-0.049^{***}	-0.057^{***}
	(0.009)	(0.006)	(0.008)
3+ Opp. Fum. Lost	-0.141^{***}	-0.099^{***}	-0.119^{***}
	(0.020)	(0.014)	(0.016)
1 Pos. Fum. Lost × Lag Pass OE	0.0003^{***}	0.0001^{***}	0.0002^{***}
	(0.0001)	(0.00002)	(0.00004)
2 Pos. Fum. Lost × Lag Pass OE	0.001^{***}	0.0001^{***}	0.0004^{***}
	(0.0001)	(0.00005)	(0.0001)
3+ Pos. Fum. Lost × Lag Pass OE	0.0003	0.0001	0.0002
	(0.0003)	(0.0001)	(0.0002)
1 Opp. Fum. Lost × Lag Pass OE	0.0002^{***}	0.0001^{***}	0.0001^{***}
	(0.0001)	(0.00002)	(0.00004)
2 Opp. Fum. Lost × Lag Pass OE	0.0005^{***}	0.0001^{**}	0.0003^{***}
	(0.0001)	(0.00005)	(0.0001)
3+ Opp. Fum. Lost × Lag Pass OE	0.001^{*}	-0.00003	0.0003
	(0.0003)	(0.0001)	(0.0002)
Lag Pass OE	-0.001^{***}	-0.0005^{***}	-0.001^{***}
	(0.00003)	(0.00001)	(0.00002)
Observations	344,148	343,203	343,203
R^2	0.028	0.060	0.044
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

D Multicollinearity

D.1 Multicollinearity and IV Estimates

Table D1 presents variance inflation factors in the OLS and 2SLS regressions. Variance inflation factors tend to be high among the three coefficients we estimate in our usage curve. This is true in the OLS estimates, where the VIF for $\hat{\beta}$ and $\hat{\delta}$ are above 10. However, they multiply in the 2SLS regression: all the VIFs are above 10 and $\hat{\delta} = 72.1$, considerably higher than is recommended. This is consistent with other cases where simultaneous equations are used to model systems with high degrees of correlation (e.g., fluvial systems as described in Rhoads, 1991). Therefore, we should expect some variance around the individual coefficients, and indeed, the coefficients we find for IV regression are quite noisy.

We are limited in our ability to manage multicollinearity to obtain precise estimates of the coefficients. We are already using a large sample, and indeed it is large enough that multicollinearity does not matter much to the variance of OLS estimates, despite high variance inflation factors there. However, we will need more years of data before we have sufficient sample size in the 2SLS regressions. Likewise, techniques like principal components analysis (that create orthogonal indices) are not appropriate for this case.

Table D1: Variance Inflation Factors for OLS and IV Regressions

Coefficient	OLS	IV
$\hat{\beta}$ (Pass)	11.2	36.2
$\hat{\gamma}$ (XPass)	2.6	18.5
$\hat{\delta}$ (Pass × XPass)	15.4	72.1

D.2 Multicollinearity and Optimal Pass Rates

While we cannot obtain precise estimates of the coefficients, we are able to obtain precise estimates of the optimal pass rate. One reason for this is the correlations between regression coefficients. Table D2 presents the correlations from the Jackknife regressions, i.e., how coefficients vary together as we drop clusters. The correlation in coefficients is quite high, and explains the increase in precision in our estimate. Recall that

$$\hat{p}^* = -\frac{1}{2}\left(\frac{\hat{\gamma} + \hat{\beta}}{\hat{\delta}}\right). \tag{1}$$

Specifically, $\hat{\gamma}$ and $\hat{\beta}$ which make up the numerator are positively correlated with each other. However, they are both negatively correlated with $\hat{\delta}$ at higher rates. As $\delta < 0$, this translates to an increase in magnitude in both the numerator and denominator of $\hat{p}^*$.

The coefficients increase to some degree from OLS to IV. One concern is that if this inflation is due to high variance, given the correlation structure of the coefficients, this might reduce the estimate of p^*. If so, we might interpret $\hat{p}^*$ from the IV estimates as a lower bound of p^*.

Table D2: Correlations Between Jackknife 2SLS Coefficients

	$\hat{\beta}$ (Pass)	$\hat{\gamma}$ (XPass)	$\hat{\delta}$ (Pass × XPass)
$\hat{\beta}$ (Pass)	1.00	0.89	-0.99
$\hat{\gamma}$ (XPass)		1.00	-0.93
$\hat{\delta}$ (Pass × XPass)			1.00

D.3 Limited Information Maximum Likelihood

One way to address this might be to use limited information maximum likelihood (LIML) (Stock and Yogo, 2005). This approach reintroduces some of the endogenous variation used by OLS, and is typically used as a solution to reduce the bias of weak instruments, when this occurs. It is plausible that it could be useful for our case, however, it ends up yielding very similar results. LIML is controlled by a parameter κ, found using a known procedure. When $\kappa = 1$, LIML will yield the same results as 2SLS. However, when we run this (using the ivregress command in Stata), $\hat{\kappa} \approx 1$, meaning that these estimates would be redundant.

E Instrumental Validity: Passing Aggression

Table E1 presents results of regressing passing strategy and efficiency on our instruments, controlling for fumbles. We measure passing aggression via the share of deep passes thrown on dropbacks and passing EPA. However, we find that on early down pass attempts, cumulative fumbles lost and lagged passing over expected do not increase the share of deep passes, nor do they systematically increase the EPA per dropback on called passes. Among the instruments, we find two coefficients significant at the 5% level, about what one would expect due to random noise. This suggests that the exclusion restriction is not violated by coaches adjusting passing aggression in addition to passing.

Table E1: Effect of Cumulative Fumbles Lost on Passing Aggression. First and second down plays outside of the two minute warning from 2006-2020. Controls omitted from table.

	Dependent variable:			
	Deep Pass		Dropback EPA	
	(1)	(2)	(3)	(4)
1 Pos. Fum. Lost	−0.002	−0.002	0.001	0.003
	(0.003)	(0.003)	(0.012)	(0.011)
2 Pos. Fum. Lost	0.001	−0.001	0.002	0.006
	(0.006)	(0.006)	(0.022)	(0.022)
3+ Pos. Fum. Lost	0.001	−0.003	0.047	0.072
	(0.013)	(0.013)	(0.046)	(0.046)
1 Opp. Fum. Lost	−0.004	−0.005	0.007	0.011
	(0.003)	(0.003)	(0.011)	(0.011)
2 Opp. Fum. Lost	−0.003	−0.003	0.002	0.011
	(0.007)	(0.007)	(0.023)	(0.023)
3+ Opp. Fum. Lost	−0.008	−0.003	0.024	0.021
	(0.016)	(0.016)	(0.056)	(0.057)
Lag Pass OE	−0.00002	−0.00001	-0.0002^{**}	-0.0003^{***}
	(0.00002)	(0.00002)	(0.0001)	(0.0001)
Constant	0.181^{***}		0.097^{***}	
	(0.002)		(0.005)	
Fixed Effects	No	Yes	No	Yes
Observations	161,971	161,971	178,210	178,210
R^2	0.0002	0.007	0.0003	0.007
Note:				*p<0.1; **p<0.05; ***p<0.01