14TH ANNUAL MIT SLOAN SPORTS ANALYTICS CONFERENCE
MARCH 6 - 7, 2020 | BOSTON, MA
presented by ESPN

Pulling Starters

Duncan Finigan Brian M. Mills Daniel F. Stone

Paper Track: Baseball

Paper ID 1548770

1. Introduction

Bayesian updating about an unobserved state of the world is hard. Especially when the state of the world (may) also be changing. While various systematic biases have been documented, many questions remain regarding variation in behavior within and across contexts. An important issue that has received relatively little attention is to what extent experts with strong incentives are subject to the most well-known systematic biases.[1]

We address this question by studying one of the most important strategic decisions in baseball: when (if at all) to make the "call to the bullpen" and relieve the starting pitcher. In baseball, there are two types of pitchers, starters and relievers. Starters indeed start the game and usually pitch the majority of the game. Managers decide when to "pull" the starter (replace him with a reliever) based on evidence that the starter is tiring and other new information used to update beliefs throughout the game. Managers must be careful not to pull the starter too soon (replacing him with an inferior reliever and/or depleting reliever resources) or too late (after the damage has been done).[2]

Various psychological factors could cause biases in these decisions. A conformity heuristic could cause decisions to be systematically off-base (in either direction); for an example of this type of behavior occurring in another sports setting see Romer (2006). The omission bias— the bias toward favoring errors of omission and not commission (in our context, pulling the starter) (Ritov and Baron, 1990; Bar-Eli, Azar, Ritov, Keidar-Levin, and Schein, 2007)—could lead starters to be generally left in games too long. Inattention could cause decisions to be insufficiently sensitive to available information (Gabaix, 2017). A recency bias, in particular the hot-hand bias, a bias toward overestimation of persistence of particularly "hot" or "cold" performance, could cause overreaction to recent information (Benjamin, 2018).[3] Confirmation or primacy bias could cause managers to pay

[1] See Benjamin (2018) for an excellent review of the literature on biases in formation and updating of beliefs under uncertainty. The literature on beliefs of highly experienced and incentivized agents outside of the lab is still relatively limited. See Green and Daniels (2018) for evidence of Bayesianism in baseball by another type of expert agent, umpires. However, their paper studies split-second judgments about the relatively simple issue of calling balls and strikes; the choices that we study (pulling starters) are both more complex and reflective.

[2] We provide a brief summary of the rules of baseball, and key terminology, in Section 2.

[3] The hot hand bias is closely related to biases causing extrapolative expectations in other contexts, such as finances, e.g. diagnostic expectations, as also discussed by Benjamin (2018). It is also worth noting that the fact that managers pull starters early when performing poorly and leave them in longer when performing well is prima facie evidence of belief in cold and hot hands in this context, a controversial topic in other settings (Benjamin, 2018), and that Stone and Arkes (2018) find evidence of underreaction to hot and cold hands.

too much heed to the starter's typical quality or quality at the start of the game (Zhou, Liu, and Ho, 2015).

The question of whether managers pull starters optimally has received surprisingly little attention in prior academic literature. We are aware of only one earlier study, Ganeshapillai and Guttag (2014).[4] They estimate a model of when to pull a starter so as to minimize the probability of giving up at least one run in the current inning, and find that managers' decisions differ from estimated optimal choices in 48% of late close game situations. Approximately 90% of these mistakes were leaving the pitcher in when he should be removed; i.e., starters seem to usually be pulled too late.

We build on this work by examining the optimality of pulling starters decisions with respect to what is typically the ultimate objective for each game—winning—in addition to runs allowed in the current inning. We do this by first, in Section 3, showing that in a simple model, optimal decisions for pulling starters imply that expected runs should decline in the inning that starters are pulled. That is, managers should not pull the starter when the manager is just indifferent between the starter and reliever's expected performance for the duration of the current inning, but when the reliever's performance is, in expectation, strictly better for that inning. This is because using a reliever in the current inning yields a benefit of lower expected runs in the current inning at the cost of a depleted bullpen and greater expected runs later in the game. Managers should wait to pull until the marginal benefit to the current inning from pulling is as large as the marginal benefit to future innings from delay. These results imply that Ganeshapillai and Guttag's results are potentially consistent with optimal managerial behavior.

We then empirically examine the effects of bullpen decisions on both outcomes in Section 4. We use linear probability models to estimate the effect of pulling the starter on the probability of winning the game using detailed controls for game situation and other relevant factors. We use similar models to estimate this effect for the outcome of giving up at least one run in the inning. While our empirical setting does not offer ideal random variation in managerial decisions, it does offer a rich array of observables for a large sample (all games from the 2008- 2017 seasons). This allows us to adjust precisely for key situational factors affecting both win and bullpen usage probability with score-inning-baserunner-out interactions, and to include proxies for a large array of additional such factors. We argue that the treatment, pulling the starter, is plausibly "as good as random" given these controls, and so if teams win more (less) often when starters are pulled, ceteris paribus, this would imply that starters are generally pulled too late (too soon). We therefore refer to estimating the effect of pulling the starter on outcomes throughout the paper. However, given the lack of randomization, we acknowledge that this interpretation is ambiguous. Our estimates can of course more conservatively be interpreted as partial correlations, which we think are still of interest.[5]

We find essentially no significant evidence that teams make such systematic mistakes in pulling starter decisions. Our results for the full sample are precisely estimated near-zero point

[4] See, e.g., Carleton (2017) and Houston (2018) for discussion and analysis of this topic outside of the academic literature.

[5] In addition to examining wins and not just runs, we also expand on Ganeshapillai and Guttag's work by using different empirical methods, additional control variables, examine different game situations, and obtain numerical estimates of the effects of pulling as starter on runs and win probability. We also use a larger data set with more recent seasons (2008-2017 vs. 2006-2010). There is evidence that analytics have affected decision-making in the last two seasons of our sample as starters are pulled somewhat earlier on average than in the rest of the sample. All of our results are similar when we restrict attention just to the prior seasons (2008-2015).

estimates, with our preferred model yielding a 95% confidence interval for the marginal effect of pulling a starter on win probability of -1.4 to +1.8 percentage points.[6] Moreover, we find the results are largely stable when various proxy controls are removed, indicating lack of confounding by omitted variables, and that pulling the starter does significantly decrease the chance of allowing a run in the inning, consistent with the theory we present and Ganeshapillai and Guttag's earlier results.

It is possible that an average effect on win probability of zero could mask off-setting biases in different game situations. Our results largely suggest that this is not the case, as estimates for subsamples restricted to various game situations and contexts (score difference, outs, men on base, and league) are typically small and insignificant. We also fail to find significant evidence of recency biases (managers do not seem to jump the gun and pull starters too soon after giving up walks, hits, runs, or even a measure of lucky runs that we construct). We corroborate these results using an alternative measure of managerial quality, votes for the Manger of the Year award.

Our results might therefore seem to support the conclusion that managerial experience does largely reduce or eliminate bias. We note that bias could affect both belief updating, the focus of our discussion above, and other factors that could affect the optimality of the complex dynamic problem of when to call the bullpen, discussed further in Section 4.1. However, there was a steady downward trend in the mean time that starters were pulled from approximately 1970 through the early 2000s, most of which occurred before the so-called "sabermetrics" revolution (Carleton, 2017; Hakes and Sauer, 2006). Thus, even if managers have learned to make decisions optimally, this learning did not take place over the course of just their careers, but also the careers of those before them. This learning was made possible largely due to the stability of the sport over time—there have not been any substantial rule changes since 1973 when the American League introduced the designated hitter.[7] Our interpretation is therefore that yes, experience reduces bias, but the quantity of experience required for this learning to occur can be very large, and is likely context-dependent.[8] Similarly, it has taken decades for NBA teams to learn the value of the three pointer, and NFL teams have gradually become more likely to "go for it" on fourth down since Romer (2006).[9]

Moreover, we do obtain two somewhat more puzzling results. One is well known to baseball fans, but rarely remarked upon: in 99.3% of cases starters pulled in our sample, this occurred in between, rather than during, at bats. This is surprising since managers could obtain information during an at bat (from individual pitches) that indicates that pulling the starter is optimal mid-at bat. We speculate (but lack sufficient data to test) that this regularity is due at least partly to convention. The second puzzling result reflects another type of conformity: we find that the probability of pulling the starter increases by 5.3 percentage points when the opposition team has pulled their starter first

[6] See Abadie (2018) for recent work on the informativeness of results in which the null is not rejected.

[7] Another factor (beyond managerial learning from experience and sabermetrics/statistical analysis) that has contributed to changes in the usage of pitchers over time is that pitchers have become more specialized, including relief pitchers specializing in different roles within the bullpen (however, learning has also driven changes in specialization Carleton (2018b)).

[8] Starters were pulled even earlier on average in the last two years of our sample and this trend has continued since then (Carleton, 2018a). We discuss this issue further as we proceed. See, https://shottracker.com/articles/the-3-point-revolution and https://www.nationalreview.com/2019/02/nfl-football-teams-fourth-down-plays/. See Fudenberg and Levine (2016) for discussion of "slow learning" in a broader range of game theoretic situations.

[9] See Weinberg (2015) for more detail on the rules of baseball in general and on the strategic question of when to pull starters in particular.

(conditional on the full set of other controls). However, we do not find significant interaction effects between the opposition's decision and the team's own decision on game outcomes. This suggests that, on average, small changes in when the starter is pulled do not have substantial effects on win probability, which could also help to explain our failure to find substantial biases in these decisions.

2. Data and Empirical Context

We provide a cursory and simplistic summary of the rules and terminology of baseball most relevant to our analysis to assist readers who are unfamiliar with the sport.[10] Baseball games consist of nine innings, each consisting of two half innings. In the first (top) half of each inning, the home team pitches while the away team bats, and vice versa in the second (bottom) half inning. Each batter is pitched to repeatedly until he either records an out, gets on base (usually via a hit or walk) or hits a home run (a ball hit out of the park causing the batter and any men on base at the start of the at bat to score). Base-runners (men on first, second, and/or third base) may also either be caused to advance bases or score (record a run by advancing from third), or record an out, when the batter records a hit, walk, or out. A half inning is over after three outs are recorded. The team with more runs after nine innings wins the game; the game goes to extra innings if there is a tie (continuing as long as the game is tied after each complete additional inning, this happens in less than 10% of games).[11]

We obtain detailed pitch-level data for all games (162 per season for each of the 30 teams) for the 2008-2017 seasons from Baseball Savant (https://baseballsavant.mlb.com/). The data include pitch location, type (fastball, curveball, etc.), and velocity for the large majority of observations, in addition to the more widely available data on the outcomes of each at bat (hits, walks, outs, etc.). We aggregate the data to the at bat unit of observation since as noted earlier over 99% of pitching changes in our sample are made in between at bats, but still exploit the data we have on individual pitches, using pitch count and velocity in particular to account for the starting pitcher's actual or perceived physical condition.

Given our goal of assessing the effects of pulling starters on game win-loss outcomes, it is helpful to restrict our sample in several ways. Starters are most often pulled in the 6th and 7th innings. We restrict our sample to these innings, plus the 8th, as decisions in that inning have relatively large effects on which team wins the game. We also limit the sample to at bats in which the current score difference is at most one run to maximize variation in win probability and potential effects of bullpen decisions on win probability.[12]

While winning the current game is likely the main objective for most teams in most games throughout the season, it is certainly plausible that teams have other objectives in some games (e.g., development

[10] See Weinberg (2015) for more detail on the rules of baseball in general and on the strategic question of when to pull starters in particular.

[11] See https://www.beyondtheboxscore.com/2017/8/5/16093390/extra-innings-time-how-long-how-many-average-rule-change.

[12] Restricting the sample to observations from the 6th inning and later also means that, for the vast majority of our observations, pitchers will be facing batters for at least the third time in the game, so we do not account directly for "third-time-through-the-order" in our regressions since this variable would have almost zero variation.

of younger players for teams out of contention). This is most likely true late in the season; moreover, team rosters can change substantially late in the season, and starting September 1 the roster size expands, so we restrict our sample to pre-September games to minimize the influence of both of these issues.[13] Since the season begins in late March or early April and ends in late September, this only causes us to lose around one sixth of our sample. Given the large number of teams and games per season, our final sample still has over 85,000 plate appearance level observations despite the various restrictions.

As noted in Section 1, there was a steady decline in the mean number of innings that starters are kept in per game beginning in the 1970s through the early 2000s (see Figure 1). This change has been driven by a variety of factors including increased use of data and statistics, "learning by doing" and social learning, and changes in specialization and development of players (Carleton, 2018a).[14] The trend is noisy during our sample time-frame but does seem to dip substantially downward in the last two years; we use the full sample throughout our analysis but do check robustness of our main results to dropping the last two seasons.

3. Model

As noted earlier, the basic trade-off that managers face when deciding when to pull the starter at "time" t (denoted in this section by PS_t, with $PS_t = 1$ $(= 0)$ referring to (not) pulling the starter at t) is the benefit of a relatively fresh reliever versus the cost of either depleting the bullpen or using a relief pitcher who is lower quality than the starter. Thus, if managers tend to pull starters too soon, we would expect that those who do pull starters later would have both a relatively high chance of winning the game and (relatively) less expected runs allowed in the inning the starter is pulled. Similarly, if managers tend to pull too late, we would expect a higher win probability and less runs for cases where the starter is pulled sooner.

However, even if pulling the starter reduces (in expectation) runs scored in the current inning, this could either increase or decrease the chance of winning the game. This is best seen with a simple model. Suppose time, t, is continuous and ranges from 0 to 1, and the starting pitcher has a "runs density function" of $f(t)$, meaning that the expected runs the starter gives up from t_1 to t_2 is $\int_{t_2}^{t_1} f(t)dt$. We assume $f'(t) > 0$ (starters tire throughout the game, on average, or are more likely to see performance decline when facing the batting order multiple times), and normalize $f(0)$ to 0. It is natural to assume that win probability is strictly decreasing in runs allowed for the game, and so we use runs allowed as a proxy for win probability.

The quality of (i.e., runs allowed by) relievers may decline or improve throughout games for various reasons. The relative quality of a reliever brought in at any particular point in the game is ambiguous. However, the average quality of the bullpen for the remainder of the game likely declines when the

[13] We also therefore exclude postseason games; strategic decisions in these games are fundamentally different than regular season games due to the limited horizon, and while it would be ideal to examine them separately, we are not able to do so due to relatively small sample size.

[14] Changes in runs scored per game could also contribute to this trend, but while average runs did increase from the 70s through the 90s, this has decreased since then and is now similar to what it was in the 70s (see, e.g., https://tht.fangraphs.com/the-height-of-the-hill/).

bullpen is called on earlier, both due to fatigue and the depletion of supply of higher quality relievers. For example, if a reliever is brought in in the 6th inning and left in until the 8th, his performance in the 8th would be expected to be worse as compared to the same reliever being first used in the 7th inning. Or perhaps he would be relieved in the 7th with a lower quality option than a fresh version of himself. Thus, we assume that it is the average runs per inning of bullpen usage (for the remainder of the game) that declines when the bullpen is called on later in the game.

We therefore use $g(t)$ to denote the average number of runs per unit of time allowed by the bullpen when the bullpen is called at time t, and assume that $g'(t) < 0$, reflecting that the quality of the bullpen (for the remainder of the game) increases as t increases (i.e., when the starter is pulled later). Given the time normalization, when the bullpen is called at *t*, it is used for $(1 - t)$ units of time, so the expected runs allowed by the bullpen equals $g(t)(1 - t)$. It is technically useful and without loss of generality to assume $g''(t) < 0$.

Consequently, expected runs for the game when $PS_{\hat{t}} = 1$ is equal to $\int_0^{\hat{t}} f(t)dt + g(\hat{t})(1 - \hat{t})$. Differentiating with respect to $\hat{t}$ is $f(\hat{t}) + g'(\hat{t})(1 - \hat{t}) - g(\hat{t})$, with second derivative $f'(\hat{t}) + g''(\hat{t})(1 - \hat{t}) - 2g'(\hat{t})$, which is weakly positive, and so expected runs are minimized at t^* such that: $f(t^*) + g'(t^*)(1 - t^*) = g(t^*)$. Since $g'() < 0$, the following is implied:

Proposition 3.1. *At the expected-runs minimizing time to pull the starter, t^*, $f(t^*) > g(t^*)$.*

This result means that we should expect a strict improvement in current pitching quality at the optimal time that the starter is pulled. Empirically, this would imply lower runs in innings where starters are pulled as compared to when they are not pulled, ceteris paribus. This is because of the future cost of pulling the starter due to depleting the bullpen. The manager should wait to pull the starter until there is a marginal benefit in the present (improvement of reliever over the starter in the current inning) to compensate for the future cost.

Specifically, given that $f(t) \geq g(t)$, the marginal cost of delay in $PS_t = 1$ is the marginal increase in current runs, $f(t) - g(t)$, while the marginal benefit of delay (decrease in future runs) is $g'(t)(t - 1)$. If the starter is pulled at the t such that he and the reliever are the same quality ($f(t) = g(t)$), then the marginal cost of delay is zero, while the marginal benefit of delay is always strictly positive (if $t > 1$). Thus, the total marginal effect of a delay would be strictly positive, and so t could not be optimal. Waiting at least somewhat longer and incurring a small cost of higher runs in the current inning is outweighed by the benefit of a slightly fresher bullpen since this applies to the entire remainder of the game.

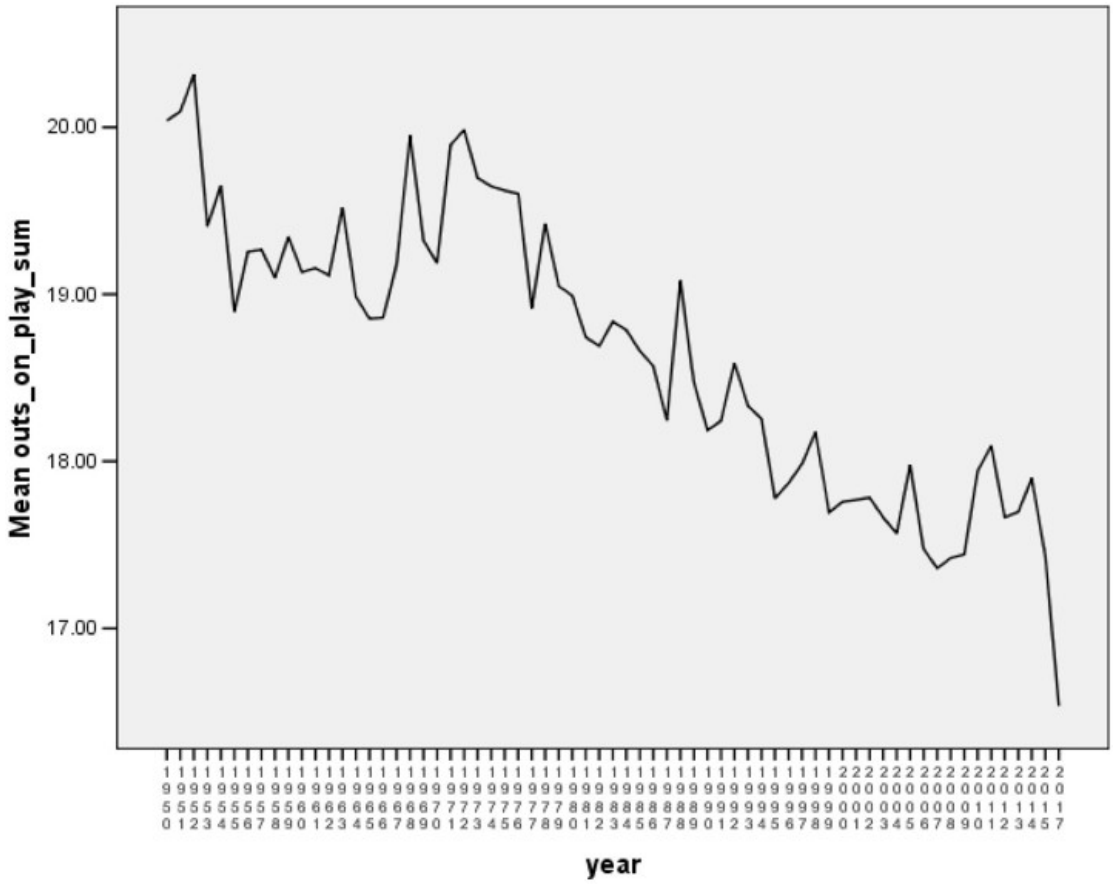
Figure 2 illustrates these points for the simple case of $f(t) = t$ and $g(t) = 1 - t$. The top graph shows the case of the optimal choice. On the margin (at the time starters are pulled), $PS_t = 1$ does not affect win probability (total expected runs for the game is exactly equal for $PS_t = 0$ and $PS_t = 1$), but does affect instantaneous expected runs at *t*. In the second graph, the starter is pulled too early, when

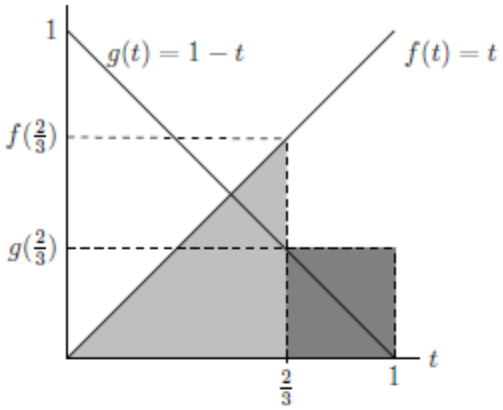
$f(t) = g(t)$. Runs do not change at the instant the starter is pulled, but are higher for the rest of the game. The third graph shows the case of a starter pulled too late.[15]

The key assumption driving the proposition is the decline in average runs per bullpen inning, and not marginal (current) runs. If $g(t)$ were defined to be analogous to $f(t)$ as current expected runs density, and declined over time, then at the optimal time to pull starters $f(t)$ and $g(t)$ would be equal. This alternative case is also plausible as teams may tend to use weaker relievers earlier in the game in general. However, if longer usage causes a reliever's future performance in the game to worsen, this would support the original (average performance) version of $g(t)$. It is also worth noting that teams may pull starters for relatively stronger relievers in more difficult situations (i.e., with men on base). This would exacerbate the result stated in the proposition (we would expect an even greater discontinuous improvement in performance at the optimal time that $PS_t = 1$).

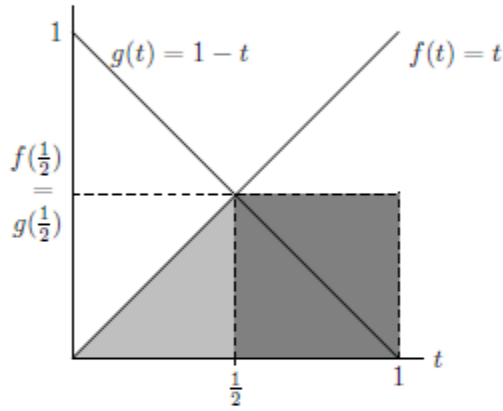
[15] As an alternative illustration of ideas, one could consider an even simpler case in which the starter's expected runs equal 0.5 in the 6th inning and 1.0 in the 7th inning. Suppose the bullpen's average runs per inning for the duration of the game equals 0.5 if called in the 6th and 0.25 if called in the 7th. If the bullpen is called in the 6th, then there are two expected runs for the remainder of the game for innings 6-9 (0.5 per inning and four innings). If the bullpen is called in the 7th, then expected runs in innings 6-9 are equal to 0.5+0.25 3 = 1.25. Expected runs decline in the current inning if the starter is pulled in the 7th, which is optimal (current inning expected runs would not decline if the starter is pulled in the 6th).

Figure 1: Mean number of outs per game recorded by starting pitchers, reproduced from Carleton (2017)

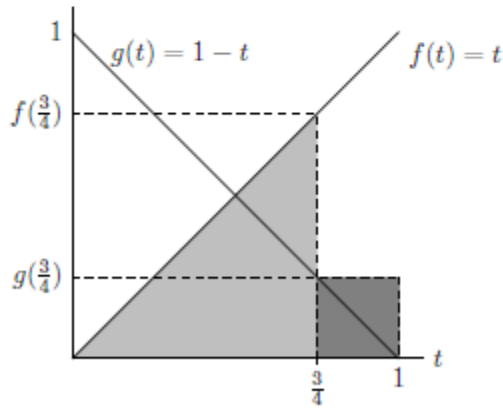




$PS_t = 1$ at $t^* = \frac{2}{3}$ (optimal): expected runs $= (1/2)(2/3)^2 + (1/3)^2 = 1/3$



$PS_t = 1$ too soon $(t = 1/2)$: expected runs $= (1/2)(1/2)^2 + (1/2)^2 = 3/8$



$PS_t = 1$ too late $(t = 3/4)$: expected runs $= (1/2)(3/4)^2 + (1/4)^2 = 11/32$

Figure 2: Expected runs (areas of light gray regions equal runs allowed by starter and darker regions are runs by bullpen) for optimal, too early, and too late decisions for PS=1.

Figure 2: Expected runs (areas of light gray regions equal runs allowed by starter and darker regions are runs by bullpen) for optimal, too early, and too late decisions for PS=1.

4. Empirical Strategy

4.1. Empirical Model and Identification

The model implies that if starters are pulled at the optimal time on average, then the marginal effect of pulling the starter on win probability is zero, but runs in the current inning may decline (ceteris paribus). If starters are pulled too soon on average, then teams that (randomly) pull starters later would be more likely to win their games. Consequently, not pulling the starter would, ceteris paribus, empirically predict a higher win probability. Conversely, if starters are pulled too late on average, then the "empirical marginal effect" of pulling the starter on win probability would be positive.

We test these implications using straightforward linear probability models in which we estimate the effect of pulling the starter on the conditional mean probability of 1) winning the game, and 2) giving up at least one run in the inning, for both our full sample and various subsamples for different game situations, controlling for a large array of characteristics of the teams and game situation.[16] Determining the appropriate controls is the key challenge, as we discuss below. Our baseline models are of the form:

$$Y_i = \beta_{PS} PS_i + \beta_X X_i + \beta_{PS\times X}(PS_i \times X_i) + \beta_Z Z_i^Y + FE_i. \qquad (1)$$

Y_i is one of the two outcomes for the pitching team (for at bat i), either Win ($Y_i = 1$ if the pitching team wins the game) or Runs ($Y_i = 1$ if one or more runs are scored in the inning during or after the current at bat).[17] PS_i is a dummy for whether the starter is pulled at the start of at bat i (so $PS_i = 1$ for at most one observation per game), X_i is a set of covariates interacted with PS_i as their effects directly depend on the value of PS_i (explained further below), Z_i^Y is a vector of additional covariates (whose composition depends on the outcome Y), and FE_i a fully saturated set of score-inning-baserunner-outs fixed effects (*FEs*).[18] We cluster standard errors by game to account for correlation

[16] These implications are motivated by the model of Section 3 but not derived directly. An alternative empirical approach would be a dynamic structural analysis of the pitching change decision, which can be thought of as an optimal stopping problem. Provencher (1997) discuss the relationship between structural and reduced form analysis of optimal stopping problems, and conclude that "for a large class of optimal stopping problems a reduced-form model which closely approximates the statistical performance of its structural counterpart is readily found." Still, a structural model of pulling starters is likely a worthwhile topic for future research; we note this again in our concluding remarks. See Goldman and Rao (2017) for a structural model of optimal stopping in sports (the decision of when to shoot in basketball).

[17] We obtain much more precise and interpretable results when examining a binary runs variable rather than a variable equal to the total number of runs allowed, and since we examine only very close games, the marginal effect on win probability of the first run given up is typically largest, so the dummy for runs is the key runs outcome regardless. Moreover, in unreported multinomial logit analysis we confirm that effects of PS_i are largest on the first run allowed.

[18] Given the sample restrictions, there are three score differences, three innings, eight base- runner situations (none on, one on 1st, 2nd, or 3rd, two on (1st and 2nd, 2nd and 3rd, 1st and 3rd), and bases loaded), and three outs, there are 3 3 8 3 total permutations of these variables.

of observations within these groups. Implicitly, we use $PS = 0$ observations as counterfactuals for $PS = 1$ observations (conditional on the covariates), and assume that unobservables randomly drive variation in the *PS* decision, and are not systematically correlated with both *PS* and *Y*. Thus, our primary co- efficient of interest is β_{PS}; if $\beta_{PS} > 0$ for $Y = Win$, this would imply that pulling the starter increases win probability on the margin, so managers wait too long to pull starters on average, and if $\beta_{PS} < 0$, this would imply managers pull starters too soon. However, the coefficients for some interaction terms, $\beta_{PS \times X}$, are also of interest for some of our hypotheses, as we discuss below. We examine the effects of removing key "proxy controls" (discussed below) to assess the importance of unobservables that are likely correlated with observables, i.e. we use "selection on observables as a guide to selection on unobservables" (Altonji, Elder, and Taber, 2005). We again acknowledge the lack of true randomization of the *PS* treatment, and discuss this issue further below.

4.2. Covariates

FE fully accounts for four key confounding variables. Other categories of variables that could confound the estimated effect of *PS* on both *Win* and *Runs* are as follows: starter's overall "quality", current condition, and matchups with upcoming hitters, the starter's team's bullpen quality and condition, and the overall quality of the opponent's hitting and that of upcoming batters in particular. Runs are more likely scored (and winning the game is less likely), and pulling the starter more (less) likely, when the starter (bullpen) is weaker X_i includes measures of just the starter and starter's team bullpen characteristics, as the effects of each of these variables obviously depend directly on whether the starter has been pulled yet or not ($PS_i = 1$ or $PS_i = 0$). Thus, it is crucial to include interactions of these variables with PS_i, and the coefficient of these interaction terms are not expected to be zero even under a null hypothesis of no managerial mistakes.

Variables used to control for the starter's quality and condition are as follows. We use WHIP from the last three months for games in July and August, and the previous season's *WHIP* otherwise, to control for the starter's typical quality across games (*Starter WHIP*).[19] We use *WHIP* from the first four innings of the game (*First 4 WHIP*) and runs allowed prior to the start of the current inning (*Pre-Inning Runs*) to account for actual or perceived variation in starter quality for the current game. We use *Pitch Count* and *Recent Fastball* to control for the starter's current physical condition. *Pitch Count* is equal to current pitch count minus the starter's average pitch count for the season when pulled, which accounts for variation in general durability across starters. *Recent Fastball* is equal to the average fastball thrown in the last six at bats minus average fastball velocity from the first four innings of the game, which accounts for the extent to which the starter appears to be tiring in the given game.[20] Starter-hitter handedness match-ups (whether the pitcher pitches with the same hand that the batter hits from) are often considered a key factor in baseball strategy (Weinberg, 2015). We account for this with *Next 3 Same Hand*, equal to the number of same-hand matchups the starter has with the next three scheduled hitters.

[19] WHIP (walks and hits per inning pitched is a standard statistic used to measure pitching quality; see, e.g., Lee (2014). We replace the previous season's WHIP with the current season's value for the 2008 season since the previous season is not in our data set.

[20] We use the last six at bats because there are numerous at bats where no fastballs are thrown and so we lose more observations when we use a smaller number of recent at bats. We choose six as it is the number of at bats that would occur in the last two innings if there were no men who reached base, but admittedly this number is somewhat arbitrary.

To account for the overall quality of the pitching team's bullpen we use season-long pre- ninth inning *WHIP* for non-starting pitchers excluding the current game. We cannot use *WHIP* of the actual reliever used since this is unobserved for all observations with $PS_i = 0$ and it cannot be assumed that the eventual reliever used is the same one that would be used at earlier points in the game. To account for the condition of the bullpen, we include variables for bullpen pitch count in the previous two days (*Lag BP Pitch Ct*, *2nd Lag BP Pitch Ct*).

We de-mean all of these variables that are not calculated as differences (i.e., all except *Recent Fastball* and *Pitch Count*) using current season means to account for secular changes. This allows β_{PS} to be interpreted as the marginal effect of PS_i when the variables comprising X_i are equal to their season-means or baselines for the difference variables.[21] An alternative way to account for a number of confounding factors referred to above, and some additional factors discussed below, is to use additional fixed effects. Using *FEs* for team-seasons, opponent team-seasons, and starting pitcher is appealing because it avoids many measurement issues. This leads to a very large number of fixed effects (over 1,300), and so including interactions of these *FEs* with PS_i is computationally difficult and makes interpretation of the non-interacted PS_i term more difficult. It is more straightforward to use specifications that include these *FEs* as additional controls, but excludes interactions of these *FEs* with PS_i; we do examine these specifications.

The vector Z^Y includes controls for factors not interacted with *PS*. For both *Y* outcomes, Z^Y includes controls for two variables accounting for quality of opponent hitting: season-mean *OPS* for the next three (scheduled) hitters and the opponent's (overall) *OPS* for the season.[22] Z^Y also includes controls for home status, year fixed effects, and month fixed effects, for both *Y*.

For the $Y = Win$ outcome, there are several other categories of confounders: the starter team's hitting quality and end of game (closer) bullpen quality, and the other team's starting pitching, bullpen, and closer quality. We again use team *OPS* for the season to account for team hitting quality, and for opponent pitching we use relevant X_i variables for the opponent (*Starter WHIP* and *Bullpen WHIP* (both interacted with PS_i), *Lag Bullpen Pitch Counts*, and *Closer WHIP*).

To account for biases potentially varying by game situation we present results for the main regressions for subsamples defined by outs, inning, men on base, and score. We also estimate the regressions for various subsamples to examine how effects may depend on several other important contextual factors: subsamples for each half of the season (pitchers may be rested more in either half), a subsample of seasons prior to 2016 (as noted above, there seems to be a shift in when starters are pulled in the 2016 and 2017 seasons), and subsamples for games played by National and American League teams (pitchers hit in the National League only, and are replaced by the designated hitter in the American League).

[21] We de-mean the three main variables that vary substantially by subsample (*WHIP*, *First WHIP*, *Pre-Inning Runs*) when we conduct different analyses by subsample.

[22] *OPS* = On-base Plus Slugging percentage is a standard baseball statistic used to measure hitting quality, again see Lee (2014). All statistics are constructed separately by home/away status and use data only from games prior to September.

4.3. Recency

To account for recent events potentially driving biases, we examine interactions of *PS* with several variables measuring events that occurred in recent at bats: *Last T WH* (walks and hits), *Last T Pitch Ct*, *Last T Runs*, and *Last T Lucky Runs*.[23] *Last T* refers to the previous T at bats; we present results for $T = 3$, but results for other values are similar. Further, we include a variable indicating whether the opponent has already pulled its starter (*Opponent PS*); examining this variable allows us to assess possible (undue) influence of the opponent's decision.

The "lucky" variable is the deviation between change in expected runs and the predicted change to expected runs based on pitch location and several other variables, summed over pitches that occurred during the at bat. This variable is intended to capture the possibility that hitters or pitchers may get lucky via incorrect umpire calls or good/bad hit outcomes relative to pitch quality, and that managers may not properly account for luck due to outcome bias (Brownback and Kuhn, 2019). Details of how this variable is constructed are provided in the appendix.

To interpret this variable, consider the following example. Suppose a two-strike pitch is thrown to an area of the plate that is usually advantageous to the pitcher (perhaps low and outside in the strike zone) and the batter manages to make contact and get a single. Now compare this outcome to a two-strike pitch in which the pitcher threw a pitch down the middle of the plate that resulted in a hitter getting a single. The latter pitch had a much higher expected runs effect than the former low-outside example. In this case, the low-outside pitch single would be considered "luckier" than the pitch down the middle, as the expected runs were much lower for the better pitch. In other words, the pitcher made a good pitch and the batter was lucky enough to either guess correctly or throw the bat at the pitch and put it in play, rather than get out. Managers may have the ability to distinguish whether pitchers are fatiguing and performing worse or whether they were unlucky on a recent at bat or pitch, and integrate this information into their *PS* decision. If managers do not distinguish between deserved and lucky outcomes, they may pull starters too soon who have gotten unlucky, or fail to pull starters who have been lucky.

4.4. Alternative Analysis

In a complementary final analysis, we use an alternative measure of managerial quality, votes for the Manager of the Year award (3-year moving average), *MoY*. We regress *PS* on *MoY* and other control variables for the full sample and key subsamples, estimating

$$PS_i = \beta_{MoY} MoY_i + \beta_X X_i + \beta_Z Z^{Win} + FE_i. \qquad (2)$$

The coefficient β_{MoY} is the estimate of interest. If β_{MoY} is not equal to zero, this would suggest that higher quality managers are more or less likely to pull starters in general or in particular situations, which would in turn suggest that other managers tend to pull starters too late or early. We use the X and $Y = Win$ control variables since they could all plausibly affect *PS*. Similarly, we use the full set of score-inning-baserunner-outs FEs, and consider the expanded set of *FEs* (which we discuss further below). We do not interact *MoY* with covariates, just using subsamples to examine how effects may

[23] We use interactions, rather than subsamples, for analysis of these effects because changes in effects of the other covariates are less plausible (and less of interest) in this context.

vary by context. In addition to allowing us to address our paper's main question in an alternative way, this analysis also allows us to assess heterogeneity in pulling starter decision-making.

5. Results

5.1. Preliminary Analysis

Rather than present summary statistics, we present results from the following regression in Table 1:

$$PS_i = \beta_X X_i + \beta_Z Z^{Runs} + \beta_{Recent} Recent_i + FE_i \qquad (3)$$

with $Recent_i$ denoting the vector of "recent events" variables discussed in Section 4.2.[24] This regression provides insight into the information used by the manager to make the *PS* decision, and validation of the covariates used for the main analysis. We use Z^{Runs} here rather than Z^{Win} because the variables affecting Runs are most relevant to the PS decision, and in the interest of limiting the number of variables included in the table. The table shows that nearly all of the variables are significant with the expected sign. Managers are sensitive to the relatively subtle variables of previous games' bullpen pitch counts, recent pitch count, current/next batter quality and handedness—and even the seemingly irrelevant variable of *Opponent PS*. The only non-significant variable is *Bullpen WHIP*. At first glance *Last 3 Lucky Runs* has a seemingly questionable coefficient sign: the negative sign indicates that starters are less likely to be pulled when recent hitters got more lucky runs. However, this sign is appropriate (consistent with rational managerial choice) conditional on the inclusion of (total) *Last 3 Runs*, which has a positive sign. The combination of these effects indicates that managers have some ability to discern whether or not negative outcomes on the field are related to pitcher fatigue or underlying performance, and showing restraint when negative outcomes may be more due to luck.

As an additional preliminary analysis, we regress *Win* and *Runs* on (just) *PS* in various game situations to get a sense of basic correlations and how these vary by situation.[25] We present these results in Table 2. Results with a man on first and other baserunner situations (other than none on) are similar, and we therefore pool these situations going forward. Correlations between *PS* and outcomes are low for these situations. There are larger and statistically significant correlations with none on: in general, $PS = 1$ predicts a team being less likely to win and, consistent with Ganeshapillai and Guttag (2014), more likely to give up runs (with the exception of one or two out situations, where the correlation is positive for *Win*). There are substantial differences in estimates by game score and inning, so we examine these situations separately below as well.

[24] For this regression, we cluster standard errors by team-game since correlation across teams and within games is less of an issue.

[25] Some situations are pooled to smooth samples sizes. For example, the number of observations declines by inning, so we pool the seventh and eighth innings.

Table 1: OLS estimates with dependent variable = PS

Variable Type	Variable Name	Definition	Coefficient
Starter and starter's bullpen (X)	Starter WHIP	Starter's WHIP in last 3 months	0.056*** (0.006)
	First 4 IP WHIP	Starter's WHIP in first 4 innings of current game	0.031*** (0.003)
	Pre-Inning Runs	Opponent's runs scored in game at start of current inning	0.021*** (0.001)
	Pitch Count	Current pitch count minus starter's mean pitch count (per game) for season	0.004*** (0.000)
	Fastball Speed	Fastball speed in last 6 at bats minus mean FB speed from 1st 4 innings	-0.053*** (0.001)
	Next 3 Same Hand	Current + next 2 hitters who bat with same hand as pitcher	-0.088*** (0.001)
	Bullpen WHIP	Team's pre-9th inning bullpen WHIP for season	-0.013 (0.007)
	Lag BP pitch ct	Previous day's pitch count for team's pre 9th inning bullpen	-0.182*** (0.030)
	2nd lag BP pitch ct	Two day prior pitch count for team's pre 9th inning bullpen	-0.139*** (0.030)
Opponent hitting (Z)	Opponent OPS	Opponent's OPS for season	-0.124*** (0.024)
	Next 3 OPS	Next 3 (including current) hitters' season OPS	0.154*** (0.013)
Recent events	Last 3 WH	Walks + hits by opponent in last 3 at bats	0.029*** (0.003)
	Last 3 Pitch Ct	Pitch count for last 3 batters	0.007*** (0.000)
	Last 3 Runs	Runs scored over the last 3 batters	0.039*** (0.003)
	Last 3 Lucky Runs	Lucky (defined in text) runs by opponent in last 3 at bats	-0.010*** (0.003)
	Opp. PS	Opponent PS	0.053*** (0.003)
		Adj R^2	0.376
		N	86868

Notes: Model includes year and score-outs-baserunners-inning fixed effects. Standard errors in parentheses clustered by team-game. See text for more detail on definitions of some variables. *** denotes 1% significance.

Table 2: Preliminary regressions

	LHS = Win			LHS = Runs		
	None on	Man on 1st	Other	None on	Man on 1st	Other
Full sample						
PS	-0.049***	0.011	0.010	0.060***	-0.030*	0.008
	(0.006)	(0.015)	(0.011)	(0.005)	(0.014)	(0.011)
N	56848	15208	18147	56848	15208	18147
6th inning						
PS	-0.045***	0.009	0.015	0.061***	-0.025	0.003
	(0.013)	(0.025)	(0.016)	(0.010)	(0.022)	(0.016)
N	31593	9451	11755	31593	9451	11755
7th or 8th inning						
PS	-0.058***	0.013	0.001	0.072***	-0.028	0.024
	(0.007)	(0.020)	(0.015)	(0.006)	(0.017)	(0.015)
N	25255	5757	6392	25255	5757	6392
No outs						
PS	-0.031***	-0.036	-0.024	-0.020***	-0.049	0.029
	(0.007)	(0.031)	(0.023)	(0.006)	(0.030)	(0.023)
N	29395	4856	3699	29395	4856	3699
1 or 2 outs						
PS	0.078***	0.019	0.018	0.012	0.000	0.004
	(0.024)	(0.018)	(0.012)	(0.015)	(0.014)	(0.012)
N	27453	10352	14448	27453	10352	14448
Down 1						
PS	-0.063***	0.008	-0.000	0.081***	-0.031	0.008
	(0.009)	(0.025)	(0.016)	(0.008)	(0.025)	(0.019)
N	17073	4514	5824	17073	4514	5824
Tied						
PS	-0.044***	-0.043	-0.036	0.047***	-0.031	0.031
	(0.010)	(0.030)	(0.019)	(0.008)	(0.026)	(0.019)
N	20300	5377	6751	20300	5377	6751
Up 1						
PS	-0.015	-0.002	0.041*	0.053***	-0.030	-0.013
	(0.009)	(0.022)	(0.018)	(0.008)	(0.021)	(0.019)
N	19475	5317	5572	19475	5317	5572

Notes: Each estimate (standard error in parentheses) is coefficient from OLS regression of Win or Runs on PS for subsample. "Other" = all other baserunner situations (two or three men on or one on second or third).

5.2. Main Results

In Table 3 we present results for regressions using our full sample for both outcomes. For each outcome we present the baseline specification, equation (1), and a specification with the additional *FEs* (team-season, opponent-season, and starter) discussed in Section 4.2. We exclude estimates for the Z variables and the lagged bullpen pitch count variables given limited space, and to make some attempt to limit the burden imposed on the reader.

PS has a small insignificant and precisely estimated effect on *Win* in both models. The model with additional *FEs* has slightly smaller standard errors, and yields a 95% confidence interval for the marginal effect of PS_i on win probability when other covariates are at their mean or baseline of $[-0.014, 0.018]$. Both specifications also yield estimates significant at the 5% level, implying that pulling the starter reduces the probability of giving up at least one run in the remainder of the inning by approximately two percentage points. All of these results are consistent with optimal *PS* choices.[26] For the remainder of the analysis we exclude the additional *FEs* to ease interpretation of other covariates, for computational simplicity, and because including them does not substantially affect results (for any of the analyses).

Most X_i variables are insignificant and those that are significant have the expected sign. (Given obvious multiple testing issues we ignore results significant at 10% unless they are at least robust across specifications.) For the Win models, the sum of coefficients for both Starter WHIP and $PS_i \times Starter\ WHIP$, and *Pitch Count* and $PS_i \times Pitch\ Count$, are less than zero, implying that these variables (that increase as the quality of the starter or his condition decline) have negative effects on win probability even when $PS_i = 1$, but this may be because teams are forced to remove starters earlier or use lower quality relievers in these cases.

Table 4 shows the effects of removing various proxy controls. Each specification assesses sensitivity of the estimates (of particular interest is the estimate of β_{PS}) to a different potentially confounding factor approximated by one or more of the included covariate(s). These factors are, in order corresponding to the models presented in the table: pitcher's current physical condition, starter's quality in the current game, starter's overall quality, bullpen status, bullpen quality, and upcoming hitter quality. If results are stable when these proxy controls are removed, this would suggest that results would also not change substantially if unobservable measures of these factors were also accounted for. Note the proxy controls do not include the score-outs-baserunners-inning fixed effects, since these are all observed (and controlled for) without any measurement error. The point estimates of β_{PS} indeed vary minimally, ranging from -0.009 to 0.005 across specifications, suggesting that bias due to omission of better measures of these factors is minimal.

[26] Sample sizes are slightly smaller for the *Win* models because there is a small amount of missing data for Z^{Win} as compared to Z^{Runs}. Note that *Bullpen WHIP* is not collinear with team fixed effects because it excludes the current game's performance, and because it is calculated separately for home and away games (for a given team) in each season.

Table 3: Full sample estimates

	LHS = Win		LHS = Runs	
	(1)	(2)	(3)	(4)
PS	0.004	0.002	-0.019**	-0.022**
	(0.009)	(0.008)	(0.009)	(0.009)
Starter WHIP	-0.095***	-0.122***	0.135***	0.113***
	(0.024)	(0.033)	(0.015)	(0.021)
First 4 IP WHIP	0.011	0.002	-0.010	-0.003
	(0.010)	(0.011)	(0.007)	(0.007)
Pre-Inning Runs	0.003	0.004	0.002	0.003
	(0.003)	(0.003)	(0.003)	(0.003)
Pitch Count	-0.001***	-0.001*	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Fastball Speed	-0.004	-0.005	0.002	0.003
	(0.003)	(0.003)	(0.002)	(0.002)
Next 3 Same Hand	0.001	-0.000	-0.000	-0.002
	(0.004)	(0.003)	(0.003)	(0.003)
Bullpen WHIP	-0.082***	-0.051	-0.002	-0.031
	(0.030)	(0.045)	(0.018)	(0.028)
$PS\times$ Starter WHIP	0.070***	0.055***	-0.090***	-0.062**
	(0.022)	(0.021)	(0.025)	(0.025)
$PS\times$ First 4 IP WHIP	-0.016*	-0.015	0.007	0.004
	(0.009)	(0.009)	(0.010)	(0.010)
$PS\times$ Pre-Inning Runs	0.002	0.001	0.004	0.005
	(0.003)	(0.003)	(0.004)	(0.004)
$PS\times$ Pitch Count	0.000	0.000	0.000	0.001
	(0.000)	(0.000)	(0.000)	(0.001)
$PS\times$ Fastball Speed	0.006*	0.003	-0.006*	-0.005
	(0.004)	(0.004)	(0.003)	(0.003)
$PS\times$ Next 3 Same Hand	-0.007	-0.002	-0.001	0.001
	(0.009)	(0.008)	(0.009)	(0.009)
$PS\times$ Bullpen WHIP	-0.012	-0.010	0.064**	0.059*
	(0.026)	(0.025)	(0.031)	(0.031)
Adj R^2	0.152	0.216	0.125	0.150
N	85749	85721	86868	86840
Z^{Win}	✓	✓		
Z^{Runs}			✓	✓
Other FEs		✓		✓

Notes: At-bat level data for at bats from 6th-8th innings, with score difference of at most one run at start of at bat. PS = 1 if starter is pulled at start of current at bat, = 0 otherwise. Win = 1 if pitching team wins the game, = 0 otherwise. Runs = 1 if pitching team allows one or more runs in the current inning, = 0 otherwise. All models include score-innings-outs-baserunner FEs. Other FEs are team-season, opponent-season, and starter. Z^{Win} and Z^{Runs} are additional controls described in text. Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.

Table 4: OLS estimates with dependent variable = Win; sensitivity to removing proxy controls

Sample	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PS	-0.009	0.004	0.005	0.005	0.004	0.003	0.004
	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Starter WHIP	-0.109***	-0.095***		-0.094***	-0.103***	-0.095***	-0.095***
	(0.023)	(0.024)		(0.024)	(0.024)	(0.024)	(0.024)
First 4 IP WHIP	0.003		0.011	0.009	0.011	0.012	0.010
	(0.010)		(0.011)	(0.010)	(0.011)	(0.010)	(0.010)
Pre-Inning Runs	0.002		0.003	0.004	0.003	0.003	0.002
	(0.003)		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Next 3 Same Hand	0.001	0.002	0.001	0.001	0.002	0.001	0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Bullpen WHIP	-0.074**	-0.080***	-0.096***	-0.081***		-0.083***	-0.083***
	(0.029)	(0.030)	(0.030)	(0.030)		(0.030)	(0.030)
Pitch Count		-0.001***	-0.002***	-0.001***	-0.001***	-0.001***	-0.001***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fastball Speed		-0.004	-0.004	-0.003	-0.004	-0.004	-0.004
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
PS× Starter WHIP	0.073***	0.069***		0.065***	0.069***	0.071***	0.069***
	(0.022)	(0.022)		(0.022)	(0.022)	(0.022)	(0.022)
PS× First 4 IP WHIP	-0.011		-0.015	-0.013	-0.016*	-0.017*	-0.016*
	(0.009)		(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
PS× Pre-Inning Runs	0.001		0.002	0.001	0.002	0.002	0.002
	(0.003)		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
PS× Next 3 Same Hand	-0.007	-0.007	-0.007	-0.006	-0.007	-0.007	-0.007
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
PS× Bullpen WHIP	-0.010	-0.013	-0.003	-0.014		-0.013	-0.012
	(0.026)	(0.026)	(0.026)	(0.026)		(0.026)	(0.026)
PS× Pitch Count		0.000	0.001	0.000	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PS× Fastball Speed		0.006*	0.006*	0.006	0.007*	0.006*	0.006
		(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Opp. OPS	-0.221**	-0.214**	-0.228**	-0.233**	-0.234**	-0.273***	
	(0.098)	(0.099)	(0.099)	(0.098)	(0.099)	(0.097)	
Next 3 OPS	-0.071**	-0.077***	-0.072**	-0.068**	-0.072***		-0.105***
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)		(0.029)
Adj R^2	0.151	0.152	0.151	0.151	0.151	0.152	0.152
N	87897	85749	85749	86846	85760	85749	85749

Notes: All models include full set of score-outs-baserunners-inning FEs and Z^{Win}. See article text for explanation of the different specifications (note model (4) removes bullpen status (lagged pitch count) controls). Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.

In Table 5 we estimate the Win models separately by subsamples defined by key values of the *FE* variables (score, inning, outs, men-on). Most *PS* estimates are insignificant. Two are significant at the 5% level (for subsamples with at least one man on base, and with at least one out) but the point estimates are still somewhat small: each is less than four percentage points. *Starter WHIP* and *Pitch Count* are again the most important covariates. It is difficult to know how to interpret the other significant estimates due to their inconsistency. Using a stronger standard of 0.5% for significance as is perhaps now becoming more common (given multiple testing issues and another concerns about type I errors), we would simply ignore all of these. In unreported results, we examine more narrowly defined subsamples motivated by these results and find limited additional significant evidence. We also present results for the additional subsamples discussed in Section 4.2 (defined by league and seasons prior to 2016) and find results are mostly consistent with those of the full sample.

In Table 6, we replicate the models from Table 5 replacing Win with the Runs dependent variable. The results are strongest for the same subsamples. There is one other estimate of interest with a p-value of approximately 0.005, the coefficient on $PS_i \times Next\ 3\ Same\ Hand$, which is estimated to be -0.041. This means that pulling the starter reduces run probability more when the upcoming batters have the same hand as the starter, which is supposed to be to the starter's advantage, indicating that managers overestimate this advantage (they should pull starters more often even when they have this advantage).

In Table 7 we examine whether the *PS* decision is unduly influenced by recent events in the game (results from the last three batters and the opponent's value of *PS* each interacted with *PS*). We present results for models with the various interactions included separately, and one model that includes all *Last 3* interactions together, to account for possible correlations between these variables. Results are almost entirely insignificant, and point estimates and standard errors are again small. It is especially interesting to note that the *Opp PS* interaction is insignificant, since we know that this variable does influence the *PS* decision. Apparently, this influence does not affect win probability. Table 8, which presents analogous models for *Runs*, provides some evidence that recent hitting performance is overreacted to. Specifically, teams that pull starters when the opponent has drawn more walks or hits are significantly (5%) more likely to give up a run in the current inning in the model with all interactions, but this is offset by a negative coefficient for the *Last 3 Lucky Runs* variable. The interpretation that these results suggest is that managers actually overreact more to non-lucky hits (as evidence of decline in the starter's ability). But again, the significance of these results is somewhat marginal and magnitudes are small.

In the final table, Table 9, we examine Manager of the Year (*MoY*) votes as an alternative measure of managerial quality. *MoY* has no significant effects on *PS*. We also present models with a dependent variable of Win (and include the interactions of X_i and PS_i) and show that the point estimates for *MoY* are consistently positive and significant, supporting the validity of *MoY* as a measure of decision-making quality. These results also demonstrate the ability of an additional regressor to predict changes in win probability despite the large set of additional controls, the analog to a standard placebo test for an analysis with null results for the main effect.

Table 5: OLS estimates with dependent variable = Win for subsamples

Sample	None on	1+ on	6th Inn.	7th or 8th	No outs	$\geq$ 1 out	Down 1	Tied	Up 1
PS	-0.002	0.024	0.024	-0.003	-0.010	0.036**	0.011	-0.001	0.008
	(0.011)	(0.016)	(0.018)	(0.010)	(0.010)	(0.018)	(0.014)	(0.017)	(0.015)
Starter WHIP	-0.107***	-0.073***	-0.090***	-0.102***	-0.107***	-0.088***	-0.048	-0.073*	-0.159***
	(0.025)	(0.028)	(0.024)	(0.034)	(0.025)	(0.025)	(0.036)	(0.040)	(0.037)
First 4 IP WHIP	0.014	0.008	0.010	0.013	0.009	0.012	0.024	-0.000	0.011
	(0.011)	(0.013)	(0.011)	(0.015)	(0.011)	(0.011)	(0.016)	(0.017)	(0.016)
Pre-Inning Runs	0.003	0.003	0.001	0.006	0.004	0.002	-0.002	0.011**	-0.001
	(0.003)	(0.004)	(0.003)	(0.005)	(0.003)	(0.003)	(0.006)	(0.005)	(0.006)
Pitch Count	-0.002***	-0.000	-0.001***	-0.001**	-0.001***	-0.001***	-0.002***	-0.001**	-0.000
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
Fastball Speed	-0.004	-0.004	-0.005	-0.002	-0.004	-0.004	0.005	-0.012**	-0.002
	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.005)	(0.006)	(0.005)
Next 3 Same Hand	0.001	0.003	0.002	0.002	0.004	0.000	0.003	0.003	-0.001
	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)	(0.006)	(0.006)	(0.006)
Bullpen WHIP	-0.088***	-0.076**	-0.085***	-0.079*	-0.078**	-0.085***	-0.122***	-0.130***	-0.002
	(0.031)	(0.037)	(0.029)	(0.040)	(0.031)	(0.031)	(0.046)	(0.049)	(0.046)
PS$\times$ Starter WHIP	0.071**	0.073*	0.078*	0.066**	0.079***	0.072	0.082**	0.035	0.088**
	(0.030)	(0.043)	(0.045)	(0.033)	(0.029)	(0.044)	(0.040)	(0.045)	(0.040)
PS$\times$ First 4 IP WHIP	-0.016	-0.015	-0.023	-0.015	-0.012	-0.014	-0.022	0.008	-0.036**
	(0.012)	(0.019)	(0.018)	(0.014)	(0.012)	(0.018)	(0.018)	(0.018)	(0.017)
PS$\times$ Pre-Inning Runs	0.005	-0.005	-0.000	0.001	0.003	-0.002	0.015**	-0.013**	0.003
	(0.004)	(0.007)	(0.006)	(0.005)	(0.004)	(0.007)	(0.007)	(0.006)	(0.007)
PS$\times$ Pitch Count	0.001*	-0.001	0.001	0.000	0.000	-0.000	0.002**	-0.001	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PS$\times$ Fastball Speed	0.006	0.008	0.002	0.007	0.006	0.007	-0.007	0.016**	0.008
	(0.004)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)	(0.007)	(0.006)
PS$\times$ Next 3 Same Hand	-0.014	0.010	0.018	-0.016	-0.015	0.015	0.006	-0.011	-0.012
	(0.011)	(0.017)	(0.018)	(0.011)	(0.011)	(0.017)	(0.014)	(0.017)	(0.015)
PS$\times$ Bullpen WHIP	-0.017	-0.004	0.002	-0.013	-0.015	-0.005	0.045	0.012	-0.092*
	(0.035)	(0.053)	(0.061)	(0.038)	(0.035)	(0.053)	(0.048)	(0.055)	(0.049)
Adj R^2	0.154	0.142	0.126	0.188	0.145	0.154	0.041	0.038	0.041
N	53990	31759	50202	35547	36104	49645	25943	30920	28886

Notes: All models include full set of score-outs-baserunners-inning FEs and Z^{Win}. Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.

Table 6: OLS estimates with dependent variable = Runs

	None on	1+ on	6th Inn.	7th or 8th	No outs	$\geq$ 1 out	Down 1	Tied	Up 1
PS	-0.012	-0.036**	-0.021	-0.015	-0.010	-0.045***	-0.015	-0.027*	-0.012
	(0.011)	(0.016)	(0.017)	(0.010)	(0.011)	(0.016)	(0.015)	(0.016)	(0.015)
Starter WHIP	0.131***	0.138***	0.140***	0.121***	0.169***	0.112***	0.148***	0.131***	0.125***
	(0.014)	(0.022)	(0.019)	(0.024)	(0.019)	(0.015)	(0.027)	(0.025)	(0.025)
First 4 IP WHIP	-0.011*	-0.008	-0.016*	0.002	-0.012	-0.009	-0.008	-0.010	-0.009
	(0.006)	(0.010)	(0.008)	(0.011)	(0.009)	(0.007)	(0.012)	(0.011)	(0.011)
Pre-Inning Runs	0.003	0.001	0.001	0.004	0.002	0.002	0.000	0.003	0.002
	(0.002)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.005)	(0.004)	(0.004)
Pitch Count	0.000	-0.000	0.000	-0.000	-0.000	-0.000	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fastball Speed	0.003	0.002	0.004	0.001	0.002	0.003	0.000	0.002	0.005
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.004)	(0.004)	(0.004)
Next 3 Same Hand	-0.000	0.001	-0.002	0.004	-0.004	0.003	0.006	-0.008*	0.003
	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.005)	(0.004)	(0.005)
Bullpen WHIP	-0.000	-0.002	0.023	-0.041	0.007	-0.006	0.053	-0.064**	0.015
	(0.017)	(0.027)	(0.023)	(0.029)	(0.023)	(0.018)	(0.033)	(0.031)	(0.031)
PS$\times$ Starter WHIP	-0.062**	-0.145***	-0.153***	-0.040	-0.112***	-0.093**	-0.091**	-0.101**	-0.070*
	(0.030)	(0.043)	(0.040)	(0.033)	(0.031)	(0.042)	(0.044)	(0.043)	(0.042)
PS$\times$ First 4 IP WHIP	0.014	-0.012	-0.009	0.008	0.019	-0.018	0.005	-0.003	0.014
	(0.012)	(0.018)	(0.017)	(0.014)	(0.014)	(0.017)	(0.019)	(0.018)	(0.017)
PS$\times$ Pre-Inning Runs	-0.000	0.014**	0.010	-0.000	-0.000	0.015**	0.004	0.004	0.005
	(0.005)	(0.007)	(0.006)	(0.005)	(0.005)	(0.006)	(0.007)	(0.007)	(0.007)
PS$\times$ Pitch Count	0.000	0.000	-0.000	0.000	0.000	0.000	-0.001	0.002**	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PS$\times$ Fastball Speed	-0.005	-0.007	0.001	-0.007*	-0.004	-0.009**	0.001	-0.005	-0.011**
	(0.003)	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.006)	(0.005)	(0.005)
PS$\times$ Next 3 Same Hand	0.010	-0.025	-0.017	0.003	0.017	-0.037**	-0.017	0.003	0.010
	(0.010)	(0.016)	(0.016)	(0.011)	(0.011)	(0.016)	(0.015)	(0.015)	(0.014)
PS$\times$ Bullpen WHIP	0.084**	0.026	0.061	0.092**	0.055	0.069	0.002	0.117**	0.056
	(0.037)	(0.052)	(0.056)	(0.040)	(0.039)	(0.051)	(0.054)	(0.055)	(0.051)
Adj R^2	0.053	0.159	0.137	0.107	0.087	0.124	0.132	0.129	0.120
N	54694	32174	50883	35985	36577	50291	26346	31272	29250

Notes: All models include full set of score-outs-baserunners-inning FEs and Z^{Runs}. Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.

6. Concluding Remarks

We show that, under reasonable assumptions, current inning pitching performance immediately improves when starters are pulled at the optimal time with respect to maximizing the probability of winning the current game. We provide empirical evidence consistent with this prediction. We also show that, empirically, win probability does not change substantially when starters are pulled in general, and in a range of situations. Finally, we show that managers more successful in manager of the year voting do not tend to pull starters particularly early or late.

We interpret these results to imply that managers make decisions to pull starters approximately optimally in between at bats. Richard Thaler's take ("Hindsight bias illustrated") regarding the claim made in the tweet quoted at the start of our paper (that managers typically remove starters too early) turns out to indeed be consistent with our results. However, a caveat is that the randomization of our treatment (pulling starters) is unclear, and to the extent that these decisions have been approximately optimal in recent years, this is likely due to a period of learning over many years. Moreover, our analysis also suggests that decisions may be improved by making more within-at bat pitching changes. In relation to past work on manager decision making, we also show that doing so is consistent with allowing run expectations to be higher in the current inning, making clear the importance of properly defining a decisionmaker's objective function. Finally, we note that a dynamic structural analysis of pitching changes may uncover subtleties that our reduced form analysis misses.

Table 7: Interactions of PS with recent outcomes, dependent variable = Win

	(1)	(2)	(3)	(4)	(5)	(6)	
PS	0.006	-0.002	0.005	0.002	0.003	0.003	
	(0.010)	(0.011)	(0.009)	(0.009)	(0.020)	(0.023)	
Opp. *PS*	0.012						
	(0.010)						
PS× Opp. *PS*	-0.010						
	(0.011)						
Last 3 WH		0.002				0.007	
		(0.004)				(0.008)	
PS× Last 3 WH		0.005				0.003	
		(0.006)				(0.012)	
Last 3 Runs			-0.002			-0.002	
			(0.005)			(0.006)	
PS× Last 3 Runs			-0.002			-0.006	
			(0.008)			(0.009)	
Last 3 Lucky Runs				0.000		-0.004	
				(0.004)		(0.008)	
PS× Last 3 Lucky Runs				0.004		0.004	
				(0.006)		(0.013)	
Last 3 Pitch Ct					0.002**	0.002**	
					(0.001)	(0.001)	
PS× Last 3 Pitch Ct					-0.000	-0.000	
					(0.001)	(0.001)	
Adj R^2		0.152	0.152	0.152	0.152	0.152	0.152
N		85749	85749	85749	85749	85749	85749

Notes: All models include full set of score-outs-baserunners-inning FEs and Z^{Win}. Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.

Table 8: Interactions of PS with recent outcomes, dependent variable = Runs

	(1)	(2)	(3)	(4)	(5)	(6)	
PS	-0.020**	-0.026**	-0.020**	-0.018**	0.007	-0.020	
	(0.010)	(0.011)	(0.009)	(0.009)	(0.019)	(0.023)	
Opp. *PS*	-0.008						
	(0.007)						
PS× Opp. *PS*	0.009						
	(0.010)						
Last 3 WH		-0.005				-0.010	
		(0.003)				(0.006)	
PS× Last 3 WH		0.008				0.025**	
		(0.006)				(0.011)	
Last 3 Runs			-0.002			-0.001	
			(0.004)			(0.005)	
PS× Last 3 Runs			0.008			0.010	
			(0.007)			(0.009)	
Last 3 Lucky Runs				-0.002		0.006	
				(0.003)		(0.006)	
PS× Last 3 Lucky Runs				0.002		-0.024*	
				(0.006)		(0.012)	
Last 3 Pitch Ct					0.000	-0.000	
					(0.001)	(0.001)	
PS× Last 3 Pitch Ct					-0.002	-0.002	
					(0.001)	(0.001)	
Adj R^2		0.125	0.125	0.125	0.125	0.125	0.125
N		86868	86868	86868	86868	86868	86868

Notes: All models include full set of score-outs-baserunners-inning FEs and Z^{Runs}. Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.


14TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 6 - 7, 2020 | BOSTON, MA
presented by ESPN

Table 9: Manager of Year (MoY) estimates

Sample	All	None on	1+ on	6th Inn.	7th or 8th	0 Outs	1-2 Outs	Down 1	Tied	Up 1
LHS = PS										
MoY	0.009	-0.004	0.031	0.005	0.019	0.003	0.020	0.030	-0.005	0.001
	(0.013)	(0.016)	(0.021)	(0.014)	(0.025)	(0.022)	(0.013)	(0.026)	(0.021)	(0.023)
Adj R^2	0.362	0.422	0.234	0.199	0.412	0.432	0.198	0.375	0.367	0.348
N	85749	53990	31759	50202	35547	36104	49645	25943	30920	28886
LHS = Win										
PS	0.003	0.004	0.022	0.023	-0.003	-0.010	0.034**	0.008	-0.003	0.005
	(0.009)	(0.008)	(0.016)	(0.018)	(0.010)	(0.010)	(0.017)	(0.015)	(0.017)	(0.015)
MoY	0.138**	0.177***	0.133**	0.164***	0.102	0.147**	0.131**	0.127	0.071	0.230***
	(0.057)	(0.058)	(0.068)	(0.057)	(0.071)	(0.057)	(0.060)	(0.085)	(0.094)	(0.082)
Adj R^2	0.152	0.150	0.143	0.127	0.189	0.146	0.154	0.042	0.038	0.042
N	85749	54694	31759	50202	35547	36104	49645	25943	30920	28886

Notes: All models include full set of score-outs-baserunners-inning interactions, X, and Z^{Win}. Models with $Y = Win$ include $X \times PS$ as regressors. Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.

References

[1] Abadie, A. (2018): "Statistical non-significance in empirical economics," Discussion paper, National Bureau of Economic Research.

[2] Albert, J. (2010): "Using the count to measure pitching performance," Journal of Quantitative Analysis in Sports, 6(4).

[3] Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," Journal of Political Economy, 113(1), 151–184.

[4] Bar-Eli, M., H. H. Azar, I. Ritov, Y. Keidar-Levin, and G. Schein (2007): "Action bias among elite soccer goalkeepers: The case of penalty kicks," Journal of Economic Psychology, 28(5), 606–621.

[5] Benjamin, D. J. (2018): "Errors in Probabilistic Reasoning and Judgment Biases," Discussion paper, National Bureau of Economic Research.

[6] Brownback, A., and M. A. Kuhn (2019): "Understanding outcome bias," Games and Economic Behavior.

[7] Carleton, R. A. (2017): "Baseball Therapy: Slouching Toward Bullpenning," Baseball Prospectus.

[8] Carleton, R. A. (2018a): "Baseball Therapy: How Much Bullpenning Could a Team Do?," Baseball Prospectus.

[9] Carleton, R. A. (2018b): "Baseball Therapy: The Surprising Evolution of the Bullpen," Baseball Prospectus.

[10] Fudenberg, D., and D. K. Levine (2016): "Whither game theory? Towards a theory of learning in games," Journal of Economic Perspectives, 30(4), 151–70.

[11] Gabaix, X. (2017): "Behavioral inattention," Discussion paper, National Bureau of Economic Research.

[12] Ganeshapillai, G., and J. Guttag (2014): "A data-driven method for in-game decision making in MLB," MIT SSAC.

[13] Goldman, M., and J. M. Rao (2017): "Optimal stopping in the NBA: Sequential search and the shot clock," Journal of Economic Behavior & Organization, 136, 107–124.

[14] Green, E., and D. Daniels (2018): "Bayesian Instinct," Available at SSRN 2916929.

[15] Hakes, J. K., and R. D. Sauer (2006): "An economic evaluation of the Moneyball hypothesis," Journal of Economic Perspectives, 20(3), 173–186.

[16] Houston, R. (2018): "Call to the Bullpen: More Often and More Effective," https://www.samford.edu/sports-analytics/fans/2018/Call-to-the-Bullpen.

[17] Lee, Y. H. (2014): "Stochastic Frontier Models in Sports Economics.," International Journal of Sport Finance, 9(4).

[18] Marchi, M., and J. Albert (2013): Analyzing baseball data with R. CRC Press.

[19] Mills, B. M. (2014): "Social pressure at the plate: Inequality aversion, status, and mere exposure," Managerial and Decision Economics, 35(6), 387–403.

[20] Mills, B. M. (2017a): "Policy Changes in Major League Baseball: Improved Agent Behavior and Ancillary Productivity Outcomes," Economic Inquiry, 55(2), 1104–1118.

[21] Mills, B. M. (2017b): "Technological innovations in monitoring and evaluation: Evidence of performance impacts among Major League Baseball umpires," Labour Economics, 46, 189–199.

[22] Mills, B. M., and S. Salaga (2018): "A natural experiment for efficient markets: Information quality and influential agents," Journal of Financial Markets, 40, 23–39.

[23] Provencher, B. (1997): "Structural versus reduced-form estimation of optimal stopping problems," American Journal of Agricultural Economics, 79(2), 357–368.

[24] Ritov, I., and J. Baron (1990): "Reluctance to vaccinate: Omission bias and ambiguity," Journal of Behavioral Decision Making, 3(4), 263–277.

[25] Romer, D. (2006): "Do firms maximize? Evidence from professional football," Journal of Political Economy, 114(2), 340–365.

[26] Stone, D. F., and J. Arkes (2018): "March Madness? Underreaction to hot and cold hands in NCAA basketball," Economic Inquiry, 56(3), 1724–1747.

[27] Weinberg, N. (2015): "The Beginner's Guide to Pulling A Starting Pitcher," https://library.fangraphs.com/the-beginners-guide-to-pulling-a-starting-pitcher/.

[28] Wood, S. N. (2011): "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(1), 3–36.

[29] Wood, S. N. (2017): Generalized additive models: an introduction with R. Chapman and Hall/CRC.

[30] Zhou, X., Y. Liu, and B. Ho (2015): "The cultural transmission of cooperative norms," Frontiers in psychology, 6, 1554.

Appendix

A.1 Construction of Lucky Runs Variables

Expected runs created for a given pitch, RC_i, are derived from Marchi and Albert (2013) using 2008 play-by-play files from retrosheet.org and called pitch values from Albert (2010). This method calculates the change in expected runs score for the remainder of the inning from before a given play (pitch) to after the result from this outcome. From this calculation, the expected run contribution is equal to the ending run expectancy state – determined by men on base, the number of outs, and/or the count – minus the starting run expectancy state, plus the number of runs actually scored as a result of the play. For example, with a man on first base and two outs, the average number of runs scored in the rest of the inning across 2008 MLB games was about 0.214, or approximately one run scored every five times this situation occurs. However, the expected runs scored will change depending on the batter's output at the plate in the subsequent pitch or at bat. If the batter causes the third out, the expected remaining runs scored will move to zero, but if they hit a home run, 2 runs have scored, plus the inning continues with an additional expectation equal to the average runs scored in bases empty, two out situations (about 0.095 runs). The contribution for the batter causing the third out would then be −0.214 (RC = 0 − 0.214), while the contribution when hitting a home run would be 1.881 (RC = 2 + 0.095 − 0.214). Albert (2010) performs a similar exposition at the ball-strike count and pitch level, and we use changes to these values as given by Mills (2014) for pitches not put in play.

We identify "lucky" outcomes for hitters using a spatial model of the strike zone, estimating the expected changes to expected additional runs created dependent on pitch location, pitch type, batter stance, pitch velocity, and the current ball-strike count. The spatial model is estimated as a semiparametric generalized additive model (GAM) with restricted maximum likelihood and two-dimensional penalized regression splines for the horizontal-vertical location of the pitch as it crosses home plate (Wood, 2011; Wood, 2017). These models have been used in past work to evaluate spatial baseball data and deviations from expected outcomes across the strike zone plane (Mills and Salaga, 2018). We interact the regression splines for coordinate location with the season year, batter stance, and ball-strike count, due to changes to the called strike zone and hitter productivity across time and the strike zone space (Mills, 2017a; Mills, 2017b). The model is estimated using all pitches in the data, but separately for umpire-called pitches and pitches at which the batter swings as:

$$RC_i = f_h(X_i, Y_i) + f_c(X_i, Y_i) + f_t(X_i, Y_i) + \sum_{k=1}^{7} \gamma_k Z_i + \beta_1 Velo_i + \varepsilon_i.$$

RC_i represents the expected runs resulting from the pitch, and $f_h()$, $f_c()$, and $f_t()$ represent the pitch-location-interacted smooth parameters for the batter's handedness, ball- strike count, and season, respectively. $\gamma_k Z_i$ identifies the various pitch types and their respective effects on expected changes to runs created, while $Velo_i$ is the velocity of the pitch in miles per hour. We note that for computational simplicity, this specification assumes run expectations shift linearly with pitch type and pitch velocity once controlling for other factors in the spatial estimation, and that additional pitch type, handedness, count, season, velocity, and locational interactions provide only marginal gains in model performance. ε_i is the error term for each pitch. We take the difference in the actual change in RC_i and the estimated RC_i for each pitch to calculate the *Last T Lucky Runs* variable. Positive values, in which the actual runs created are larger than expected, are considered luckier outcomes than negative values, where actual runs created are lower than expected. These values are then aggregated

over the last T at bats for the starting pitcher, and is included in the central regression models to detect and control for managerial ability to identify luckiness in recent pitcher outcomes.

A.2 Supplemental Tables

Table A1: OLS estimates with dependent variable = Win for 2008-2015 seasons only

	None on	1+ on	6th Inn.	7th or 8th	No outs	$\geq$ 1 out	Down 1	Tied	Up 1
PS	-0.000	0.018	0.034	-0.004	-0.009	0.034*	0.010	-0.017	0.021
	(0.012)	(0.018)	(0.021)	(0.011)	(0.011)	(0.020)	(0.016)	(0.019)	(0.017)
Starter WHIP	-0.091***	-0.069**	-0.079***	-0.089**	-0.094***	-0.078***	-0.016	-0.061	-0.160***
	(0.028)	(0.031)	(0.026)	(0.038)	(0.028)	(0.028)	(0.039)	(0.045)	(0.040)
First 4 IP WHIP	0.014	0.007	0.008	0.015	0.012	0.011	0.017	0.005	0.010
	(0.012)	(0.014)	(0.012)	(0.016)	(0.012)	(0.012)	(0.018)	(0.019)	(0.018)
Pre-Inning Runs	0.004	0.004	0.003	0.004	0.005	0.003	-0.001	0.010*	0.001
	(0.004)	(0.005)	(0.003)	(0.005)	(0.004)	(0.004)	(0.007)	(0.005)	(0.007)
Pitch Count	-0.002***	-0.000	-0.001**	-0.002**	-0.001**	-0.001***	-0.001*	-0.002**	-0.001
	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
Fastball Speed	-0.007	-0.008*	-0.008*	-0.006	-0.008*	-0.007*	0.005	-0.016**	-0.008
	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)	(0.006)	(0.006)	(0.006)
Next 3 Same Hand	-0.002	0.003	0.001	-0.001	0.003	-0.002	0.002	-0.003	0.001
	(0.004)	(0.005)	(0.005)	(0.006)	(0.005)	(0.004)	(0.006)	(0.007)	(0.007)
Bullpen WHIP	-0.081**	-0.080**	-0.081**	-0.077*	-0.083**	-0.078**	-0.129**	-0.140**	0.017
	(0.034)	(0.041)	(0.033)	(0.045)	(0.035)	(0.034)	(0.051)	(0.054)	(0.051)
PS× Starter WHIP	0.051	0.076	0.055	0.058	0.067**	0.061	0.050	0.007	0.112**
	(0.033)	(0.048)	(0.050)	(0.036)	(0.033)	(0.049)	(0.044)	(0.050)	(0.045)
PS× First 4 IP WHIP	-0.016	-0.007	-0.030	-0.012	-0.012	-0.009	-0.022	0.015	-0.034*
	(0.013)	(0.021)	(0.021)	(0.015)	(0.013)	(0.021)	(0.019)	(0.020)	(0.019)
PS× Pre-Inning Runs	0.006	-0.013*	-0.006	0.003	0.003	-0.007	0.014*	-0.014**	0.002
	(0.005)	(0.008)	(0.007)	(0.006)	(0.005)	(0.007)	(0.008)	(0.007)	(0.007)
PS× Pitch Count	0.001**	-0.001	0.001	0.001	0.001	-0.001	0.002**	-0.001	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PS× Fastball Speed	0.010**	0.010	0.004	0.011**	0.011**	0.009	-0.006	0.019***	0.014**
	(0.005)	(0.006)	(0.007)	(0.005)	(0.005)	(0.006)	(0.007)	(0.007)	(0.007)
PS× Next 3 Same Hand	-0.012	-0.007	0.011	-0.017	-0.017	0.003	0.013	-0.032*	-0.012
	(0.012)	(0.019)	(0.020)	(0.012)	(0.012)	(0.019)	(0.016)	(0.019)	(0.017)
PS× Bullpen WHIP	-0.039	0.025	0.026	-0.032	-0.024	0.007	0.062	-0.009	-0.107*
	(0.038)	(0.060)	(0.071)	(0.042)	(0.039)	(0.061)	(0.054)	(0.061)	(0.055)
Adj R^2	0.152	0.142	0.125	0.185	0.144	0.152	0.041	0.039	0.040
N	44822	26701	41231	30292	29961	41562	21605	25743	24175

Notes: All models include full set of score-outs-baserunners-inning FEs and Z^{Win}. Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.

Table A2: OLS estimates with dependent variable = Win for American League only

Sample	None on	1+ on	6th Inn.	7th or 8th	No outs	$\geq$ 1 out	Down 1	Tied	Up 1
PS	0.014	0.018	0.011	0.017	-0.002	0.044*	0.031	-0.014	0.028
	(0.017)	(0.022)	(0.027)	(0.016)	(0.016)	(0.024)	(0.023)	(0.025)	(0.022)
Starter WHIP	-0.118***	-0.074*	-0.090***	-0.126***	-0.123***	-0.091**	-0.086	-0.059	-0.145***
	(0.037)	(0.042)	(0.034)	(0.048)	(0.036)	(0.037)	(0.053)	(0.058)	(0.054)
First 4 IP WHIP	0.028*	0.007	0.012	0.033	0.013	0.024	0.025	-0.007	0.044*
	(0.016)	(0.018)	(0.015)	(0.020)	(0.016)	(0.016)	(0.023)	(0.024)	(0.024)
Pre-Inning Runs	0.002	0.004	0.001	0.006	0.005	0.002	-0.004	0.014**	-0.002
	(0.004)	(0.006)	(0.004)	(0.007)	(0.005)	(0.004)	(0.009)	(0.007)	(0.009)
Pitch Count	-0.002***	-0.001	-0.001**	-0.002*	-0.001*	-0.002***	-0.002*	-0.002**	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Fastball Speed	-0.004	-0.003	-0.002	-0.005	-0.004	-0.003	0.010	-0.019**	0.004
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.008)	(0.008)	(0.008)
Next 3 Same Hand	0.002	0.002	0.001	0.002	0.003	0.001	0.004	0.002	-0.002
	(0.006)	(0.008)	(0.006)	(0.008)	(0.007)	(0.006)	(0.009)	(0.009)	(0.009)
Bullpen WHIP	-0.076*	-0.086	-0.076*	-0.084	-0.056	-0.091**	-0.159**	-0.112	-0.001
	(0.045)	(0.054)	(0.042)	(0.058)	(0.045)	(0.045)	(0.068)	(0.070)	(0.068)
PS$\times$ Starter WHIP	0.108**	0.111*	0.139**	0.100**	0.117***	0.117*	0.041	0.117*	0.132**
	(0.044)	(0.059)	(0.064)	(0.047)	(0.044)	(0.060)	(0.062)	(0.064)	(0.057)
PS$\times$ First 4 IP WHIP	-0.020	-0.006	-0.031	-0.018	-0.002	-0.024	0.009	0.034	-0.087***
	(0.018)	(0.025)	(0.026)	(0.020)	(0.018)	(0.025)	(0.025)	(0.027)	(0.025)
PS$\times$ Pre-Inning Runs	0.008	-0.008	0.008	-0.003	0.005	-0.001	0.012	-0.017*	0.013
	(0.006)	(0.009)	(0.009)	(0.007)	(0.007)	(0.009)	(0.010)	(0.009)	(0.009)
PS$\times$ Pitch Count	0.001	-0.000	0.001	0.000	0.000	-0.000	0.002*	-0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PS$\times$ Fastball Speed	0.007	0.004	-0.006	0.010	0.007	0.004	-0.008	0.018**	0.002
	(0.006)	(0.008)	(0.008)	(0.007)	(0.006)	(0.007)	(0.009)	(0.009)	(0.009)
PS$\times$ Next 3 Same Hand	0.010	0.003	0.019	0.002	0.006	0.016	0.021	-0.009	0.008
	(0.017)	(0.023)	(0.026)	(0.016)	(0.016)	(0.024)	(0.022)	(0.025)	(0.022)
PS$\times$ Bullpen WHIP	-0.072	-0.002	-0.064	-0.024	-0.082	0.007	-0.019	0.034	-0.117
	(0.053)	(0.073)	(0.087)	(0.054)	(0.053)	(0.074)	(0.071)	(0.081)	(0.072)
Adj R^2	0.130	0.121	0.103	0.168	0.122	0.132	0.045	0.043	0.041
N	26233	15366	23939	17660	17375	24224	12765	14928	13906

Notes: All models include full set of score-outs-baserunners-inning FEs and Z^{Win}. Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.

Table A3: OLS estimates with dependent variable = Win for National League only

	None on	1+ on	6th Inn.	7th or 8th	No outs	$\geq$ 1 out	Down 1	Tied	Up 1
PS	-0.011	0.029	0.036	-0.019	-0.013	0.029	-0.004	0.011	-0.011
	(0.014)	(0.024)	(0.025)	(0.014)	(0.014)	(0.026)	(0.019)	(0.024)	(0.021)
Starter WHIP	-0.095***	-0.073*	-0.090***	-0.082*	-0.089***	-0.086**	-0.024	-0.091	-0.152***
	(0.035)	(0.039)	(0.033)	(0.049)	(0.034)	(0.035)	(0.049)	(0.057)	(0.050)
First 4 IP WHIP	-0.002	0.006	0.007	-0.011	0.003	-0.000	0.024	0.007	-0.023
	(0.015)	(0.018)	(0.015)	(0.021)	(0.016)	(0.015)	(0.023)	(0.025)	(0.023)
Pre-Inning Runs	0.004	0.002	0.002	0.004	0.004	0.002	-0.002	0.008	0.002
	(0.005)	(0.006)	(0.004)	(0.007)	(0.005)	(0.005)	(0.009)	(0.007)	(0.009)
Pitch Count	-0.001**	0.000	-0.001	-0.001	-0.001*	-0.001	-0.002*	-0.001	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Fastball Speed	-0.003	-0.006	-0.007	0.001	-0.004	-0.004	-0.002	-0.002	-0.008
	(0.005)	(0.006)	(0.005)	(0.007)	(0.006)	(0.005)	(0.008)	(0.008)	(0.007)
Next 3 Same Hand	-0.000	0.003	0.001	-0.001	0.005	-0.001	0.002	0.002	-0.003
	(0.005)	(0.007)	(0.005)	(0.007)	(0.006)	(0.005)	(0.008)	(0.008)	(0.008)
Bullpen WHIP	-0.101**	-0.059	-0.091**	-0.073	-0.101**	-0.074*	-0.088	-0.150**	-0.003
	(0.043)	(0.051)	(0.041)	(0.056)	(0.043)	(0.043)	(0.064)	(0.070)	(0.063)
PS$\times$ Starter WHIP	0.041	0.033	0.022	0.035	0.046	0.021	0.112**	-0.052	0.045
	(0.041)	(0.065)	(0.064)	(0.048)	(0.040)	(0.066)	(0.053)	(0.064)	(0.059)
PS$\times$ First 4 IP WHIP	-0.010	-0.020	-0.017	-0.006	-0.016	-0.008	-0.054**	-0.015	0.012
	(0.017)	(0.029)	(0.026)	(0.020)	(0.017)	(0.028)	(0.025)	(0.026)	(0.024)
PS$\times$ Pre-Inning Runs	0.002	-0.005	-0.006	0.004	0.002	-0.004	0.019*	-0.010	-0.008
	(0.006)	(0.011)	(0.009)	(0.008)	(0.006)	(0.011)	(0.010)	(0.009)	(0.010)
PS$\times$ Pitch Count	0.001	-0.002	0.001	0.000	0.000	-0.001	0.001	-0.001	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
PS$\times$ Fastball Speed	0.005	0.012	0.009	0.002	0.005	0.011	-0.002	0.009	0.014
	(0.006)	(0.009)	(0.009)	(0.007)	(0.006)	(0.008)	(0.008)	(0.010)	(0.008)
PS$\times$ Next 3 Same Hand	-0.030**	0.018	0.018	-0.028**	-0.032**	0.016	-0.008	-0.018	-0.029
	(0.014)	(0.024)	(0.024)	(0.014)	(0.014)	(0.025)	(0.018)	(0.023)	(0.021)
PS$\times$ Bullpen WHIP	0.027	-0.007	0.058	-0.010	0.037	-0.017	0.092	-0.011	-0.067
	(0.045)	(0.078)	(0.085)	(0.053)	(0.046)	(0.078)	(0.067)	(0.075)	(0.066)
Adj R^2	0.177	0.164	0.149	0.211	0.168	0.174	0.040	0.040	0.046
N	27757	16393	26263	17887	18729	25421	13178	15992	14980

Notes: All models include full set of score-outs-baserunners-inning FEs and Z^{Win}. Standard errors clustered by game. *, **, *** denotes 10%, 5%, 1% significance, respectively.