DE GRUYTER

J. Quant. Anal. Sports 2023; 19(4): 245–262

Research Article

Ryan S. Brill*, Sameer K. Deshpande and Abraham J. Wyner

A Bayesian analysis of the time through the order penalty in baseball

Abstract: As a baseball game progresses, batters appear to perform better the more times they face a particular pitcher. The apparent drop-off in pitcher performance from one time through the order to the next, known as the Time Through the Order Penalty (TTOP), is often attributed to within-game batter learning. Although the TTOP has largely been accepted within baseball and influences many managers' in-game decision making, we argue that existing approaches of estimating the size of the TTOP cannot disentangle continuous evolution in pitcher performance over the course of the game from discontinuities between successive times through the order. Using a Bayesian multinomial regression model, we find that, after adjusting for confounders like batter and pitcher quality, handedness, and home field advantage, there is little evidence of strong discontinuity in pitcher performance between times through the order. Our analysis suggests that the start of the third time through the order should not be viewed as a special cutoff point in deciding whether to pull a starting pitcher.

Keywords: baseball; Bayesian statistics; mathematical modeling; pitching; time through the order penalty

1 Introduction

In Game 6 of the 2020 World Series, the Tampa Bay Rays' manager, Kevin Cash, pulled his starting pitcher, Blake Snell, midway through the sixth inning. When he was pulled, Snell had been pitching extremely well; he had allowed just two hits and struck out nine batters on 73 pitches. Moreover, the Rays had a one run lead. Snell's replacement, Nick Anderson, promptly gave up two runs, which ultimately proved decisive: the Rays went on to lose the game and the World Series. After the game, Cash justified his decision to pull Snell, remarking that he "didn't want Mookie [Betts] or [Corey] Seager seeing Blake a third time" (Rivera 2020).

In his justification, Cash cites the third *Time Through the Order Penalty* (TTOP), which was first formally identified in Tango, Lichtman, and Dolphin (2007, pp. 187–190) and recently popularized by Lichtman (2013). It has long been observed that, on average, batters tend to perform better the more times they face a pitcher; for instance, they tend to get on base more often on their third time facing a pitcher than on their second. Tango, Lichtman, and Dolphin (2007) quantified the corresponding drop-off in pitcher performance as increases in *weighted on-base average* (wOBA; see Section 2.4 for details). They observed that the average wOBA of a plate appearance in the first time through the order (1TTO) is about 9 wOBA points less than that in the second TTO (2TTO). Further, the average wOBA of a plate appearance in the second TTO is about 8 wOBA points less than that in the third TTO (3TTO) (Tango, Lichtman, and Dolphin 2007, Table 81).

The TTOP is considered canon by much of the baseball community. Announcers routinely mention the 3TTOP during broadcasts and several managers regularly use the 3TTOP to justify their decisions to pull starting pitchers at the start of the third TTO. For instance, A. J. Hinch, who managed the Houston Astros from 2015 to 2019, noted "the third time through is very difficult for a certain caliber of pitchers to get through." Brad Ausmus, who managed the Detroit Tigers from 2014 to 2017, explained "the more times a hitter sees a pitcher, the more success that hitter is going to have" (Laurila 2015).

Tango, Lichtman, and Dolphin (2007) attribute the increased average wOBA from one TTO to the next to within-game batter learning. According to them, batters learn the tendencies of a pitcher as the game progresses. In fact, they observe "pitchers hitting a wall after 10 or 11 batters" rather than a "steady degradation in [pitcher] performance"

*Corresponding author: Ryan S. Brill**, Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania, Philadelphia, PA, USA, E-mail: ryguy123@sas.upenn.edu.
https://orcid.org/0000-0001-6387-7713
Sameer K. Deshpande, Department of Statistics, University of Wisconsin–Madison, Madison, USA, E-mail: sameer.deshpande@wisc.edu
Abraham J. Wyner, Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, USA, E-mail: ajw@wharton.upenn.edu

(Tango, Lichtman, and Dolphin 2007, pg. 189). Lichtman (2013) agrees and goes further, stating "the TTOP is not about fatigue. It is about [batter] familiarity."

We argue that Tango, Lichtman, and Dolphin (2007)'s analysis is insufficient to justify such sweeping conclusions. Tango, Lichtman, and Dolphin (2007) estimated the 2TTOP and 3TTOP by first binning plate appearances by lineup position and TTO. They then computed the average wOBA within each bin. Their analysis, by design, cannot disentangle continuous evolution in pitcher performance over the course of a game (e.g., from pitcher fatigue) from discontinuities between successive TTOs (e.g., from batter learning). Further, they provide no uncertainty quantification about their estimated TTOPs.

We conduct a more rigorous statistical analysis of the trajectory of pitcher performance over the course of a baseball game. Specifically, we fit a Bayesian multinomial logistic regression model to predict the outcome of each plate appearance as a function of the *batter sequence number*, batter quality, pitcher quality, handedness match, and home field advantage. The batter sequence number simply counts how many batters the pitcher has faced up to and including the current plate appearance. We find that the expected wOBA forecast by our model increases steadily over the course of a game and does not display sharp discontinuities between times through the order. Based on these results, we recommend managers cease pulling starting pitchers at the beginning of the 3TTO.

The remainder of this paper is organized as follows. We introduce our Bayesian multinomial logistic regression model of a plate appearance outcome in Section 2. We present our main findings in Section 3 and conclude by discussing implications of our results in Section 4.

2 Data and model specification

We begin with a brief overview of our MLB plate appearance dataset and identify several variables that may be predictive of the outcome of a plate appearance. We then introduce our Bayesian multinomial logistic regression model.

2.1 Retrosheet data

We scraped every plate appearance from 1990 to 2020 from the Retrosheet database. For each plate appearance, we record the outcome (e.g., out, single, etc.), the event wOBA, the handedness match between the batter and pitcher, and whether the batter is at home. We further compute measures of batter and pitcher quality for each plate appearance (see Section 2.5 for details). We include our final dataset, along with all pre-processing and data analysis scripts in Appendix A. We used R (R Core Team 2020) for all analyses.

We restrict our analysis to every plate appearance from 2012 to 2019 featuring a starting pitcher in one of the first three times through the order, using the 2017 season as our primary example. We remove plate appearances featuring switch hitters from our dataset. Our 2017 dataset consists of 108,519 plate appearances, 691 unique batters, and 315 unique starting pitchers.

There are $K = 7$ possible outcomes of a plate appearance: out, unintentional walk (uBB), hit by pitch (HBP), single (1B), double (2B), triple (3B), and home run (HR). For each $i = 1, \dots, n$, let y_i be the categorical variable indicating the outcome of the ith plate appearance. Notationally, we write

$$y_i \in \{1, 2, \dots, 7\} = \{\text{Out, uBB, HBP, 1B, 2B, 3B, HR}\}. \quad (1)$$

In predicting the probability of each plate appearance outcome, we need to control for several factors. We introduce the *batter sequence number* $t \in \{1, \dots, 27\}$, which records how many batters the pitcher has faced up to and including that plate appearance. We additionally construct indicators of being in the 2TTO and 3TTO, $\mathbb{I}(10 \le t \le 18)$ and $\mathbb{I}(19 \le t \le 27)$.

Intuitively, we expect that most pitchers are more likely to give up base hits and home runs to elite batters than they are to strike out elite batters. Similarly, we expect elite pitchers would have more plate appearances ending in outs than base hits against most batters. Accordingly, when modeling the outcome of a plate appearance, we adjust for the quality or skill of the batter and pitcher. To this end, let $x^{(p)}$ and $x^{(b)}$ denote the estimates of pitcher and batter quality, respectively. We discuss the computation of both quality measures in Section 2.5.

Additionally, we expect that a pitcher whose handedness matches that of the batter (e.g., the pitcher and batter are both right handed) is less likely to give up base hits and home runs than a pitcher whose handedness doesn't match the batter's. To this end, we define hand, an indicator that is equal to one when the batter and pitcher have matching handedness and zero otherwise. Finally, we expect that a pitcher on the road is more likely to give up base hits and home runs than a pitcher at home. Thus we define home, an indicator that is equal to one when the batter is at home and zero otherwise.

Table 1 summarizes the variables that we record from plate appearance *i*.

Table 1: Summary of variables measured for each at-bat that are used in our analysis.

Covariate symbol	Covariate description
y_i	Outcome of the ith plate appearance $\in \{1, \ldots, K = 7\}$
t_i	The batter sequence number $\in \{1, \ldots, 27\}$
$\mathbb{I}(t_i \in 2\text{TTO})$	Binary variable indicating whether the pitcher is in his second TTO
$\mathbb{I}(t_i \in 3\text{TTO})$	Binary variable indicating whether the pitcher is in his third TTO
$x_i^{(b)}$	Running-average estimator of batter quality
$x_i^{(p)}$	Running-average estimator of pitcher quality
hand_i	Binary variable indicating handedness match between batter and pitcher
home_i	Binary variable indicating whether the batter is at home
$\boldsymbol{x}_i$	$\boldsymbol{x}_i = (x_i^{(b)}, x_i^{(p)}, \text{hand}_i, \text{home}_i)$

2.2 A multinomial logistic regression model

We fit a Bayesian multinomial logistic regression model to predict the outcome of each plate appearance. For each non-out result ($k \neq 1$), we model

$$\log\left(\frac{\mathbb{P}(y_i = k)}{\mathbb{P}(y_i = 1)}\right) = \alpha_{0k} + \alpha_{1k}t_i + \beta_{2k}\mathbb{I}(t_i \in 2\text{TTO})$$
$$+ \beta_{3k}\mathbb{I}(t_i \in 3\text{TTO}) + \boldsymbol{x}_i^{\top}\eta_k, \qquad (2)$$

where the vector $\boldsymbol{x}_i$ concatenates our batter and pitcher quality and indicators for handedness and home team: $\boldsymbol{x}_i^{\top} = (x_i^{(b)}, x_i^{(p)}, \text{hand}_i, \text{home}_i)$.

The parameters α_{0k} and α_{1k} control the continuous evolution of the probability of each plate appearance outcome throughout the game. In contrast, the parameters β_{2k} and β_{3k} allow for discontinuities in these probabilities between different times through the order. Pitchers face each of the opposing team's batters, and so we interpret the term $\alpha_{0k} + \alpha_{1k}t$ as the *continuous* effect of a change in pitcher performance on the probability of each outcome. Batters, on the other hand, take turns facing the opposing team's pitcher, and so we interpret β_{2k} and $\beta_{3k} - \beta_{2k}$ as the respective *discontinuous* effects of a change in batter performance between the first and second times through the order and between the second and third times through the order. Observe that for $k \neq 1$, a large positive value of β_{2k} suggests that the non-out outcome k is systematically more likely to occur in the second time through the order than the first. Similarly, a large positive value of $\beta_{3k} - \beta_{2k}$ suggests that the outcome is more likely to occur in the third time through the order than the second. Consequently, based on our model parametrization, we would anticipate the 2TTOP and 3TTOP to manifest as positive values β_{2k} and $\beta_{3k} - \beta_{2k}$.

Our model allows the log-odds of each non-out plate appearance outcome to evolve linearly with batter sequence number. A more flexible model would not enforce a particular functional form on the change in pitcher performance

over the course of a game. Additionally, our model assumes that the trajectory of within-game pitcher deterioration is the same across all pitchers and batters. A more elaborate model would allow within-game performance to change at different rates for different players. We find that using these more elaborate models doesn't change the qualitative results of our study (see Appendix E).

Moreover, previous research suggests that pitchers decline continuously over the course of the game; Greenhouse (2011), for instance, documented continuous decreases in pitch velocity. On this view, the longer a pitcher stays in the game, the more likely he is to allow non-out outcomes in a plate appearance due to his continuous deterioration. We encode our intuition in Model (2) by constraining the slopes α_{1k} to be positive with a truncated prior:

$$\alpha_{1k} \sim \text{half } t_7. \qquad (3)$$

We specify standard normal priors to all of our other coefficients,

$$\alpha_{0k}, \ \beta_{2k}, \ \beta_{3k}, \ \eta_{\ell k} \sim \mathcal{N}(0, 1). \qquad (4)$$

Note that the qualitative results of our study remain the same when we use a diffuse prior $\mathcal{N}(0, 25)$ and drop the positive-slope constraint (see Appendix E.2).

Because the posterior distribution of (α, β, η) is not analytically tractable, we use Markov Chain Monte Carlo (MCMC) to draw approximate samples from the posterior distribution. We implement our sampler in Stan (Carpenter et al. 2017) and perform our MCMC simulation using the **rstan** package (Stan Development Team 2022). We use a high-performance computing cluster to run all of our computations.

Additionally, in Appendix B we conduct a simulation study to assess the capacity of our model to estimate time through the order penalties of various sizes. Specifically, we simulate data consistent with different TTOPs and verify that our posterior estimates are close to the data generating parameters.

2.3 Selection bias

We are primarily interested in understanding the magnitude and significance of discontinuous pitcher decline between times through the order. Formally, we are interested in the parameters β_{2k} and β_{3k} from our Model (2). Ideally we want to estimate β in the counterfactual scenario in which each pitcher faces each of the first 27 opposing batters. But, we cannot conduct a randomized controlled experiment; rather, we use observational data which is subject to the selection process of a baseball manager removing his starting pitcher. We visualize this selection process in Figure 1, which shows that worse pitchers (pitcher quality larger than, say, 0.34) are slightly more likely to be removed earlier in the game, as the corresponding histogram is shifted slightly to the left. Note that the six pitcher quality bins in Figure 1 are six evenly sized quantiles of pitcher quality.

Because our dataset is missing some 3TTO batting observations against worse pitchers, fitting Model (2) on our raw dataset may lead to a lower estimated probability of each non-out plate appearance outcome in 3TTO. To combat this, we remove all games from our dataset in which the starting pitcher is pulled prior to 3TTO. In 2017, for instance, this reduces our dataset by 8 % from 4860 games to 4469 games. Then, we fit our Model (2) on the reduced dataset, and we interpret our results as a TTOP (or lack thereof) *conditional on getting through* 2TTO. Conditional on getting through 2TTO, our dataset of all starting pitcher at-bats in the first three TTOs is balanced on the pitcher quality covariate, and so the TTOP discontinuity parameters β will not be biased due to the selection process. In other words, after our data truncation, the distribution of pitcher quality is similar for each batter sequence number t.

Even after truncating our dataset, since most starting pitchers are removed during 3TTO, our dataset is missing observations at the end of 3TTO. If each pitcher were allowed to pitch to the end of 3TTO, it is plausible that he would perform even worse than he did earlier in the game due to, for instance, additional fatigue. Therefore, we still underestimate the continuous pitcher decline parameters α. Nonetheless, as we are primarily interested in the discontinuity parameters β and not the continuity parameters α, we leave a more elaborate estimation of continuous pitcher decline to future work.

2.4 Measuring pitcher performance via wOBA

2.4.1 Weighted on-base average

Although Model (2) allows us to examine potential TTOPs for each plate appearance outcome, such multivariate measures are somewhat difficult to interpret and compare. We instead focus on quantifying the TTOP using a much more interpretable quantity, *weighted on-base average* (wOBA), which was first introduced in Tango, Lichtman, and Dolphin (2007).

wOBA overcomes many limitations of traditional metrics like batting average, on-base percentage, and slugging percentage. Briefly, batting average and on-base percentage treat all hits equally, with singles being worth as much as triples. Slugging percentage attempts to reward different types of hits differently, but does so in too simplistic of a fashion: in computing slugging percentage, a triple is worth three times what a single is worth. Such weighting is arbitrary, and is not tied to the relative impact of a triple over
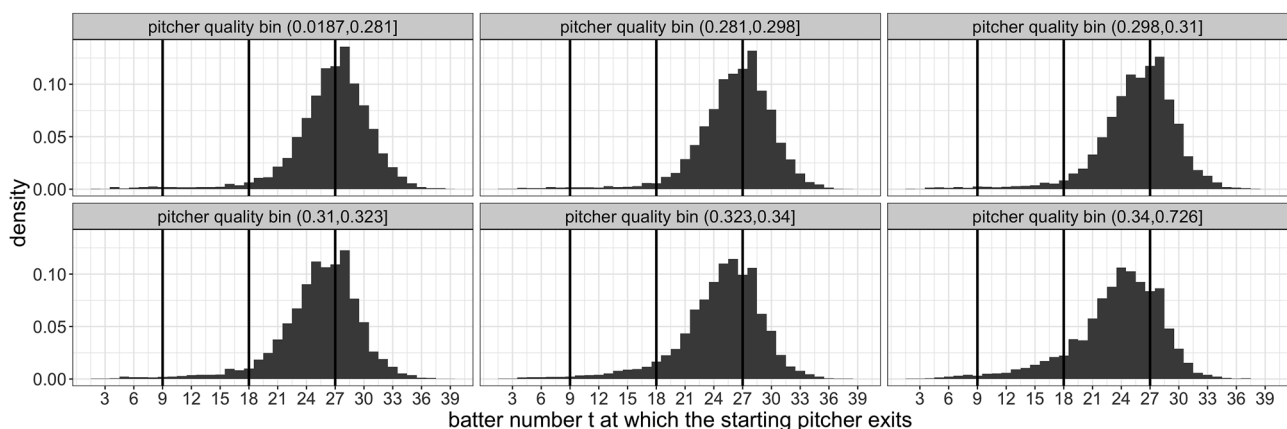

Figure 1: Histogram of the batter sequence number t at which a starting pitcher exits the game for different bins of pitcher quality.

a single with regard to, say, run scoring or win probability. wOBA combines the different aspects of offensive production into one metric, weighing each offensive action in proportion to its actual run value (Slowinski 2010). The wOBA of a plate appearance is simply the weight associated with the offensive action of the outcome. Specifically, the 2019 wOBA weight of each offensive action in decreasing order is 1.940 for a home run (HR), 1.529 for a triple (3B), 1.217 for a double (2B), 0.870 for a single (1B), 0.719 for hit-by-pitch (HBP), 0.690 for unintentional walks (uBB), and 0 for an out (OUT) Fangraphs (2021). wOBA is rescaled so that the league average wOBA equals the league average on-base percentage. Throughout this paper, we use 2019 wOBA weights for each season. Additionally, we usually refer to *wOBA points*, which is wOBA multiplied by 1000, to be consistent with the baseball community's use of wOBA.

To understand the effect size of a potential time through the order penalty, it is important to understand the distribution of wOBA points across batters and pitchers. In Figure 2, we plot the distribution of end-of-season mean plate appearance wOBA points for all batters and for all pitchers in 2017 who have over 100 plate appearances. Both batters and pitchers have a median wOBA points of 315. The standard deviation of wOBA points for batters is 41.5, and for pitchers is 36.7.

2.4.2 Expected weighted on-base average

Using Model (2), we can predict the probability of each plate appearance outcome. We can use these predicted probabilities to derive an *expected wOBA* for each plate appearance. We use expected wOBA to examine the trajectory of pitcher performance throughout the game.

To this end, let $k \in \{1, \ldots, K\}$ denote the outcome of a plate appearance and let $t \in \{1, \ldots, 27\}$ denote the tth batter a pitcher faces in a game. Also, let $x^{(b)}$ be the logit-transformed quality of the batter, $x^{(p)}$ the logit-transformed quality of the pitcher, hand be the binary indicator of the handedness match between the batter and pitcher, and home be the binary indicator of home field advantage. Define the *plate-appearance-state vector x* by

$$x^{\top} = (x^{(b)}, x^{(p)}, \text{hand}, \text{home}). \qquad (5)$$

Then, according to Model (2), the probability that a plate appearance involving the tth batter of a game and plate-appearance-state vector x results in outcome k is

$$\mathbb{P}(y = k | t, x) = \frac{\lambda_k(t, x)}{\sum\limits_{j=1}^{K} \lambda_j(t, x)}, \qquad (6)$$

where

$$\lambda_k(t, x) = \exp\big(\alpha_{0k} + \alpha_{1k}t + \beta_{2k}\mathbb{I}(t \in 2\text{TTO})$$
$$+ \beta_{3k}\mathbb{I}(t \in 3\text{TTO}) + x^{\top}\eta_k\big) \qquad (7)$$

when $k \neq 1$ and $\lambda_k(t, x) = 1$ when $k = 1$. From this, we define the *expected wOBA points* of a plate appearance involving the tth batter of a game and plate-appearance-state vector x by

$$\text{xwOBA}(t, x) = \sum_{k=1}^{K} 1000 \cdot w_k \cdot \mathbb{P}(y = k | t, x), \qquad (8)$$

where w_k is the wOBA weight of the kth plate appearance outcome.

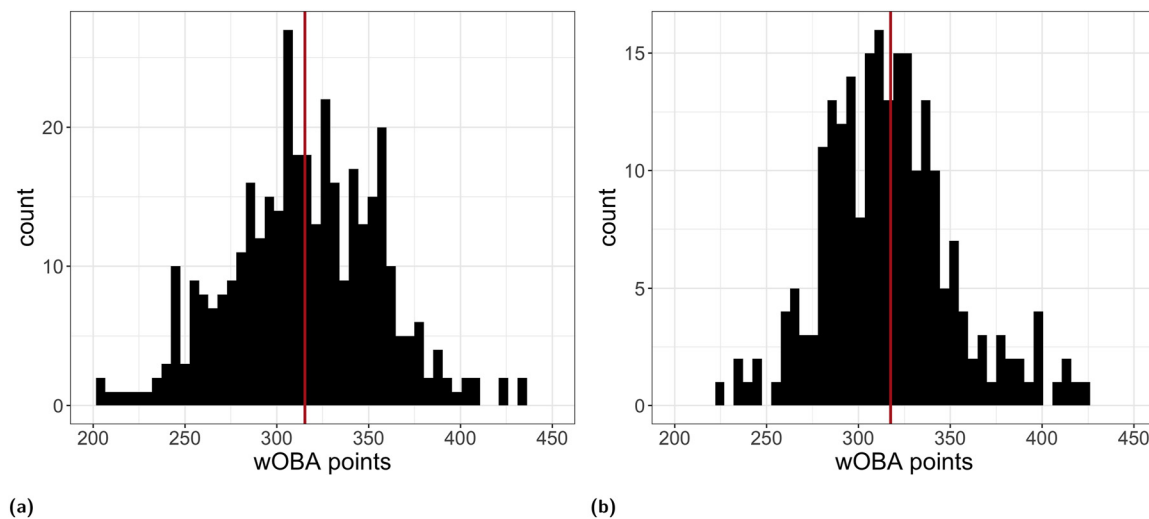**(a)**   **(b)**

Figure 2: The distribution of end-of-season mean wOBA points for all batters in 2017 (a) and all pitchers in 2017 (b) with over 100 plate appearances, using 2019 wOBA weights. The red line denotes the mean.

To visualize the nature of within-game pitcher decline implied by Model (2), we in Section 3 plot the trajectory of the expected wOBA of a plate appearance over the course of a game,

$$\{\text{xwOBA}(t, \boldsymbol{x})\}_{t=1}^{27}, \tag{9}$$

holding the plate-appearance-state vector $\boldsymbol{x}$ constant.

2.5 Definitions of pitcher and batter quality

To measure batter quality, we could use a batter's end-of-season average wOBA. Doing so, however, introduces a form of *data bleed* into our analysis: y_i, the wOBA of the ith plate appearance, is used to compute the batter's end-of-season average wOBA, so to use it as a covariate is to use y_i to help predict y_i. To avoid data bleed, we could instead use a batter's average wOBA over all prior plate appearances during the current season. Early in the season, however, this metric is extremely noisy. Hence we introduce a normal-normal conjugate running-average estimator that early in the season is close to a batter's average wOBA from the end of his previous season and that is closer to his current average wOBA later in the season.

Specifically, let x_{bj} be batter b's wOBA in his jth plate appearance of this season, and let θ_b represent batter b's unobservable "true quality" (the expected wOBA of a plate appearance with batter b this season). After observing j plate appearances, we model

$$x_{b1}, \ldots, x_{bj} | \theta_b \sim \mathcal{N}\left(\theta_b, \tau^2\right)$$
$$\theta_b \sim \mathcal{N}\left(\theta_{b0}, \nu^2\right). \tag{10}$$

Here, θ_{b0} represents batter b's prior "true quality." For non-rookies, we set θ_{b0} as the average wOBA of a plate appearance with batter b from his most recent previous season, and for rookies, we use the median $\theta_{b'0}$ over all other non-rookie batters b'. Additionally, ν represents the season-by-season standard deviation in a batter's average plate-appearance wOBA, and τ represents the within-season standard deviation of the wOBA of a batter's plate appearances.

Then, to measure batter b's quality through j plate appearances this season, we introduce the running-average estimator $\hat{\theta}_{bj}$ as the posterior mean $\mathbb{E}[\theta_b | x_{b1}, \ldots, x_{bj}]$ of θ_b, which as a result of our normal-normal conjugate model (10) is given by

$$\hat{\theta}_{bj} = \frac{\tau^{-2} \sum_{i=1}^{j} x_{bi} + \nu^{-2} \theta_{b0}}{j \tau^{-2} + \nu^{-2}}. \tag{11}$$

We then set $x_j^{(b)} = \text{logit}(\hat{\theta}_{bj})$. We use the logit-transformed estimates of batter quality because we felt it was more natural to allow the log-odds of each plate appearance outcome to evolve non-linearly with respect to these quality metrics. Specifically, we find it plausible that there are diminishing returns at both extremes of player quality. That is, we did not expect a small change in pitcher quality to manifest the same changes in the log-odds of a particular plate appearance outcome for a mediocre pitcher, an average pitcher, or an elite pitcher (keeping all else constant). The logit transformation allows us to capture this phenomenon. While this choice may appear somewhat unusual, we have found that it also yields a model with better predictive accuracy than a model that uses the raw quality covariates (see Appendix D.2).

We similarly construct a running estimate of pitcher p's quality through j plate appearances of the season with an analogous normal-normal model. For simplicity, we used the same values of ν and τ for batters and pitchers. To set ν, we first compute the event wOBA for each player-season from 2006 to 2019. Then we compute the standard deviation of these seasonal averages for each player. The median of these player-specific standard deviations was 0.0396 for pitchers and 0.0586 for batters. We finally set $\nu = 0.05$ to be the average of these values. To set τ, we compute the standard deviation of event wOBA for each player-season from 2006 to 2016. Across player-seasons, the median of these standard deviations was 0.509 for pitchers and 0.489 for batters. We set $\tau = 0.5$ as a simple compromise between these values.

3 Results

We fit our model to the data from each season in our dataset. In this section, we discuss our modeling results for the 2017 season. We observe qualitatively similar results in each other season.

To obtain our posterior samples, we run four MCMC chains for 1500 iterations. After discarding the first 750 iterations of each chain as "burn-in", the Gelman–Rubin $\hat{R}$ statistic is less than 1.1, suggesting convergence (Gelman and Rubin 1992). Additionally, the effective sample size of each parameter exceeds 1172 and the average effective sample size across all parameters is 2852. It took about eight hours to run each chain.

We begin in Section 3.1 by examining the marginal posterior distributions of β_{2k} and $\beta_{3k} - \beta_{2k}$, which quantify discontinuity in pitcher performance between successive times through the order. As noted in Section 2.2, large 2TTOP or 3TTOP would correspond to large, positive values of β_{2k} or $\beta_{3k} - \beta_{2k}$. We find, however, that the posterior distributions of these parameters are not tightly concentrated on positive values. Instead, we find that these distributions are,
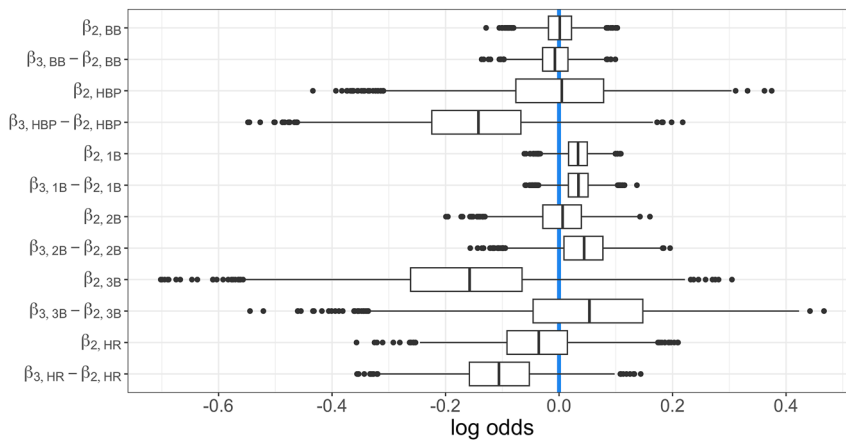
Figure 3: Posterior boxplots of the TTOP discontinuity parameters from Model (2), fit on data from 2017. The blue line denotes 0. We see that each posterior distribution covers both positive and negative values.

for the most part, centered near zero and place substantial probability on both positive and negative values. We also see that fitted xwOBA values increase steadily over the course of the game without discontinuity in the second or third time through the order. Taken together, these findings suggest that our model finds little evidence of strong discontinuity between successive times through the order.

At first glance, our results appear to contradict the findings of Tango, Lichtman, and Dolphin (2007). In Section 3.2, however, we discuss how the conclusions of Tango, Lichtman, and Dolphin (2007) actually fit within the framework of our model. We further find that pitcher and batter quality are much stronger predictors of xwOBA than the within-game change in pitcher performance.

3.1 Little evidence of strong discontinuity between successive times through the order

First, we examine the posterior distributions of the parameters β from our model (Equation (2)) which control discontinuous changes in pitcher performance. In Figure 3 we show boxplots of the posterior distributions of the discontinuity parameters[1] β_{2k} and $\beta_{3k} - \beta_{2k}$ from our model fit on data from 2017. Immediately we observe that none of these posterior distributions is tightly concentrated around a large positive value, which is what we would expect in the presence of a large 2TTOP or 3TTOP. Instead, most of these place considerable probability on both positive and negative values. The only exceptions are the posterior

distributions of $\beta_{2,1B}$ and $\beta_{3,1B} - \beta_{2,1B}$, which measure discontinuities in the log-odds of a single between times through the order. Although they both place over 80 % posterior probability on the positive axis, these distributions are supported on relatively small values. For instance, the posterior mean of $\beta_{3,1B} - \beta_{2,1B}$ is about 0.03 on the log-odds scale, which corresponds to a change in probability no greater than 0.75 percentage points. We additionally observe the posterior distributions corresponding to some outcomes like triples and hit-by-pitches are much more diffuse than those corresponding to other outcomes like walks and singles. This is not entirely unexpected: there are considerably more singles and walks in the dataset than triples and hit-by-pitches and the relative uncertainties about the corresponding β_{2k} and $\beta_{3k} - \beta_{2k}$ values closely track the frequencies of these outcomes. Ultimately, we do not find the posterior distributions in Figure 3 to be indicative of large, systematic time through the order penalties. We obtain similar findings in each season from 2012 to 2019 (see Figure 12 in Appendix D.3).

Furthermore, we plot the trajectory of a pitcher's expected wOBA over the course of the game according to our model, fit on data from 2017. Specifically, in Figure 4, we plot the posterior distribution of the sequence of xwOBA$(t, \tilde{x})$, where $\tilde{x}$ corresponds to an average batter facing an average pitcher of the same handedness on the road,

$$\tilde{x}^{\top} = (\overline{x^{(b)}}, \overline{x^{(p)}}, 1, 0). \tag{12}$$

The white dots, thick black bars, and thin black bars denote the posterior mean, 50 % credible interval, and 95 % credible interval of xwOBA$(t, \tilde{x})$. For now, ignore the blue lines, blue shaded regions, and gray shaded regions, which we explain the next Section 3.2. We see that expected wOBA increases steadily over the course of a game, without

---

[1] To better interpret the effect sizes of these parameters, which are on the log odds scale, we translate these values to the probability scale and the expected wOBA scale in Appendix D.1.
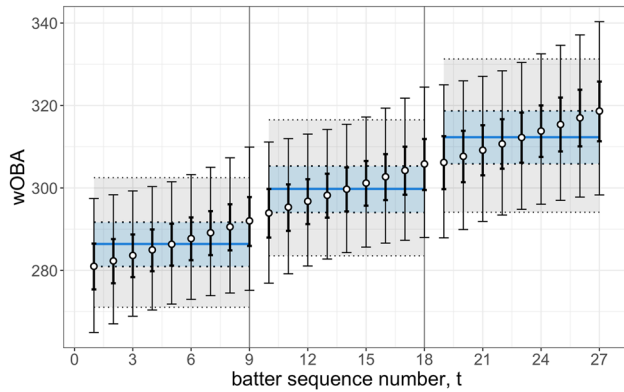
Figure 4: Trend in expected wOBA over the course of a game in 2017 for an average batter facing an average pitcher of the same handedness on the road. The white dots, thick black bars, and thin black bars denote the posterior mean, 50 % credible interval, and 95 % credible interval of xwOBA$(t, \tilde{x})$. The blue lines, blue shaded regions, and gray shaded regions denote the posterior mean, 50 % credible interval, and 95 % credible interval of xwOBA$(t, \tilde{x})$ averaged within each TTO.

discontinuity in the second or third time through the order. In other words, our model finds little evidence for a strong discontinuity in the expected wOBA of a plate appearance. This trend is persistent across each year from 2012 to 2019 (see Figure 13 of Appendix D.3) and other choices of x.

3.2 Tango et al (2007)'s conclusions fit within our framework

At first glance, our results appear to contradict the findings of Tango, Lichtman, and Dolphin (2007). Recall, however, that while we carefully estimate the xwOBA for each batter faced, Tango, Lichtman, and Dolphin (2007) identified the TTOP by comparing wOBA averaged across entire times through the order. By similarly averaging xwOBA(t, x) within times through the order, it turns out that we can recover the TTOP identified by Tango, Lichtman, and Dolphin (2007).

Formally, for each plate-appearance-state vector x, consider the average difference of xwOBA(t, x) between the first and second times through the order,

$$\mathscr{D}_{12}(x) = \frac{1}{9} \sum_{t=1}^{9} \left[\text{xwOBA}(t + 9, x) - \text{xwOBA}(t, x) \right]. \quad (13)$$

Using our fitted model, we study the posterior distribution of $\mathscr{D}_{12}(x)$ and the similarly defined $\mathscr{D}_{23}(x)$, which captures the change in average xwOBA between the second and third TTO.

The posterior means of $\mathscr{D}_{12}(\tilde{x})$ and $\mathscr{D}_{23}(\tilde{x})$ are about 13 wOBA points, which are consistent with Tango, Lichtman,

and Dolphin (2007)'s findings. Also, virtually all of the posterior samples are positive, suggesting that average pitcher performance indeed declines from one TTO to the next. Specifically, our model suggests that the expected wOBA points of an average plate appearance increases by 13.4 (with a 95 % credible interval of [7.78, 19.0]) from the first TTO to the second, and by 12.5 (with a 95 % credible interval of [5.98, 18.7]) from the second TTO to the third. We show histograms of the posterior samples of $\mathscr{D}_{12}(\tilde{x})$ and $\mathscr{D}_{23}(\tilde{x})$ in Figure 11 in Appendix D.1.

Figure 4 overlays the trajectory of xwOBA$(t, \tilde{x})$ with the posterior mean (the blue lines), the 50 % credible intervals (the blue shaded regions), and the 95 % posterior credible intervals (the gray shaded regions) of the xwOBA$(t, \tilde{x})$ trajectory averaged over each TTO. We see that mean pitcher performance within a TTO declines from each TTO to the next by about 13 wOBA points. Figure 4 reveals how these declines in average performance are an artifact of continuous, not discontinuous, pitcher decline.

3.3 The impact of handedness match and home field advantage on the outcome of a plate appearance

As discussed previously, pitchers decline from one TTO to the next by about 13 wOBA points on average. Now, we compare this effect size to that of confounders like batter quality, pitcher quality, handedness match, and home field advantage. We find that batter quality and pitcher quality have a much larger impact on predicting the outcome of a plate appearance, whereas handedness and home field advantage have a similar effect size as the batter sequence number.

We begin by assessing the impact of handedness match and home field advantage on the outcome of a plate appearance. To do so, we compute the posterior mean of the expected wOBA of a plate appearance averaged over the batter sequence numbers, for different combinations of handedness and home field advantage. Mathematically, for a batter of average quality with batter-at-home value home $\in$ $\{0, 1\}$ facing a pitcher of average quality having handedness match value hand $\in \{0, 1\}$, yielding plate-appearance-state vector

$$x^\top = (\overline{x^{(b)}}, \overline{x^{(p)}}, \text{home}, \text{hand}), \quad (14)$$

we compute the posterior mean and standard deviation of

$$\frac{1}{27} \sum_{t=1}^{27} \text{xwOBA}(t, x). \quad (15)$$

Table 2: For different combinations of handedness match and home field advantage, the posterior mean (and, in parenthesis, twice the posterior standard deviation) of the expected wOBA points of a plate appearance, assuming a batter of average quality faces a pitcher of average quality, averaged over the batter sequence numbers $t = 1, \dots, 27$.

		Batter at home	
		0	**1**
Hand match	0	316 (±7.8)	328 (±7.8)
	1	298 (±6.9)	310 (±7.2)

In Table 2 we show the posterior mean ± two posterior standard deviations[2] of Formula (15) for all combinations of hand and home. Home field advantage has a similar effect size as pitcher decline across one TTO: a batter at home has about 12 more mean expected wOBA points than a batter on the road. Handedness match has a slightly larger effect: a pitcher whose handedness matches that of the batter has about 18 fewer mean expected wOBA points than one whose handedness does not match. The xwOBA intervals, given by the posterior mean ± two posterior standard deviations, overlap for a batter at home versus away but do not overlap for a batter with versus without a handedness match. In other words, we find a significant handedness effect but not a significant home field effect.

3.4 The impact of batter quality and pitcher quality on the outcome of a plate appearance

Now, we assess the impact of batter quality and pitcher quality on the outcome of a plate appearance. To do so, for different combinations of batter and pitcher quality, we compute the posterior mean of the expected wOBA of a plate appearance, averaged over the batter sequence numbers $t \in \{1, \dots, 27\}$. Mathematically, for a batter of quality $x^{(b)}$ on

---

2 Although the posterior distribution of Formula (15) is not exactly Gaussian, we find that the actual 95 % credible interval is extremely close to the interval computed as the posterior mean ± twice the standard deviation.

the road facing a pitcher of quality $x^{(p)}$ with a handedness match, yielding plate-appearance-state vector

$$x^{\top} = (x^{(b)}, x^{(p)}, 1, 0), \qquad (16)$$

we compute the posterior mean and standard deviation of

$$\frac{1}{27} \sum_{t=1}^{27} \text{xwOBA}(t, x). \qquad (17)$$

In Table 3 we show the posterior mean ± two posterior standard deviations of Formula (17) for all combinations of the 25th, 50th, and 75th quantiles of $x^{(b)}$ and $x^{(p)}$. Specifically, we take the quantiles of the empirical distributions from Figure 2 from Section 2.4. For batters, the 25th quantile represents a bad batter, the 50th an average batter, and the 75th a good batter. Conversely, for pitchers, the 25th quantile represents a good pitcher, the 50th an average pitcher, and the 75th a bad pitcher.

As shown in Table 3, the quality of the batter and pitcher has a larger impact on the outcome of a plate appearance than the batter sequence number $t \in \{1, \dots, 27\}$. For instance, fix a batter's quality. The difference in mean expected wOBA points between a good and bad pitcher is large: about 42–48 wOBA points, depending on the batter quality. To see this, consider the second row of Table 3, in which a median batter (50th quantile) faces pitchers of various quality, assuming the batter is on the road and has the same handedness as the pitcher, averaged over each lineup position. The expected wOBA points of a plate appearance against a good pitcher (25th quantile) is 288, and against a bad pitcher (75th quantile) is 333. So, for a median batter, the difference in expected wOBA points between a good and a bad pitcher is about 45 wOBA points.

Conversely, fix a pitcher's quality. Then the difference in mean expected wOBA points between a good and bad batter is also large: about 36–41 wOBA points, depending on the pitcher quality. Finally, note that these effects are significant, as the intervals given by the posterior mean ± two posterior standard deviations do not overlap.

Therefore, pitcher quality and batter quality have a much larger impact on the outcome of a plate appearance than within-game pitcher decline.

Table 3: For different combinations of batter quality and pitcher quality (in terms of wOBA points) – in particular, the 25th, 50th, and 75th quantile – the posterior mean (and, in parenthesis, twice the posterior standard deviation) of the expected wOBA points of a plate appearance, assuming batters are on the road and have the same handedness as the pitcher, averaged over the batter sequence numbers $t = 1, \dots, 27$.

		Pitcher quality		
		25th quantile	**50th quantile**	**75th quantile**
Batter quality	25th quantile	270 (±6.7)	291 (±6.9)	313 (±7.5)
	50th quantile	288 (±7.0)	310 (±7.1)	333 (±7.7)
	75th quantile	306 (±7.7)	329 (±7.8)	354 (±8.5)

4 Discussion

It has long been observed that batters tend to perform better the more times they face a particular pitcher. Tango, Lichtman, and Dolphin (2007) first quantified the corresponding drop-off in pitcher quality and attributed the apparent time through the order penalty to batter learning. Their analysis, however, does not attempt to disentangle continuous evolution in pitcher performance over the course of the game from discontinuities between successive times through the order. We instead model the outcome of a plate appearance in a way that accommodates both of these. Our analysis reveals the expected wOBA of a plate appearance increases steadily over the course of the game, over average, without significant discontinuity between each time through the order. Additionally, the posterior distributions of the model parameters that quantify discontinuous pitcher decline cover both positive and negative values. These results suggest there is little evidence of strong discontinuity in pitcher performance between successive times through the order. Based on our analysis, we do not believe it always appropriate to pull pitchers at the start of the third time through the order. Rather, we recommend managers base their decisions to pull a pitcher on a pitcher's quality and continuous decline throughout the game.

Although Tango, Lichtman, and Dolphin (2007) attribute within-game pitcher decline to batter learning, we hesitate to make conclusions about the potential causes of within-game pitcher decline. Nonetheless, we offer potential interpretations of the parameters of our model from Equation (2). Because a batter faces the opposing team's pitcher at most once in each TTO, it is natural to interpret the parameters β_{2k} and $\beta_{3k} - \beta_{2k}$ which quantify discontinuous pitcher evolution as batter learning parameters. A pitcher, on the other hand, faces each opposing batter. Thus it is natural to interpret the parameters α_{0k} and α_{1k} which quantify continuous pitcher decline as pitcher fatigue parameters. In particular, it is known that pitchers fatigue continuously over the course of a game (e.g., Greenhouse 2011). Nonetheless, there are other potential mechanisms of pitcher decline (e.g., a changing pitch selection, discussed below), and we don't explicitly adjust for pitcher fatigue. Hence we hesitate to make causal conclusions from our model.

Furthermore, although our analysis is more nuanced than Tango, Lichtman, and Dolphin (2007)'s, our analysis is not without limitations. Recall that our model allows the log-odds of each non-out plate appearance outcome to evolve linearly with batter sequence number. A more flexible model would not force a particular functional form on the change in pitcher performance over the course of a game. We find that using a more flexible model doesn't change the qualitative results of our study (see Appendix E.1). Additionally, our model assumes that the trajectory of within-game pitcher deterioration is the same across all pitchers and batters. A more elaborate model would allow within-game performance to change at different rates for different players. We find that using this more elaborate model doesn't change the qualitative results of our study (see Appendix E.2).

Additionally, we note that there is enormous variation in pitching performance on a game-by-game basis. Although Tango, Lichtman, and Dolphin (2007, Chapter 7) believe this is due to randomness rather than pitcher "hotness", a more flexible model may use an estimate of pitcher quality which updates as a game evolves. For those who believe in pitcher "hotness", omitting a measure of within-game pitcher quality contributes further to selection bias. In particular, whether we observe a pitcher in 3TTO depends on his performance earlier in the game, as a pitcher who "bombs" or begins pitching poorly is more likely to be removed earlier in the game. We visualize this survival process in Figure 5, which shows that pitchers who have a bad pitching day (mean game wOBA larger than, say, 0.437) are much more likely to be removed earlier in the game. Note that the six mean game wOBA bins in Figure 5 are six evenly sized quantiles of mean game wOBA. So, a starting pitcher who remains in 3TTO pitched better that day over average than one who is pulled prior to 3TTO, and it is plausible that the former pitcher would be better in 3TTO than the latter pitcher. On this view, our approach underestimates the magnitude of continuous pitcher decline. But, as discussed in Section 2.3, our goal is to estimate the discontinuous decline parameters β, which our approach does a reasonable job of; we leave a more elaborate estimation of continuous pitcher decline to future work.

Furthermore, our analysis does not account for pitch selection, which, for some pitchers, evolves over the course of the game. Changes in pitch selection may be a response to pitcher fatigue: for instance, the more tired a pitcher becomes, the more difficult it may be to throw a fastball. Alternatively, pitchers might change their pitches in response to perceived batter learning: to prevent batters from learning his tendencies, a pitcher can perhaps be more unpredictable by changing his pitch selection over the game. A more fine-grained analysis would capture this within-game change in pitch selection, perhaps by modeling pitcher quality as a function of pitch selection. Nonetheless, modeling a pitcher's continuous change over the game may
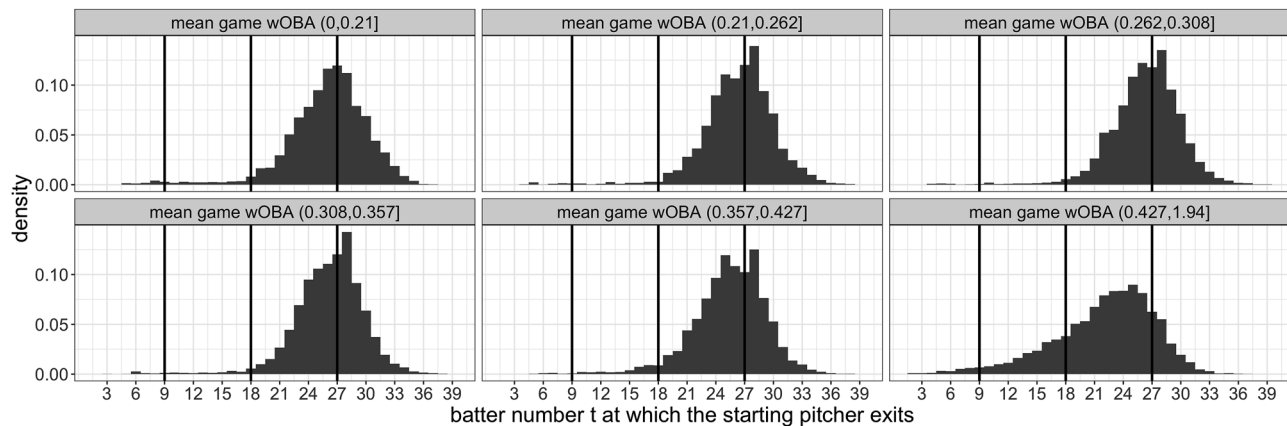
Figure 5: Histogram of the batter sequence number *t* at which a starting pitcher exits the game, for different bins of mean game wOBA.

simultaneously adjust for pitcher fatigue and an evolving pitch selection.

Additionally, recall that we use an empirical Bayes approach to quantify batter and pitcher quality. Specifically, early in the season we let a player's quality be close to his average wOBA from the end of his previous season, and later in the season be closer to his current average wOBA. Our current analysis shrinks to the prior season's average wOBA similarly for all players (e.g., the prior variance is constant). But, the more we've observed a player in the past, the more confident we should be in the player's ability this season. Thus a more fine-grained analysis would employ a more flexible empirical Bayes approach which allows the prior variance to vary in the number of last season's plate appearances (e.g., see Brown 2008). Additionally, a more elaborate approach may shrink to some combination of a pitcher's previous season mean wOBA and the overall mean of pitcher quality from the previous season, rather than shrinking to just the former.

Appendix A: Our code and data

Our code is available on Github.[3] The `data_wrangling` folder of the Github repository contains our dataset processing, including the Retrosheet data scraper. The `data` folder further processes the full dataset into a smaller dataset relevant for this paper. Finally, the `model_positive_slope_prior` folder contains our data analysis, including our `Stan` model.

The final datasets used in this paper are available for download.[4] The cleaned dataset of all MLB plate appearances from 1990 to 2020 is `retro_final_PA_1990-2020d.csv`. The datasets `TTO_dataset_410.csv` and `TTO_dataset_510.csv` are processed subsets of the large dataset which we use to fit our models.

Appendix B: Model simulation study

We conduct a simulation study to assess the capacity of our model (Equation (2)) to estimate time through the order penalties of various sizes. Specifically, we simulate data consistent with different TTOPs and verify that our posterior estimates are close to the data generating parameters.

B.1 Simulation setup

For our first simulation, we generate data consistent with continuous pitcher fatigue and no TTOP for any of the

---

3 https://github.com/snoopryan123/TTO_.
4 https://upenn.app.box.com/folder/144635702840?v=retrosheet-pa-1990-2000.

plate appearance outcomes by setting $\beta_{2k} = \beta_{3k} = 0$ for each $k \neq 1$. In our second simulation, for each $k \neq 1$, we set the β_{2k} and β_{3k} so that the resulting xwOBA curves display TTOPs consistent with Tango, Lichtman, and Dolphin (2007)'s findings of about 10 expected wOBA points between successive times through the order. Finally, for our third simulation, we set β_{2k} and β_{3k} so that there is no 2TTOP (in terms of xwOBA) but a large 3TTOP of about 50 wOBA points. For each simulation, we set the values of the α_{0k}'s, α_{1k}'s, and η_k's in a way that is consistent with observed data. Additional details about the simulation setup, including the data generating parameter values, are available in Appendix C.

For each simulation, we generate 225 full seasons worth of data. We fit our model to 80 % of the data from each simulated season and evaluate our fitted model's predictive performance on the remaining 20 %. We further assess how well our fitted model recovers the function xwOBA(t, x) for a set of average confounder values.

B.2 Simulation results

In all three simulation studies, we reliably recover the data generating parameters: averaged across all parameters, the estimated frequentist coverage of the marginal 95 % posterior credible intervals exceeds 92 % in each study. Importantly, the coverage of the 95 % posterior credible intervals for the discontinuity parameters β_{2k} and β_{3k} exceeds 91 % in each study. That is, for each simulated dataset, the 95 % credible intervals for the β_{2k}'s and β_{3k}'s usually contain the true data generating parameters. Furthermore, our model demonstrates good predictive capabilities (see Appendix C for details).

B.3 Simulation visualization

In each simulation, we visualize the trajectory of posterior expected wOBA over the course of the game for an average

batter on the road facing an average pitcher with the same handedness. That is, we plot the sequence $\{\text{xwOBA}(t, \tilde{x})\}_{t=1}^{27}$ where

$$\tilde{x}^\top = (\overline{x^{(b)}}, \overline{x^{(p)}}, 1, 0). \tag{18}$$

Figure 6 shows the sequence of posterior means, 50 %, and 95 % credible intervals of xwOBA$(t, \tilde{x})$ based on a single simulated dataset from each simulation setting. We overlay the true values of xwOBA$(t, \tilde{x})$, computed from the data generating parameters, to each plot. We see that in each of the three simulation studies, we recover the true underlying expected wOBA trajectory.

Appendix C: Simulation details

C.1 Data generating parameters

The exact data generating parameter values of β_{2k} and β_{3k} for our three simulation studies are shown in Table 4.

Furthermore, in each of our simulation studies, we assume that pitchers fatigue linearly over the course of a game. The particular true parameter values of α_{0k} and α_{1k} used in each of our simulation studies are shown in Table 5.

Finally, in each of our simulation studies, we set the value of η to mimic fitted values from observed data. The particular true parameter values of η used in each of our simulation studies are shown in Table 6.

C.2 Predictive performance on simulated data

Our model demonstrates good predictive capabilities. To get a general sense of our model's performance, we use out-of-sample cross entropy loss, given by

$$-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{7} 1\{y_i = k\} \cdot \log\bigl(\mathbb{P}(y_i = k)\bigr). \tag{19}$$
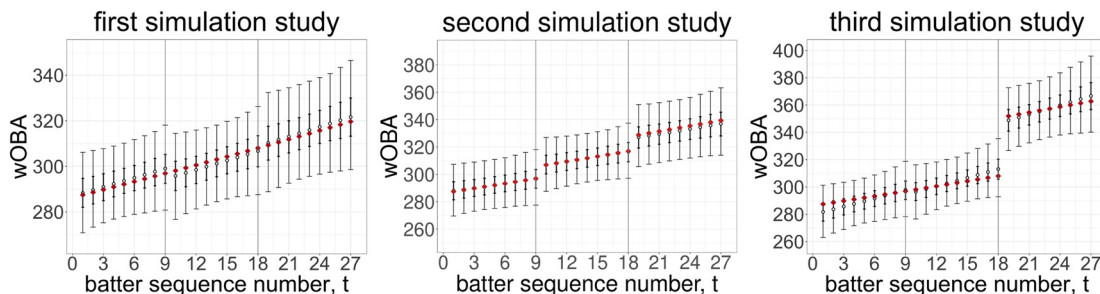

Figure 6: Trend in xwOBA over the course of a game from our first, second, and third simulation studies. The red dots indicate the true underlying expected wOBA values, the white dots indicate the posterior means of the xwOBA values, the thick black error bars denote the 50 % posterior credible intervals, and the thin black error bars denote the 95 % posterior credible intervals.

Table 4: The data generating parameter values of β_{2k} and β_{3k} in each of our three simulations.

	$k =$ BB	$k =$ HBP	$k =$ 1B	$k =$ 2B	$k =$ 3B	$k =$ HR
β_{2k} for sim 1	0	0	0	0	0	0
β_{3k} for sim 1	0	0	0	0	0	0
β_{2k} for sim 2	2/65	0	4/65	2/65	0	2/65
β_{3k} for sim 2	1/15	0	2/15	1/15	0	1/15
β_{2k} for sim 3	0	0	0	0	0	0
β_{3k} for sim 3	1/10	1/10	3/10	1/10	1/10	3/20

Table 5: The data generating parameter values of α_{0k} and α_{1k} in each of our three simulations.

	$k =$ BB	$k =$ HBP	$k =$ 1B	$k =$ 2B	$k =$ 3B	$k =$ HR
α_{0k}	−0.601	−1.804	−0.475	−0.943	−1.510	−0.565
α_{1k}	0.00271	0.0122	0.00354	0.00635	0.0223	0.00926

Table 6: The data generating parameter values of η in each of our three simulations.

	$k =$ BB	$k =$ HBP	$k =$ 1B	$k =$ 2B	$k =$ 3B	$k =$ HR
$\eta_{\text{bat\_quality}}$	0.865	1.408	0.371	0.856	1.399	1.525
$\eta_{\text{pit\_quality}}$	1.128	1.987	1.050	1.472	3.286	1.850
η_{hand}	−0.201	0.166	−0.0164	−0.0420	−0.462	−0.0958
η_{home}	0.0792	−0.0776	0.0245	−0.00103	0.107	0.0230

For each of our three simulations, the average cross entropy loss over each of our 25 datasets is 1.05, 1.06, and 1.07, respectively. Using the empirical outcome probabilities yields an average out-of-sample cross-entropy loss of 1.06, 1.08, and 1.08, respectively, for each of our three simulations. It is reassuring that our model (barely) outperforms the observed base rates.

Appendix D: Observed model fit details

D.1 The impact of pitcher decline on the outcome of a plate appearance

In this Section, we quantify the effect size of pitcher decline over the course of a game, again using the 2017 season as our primary example.

In particular, we examine how the probability of each outcome of a plate appearance changes over the course of a game. Specifically, we use the posterior distribution of $\mathbb{P}(y = k|t, \boldsymbol{x})$, defined in Formula (6), to characterize the amount by which pitchers decline within a game. In particular, we compute the posterior distribution of the change in the probability of outcome $k \neq 1$ from 1TTO to 2TTO, over average,

$$\mathscr{D}_{12}(k, \boldsymbol{x}) = \frac{1}{9}\sum_{t=10}^{18} \mathbb{P}(y = k|t, \boldsymbol{x}) - \frac{1}{9}\sum_{t=1}^{9} \mathbb{P}(y = k|t, \boldsymbol{x}), \quad (20)$$

and the similarly defined $\mathscr{D}_{23}(k, \boldsymbol{x})$, which captures the change in the probability of outcome $k \neq 1$ from 2TTO to 3TTO, over average.

In Figure 7 we plot the posterior distribution of $\mathscr{D}_{12}(k, \tilde{\boldsymbol{x}})$, using plate-appearance-state vector $\tilde{\boldsymbol{x}}$ from Formula (12). From 1TTO to 2TTO, the probability of a single increases by about 0.005, the probability of a home run increases by about 0.003, and the probability of the other non-out categories change negligibly. With this, the probability of an out decreases by about 0.01. So, there is a small decrease in pitcher performance on average from 1TTO to 2TTO.

Similarly, in Figure 8 we plot the posterior distribution of $\mathscr{D}_{23}(k, \tilde{\boldsymbol{x}})$. From 2TTO to 3TTO, the probability of a single increases by about 0.005, the probability of a double increases by about 0.004, and the probability of the other non-out categories change negligibly. With this, the probability of an out decreases by about 0.01. So, there is a small decrease in pitcher performance on average from 2TTO to 3TTO.

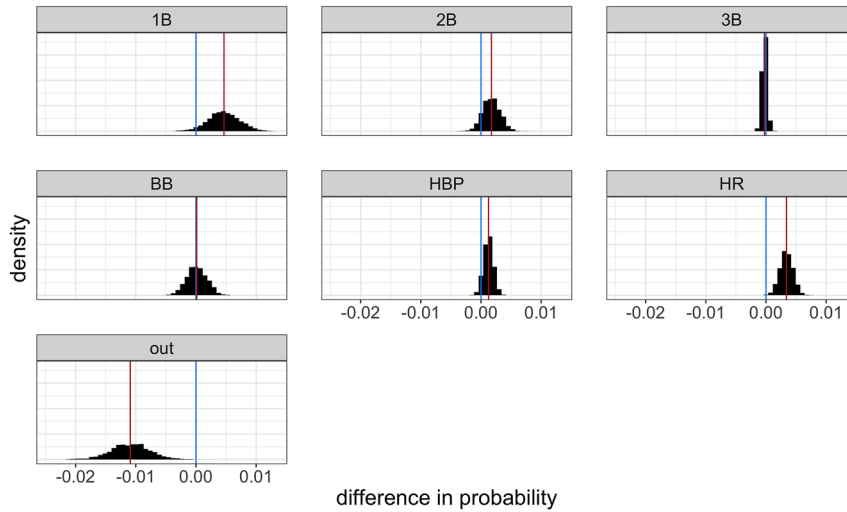Additionally, we examine how the expected wOBA of each outcome of a plate appearance changes over the course
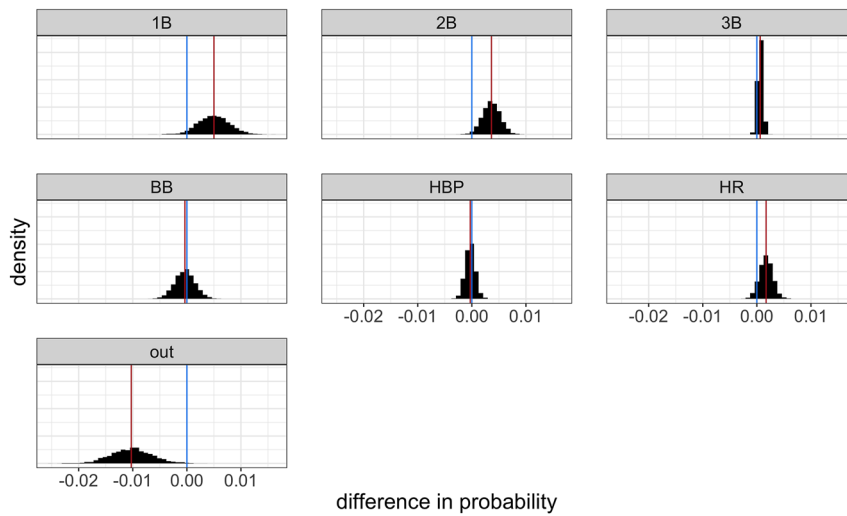
Figure 7: The difference in probability of each plate appearance outcome between 2TTO and 1TTO on average (assuming a batter of average quality on the road faces a pitcher of average quality with a handedness match during each plate appearance). Equivalently, the posterior distribution of $\mathscr{D}_{12}(k, \tilde{x})$. The red line denotes the mean, and the blue line denotes 0.

Figure 8: The difference in probability of each plate appearance outcome between 3TTO and 2TTO on average (assuming a batter of average quality on the road faces a pitcher of average quality with a handedness match during each plate appearance). Equivalently, the posterior distribution of $\mathscr{D}_{23}(k, \tilde{x})$. The red line denotes the mean, and the blue line denotes 0.

of a game. In particular, we compute the posterior distribution of the change in the expected wOBA of outcome $k \neq 1$ from 1TTO to 2TTO, over average,

$$\mathscr{D}_{12}'(k, x) = \frac{1}{9} \sum_{t=10}^{18} 1000 \cdot w_k \cdot \mathbb{P}(y = k|t, x)$$

$$- \frac{1}{9} \sum_{t=1}^{9} 1000 \cdot w_k \cdot \mathbb{P}(y = k|t, x), \quad (21)$$

where w_k is the wOBA weight for outcome k as discussed in Section 2.4. Similarly, we define $\mathscr{D}_{23}'(k, x)$, which captures

the change in the expected wOBA of outcome $k \neq 1$ from 2TTO to 3TTO, over average.

In Figure 9 we plot the posterior distribution of $\mathscr{D}_{12}'(k, \tilde{x})$, using plate-appearance-state vector $\tilde{x}$ from Formula (12). From 1TTO to 2TTO, the expected wOBA points of a home run increases by about six, the expected wOBA points of a single increases by about four, and the other non-out categories change negligibly. Note that the expected wOBA of an out doesn't change because an out is worth zero wOBA.

Similarly, in Figure 10 we plot the posterior distribution of $\mathscr{D}_{23}'(k, \tilde{x})$. From 2TTO to 3TTO, the expected wOBA of a
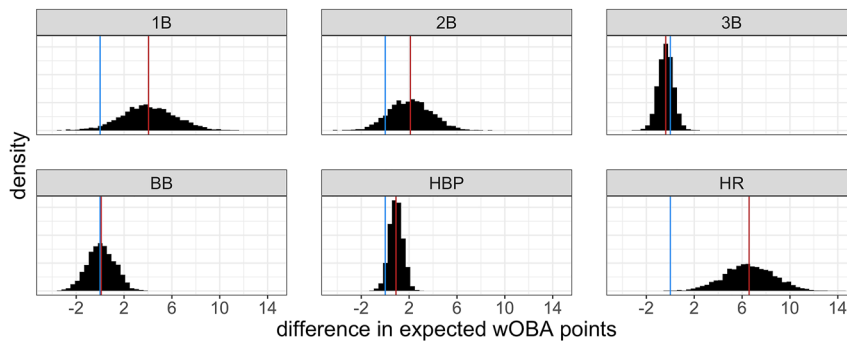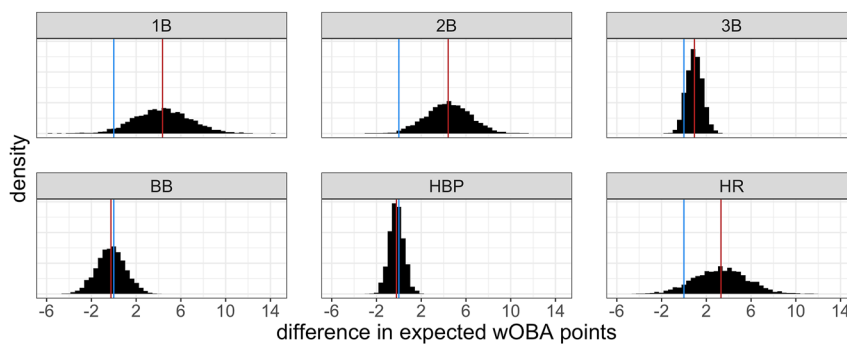
Figure 9: The difference in xwOBA of each plate appearance outcome between 2TTO and 1TTO on average (assuming a batter of average quality on the road faces a pitcher of average quality with a handedness match during each plate appearance). Equivalently, the posterior distribution of $\mathscr{D}'_{23}(k, \tilde{x})$. The red line denotes the mean, and the blue line denotes 0.

Figure 10: The difference in xwOBA of each plate appearance outcome between 3TTO and 2TTO on average (assuming a batter of average quality on the road faces a pitcher of average quality with a handedness match during each plate appearance). Equivalently, the posterior distribution of $\mathscr{D}'_{23}(k, \tilde{x})$. The red line denotes the mean, and the blue line denotes 0.

double and single increases by about five, the xwOBA of a home run increases by about three, and the other categories change negligibly.

Furthermore, we aggregate the increase in the probability of each non-out plate appearance outcome k from one TTO to the next via expected wOBA, defined in Equation (8). In particular, recall from Section 3.2 that a pitcher declines by about 13 wOBA points from one TTO to the next, over average, which is consistent with the effect sizes from Figures 7 and 8. Figure 11 illustrates this via a histogram of the posterior samples of $\mathscr{D}_{12}(\tilde{x})$ and $\mathscr{D}_{23}(\tilde{x})$. We see that virtually all of these samples are positive, suggesting that average pitcher performance declines from one TTO to the next, and that the means of these distributions are around 13 wOBA points, which are consistent with Tango, Lichtman, and Dolphin (2007)'s findings. Specifically, our model suggests that the expected wOBA points of an average plate appearance increases by 13.4 (with a 95 % credible interval of [7.78, 19.0]) from the first TTO to the second, and by 12.5 (with a 95 % credible interval of [5.98, 18.7]) from the second TTO to the third.

D.2 Predictive performance on observed data

To get a general sense of our model's performance on observed data, we run a five-fold cross validation to predict the probability of each plate appearance outcome for each plate appearance in 2017. The out-of-sample cross entropy loss, given by Formula (19), is 1.035. We compare our model's cross entropy loss to that of other prediction strategies to better understand its performance. Consider a five-fold cross validation using the base rates of each plate appearance outcome. So, for each fold, find the proportion of plate appearances in which each outcome occurs, and compute the cross entropy loss using these base rates on the remaining out-of-sample plate appearances. For reference, in 2017, an out occurs in 67.6 % of plate appearances, an uBB 7.8 %, an HBP 0.9 %, a 1B 14.9 %, a 2B 4.8 %, a 3B 0.45 %, and an HR in 3.5 % of plate appearances. The out-of-sample cross entropy loss of the base rates of each outcome is 1.042. So, our model very slightly outperforms the base rates. Finally, note that our model using raw batter and pitcher quality
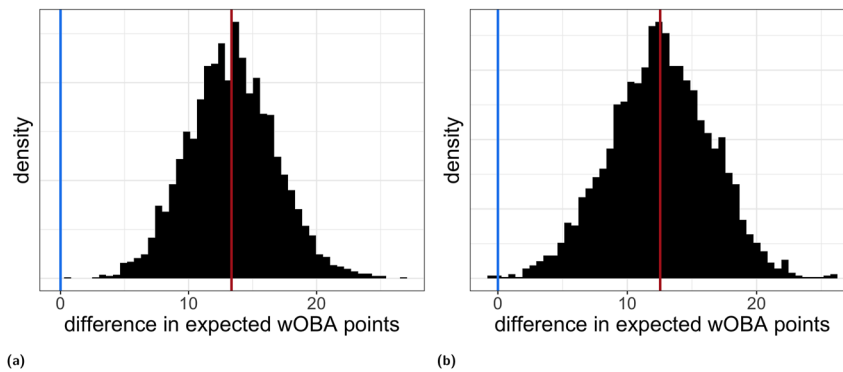
(a) **(b)**

Figure 11: The posterior distribution of the mean batter improvement, or mean pitcher decline in xwOBA, from 1TTO to 2TTO (left) and from 2TTO to 3TTO (right). Equivalently, the posterior distributions of $\mathcal{D}_{12}(\tilde{\boldsymbol{x}})$ (left) and $\mathcal{D}_{23}(\tilde{\boldsymbol{x}})$ (right) (see Formula (13)). The red line denotes the mean, and the blue line denotes 0. We see that batters improve relative to the pitcher by about 13 wOBA points on average from one TTO to the next.

covariates, rather than logit-transformed batter and pitcher quality covariates, has a cross-validated out-of-sample cross entropy loss of 1.040. That the logit-transformed player quality covariates have better out-of-sample predictive performance helps justify using the logit transform.

D.3 The trend is persistent across years

In Figure 12 we show boxplots of the posterior distributions of the discontinuity parameters β_{2k} and $\beta_{3k} - \beta_{2k}$ from our model (Equation (2)) fit separately on data from each season from 2012 to 2019. For some outcomes (e.g., walks), the posterior distributions are tightly concentrated around 0, and for other outcomes (e.g., triples and

hit-by-pitches, which are rare events), the posterior distributions are quite wide, which is compatible with a large effect in either direction. Overall, the posterior distributions of the discontinuity parameters cover both positive and negative values, and most of them are centered around 0. In particular, we don't see what we would expect to see if there were strong evidence for a TTOP (i.e., we don't see the posterior distributions tightly concentrated around a positive number). Ultimately, we do not find the posterior distributions in Figure 12 to be consistent with large, systematic time through the order penalties.

In Figure 13 we plot the posterior distribution of xwOBA over the course of a game according to our model fit separately on data from each year from 2012 to 2019. We see that
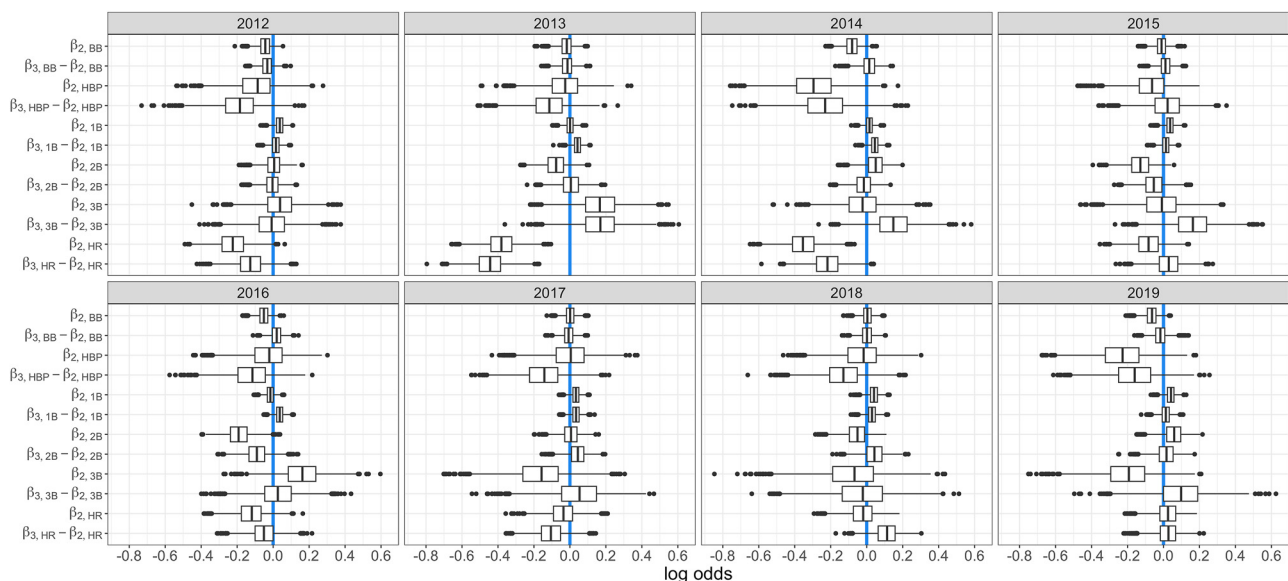

Figure 12: Posterior boxplots of the TTOP discontinuity parameters from Model (2), fit separately on data from each year from 2012 to 2019. The blue line denotes 0. We see that each posterior distribution covers both positive and negative values.
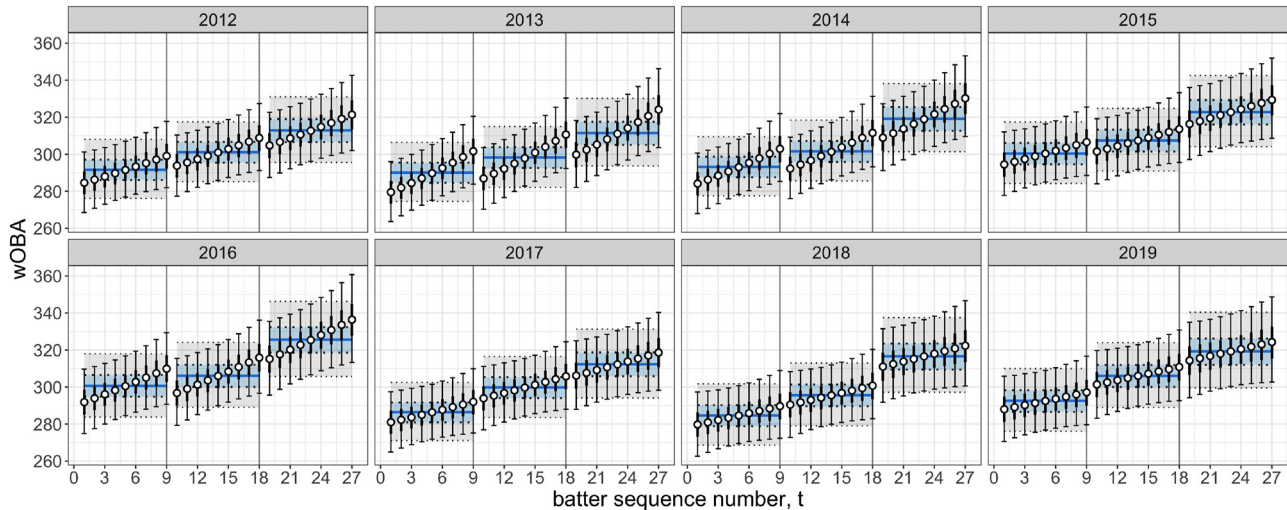
Figure 13: Trend in expected wOBA over the course of a game for an average batter facing an average pitcher of the same handedness on the road, according to the model from Equation (2) fit on separately on data from each year from 2012 to 2019. The white dots indicate the posterior means of the expected wOBA values, the thick black error bars denote the 50 % credible intervals, and the thin black error bars denote the 95 % credible intervals.

expected wOBA increases steadily over the course of a game, without significant discontinuity (in particular, significant *upward* discontinuity) between times through the order. The 2018 season is the only season in which we see an upward discontinuity in the posterior means, which occurs between 2TTO and 3TTO. This discontinuity, however, lies inside of the credible intervals and so is not significant.

Appendix E: Alternative models

E.1 A more flexible model: the indicator model

In Equation (2) we model pitcher decline over the course of a game as the combination of discontinuous decline from each TTO to the next and continuous linear pitcher decline across all the batters. A more flexible model would not enforce a particular functional form on within-game pitcher decline. In particular, the most flexible model has a separate coefficient for each batter $t \in \{1, \ldots, 27\}$,

$$\log\left(\frac{\mathbb{P}(y_i = k)}{\mathbb{P}(y_i = 1)}\right) = \sum_{t=1}^{27} \alpha_{tk} \mathbb{I}(t_i = t) + x_i^\top \eta_k. \quad (22)$$

With this more flexible model, the qualitative results of our study don't change. For instance, as in Figure 4, in Figure 14 we plot the posterior distribution of the trajectory of expected wOBA over the course of a game, according to the indicator model from Equation (22) fit on data from 2017. We do not see a significant discontinuity in pitcher
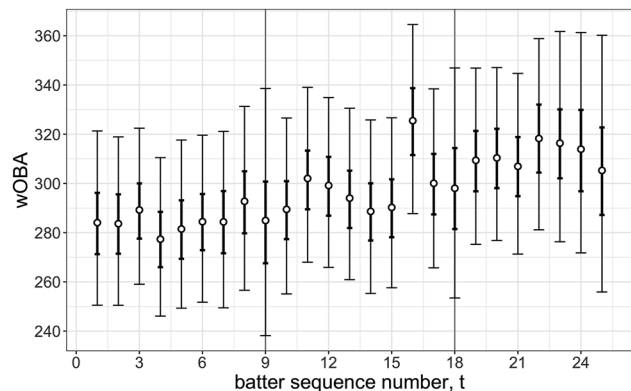

Figure 14: Trend in expected wOBA over the course of a game in 2017 for an average batter facing an average pitcher of the same handedness on the road, according to the indicator model from Equation (22). The white dots indicate the posterior means of the expected wOBA values, the thick black error bars denote the 50 % credible intervals, and the thin black error bars denote the 95 % credible intervals.

performance from one TTO to the next. In other words, we don't find evidence of a strong batter discontinuity between times through the order. This trend is persistent across each year from 2012 to 2019.

E.2 A more elaborate model: pitcher-specific and batter-specific effects

In our model from Equation (2), we make the simplifying assumption that the trajectory of within-game pitcher deterioration is the same across all pitchers and batters.

Nonetheless, it is likely that pitcher performance declines at different rates for different players. To account for such heterogeneity, we extend our model by introducing player-specific rates of decline. Specifically, we model

$$\log\left(\frac{\mathbb{P}(y_i = k)}{\mathbb{P}(y_i = 1)}\right) = \alpha_{0k\,p(i)} + \alpha_{1k\,p(i)}t_i + \beta_{2kb(i)}\mathbb{I}\left(t_i \in 2\text{TTO}\right)$$

$$+ \beta_{3kb(i)}\mathbb{I}\left(t_i \in 3\text{TTO}\right) + \boldsymbol{x}_i^\top \eta_k, \qquad (23)$$

where $p(i)$ is the index of the pitcher and $b(i)$ is the index of the batter in at-bat i. The pitcher-specific continuous decline parameters and batter-specific discontinuity parameters have Gaussian priors,

$$\begin{cases} \alpha_{0k\,p(i)} \sim \mathcal{N}\left(\alpha_{0k}, \sigma_{0k}^2\right), \\ \alpha_{1k\,p(i)} \sim \mathcal{N}\left(\alpha_{1k}, \sigma_{1k}^2\right), \\ \beta_{2kb(i)} \sim \mathcal{N}\left(\beta_{2k}, \sigma_{2k}^2\right), \\ \beta_{3kb(i)} \sim \mathcal{N}\left(\beta_{3k}, \sigma_{3k}^2\right), \end{cases} \qquad (24)$$

which themselves have priors,

$$\begin{cases} \alpha_{0k}, \alpha_{1k}, \beta_{2k}, \beta_{3k} \sim \mathcal{N}(0, 25), \\ \sigma_{0k}^2, \sigma_{1k}^2, \sigma_{2k}^2, \sigma_{3k}^2 \sim \text{half } \mathcal{N}(0, 1). \end{cases} \qquad (25)$$

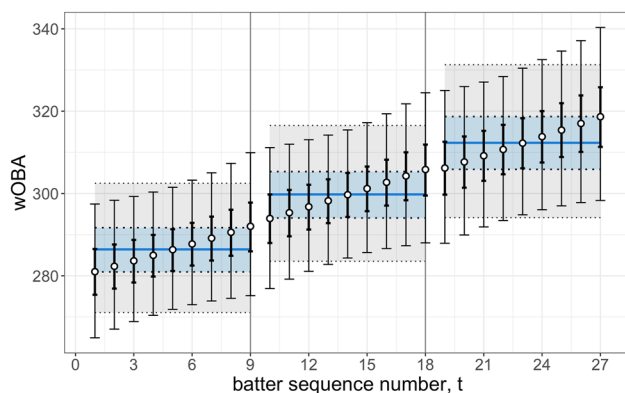With this more flexible model, the qualitative results of our study don't change. For instance, as in Figure 4, in

Figure 15: Trend in expected wOBA over the course of a game in 2017 for an average batter facing an average pitcher of the same handedness on the road, according to the model from Equation (23). The white dots indicate the posterior means of the expected wOBA values, the thick black error bars denote the 50 % credible intervals, and the thin black error bars denote the 95 % credible intervals.

Figure 15 we plot the posterior distribution of the trajectory of expected wOBA over the course of a game, according to the player-specific model from Equation (23) fit on data from 2017. In particular, we use the posterior distributions of the prior means $\alpha_{0k}, \alpha_{1k}, \beta_{2k}$, and β_{3k} to compute the xwOBA trajectory for an average pitcher facing an average batter. We do not see a significant upwards discontinuity in expected wOBA from one TTO to the next. In other words, we find little evidence for a strong batter discontinuity between times through the order. This trend is persistent across each year from 2012 to 2019.

References

Brown, L. D. 2008. "In-Season Prediction of Batting Averages: A Field Test of Empirical Bayes and Bayes Methodologies." *Annals of Applied Statistics* 2 (1): 113−52.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1−32.

Fangraphs. 2021. wOBA and FIP Constants. https://www.fangraphs.com/guts.aspx?type=cn.

Gelman, A., and D. B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7: 457−72.

Greenhouse, J. 2011. Spitballing: Fourth Time's the Harm. https://www.baseballprospectus.com/news/article/13117/spitballing-fourth-times-the-harm/.

Laurila, D. 2015. Managers on the Third Time through the Order. https://blogs.fangraphs.com/managers-on-the-third-time-through-the-order/.

Lichtman, M. 2013. Baseball ProGUESTus: Everything You Always Wanted to Know about the Times through the Order Penalty. https://www.baseballprospectus.com/news/article/22156/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rivera, J. 2020. Rays' Kevin Cash Explains Decision to Pull Blake Snell in World Series: 'I Regret it Because it Didn't Work Out'. https://www.sportingnews.com/us/mlb/news/kevin-cash-blake-snell-world-series-explained/lfnyfc4nqwys1pcncc2lnyjho.

Slowinski, P. 2010. wOBA. https://library.fangraphs.com/offense/woba/.

Stan Development Team. 2022. *RStan: The R Interaface for Stan*.

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