Sameer K. Deshpande* and Abraham Wyner

A hierarchical Bayesian model of pitch framing

https://doi.org/10.1515/jqas-2017-0027

Abstract: Since the advent of high-resolution pitch tracking data (PITCHf/x), many in the sabermetrics community have attempted to quantify a Major League Baseball catcher's ability to "frame" a pitch (i.e. increase the chance that a pitch is a called as a strike). Especially in the last 3 years, there has been an explosion of interest in the "art of pitch framing" in the popular press as well as signs that teams are considering framing when making roster decisions. We introduce a Bayesian hierarchical model to estimate each umpire's probability of calling a strike, adjusting for the pitch participants, pitch location, and contextual information like the count. Using our model, we can estimate each catcher's effect on an umpire's chance of calling a strike. We are then able translate these estimated effects into average runs saved across a season. We also introduce a new metric, analogous to Jensen, Shirley, and Wyner's Spatially Aggregate Fielding Evaluation metric, which provides a more honest assessment of the impact of framing.

Keywords: baseball; Bayesian modeling; uncertainty quantification.

1 Introduction

The New York Yankees and Houston Astros played each other in the American League Wild Card game in October 2015, with the winner continuing to the next round of the Major League Baseball playoffs. During and immediately after the game, several Yankees fans took to social media expressing frustration that home plate umpire Eric Cooper was not calling balls and strikes consistently for both teams, thereby putting the Yankees at a marked disadvantage. Even players in the game took exception to Cooper's decision making: after striking out, Yankees catcher Brian McCann argued with Cooper that he was calling strikes on similar pitches when the Astros were pitching but balls when the Yankees were pitching. Figure 1 shows two pitches thrown during the game, one by Astros pitcher Dallas Keuchel and the other by Yankees pitcher Masahiro Tanaka.

Both pitches were thrown in roughly similar locations, near the bottom-left corner of the *strike zone*, the rectangular region of home plate shown in the figure. According to the official rules, if any part of the pitched ball passes through the strike zone, the umpire ought to call it a strike. Keuchel's pitch barely missed the strike zone while Tanaka's missed by a few inches. As a result, the umpire Cooper ought to have called both pitches a ball. That Cooper did not adhere strictly to the official rules is hardly surprising; previous research has shown umpires' ball/strike decisions may be influenced by the race or ethnicity of the pitcher (see, e.g. Parsons et al. 2011; Tainsky, Mills, and Winfree 2015), player status as measured by age or ability (see, e.g. Kim and King 2014; Mills 2014), and their previous calls (Chen, Moskowitz, and Shue 2016). During the television broadcast of the game, the announcers speculated that the difference in Cooper's strike zone enforcement was the ability of Astros catcher Jason Castro to "frame" pitches, catching them in such a way to increases Cooper's chance of calling a strike (Sullivan 2015).

Though pitch framing has received attention from the sabermetrics community since 2008, it has generated tremendous interest in the popular press (see, e.g. Lindbergh 2013; Pavlidis 2014; Sanchez 2015) and among team officials (see, e.g. Drellich 2014; Holt 2014) in the last 3 or 4 years, due to its apparently large impact on team success. According to Woodrum (2014), most studies of framing, including the most recent by Judge, Pavlidis, and Brooks (2015) for the website Baseball Prospectus, estimate that a good framer can, on average, save his team as many as 25 runs more than the average catcher, over the course of the season. By the traditional heuristic of 10 average runs per win (Cameron 2008), these results suggest that the way a good framer catches a few pitches per game may be worth as many as an additional 2–3 wins, relative to the average catcher. Despite the ostensibly large impact framing may have on team success, framing itself has been overlooked and undervalued until just recently. Sanchez (2015) highlights the catcher Jonathan Lucroy, whose framing accounted for about 2 wins in the 2014 and worth about $14 M, writing that "the most impactful player in baseball today is the game's 17th highest-paid catcher."

***Corresponding author: Sameer K. Deshpande,** The Wharton School, University of Pennsylvania – Statistics, 434 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104, USA, e-mail: dsameer@wharton.upenn.edu. http://orcid.org/0000-0003-4116-5533
Abraham Wyner: University of Pennsylvania, Philadelphia, PA 19104-6243, USA

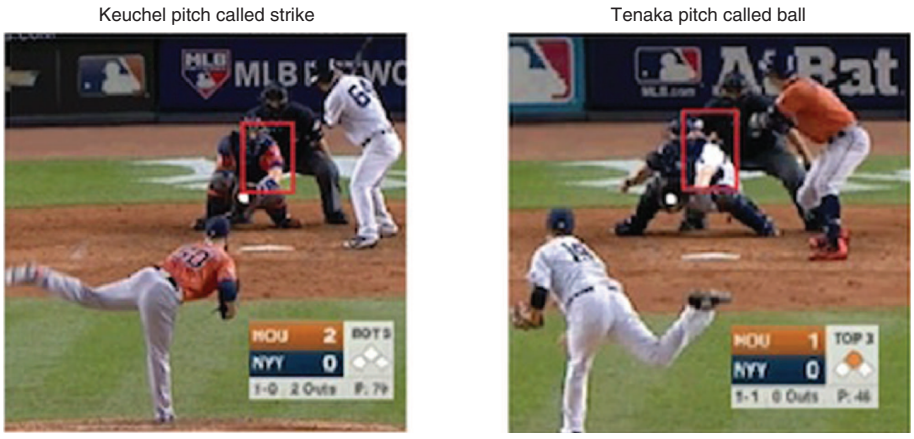

Figure 1: Both pitches missed the strike zone (outlined in red) and by rule, should have been called balls. Keuchel's pitch (left) was called a strike while Tanaka's pitch (right) was called a ball. Screenshot source: http://www.fangraphs.com/blogs/how-the-astros-wound-up-with-a-bigger-zone/

Returning to the two pitches in Figure 1, Cooper may have been more likely to call the Keuchel pitch a strike because of Castro's framing. However, looking carefully at Figure 1, we see that the two pitches are quite different, making it difficult to immediately attribute the difference in calls to Castro. First, Keuchel's pitch is much closer to the displayed strike zone than Tanaka's and it was thrown in a 1–0 count while Tanaka's was thrown in a 1–1 count. We also note that the batters, catchers, and pitchers involved in each pitch are, necessarily, different. In particular, Keuchel is a left-handed pitcher and Tanaka is a right-handed pitcher. Any of these factors may have contributed to Cooper being more likely to call Keuchel's pitch a strike. Of course, it could also be the case that Cooper was equally likely to call both pitches a strike and the different calls are simply due to noise. This raises questions: what effect did Castro have on Cooper's called strike probability, over and above factors like the pitch location, count, and the other pitch participants? And what impact does such an effect have on his team's success?

Existing attempts to answer these questions fall broadly into two categories: those that do not fit statistical models of the called strike probability and those that do. The first systematic study of framing (Turkenkopf 2008) falls into the former category. For each catcher, he counts the number of strikes called on pitches thrown outside an approximate strike zone introduced by Walsh (2007). Turkenkopf (2008) then took the counts of "extra strikes" received by each catcher and converted them into a measure of runs saved using his own valuation of 0.16 runs saved per strike. Missing from this analysis, however, is any consideration of the other players and the umpire involved in the pitch, as well as the context in which the pitch was thrown (e.g. count, run differential, inning, etc.). This omission could overstate the apparent impact of framing since it is not immediately clear that a catcher deserves all of the credit for an extra strike. More recently, Rosales and Spratt (2015) proposed an iterative method to distribute credit for a called strike among the batter, catcher, pitcher, and umpire. Unfortunately, many aspects of their model remain proprietary and thus, the statistical properties of their procedure are unknown.

The second broad category of framing studies begins by fitting a statistical model of the called strike probability that accounts for the above factors. Armed with such a model, one then estimates the predicted called strike probability with and without the catcher. The difference in these probabilities reflects the catcher's apparent framing effect on that pitch. One then estimates the impact of framing by weighting these effects by the value of "stealing a strike" and summing over all pitches caught by a catcher. Marchi (2011) fit a mixed-effects logistic regression model, expressing the log-odds of a called strike as a function of the identities of the pitch participants and interactions between them. This model does not systematically incorporate pitch location, meaning that the resulting estimates of framing effects are confounded by the location just like Turkenkopf (2008)'s. To our knowledge, the most systematic attempt to study framing to date is Judge et al. (2015). They introduce a mixed-effects probit regression model built in two stages: first, they estimate a baseline called strike probability using a proprietary model that accounts for location, count, handedness of batter, and ballpark. They then fit a probit regression model with a fixed effect for this baseline estimate and random effects

for the pitch participants. Underpinning their model is the curious assumption that the probit transformed called strike probability is *linear* in the baseline probability estimate. This assumption can over-leverage pitches with baseline probabilities close to 0 or 1 (e.g. pitches straight over home plate or several inches outside the strike zone) by arbitrarily inflating the associated intercept and slopes in the final probit model. This can potentially result in highly unstable parameter estimates. Both Judge et al. (2015)'s and Marchi (2011)'s models unrealistically assume umpires differ only in some base-rate of calling strikes and that effect of factors like pitch location, count influence, and players is constant across umpires. In light of this, we will proceed by fitting a separate model for each umpire.

Before proceeding, we introduce some notation. For a given taken pitch, let $y = 1$ if it is called a strike and let $y = 0$ if it is called a ball. Let **b**, **ca**, **co**, **p** and **u** be indices corresponding to the batter, catcher, count, pitcher, and umpire for that pitch. Further, let x and z be the horizontal and vertical coordinates of the pitch as it crosses the front plane of home plate, respectively. To accommodate a separate model for each umpire u we introduce vectors $\Theta^{u,B}$, $\Theta^{u,CA}$, $\Theta^{u,P}$, and $\Theta^{u,CO}$ to hold the *partial effect* of each batter, catcher, count, and pitcher, respectively, on umpire u's likelihood to call a strike. For each umpire u, we introduce a function of pitch location, $f^u(x, z)$, that we will specify in more detail later. At a high level, we model

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = \Theta^{\mathbf{u},B}_{\mathbf{b}} + \Theta^{\mathbf{u},CA}_{\mathbf{ca}} + \Theta^{\mathbf{u},P}_{\mathbf{p}} + \Theta^{\mathbf{u},CO}_{\mathbf{co}} + f^{\mathbf{u}}(x, z) \tag{1}$$

We leverage high-resolution pitch tracking data from the PITCHf/x system, described briefly in Section 2.1, to estimate how much a catcher influences umpires' chances of calling strikes and how large an impact such effects have on his team's success. In Section 2.3, we introduce several simplifications of the model in Equation 1 that still elicit umpire-to-umpire heterogeneity. All of these models are fit in a hierarchical Bayesian framework, which provides natural uncertainty quantification for our framing estimates. Such quantification, notably absent in previous framing studies, is vital, considering the fact that several teams are making framing-based roster decisions (Drellich 2014; Holt 2014; Sanchez 2015). We compare the predictive performances of these models in Section 3.1 and assess the extent to which incorporating umpire-specific count and player effects lead to overfitting. We then translate our estimates of catcher effects from the log-odds scale to the more conventional scale of average runs saved. We introduce two metrics in Section 4 to estimate the impact framing has on team success. We conclude with a discussion and outline several potential extensions of our modeling efforts.

2 Data and model

We begin this section with a brief overview, adapted primarily from Fast (2010) and Sidhu and Caffo (2014), of our pitch tracking dataset before introducing the hierarchical logistic regression model used to estimate each umpire's called strike probability.

2.1 PITCHf/x data

In 2006, the TV broadcast company Sportvision began offering the PITCHf/x service to track and digitally record the full trajectory of each pitch thrown using a system of cameras installed in major league ballparks. During the flight of each pitch, these cameras take 27 images of the baseball and the PITCHf/x software fits a quadratic polynomial to the 27 locations to estimate its trajectory (Sidhu and Caffo 2014). This data is transmitted to the MLB Gameday application, which allows fans to follow the game online (Fast 2010). In addition to collecting pitch trajectory data, an MLB Advanced Media employee records game-state information during each pitch. For instance, he or she records the pitch participants (batter, catcher, pitcher, and umpire) as well as the outcome of the pitch (e.g, ball, swinging strike, hit), the outcome of the at-bat (e.g. strike-out, single, home run), and any other game action (e.g. substitutions, baserunners stealing bases). The PITCHf/x system also reports the approximate vertical boundaries of the strike zone for each pitch thrown. Taken together, the pitch tracking data and game-state data provide a high-resolution pitch-by-pitch summary of the game, available through the MLB Gameday API.

Though our main interest in this paper is to study framing effects in the 2014 season, we collected all PITCHf/x data from the 2011 to 2015 regular season. In Section 2.2, we use the data from the 2011 to 2013 seasons to select the function of pitch location $f^u(x, z)$ from Equation 1. We then fit our model using the 2014 data and in Section 3.1, we assess our model's predictive performance using data from 2015. In the 2014 season, there were a total of 701,490 pitches, of which 355,293 (50.65%) were *taken* (i.e. not swung at) and of these taken pitches, 124,642 (35.08%) were called strikes. Rather than work with all of the taken pitches, we restrict our attention to those pitches that are "close enough" to home plate to be "frameable." More precisely, we first approximate a crude

"average rule book strike zone" by averaging the vertical strike zone boundaries recorded by the PITCHf/x system across all players and all pitches, and then focused on the $N = 308{,}388$ taken pitches which were within one foot of this approximate strike zone. In all, there were a total of $n_U = 93$ umpires, $n_B = 1010$ batters, $n_C = 101$ catchers, and $n_P = 719$ pitchers.

2.2 Adjusting for pitch location

Intuitively, pitch location is the main driver of called strike probability. The simplest way to incorporate pitch location into our model would be to include the horizontal and vertical coordinates (x, z) recorded by the PITCHf/x system as linear predictors so that $f^u(x, z) = \theta_x^u x + \theta_z^u z$, where θ_x^u and θ_z^u are parameters to be estimated. While simple, this forces an unrealistic left-to-right and top-to-bottom monotonicity in the called strike probability surface. Another simple approach would be to use a polar coordinate representation, with the origin taken to be the center of the approximate rule book strike zone. While this avoids any horizontal or vertical monotonicity, it assumes that, all else being equal, the probability of a called strike is symmetric around this origin.

Such symmetry is not observed empirically, as seen in Figure 2, which divides the plane above home plate into 1″ squares whose color corresponds to the proportion of pitches thrown in the 3 year window 2011–2013 that pass through the square that are called strikes. Also shown in Figure 2 is the average rule book strike zone, demarcated with the dashed line, whose vertical boundaries are the average of the top and bottom boundaries recorded by the PITCHf/x system. If the center of the pitch passes through the region bound by the solid line, then some part of the pitch passes through the approximate strike zone. This heat map is drawn from the umpire's perspective so right handed batters stand to the left of home plate (i.e. negative X values) and left-handed batters stand to the right (i.e. positive X values). We note that the bottom edge of the figure stops 6 inches off of the ground and the left and right edges end 12 inches away from the edges of home plate. Typically, batters stand an additional 12 inches to the right or left of the region displayed. Interestingly, we see that the empirical called strike probability changes rapidly from close to 1 to close to 0 in the span of only a few inches.

Rather than specifying a explicit parametrization in terms of the horizontal and vertical coordinates, we propose using a smoothed estimate of the historical log-odds of a called strike as an implicit parametrization

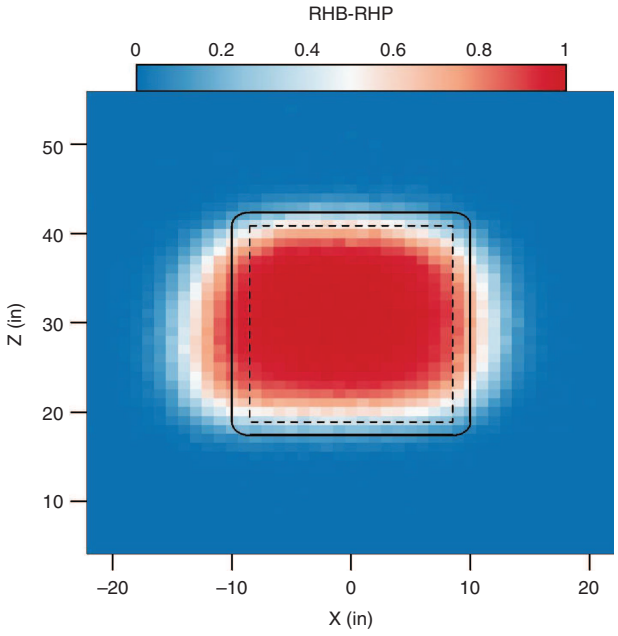

Figure 2: Heat map of empirical called strike probabilities, aggregate over the 3-year window 2011–2013. The boundary of the approximate 2014 rule book strike zone is shown in dashed line. If the center of the pitch passes through the region bounded by the solid line, some part of the pitch passes through the approximate strike zone. Red = 100% called strike probability, white = 50%, and blue = 0%.

of pitch location. This is very similar to the model of Judge et al. (2015), who included the estimated called strike probability as a covariate in their probit model.

Figure 3 plots the spatial distribution of taken pitches broken down by the batter and pitcher handedness. Once again, the plots are drawn from the umpires' perspective so that a right handed batter stands to the left side of the figure and vice versa. We see immediately that the spatial distribution of taken pitches varies considerably with the combination of batter and pitcher handedness. When the batter and pitcher are of the same handedness, we see a decidedly higher density of "low and outside" pitches near the bottom corner of the average rule book strike furthest away from the batter. In contrast, in the matchup between left-handed batters and right-handed pitchers, we see a higher density of pitches thrown to the outside edge of the strike zone further away from the batter. The differences in spatial distribution of pitches seen in Figure 3 motivate us to use a separate smoothed estimate of the historical log-odds of a called strike for each combination of batter and pitcher handedness.

Inspired by Mills (2014), we fit generalized additive models with a logistic link to the data aggregated from 2011 to 2013, one for each combination of pitcher and batter handedness, hereafter referred to as the "hGAMs"

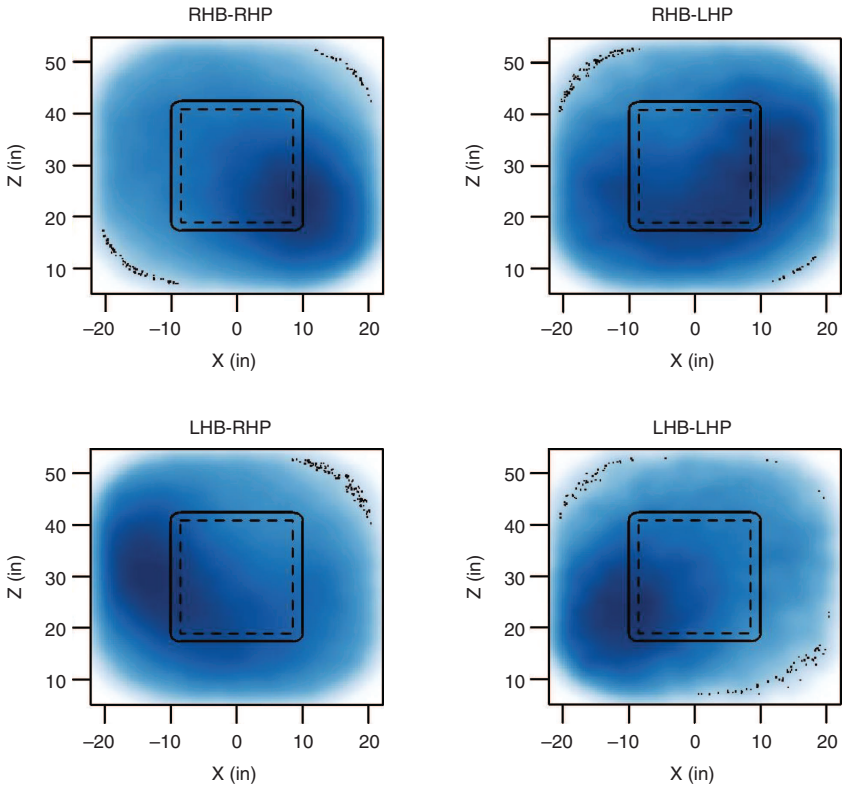

Figure 3: Kernel density estimate of pitch location based on batter and pitcher handedness. Figures drawn from the umpires perspective so right-handed batters stand to the left of the displayed strike zone. Darker regions correspond to a higher density of pitches thrown to those locations.

or "historical GAMS." These models express the log-odds of a called strike as a smooth function of the pitch location. Figure 4 shows the hGAM forecasted called strike probabilities. Interestingly, we see that for right handed pitchers, the corresponding hGAMs called strike probability surfaces very nearly align with the average rule book strike zone. For left-handed pitchers, however, the hGAMs forecast a high called strike probability several inches to the left of the average rule book strike zone. This is perhaps most prominent for the matchup between right-handed batters and left-handed pitchers. For each taken pitch in 2014 dataset, we used the appropriate hGAM to estimate the historical log-odds that the pitch was called a strike. We then use these estimates as continuous predictors in our model, so that potential player effects and count effects may be viewed as adjustments to these historical baselines.

2.3 Bayesian logistic regression models

Before fully specifying our models, we label the 93 umpires $u_1,\ldots,u_{93}$. Consider the ith called pitch and let $y_i = 1$ if it is called a strike and $y_i = 0$ if it is called a ball. Let h_i be a vector of length four, encoding the combination of batter and pitcher handedness on this pitch and let $\mathbf{LO}_i$ be a vector of length four, containing three zeros and the estimated log-odds of a strike from the appropriate historic GAM based on the batter and pitcher handedness. Letting x_i and z_i denote the PITCHf/x coordinates of this pitch, we take $f^u(x_i, z_i) = h_i^\top \Theta_0^u + \mathbf{LO}_i^\top \Theta_{LO}^u$, where Θ_{LO}^u is a vector of length four recording the partial effect of location and Θ_0^u is a vector of length four containing an intercept term, one for each combination of batter and pitcher handedness. Finally, let $u(i)$ denote which umpire called this pitch. To place all of the variables in our model on roughly similar scales, we first re-scale the corresponding historical GAM estimates for each combination of batter and pitcher handedness to have standard deviation 1.

Finally let $\mathbf{CO}_i$, $\mathbf{CA}_i$, $\mathbf{P}_i$ and $\mathbf{B}_i$ be vectors encoding the count, catcher, pitcher, and batter involved with this pitch, and let $\Theta_{CO}^u, \Theta_{CA}^u, \Theta_P^u$, and Θ_B^u be vectors containing the partial effect of count, catcher, pitcher, and batter on umpire u. For identifiability, we specify a single catcher, Brayan Pena, and count, 0–0,

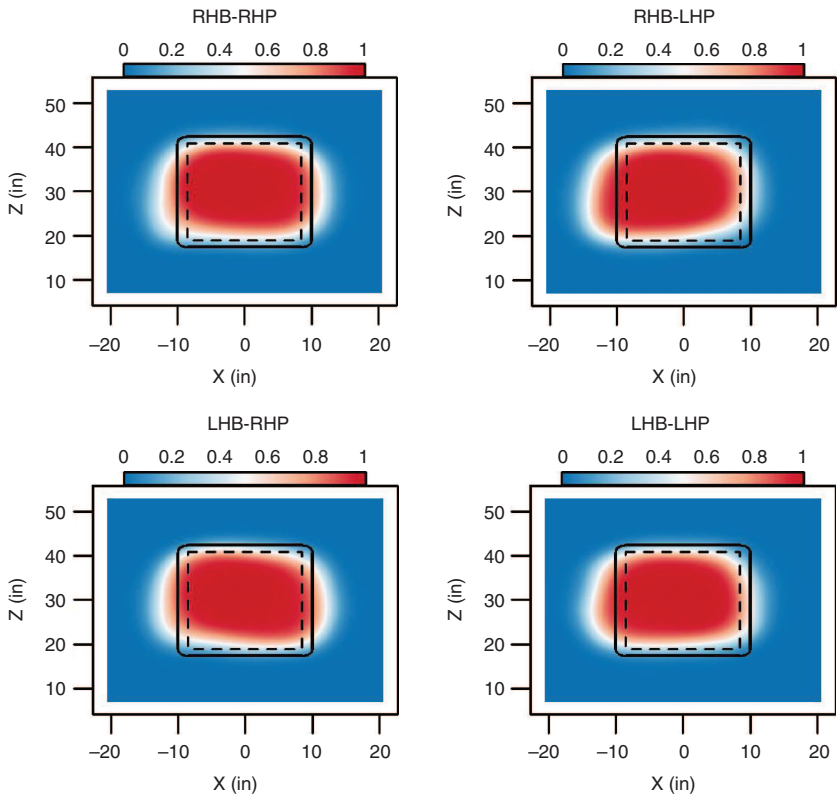

Figure 4: hGAM forecasts based on batter and pitcher handedness. Red = 100% called strike probability, white = 50%, and blue = 0%.

as baseline values. We can re-write the model from Equation 1 as

$$\log\left(\frac{P(y_i=1)}{P(y_i=0)}\right) = h_i^\top \Theta_0^{u(i)} + \mathbf{LO}_i^\top \Theta_{LO}^{u(i)} + \mathbf{CO}_i^\top \Theta_{CO}^{u(i)} + \mathbf{CA}_i^\top \Theta_{CA}^{u(i)} + \mathbf{P}_i^\top \Theta_P^{u(i)} + \mathbf{B}_i^\top \Theta_B^{u(i)}$$

We are now ready to present several simplifications of this general model of gradually increasing complexity. We begin first by assuming that the players and count have no effect on the log-odds of a called strike (i.e. that $\Theta_{CO}^u, \Theta_{CA}^u, \Theta_P^u$, and Θ_B^u are all equal to the zero vector for each umpire). This model, hereafter referred to as Model 1, assumes that the only relevant predictor of an umpire's ball/strike decision is the pitch location but allows for umpire-to-umpire heterogeneity. We model, *a priori*,

$$\Theta_0^{u_1}, \ldots, \Theta_0^{u_{93}} | \Theta_0 \sim N\left(\Theta_0, \tau_0^2 I_4\right)$$
$$\Theta_{LO}^{u_1}, \ldots, \Theta_{LO}^{u_{93}} | \Theta_{LO} \sim N\left(\Theta_{LO}, \tau_{LO}^2 I_4\right)$$
$$\Theta_0 | \sigma_0^2 \sim N\left(0_4, \sigma_0^2 I_4\right)$$
$$\Theta_{LO} | \sigma_{LO}^2 \sim N\left(\mu_{LO}, \sigma_{LO}^2 I_4\right)$$

The vector μ_{LO} is taken to be the vector of standard deviations of the hGAM forecast for each combination of batter and pitcher handedness. In this way, Model 1 centers the prior distribution of the log-odds of a strike at the hGAM forecasted log-odds. We may interpret the parameters τ_0^2 and τ_{LO}^2 as capturing the umpire-to-umpire variability in the intercept and location effects and we may view Θ_0 and Θ_{LO} as the mean intercept and location effects averaged over all umpires. By placing a further level of prior hierarchy on Θ_0 and Θ_{LO}, we render the Θ_0^u's and Θ_{LO}^u's dependent, both *a priori* and *a posteriori*. In this way, while we are fitting a separate model for each umpire, these models are "mutually informative" in the sense that the estimate of umpire u's intercept vector Θ_0^u will, for instance, be "shrunk" towards the average of all umpires' intercept vectors by an amount controlled by τ_0^2 and τ_{LO}^2. Further priors on the hyper-parameters σ_0^2 and σ_{LO}^2 introduce dependence between the components of Θ_0^u and Θ_{LO}^u as well, enabling us to "borrow strength" between the four combination of batter and pitcher handedness.

While Model 1 essentially estimates a separate called strike probability surface for each umpire, it entirely precludes the possibility of player or count effects. We now consider two successive expansions of Model 1. In Model 2, we incorporate both catcher and count effects that are assumed to be constant across umpires. That is, in Model 2

we assume that all of the the Θ^u_{CO}'s are all equal to some common value Θ_{CO} and all of the Θ^u_{CA}'s are equal to some common value Θ_{CA}. Similarly, in Model 3 we augment Model 2 with constant pitcher effects and constant batter effects. *A priori*, we model

$$\Theta_{CO}|\sigma^2_{CO} \sim N\left(0_{11}, \sigma^2_{CO} I_{11}\right)$$

and consider similar, zero-mean spherically-symmetric Gaussian priors for Θ_{CA}, Θ_P and Θ_B, while retaining the same prior specification on the Θ^u_0's and Θ^u_{LO}'s. Though they elaborate on Model 1, Models 2 and 3 still represent a vast simplification to the general model in Equation 1 as they assume that there is no umpire-to-umpire variability in the count or player effects. This leads us to consider Model 4, which builds on Model 2 by allowing umpire-specific count and catcher effects, and Model 5, which includes umpire-specific batter and pitcher effects and corresponds to the general model in Equation 1 We model

$$\Theta^{u_1}_{CO}, \ldots, \Theta^{u_{93}}_{CO}|\Theta_{CO}, \tau^2_{CO} \sim N\left(\Theta_{CO}, \tau^2_{CO} I_{11}\right)$$
$$\Theta^u_{CO}|\sigma^2_{CO} \sim N\left(0_{11}, \sigma^2_{CO} I_{11}\right)$$

and consider similarly structured prior hierarchies for $\Theta^u_{CA}, \Theta^u_B, \Theta^u_P$ in Models 4 and 5. Throughout, we place independent Inverse Gamma(3,3) hyper-priors on the top-level variance parameters σ^2_0, σ^2_{LO}, σ^2_{CO}, σ^2_{CA}, σ^2_P and σ^2_B.

It remains to specify the hyper-parameters τ^2_0, $\tau^2_{LO}, \tau^2_{CO}, \tau^2_{CA}, \tau^2_P$ and τ^2_B which capture the umpire-to-umpire variability in the intercept, location, count, and player effects. For simplicity, we fix these hyper-parameters to be equal to 0.25 in the appropriate models. To motivate this choice, consider how two umpires would call a pitch thrown at a location where the historical GAM forecasts a 50% called strike probability. According to Model 2, the difference in the two umpires' log-odds of a called strike follows a $N\left(0, 2(\tau^2_0 + \tau^2_{CO} + \tau^2_{CA})\right)$ distribution, *a priori*. Taking $\tau^2_0 = \tau^2_{CO} = \tau^2_{CA} = 0.25$ reflects a prior belief that there is less than a 10% chance that one umpire would call a strike 75% of the time while the other calls it a strike only 25% of the time. For simplicity, we take $\tau^2_{LO} = \tau^2_B = \tau^2_P = 0.25$ as well.

3 Model performance and comparison

3.1 Predictive performance

We fit each model in Stan (Carpenter et al. 2016) and ran two MCMC chains for each model. All computations were done in R (versions 3.3.2 and later) and the MCMC simulation was carried out in RStan (versions 2.14.1 and later) on a high-performance computing cluster. For each model, after burning-in the first 2000 iterations, the Gelman-Rubin $\hat{R}$ statistic for each parameter was <1.1, suggesting convergence. We continued to run the chains, after this burn-in, until each parameter's effective sample size exceeded 1000. For Models 1 and 2, we found that running the sampler for 4000 total iterations was sufficient while for Models 3, 4, and 5, we needed 6000 iterations. The run time of these samplers ranged from just under an hour (Model 1) to 50 h (Model 5).

Using the simulated posterior draws from each model, we can approximate the mean of the posterior predictive distribution of the called strike probability for each pitch in our 2014 dataset. Table 1 shows the misclassification and mean square error for Models 1–5 over all pitches from 2014 and over two separate regions, as well as the error for the historical GAM forecasts. Region 1 consists of all pitches thrown within 1.45 inches on either side of the boundary of the average rule book strike zone defined in Section 2.1. Since the radius of the ball is about 1.45 inches, the pitches in Region 1 are all "borderline" calls in the sense that only part of the ball passes through the strike zone but are, by rule, strikes. Region 2 consists of all pitches thrown between 1.45 and 2.9 inches outside the boundary of the average rule book strike zone. These

Table 1: In-sample predictive performance for several models.

		Model 1	Model2	Model 3	Model 4	Model 5	hGAM
	# Parameters	744	855	2582	11,067	171,168	–
Overall	MISS	0.103	0.100	0.099	0.096	**0.086**	0.105
	MSE	0.073	0.071	0.069	0.068	**0.061**	0.074
Region 1	MISS	0.248	0.236	0.232	0.225	**0.195**	0.258
	MSE	0.163	0.156	0.153	0.150	**0.133**	0.168
Region 2	MISS	0.214	0.209	0.205	0.203	**0.184**	0.215
	MSE	0.153	0.149	0.146	0.144	**0.129**	0.156

The smallest errors are in boldface.

pitches miss the strike zone by an amount between one and two ball's width, and ought to be called balls by the umpire. To compute misclassification error, we used 0.5 as the threshold for a strike.

We see that Models 1–5 outperform the historical GAMs overall and in both Regions 1 and 2. This is hardly surprising, given that the hGAMs were trained on data from 2011 to 2013 and the other models were trained on the 2014 data. Recall that Model 1 only accounted for pitch location. As we successively incorporating count and catcher (Model 2), and pitcher and batter (Model 3) effects, we find that the overall error drops. Finally, Model 5 has the best performance across the board. This is entirely expected as Model 5 given the tremendous number of parameters.

Of course, we would be remiss if we assessed predictive performance only with training data. Table 2 compares such out-of-sample predictive performance, by considering pitches from the 2015 season for which the associated batter, catcher, pitcher, and umpire all appeared in our 2014 dataset.

Now we see that Model 3 has the best out-of-sample performance overall and in Regions 1 and 2. The fact that Models 4 and 5 have worse out-of-sample performance, despite having very good in-sample performance is a clear indication that these two over-parametrized models have overfit the data. One could argue, however, that comparing predictive performance on 2015 data is not the best means of diagnosing overfitting. Roegelle (2014), Mills (2017a), and Mills (2017b) have documented year-to-year changes in umpires' strike zone enforcement ever since Major League Baseball began reviewing and grading umpires' decisions in 2009. In Appendix A, we report the results from a cross-validation study, in which we repeatedly re-fit Models 1–5 using 90% of the 2014 data and assessing performance on the remaining 10%, that similarly demonstrates Model 3's superiority.

Model 3's superiority over Models 1 and 5 reveals that although accounting for player effects can lead to improved predictions of called strike probabilities, we cannot reliably estimate an individual catchers catcher effects on individual umpires with a single season's worth of data.

3.2 Full posterior analysis

We now examine the posterior samples from Model 3 more carefully. Figure 5 shows box plots of the posterior distributions of catcher effects on the log-odds scale for the catchers with the top 10 posterior means, the bottom 10 posterior means, and the middle 10 posterior means.

Table 2: Out-of-sample predictive performance for several models.

		Model 1	Model2	Model 3	Model 4	Model 5	hGAM
	# Parameters	744	855	2582	11,067	171,168	–
Overall	MISS	0.107	0.105	**0.105**	0.106	0.106	0.109
	MSE	0.075	0.074	**0.074**	0.075	0.074	0.076
Region 1	MISS	0.256	0.245	**0.244**	0.248	0.246	0.267
	MSE	0.167	0.162	**0.161**	0.163	0.162	0.173
Region 2	MISS	0.236	0.232	**0.231**	0.233	0.234	0.237
	MSE	0.169	0.166	**0.165**	0.166	0.165	0.170

The smallest errors are boldfaced.

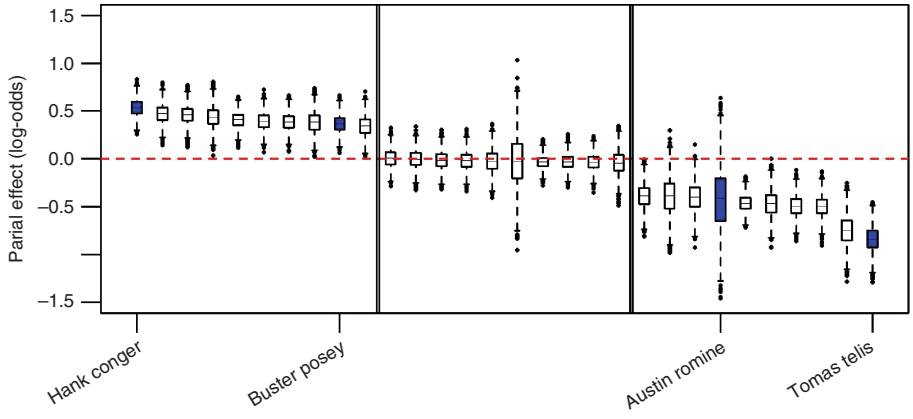

Figure 5: Comparative box plots of 30 catcher effects sorted by the posterior mean of their partial effect on the log-odds scale.

We see that there are some catchers, like Hank Conger and Buster Posey, whose posterior distributions are entirely supported on the positive axis, indicating that, all else being equal, umpires are more likely to call strikes when they are caught by these catchers as opposed to the baseline catcher, Brayan Pena. On the other extreme, there are some catchers like Tomas Tellis with distinctly negative effects. As we would expect, catchers who appeared very infrequently in our dataset have very wide posterior distributions. For instance, Austin Romine caught only 61 called pitches and his partial effect has the largest posterior variance among all catchers. It is interesting to see that all of the catcher effects, on the log-odds scale, are contained in the interval $[-1.5, 1.5]$, despite the prior placing nearly 20% of its probability outside this interval. The maximum difference on the log-odds scale between the partial effects of any two catchers is 3, with high posterior probability. For context, a change of 3 in log-odds corresponds to a change in probability from 18.24% to 81.76%. As it turns out, the posterior distribution of each count effect is also almost entirely supported in the interval $[-1.5, 1.5]$, on the log-odds scale. This would seem to suggest that catcher framing effects are comparable in magnitude to the effect of count. We explore this possibility in much greater detail in Appendix B.

Armed with our simulated posterior draws, we can create posterior predictive strike zones for a given batter-pitcher-catcher-umpire matchup. Suppose, for instance, that Madison Bumgarner is pitching to the batter Yasiel Puig, with Buster Posey catching. Figure 6 shows the 50% and 90% contours of the posterior predictive called strike probability for two umpires, Angel Hernandez and Mike DiMuro, and an average umpire in a 2–0 and 0–2 count. Note, if the center of the pitch passes within the region bounded by the dashed gray line in the figure, then some part of the ball passes through the average rule book strike zone, shown in gray. Puig is a right-handed batter, meaning that he stands on the left-hand side of the approximate rule book strike zone, from the umpire's perspective.

Across the board, Hernandez's contours enclose more area than the average umpire's contours and DiMuro's contours enclose less area than the average umpire's. For instance, on a 2–0 count, Hernandez's 50% contour covers 4.37 sq. ft., the average umpire's covers 3.87 sq. ft., and DiMuro's covers 3.53 sq. ft. The contours on a 0–2 pitch are much smaller, indicating that all else being equal, these umpires are less likely to call a strike on an 0–2 pitch than on a 2–0 pitch. Each of the 50% contours extend several inches to the left or *inside* edge of the approximate rule book strike zone. At the same time, the same contours do not extend nearly as far beyond the right or *outside* edge of the strike zone. This means that, Hernandez, DiMuro, and the average umpires are more likely to call strikes on pitches that miss the inside edge of the strike zone than they are on pitches that miss the outside edge by the same amount. Even the 90% contour on a 2–0 count extend a few inches past the inside edge of the strike zone, implying that Hernandez, DiMuro, and the average umpire will almost always call a strike that misses the inside edge of a the strike zone so long as it is not too high or low. Interestingly, we see that the leftmost extents of DiMuro's and the average umpire's 90% contours on a 0–2 pitch nearly align with the dashed boundary on the inside edge. A pitch thrown at this location will barely cross the average rule book strike zone, indicating that at least on the inside edge, DiMuro and umpires on average tend to follow the rule book prescription, calling strikes over 90% of the time.

The same is not true at the top, bottom, or outside edge. For instance, the rightmost extent of average

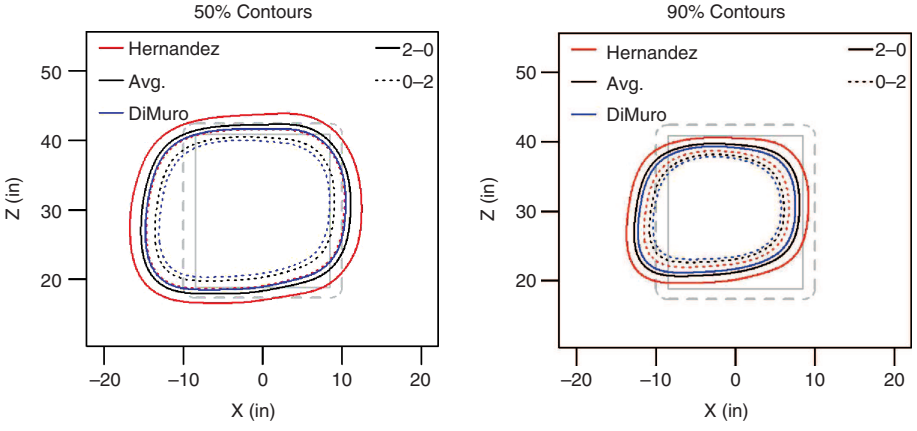

Figure 6: Fifty percent and 90% contours for called strike probability for the Bumgarner-Puig-Posey matchup for different umpires in different counts.

umpires' 90% contour on a 0–2 pitch lies several inches within the outside edge of the strike zone. So in the space of about four and a half inches, the average umpires' called strike probability drops dramatically from 90% to 50%, despite the fact that according to the rule book these pitches should be called a strike.

Figure 6 is largely consistent with the empirical observation that Hernandez tends to call a much more permissive strike zone than DiMuro: Hernandez called 42.67% of taken pitches strikes (1624 strikes to 2182 balls) and DiMuro called 39.92% of taken pitches strikes (1220 strikes to 1836 balls). On 2–0 pitches, Hernandez's strike rate increased to 51.44% (71 strikes to 67 balls) and DiMuro's increased to 48.31% (57 strikes to 67 balls).

4 Impact of framing

We now turn our attention now to measuring the impact framing has on the game. Formally, let S be a random variable counting the number of runs the pitching team gives up after the current pitch to the end of the half-inning. Using slightly different notation than that in Section 2.3, let **h** encode the handedness of the batter and pitcher and let **lo** be the estimated log-odds of a called strike from the appropriate historical GAM. Let **b**, **ca**, **co**, **p** and **u** denote the batter, catcher, count, pitcher, and umpire involved in the pitch. Finally, denote the baseline catcher Brayan Pena by ca_0. For compactness, let $\xi = (\mathbf{u}, \mathbf{co}, \mathbf{lo}, \mathbf{b}, \mathbf{p}, \mathbf{h})$ and observe that every pitch in our dataset can be identified by the combination $(\mathbf{ca}, \xi)$. For each catcher ca, let $\mathscr{P}_{ca}$ be the set of all called pitches caught by catcher ca:

$$\mathscr{P}_{ca} = \{(\mathbf{ca}, \xi) : \mathbf{ca} = ca\}\,.$$

Finally, let *TAKEN* be an indicator for the event that the current pitch was taken and let $CALL \in \{Ball, Strike\}$ be the umpire's ultimate call. We will be interested in the expected value of S, conditioned on $(\mathbf{ca}, \xi)$, the fact that pitch was taken, and the umpire's call. Assuming, conditioned on the count, the fact that the pitch was taken, and the call, S is independent of pitch location and participants, we have

$$E[S|\mathbf{ca},\xi, TAKEN] = \sum_{CALL} E[S|COUNT, TAKEN, CALL]\, P(CALL|\mathbf{ca}, \xi, TAKEN)$$

To determine the expected number of runs given up that can be attributed to a catcher ca, we may consider the counter-factual scenario in which the catcher is replaced by the baseline catcher, Brayan Pena, with all other factors remaining the same. In this scenario, the expected number of runs the fielding team gives up in the remaining of the half-inning is $E[S|\mathbf{ca} = ca_0, \xi, TAKEN, CALL]$. We may interpret the difference

$$E[S|\mathbf{ca} = ca, \xi, TAKEN, CALL] - E[S|\mathbf{ca} = ca_0, \xi, TAKEN, CALL]$$

as the average number of runs saved (i.e. negative of runs given up) by catcher ca's framing, relative to the baseline. A straightforward calculation shows that this difference is exactly equal to

$$f(ca, \xi) = (P(Strike|\mathbf{ca} = ca, \xi, TAKEN) - P(Strike|\mathbf{ca} = ca_0, \xi, TAKEN)) \times \rho(COUNT),$$

where

$$\rho(COUNT) = E[S|COUNT, TAKEN, Ball] - E[S|COUNT, TAKEN, Strike].$$

We can interpret the difference in called strike probabilities above as catcher ca's *framing effect*: it is precisely how much more the catcher adds to the umpires' called strike probability than the baseline catcher, over and above the other pitch participants, pitch location, and count. We can easily simulate approximate draws from the posterior distribution of this difference using the posterior samples from Model 3. We interpret ρ as the value of a called strike in a given count: it measures how many more runs a team is expected to give up if a taken pitch is called a ball as opposed to a strike.

To compute ρ, we begin by computing the difference in the average numbers of runs scored after a called ball and after a called strike in each count. For instance, 182,405 0–1 pitches were taken (140,667 balls, 41,738 called strikes) between 2011 and 2014. The fielding team gave up an average of 0.322 runs following a ball on a taken 0–1 pitch, while they only gave up an average of 0.265 runs following a called strike on a taken 0–1 pitch. So conditional on an 0–1 pitch being taken, a called strike saves the fielding team 0.057 runs, on average. Table 3 shows the number of runs scored after a called ball or a called strike for each count, as well as an estimate of ρ. Also shown is the relative proportion of each count among our dataset of taken pitches from 2011 to 2014. We see, for instance, that a called strike is most valuable on a 3–2 pitch but only 2.1% of the taken pitches in our dataset occurred in a 3–2 count. This calculation is very similar to the seminal run expectancy calculation of Lindsey (1963), though ours is based solely on count rather than on the number of outs and baserunner configuration. Albert (2010) also computes a count-based run expectancy, through his valuations are derived

using the linear weights formula of Thorn and Palmer (1985) rather than the simple average. See Albert (2015) for a more in-depth discussion of run expectancy.

The weighted average run value of a called strike based on Table 3 is 0.11 runs, slightly smaller than the value of 0.14 used by Judge et al. (2015) and much smaller than the 0.161 figure used by Turkenkopf (2008). The discrepancy stems from the fact that we estimated the run values based only on taken pitches while most other valuations of strikes include swinging strikes and strikes called off of foul balls. It is worth stressing at this point that in our subsequent calculations of framing impact we use the count-based run valuation as opposed to the weighted average value.

With our posterior samples and estimates of ρ in hand, we can simulate draws from the posterior distribution of $f(ca, \xi)$ for each pitch in our dataset. An intuitive measure of the impact of catcher ca's framing, which we denote RS for "runs saved" is

$$\mathrm{RS}(ca) = \sum_{(\mathbf{ca},\xi)\in\mathscr{P}_{ca}} f(ca, \xi).$$

The calculation of RS is very similar to the one used by Judge et al. (2015) to estimate the impact framing has on the game. Rather than using fixed baseline catcher, Judge et al. (2015) reports the difference in expected runs saved relative to a hypothetical average catcher. According to their model, Brayan Pena, our baseline catcher, was no different than this average catcher, so our estimates of RS may be compared to the results of Judge et al. (2015). Table 4 shows the top and bottom 10 catchers, along with the number of pitches in our dataset received by the catchers, and the posterior mean, standard deviation, and 95% credible interval of their RS values. Also shown are Judge et al. (2015)'s estimates of runs saved for the catchers, as well as the number of pitches used in their analysis.

According to our model, there is little posterior uncertainty that the framing effects of the top 10 catchers shown in Table 4 had a positive impact for their teams, relative to the baseline catcher. Similarly, with the exception of Welington Castillo and Chris Iannetta, we are rather certain that the bottom 10 catchers' framing had an overall negative impact, relative to the baseline. We estimate that Miguel Montero's framing saved his team 25.71 runs on average, relative to the baseline. That is, had he been replaced by the baseline catcher on each of the 8086 called pitches he received, his team would have given up an additional 25.71 runs, on average. Unsurprisingly, our estimates of framing impact differ from those of Judge et al. (2015)'s model. This is largely due to differences in the model construction, valuation of a called strike, and collection of pitches analyzed. Indeed, in some cases, (e.g. Montero and Rene Rivera), they used more pitches to arrive at their estimates of runs saved while in others, we used more pitches (e.g. Mike Zunino and Jonathan Lucroy). Nevertheless, our estimate are not wholly incompatible with theirs; the correlation between our estimates and theirs is 0.94. Moreover, if we re-scale their estimates to the same number of pitches we consider, we find overwhelmingly that these re-scaled estimates fall within our 95% posterior credible intervals.

4.1 Catcher aggregate framing effect

Looking at Table 4, it is tempting to say that Miguel Montero is the best framer. After all, he is estimated to have saved the most expected runs relative to the baseline catcher. We observe, however, that Montero received 8086 called pitches while Conger received only 4743. How much of the difference in the estimated number of runs saved is

Table 3: Empirical estimates of run expectancy and run value, with standard errors in parentheses.

Count	Ball	Strike	Value of called strike ρ	Proportion (%)
0–0	0.367 (0.002)	0.305 (0.002)	0.062 (0.002)	36.2
0–1	0.322 (0.002)	0.265 (0.004)	0.057 (0.004)	12.5
0–2	0.276 (0.003)	0.178 (0.007)	0.098 (0.008)	5.5
1–0	0.427 (0.003)	0.324 (0.003)	0.103 (0.005)	11.5
1–1	0.364 (0.003)	0.280 (0.004)	0.084 (0.005)	8.8
1–2	0.302 (0.003)	0.162 (0.006)	0.140 (0.006)	6.9
2–0	0.571 (0.007)	0.370 (0.006)	0.201 (0.009)	3.9
2–1	0.468 (0.005)	0.309 (0.006)	0.159 (0.008)	4.0
2–2	0.383 (0.004)	0.165 (0.006)	0.218 (0.007)	4.8
3–0	0.786 (0.013)	0.481 (0.008)	0.305 (0.015)	1.9
3–1	0.730 (0.010)	0.403 (0.009)	0.327 (0.014)	1.8
3–2	0.706 (0.008)	0.166 (0.008)	0.540 (0.011)	2.1

Table 4: Top and Bottom 10 catchers according to the posterior mean of RS.

Rank	Catcher	Runs saved (SD)	95% interval	*N*	BP
1	Miguel Montero	25.71 (5.03)	[15.61, 35.09]	8086	11.2 (8172)
2	Mike Zunino	22.72 (5.17)	[12.56, 32.31]	7615	20.4 (7457)
3	Jonathan Lucroy	19.56 (5.69)	[8.16, 30.49]	8398	16.4 (8241)
4	Hank Conger	19.34 (3.24)	[12.93, 25.65]	4743	23.8 (4768)
5	Rene Rivera	18.81 (3.69)	[11.63, 25.89]	5091	22.5 (5182)
6	Buster Posey	17.01 (4.14)	[8.79, 25.01]	6385	23.6 (6190)
7	Russell Martin	14.35 (4.41)	[5.85, 22.77]	6388	14.9 (6502)
8	Brian McCann	14.01 (3.95)	[6.18, 21.66]	6335	9.7 (6471)
9	Yasmani Grandal	12.88 (2.98)	[7.18, 18.69]	4248	14.5 (4363)
10	Jason Castro	12.61 (4.43)	[3.80, 21.08]	7065	11.5 (7261)
92	Josmil Pinto	−6.49 (1.41)	[−9.32, −3.76]	1748	−6.9 (1721)
93	Welington Castillo	−6.70 (4.28)	[−15.19, 1.78]	6667	−15.6 (6661)
94	Chris Iannetta	−7.50 (4.46)	[−16.18, 1.08]	6493	−7.3 (6527)
95	John Jaso	−7.76 (2.41)	[−12.50, −3.07]	3172	−11.3 (2879)
96	Anthony Recker	−8.37 (2.33)	[−13.29, −3.93]	2935	−13 (3102)
97	Gerald Laird	−8.68 (1.87)	[−12.29, −4.99]	2378	−9.6 (2616)
98	A. J. Ellis	−12.90 (3.79)	[−20.10, −5.38]	5476	−12.3 (5345)
99	Kurt Suzuki	−17.67 (4.25)	[−26.07, −9.35]	6811	−19.5 (7110)
100	Dioner Navarro	−18.81 (4.68)	[−28.00, −9.40]	6659	−19.8 (6877)
101	Jarrod Saltalamacchia	−23.98 (4.35)	[−32.76, −15.87]	6498	−34 (6764)

The column BP contain Judge et al. (2015)'s estimates that appeared on the Baseball Prospectus website.

due to their framing ability and how much to the disparity in the called pitches they received?

A naive solution is to re-scale the RS estimates and compare the average number of runs saved on a per-pitch basis. While this accounts for the differences in number of pitches received, it does not address the fact that Montero appeared with different players than Conger and that the spatial distribution of pitches he received is not identical to that of Conger. In other words, even if we convert the results of Table 4 to a per-pitch basis, the results would still be confounded by pitch location, count, and pitch participants.

To overcome this dependence, we propose to *integrate* $f(ca, \xi)$ over all ξ rather than summing $f(ca, \xi)$ over $\mathscr{P}_{ca}$. Such a calculation is similar to the spatially aggregate fielding evaluation (SAFE) of Jensen, Shirley, and Wyner (2009). They integrated the average number of runs saved by a player successfully fielding a ball put in play against the estimated density of location and velocity of these balls to derive an overall fielding metric un-confounded by dispraise in players' fielding opportunities. We propose to integrate $f(ca, \xi)$ against the empirical distribution of ξ and define catcher ca's "Catcher Aggregate Framing Effect" or CAFE to be

$$\text{CAFE}(ca) = 4000 \times \frac{1}{N} \sum_{\xi} f(ca, \xi), \tag{2}$$

The sum in Equation 2 may be viewed as the number of expected runs catcher ca saves relative to the baseline if he participated in every pitch in our dataset. We then re-scale this quantity to reflect the impact of his framing on 4000 "average" pitches. We opted to re-scale by CAFE by 4000 as the average number of called pitches received by catchers who appeared in more than 25 games was just over 3992. Of course, we could have easily re-scaled by a different amount.

Once again, we can use our simulated posterior samples of the Θ^u's to simulate draws from the posterior distribution of CAFE. Table 5 shows the top and bottom 10 catchers ranked according to the posterior mean of their CAFE value, along with the posterior standard deviation, and 95% credible interval for their CAFE value. Also shown is the a 95% interval of each catchers marginal rank according to CAFE.

We see that several of the catchers from Table 4 also appear in Table 5. The new additions to the top ten, Christian Vazquez, Martin Maldonado, Chris Stewart, and Francisco Cervelli were ranked 13th, 17th, 18th and 19th according to the RS metric. The fact that they rose so much in the rankings when we integrated over all ξ indicates that their original rankings were driven primarily by the fact that they all received considerably fewer pitches in the 2014 season than the top 10 catchers in Table 4. In particular, Vazquez received 3198 called pitches, Cervelli received 2424, Stewart received 2370, and Maldonado received only 1861.

Interestingly, we see that now Hank Conger ranks ahead of Miguel Montero according to the posterior mean

Table 5: Top and Bottom 10 catchers according to the posterior mean of CAFE.

Rank	Catcher	Mean (SD)	95% interval	95% rank interval
1	Hank Conger	16.20 (2.72)	[10.84, 21.50]	[1, 11]
2	Christian Vazquez	14.33 (2.94)	[8.26, 20.03]	[1, 19]
3	Rene Rivera	14.04 (2.76)	[8.75, 19.31]	[1, 18]
4	Martin Maldonado	13.24 (3.33)	[6.73, 19.68]	[1, 24]
5	Miguel Montero	12.36 (2.42)	[7.50, 16.90]	[2, 22]
6	Yasmani Grandal	11.90 (2.76)	[6.56, 17.29]	[2, 27]
7	Mike Zunino	11.78 (2.69)	[6.51, 16.74]	[2, 26]
8	Chris Stewart	11.63 (3.28)	[5.21, 18.03]	[1, 30]
9	Buster Posey	11.16 (2.73)	[5.74, 16.51]	[2, 30]
10	Francisco Cervelli	10.45 (3.21)	[4.06, 16.72]	[2, 36]
92	Jordan Pacheco	−11.73 (3.80)	[−19.26, −4.30]	[68, 98]
93	Koyie Hill	−11.79 (5.67)	[−22.48, −0.68]	[53, 100]
94	Josh Phegley	−12.05 (4.66)	[−21.40, −3.20]	[64, 99]
95	Austin Romine	−12.76 (9.78)	[−32.14, 5.81]	[30, 101]
96	Jarrod Saltalamacchia	−14.00 (2.53)	[−19.11, −9.26]	[82, 99]
97	Brett Hayes	−14.04 (4.06)	[−21.51, −5.93]	[73, 100]
98	Gerald Laird	−14.96 (3.21)	[−21.17, −8.69]	[81, 99]
99	Josmil Pinto	−15.04 (3.27)	[−21.57, −8.78]	[82, 100]
100	Carlos Santana	−22.48 (4.63)	[−31.63, −13.26]	[93, 101]
101	Tomas Telis	−25.06 (3.85)	[−32.41, −17.27]	[98, 101]

CAFE, indicating that the relative rankings in Table 4 was driven at least partially by disparities in the pitches the two received than by differences in their framing effects. Though Conger emerges as a slightly better framer than Montero in terms of CAFE, the difference between the two is small, as evidenced by the considerable overlap in their 95% posterior credible intervals.

We find that in 95% of the posterior samples, Conger had anywhere between the largest and 11th largest CAFE. In contrast, we see that in 95% of our posterior samples, Tomas Tellis's CAFE was among the bottom 3 CAFE values. Interestingly, we find much wider credible intervals for the marginal ranks among the bottom 10 catchers. Some catchers like Koyie Hill and Austin Romine appeared very infrequently in our dataset. To wit, Hill received only 409 called pitches and Romine received only 61. As we might expect, there is considerable uncertainty in our estimate about their framing impact, as indicated by the rather wide credible intervals of their marginal rank.

4.2 Year-to-year reliability of CAFE

We now consider how consistent CAFE is over multiple seasons. We re-fit our model using data from the 2012 to 2015 seasons. For each season, we restrict attention to those pitches within one foot of the approximate rule book strike zone from that season. We also use the log-odds from the GAM models trained on all previous seasons so that the model fit to the 2012 data uses GAM forecasts trained only on data from 2011 while the model fit to the 2015 data uses GAM forecasts trained only on data from 2011 to 2014. When computing the values of CAFE, we use the run values given in Table 3 for each season. There were a total of 56 catchers who appeared in all four of these seasons. Table 6 shows the correlation between their CAFE values over time.

In light of the non-stationarity in strike zone enforcement across seasons, it is encouraging to find moderate to high correlation between a player's CAFE in one season and the next. In terms of year-to-year reliability, the autocorrelations of 0.5–0.7 place CAFE on par with slugging percentage for batters. Interestingly, the correlations between 2012 CAFE and 2013 CAFE and the correlation between 2013 CAFE and 2014 CAFE are >0.7, but the correlation between 2014 CAFE and 2015 CAFE is somewhat lower, 0.58. While this could just be an artifact of noise, we do note that there was a marked uptick in awareness of framing between the 2014 and 2015 seasons, especially among fans and in the popular press. One possible reason

Table 6: Correlation of CAFE across multiple seasons.

	2012	2013	2014	2015
2012	1.00	0.70	0.56	0.41
2013	0.70	1.00	0.71	0.61
2014	0.56	0.71	1.00	0.58
2015	0.41	0.61	0.58	1.00

for the drop in correlation might be umpires responding to certain catcher's reputations as elite pitch framers by calling stricter strike zones, a possibility suggested by Sullivan (2016).

5 Discussion

We systematically fit models of increasing complexity to estimate the effect a catcher has on an umpire's likelihood of calling a strike over and above factors like the count, pitch location, and other pitch participants. We found evidence that some catchers do exert a substantially positive or negative effect on the umpires but that the magnitude of these effects are about as large as the count effects. Using the model that best balanced fit and generalization, we were able to simulate draws from the posterior predictive distribution of the called strike probability probability of each taken pitch in 2014. For each pitch, we estimated the apparent framing effect of the catcher involved and, following a procedure similar to that of Judge et al. (2015), we derived an estimate of the impact framing has on the game, RS. Our RS metric is largely consistent with previously reported estimates of the impact of catcher framing, but a distinct advantage is our natural quantification of the estimation uncertainty. We find that there is considerable posterior uncertainty in this metric, making it difficult to estimate precisely the impact a particular catcher's framing had on his team's success.

While the construction of RS is intuitive, we argue that it does not facilitate reasonable comparisons of catchers' framing since, by construction, the metric is confounded by the other factors in our model. We propose a new metric CAFE that integrates out the dependence of RS on factors like pitch location, count, and other pitch participants. CAFE compares catchers by computing the impact each catcher's framing would have had had he received every pitch in our dataset. Like RS, there is a considerable uncertainty in our CAFE estimates. While we are able to separate the posterior distributions of CAFE of good framers from bad framers, there is considerable overlap in the posterior distributions of CAFE within these groups, making it difficult to distinguish between the good framers or between the bad framers. Despite this, we find rather high year-to-year correlation in CAFE, though there is a marked drop-off between 2014 and 2015. This coincides with the increased attention on framing in the sports media and sabermetrics community following the 2014 season. One potential explanation for this drop-off is that umpires adjusted their strike zone enforcement when calling pitches caught by catchers with reputations as good framers.

Our findings may have several implications for Major League Baseball teams. The uncertainty in both RS and CAFE make it difficult to precisely value pitch framing with any reasonable degree of certainty. For instance, the 95% credible interval of Jonathan Lucroy's RS is [8.16, 30.49]. Using the heuristic of 10 expected runs per win and \$7 M per win (Cameron 2014; Pollis 2013), our model suggests that Lucroy's framing was worth anywhere between \$5.7 M and \$21.34 M. In light of the non-stationarity between seasons and the recent drop-off in correlation in CAFE, it is difficult to forecast the impact that any individual catcher's framing will have into the future. The observed overlaps in the posterior distribution of CAFE means that with a single season's worth of data, we cannot discriminate between good framers with the same certainty that we can separate good framers from bad framers. As a concrete example, our model indicates that both Miguel Montero and Hank Conger were certainly better framers than Jarrod Saltalamacchia, but it cannot tell which of Montero or Conger had a larger, positive impact.

5.1 Extensions

There are several extensions of and improvements to our model that we now discuss. While we have not done so here, one may derive analogous estimates of RS and CAFE for batters and pitchers in a straightforward manner. Our model only considered the count into which a pitch was thrown but there is much more contextual information that we could have included. For instance, Rosales and Spratt (2015) have suggested the distance between where a catcher actually receives the pitch and where he sets up his glove before the pitch is thrown could influence an umpire's ball-strike decision making. Such glove tracking data is proprietary but should it be become publicly available, one could include this distance along with its interaction with the catcher indicator into our model. In addition, one could extend our model to include additional game-state information such as the ball park, the number of outs in the half-inning, the configuration of the base-runners, whether or not the home team is batting, and the number of pitches thrown so far in the at-bat. One may argue that umpires tend to call more strikes late in games which are virtually decided (e.g. when the home team leads by 10 runs in the top of the ninth inning) and easily include measures related to the run-differential and time remaining into our model. Expanding our model in these directions may improve the overall predictive performance slightly without dramatically increasing the computational overhead.

More substantively, we have treated the umpires' calls as independent events throughout this paper. Chen et al. (2016) reported a negative correlation in consecutive calls, after adjusting for location. To account for this negative correlation in consecutive calls, we could augment our model with binary predictors encoding the results of the umpires' previous k calls in the same at-bat, inning, or game. Incorporating this Markov structure to our model would almost certainly improve the overall estimation of called strike probability and may produce slightly smaller estimates of RS and CAFE. At this point, however, it is not *a priori* obvious how large the differences would be or how best to pick k. It is also well-known that pitchers try to throw to different locations based on the count, but we make no attempt to model or exploit this phenomenon. Understanding the effect of pitch sequencing on umpires' decision making (and vice-versa) would also be an interesting line of future research.

We incorporated pitch location in a two-step procedure: we started from an already quite good generalized additive model trained with historical data and used the forecasted log-odds of a called strike as a predictor in our logistic regression model. Much more elegant would have been to fit a single semi-parametric model by placing, say, a common Gaussian process prior on the umpire-specific functions of pitch location, $f^u(x, z)$ in Equation 1. We have also not investigated any potential interactions between pitch location, player, and count effects. While we could certainly add interaction terms to the logistic models considered above, doing so vastly increases the number of parameters and may require more thoughtful prior regularization. A more elegant alternative would be to fit a Bayesian "sum-of-trees" model using Chipman, George, and McCulloch (2010)'s BART procedure. Such a model would likely result in more accurate called strike probabilities as it naturally incorporates interaction structure. We suspect that this approach might reveal certain locations and counts in which framing is most manifest.

Finally, we return to the two pitches from the 2015 American League Wild Card game in Figure 1. Fitting our model to the 2015 data, we find that Eric Cooper was indeed much more likely to call the Keuchel pitch a strike than the Tanaka pitch (81.72% vs 62.59%). Interestingly, the forecasts from the hGAMs underpinning our model were 51.31% and 50.29%, respectively. Looking a bit further, had both catchers been replaced by the baseline catcher, our model estimates a called strike probability of 77.58% for the Keuchel pitch and 61.29% for the Tanaka pitch, indicating that Astros' catcher Jason Castro's apparent framing effect (4.14%) was slightly larger than Yankee's catcher Brian McCann's (1.30%). The rather large discrepancy between the apparent framing effects and the estimated called strike probabilities reveals that we cannot immediately attribute the difference in calls on these pitches solely to differences in the framing abilities of the catchers. Indeed, we note that the two pitches were thrown in different counts: Keuchel's pitch was thrown in a 1–0 count and Tanaka's was thrown in a 1–1 count. In 2015, umpires were much more likely to call strikes in a 1–0 count than they were in a 1–1 count, all else being equal. Interestingly, had the Keuchel and Tanaka pitches been thrown in the same count, our model still estimates that Cooper would be consistently more likely to call the Keuchel pitch a strike, lending some credence to disappointed Yankees' fans' claims that his strike zone enforcement favored the Astros. Ultimately, though, it is not so clear that the differences in calls on the two pitches shown in Figure 1 specifically was driven by catcher framing as much as it was driven by random chance.

Appendix

A Model comparison with cross-validation

As mentioned in Section 3.1, Roegelle (2014), Mills (2017a), and Mills (2017b) have documented year-to-year changes in umpires' strike zone enforcement ever since Major League Baseball began reviewing and grading umpires' decisions in 2009. In other words, umpire tendencies are non-stationary across seasons and we cannot reasonably expect Models 4 and 5, which attempt to identify umpire-specific player effects, to forecast future umpire decisions particularly well. A potentially more appropriate way to diagnose overfitting issues would be to hold out a random subset of our 2014 data, say 10%, fit each model on the remaining 90% of the data, and assess the predictive performance on the held-out 10%. Table 7 shows the average misclassification rate and mean square error for Models 1–5 over 10 such holdout sets. The results in the table confirm our finding that Model 3 represents the best balance between model expressivity and predictive capability.

B Catcher and count effects

In Section 3.2, we reported that the posterior distributions of catcher and count effects on the log-odds scale were largely supported in the interval $[-1.5, 1.5]$. This would indicate that a catcher's framing effect is of roughly similar magnitude as the effect of count.

Table 7: Hold-out misclassification rate (MISS) and mean square error (MSE) for several models).

	# Parameters	Model 1 744	Model 2 855	Model 3 2582	Model 4 11,067	Model 5 171,168
Overall	MISS	0.104	0.101	**0.101**	0.107	0.101
	MSE	0.073	0.071	**0.071**	0.072	0.072
Region 1	MISS	0.251	**0.239**	0.240	0.242	0.240
	MSE	0.164	0.158	**0.157**	0.159	0.158
Region 2	MISS	0.213	0.208	**0.206**	0.208	0.207
	MSE	0.154	0.150	**0.148**	0.150	0.149

The smallest errors are boldfaced.

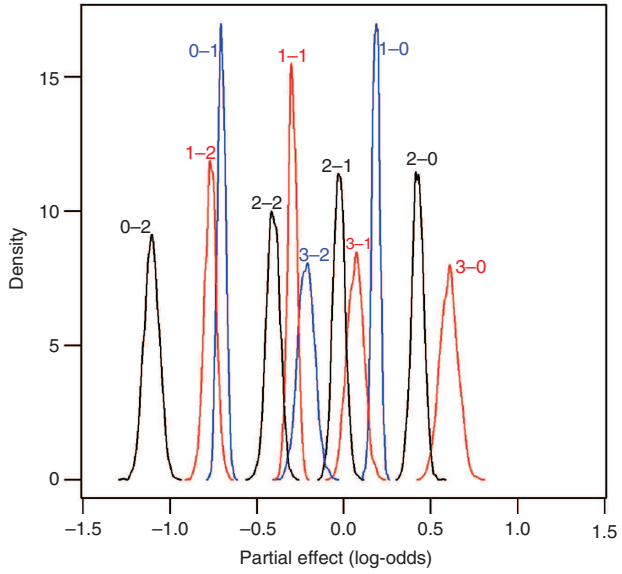

Figure 7: Posterior densities of the partial effect of count. Densities computed using a standard kernel estimator.

Figure 7 shows the approximate posterior densities of the count effects. Recall that these are the partial effects relative to the baseline count of 0–0. As we might expect, umpires are more likely to call strikes in 3–0 and 2–0 counts than in 0–0 counts and much less likely to call strikes in 0–2 and 1–2 counts, all else being equal. Somewhat interestingly, we find that umpires are slightly less likely to call strikes in a 3–1 count than they are in a 1–0 count.

We also see that the posterior distribution for the effects of a 3–0 counts are considerably wider than those for a 0–1 count, indicating that we are much more uncertain about the effect of the former two counts than the latter two. This is due the rather large disparity in the numbers of pitches taken in these counts: in our dataset, there were more than five times called pitches thrown into a 0–1 count than into a 3–0 count (37,513 versus 6162).

To compare the relative magnitudes of catcher and count effects on the probability scale, we return to the hypothetical matchup between batter Yasiel Puig, catcher Buster Posey, and pitcher Madison Bumgarner. Suppose that Bumgarner's pitch is thrown in a location where the hGAM called strike probability forecast is exactly 50%. According to our model, if this pitch were hypothetically caught by the baseline catcher, Brayan Pena, the forecasted called strike probability averaged over all 93 umpires is 54%, with the difference of 4% attributable to intercept, batter, and pitcher effects. In contrast, if the same pitch had been caught by Buster Posey, our model estimates the called strike probability to be 64%, indicating that on this pitch, Posey added an additional 10% to the forecasted called strike probability. If Posey had caught the same pitch but the count were 2–2 instead of 0–0, the forecast would be 55%. In this way, at least for this pitch, the effect of swapping the baseline catcher with Posey on an 0–0 pitch is the about the same as changing the count from 0–0 to 2–2 with Posey catching.

Table 8 elaborates on this example and shows the estimated average called strike probability for the same pitch as a function of count and catcher. That is, we forecasted

Table 8: Difference in forecasted called strike probabilities averaged over all umpires when Bumgarner pitches to Puig, relative to the baseline called strike probability of 54%, for various combinations of catcher and count.

	Hank Conger	Buster Posey	Brayan Pena	Tomas Telis
0–2	−0.14	−0.18	−0.26	−0.39
1–2	−0.06	−0.10	−0.18	−0.35
0–1	−0.04	−0.08	−0.17	−0.34
2–2	0.03	−0.01	−0.10	−0.28
1–1	0.06	0.02	−0.07	−0.26
3–2	0.07	0.0 4	−0.05	−0.25
2–1	0.12	0.08	−0.01	−0.20
0–0	0.12	0.09	0.00	−0.20
3–1	0.14	0.10	0.02	−0.18
1–0	0.16	0.13	0.04	−0.16
2–0	0.21	0.18	0.10	−0.10
3–0	0.24	0.21	0.14	−0.06

Counts are ordered as in Figure 7.

the called strike probability, averaged across umpires, on a pitch thrown by Bumgarner to Puig at a location where the hGAM forecast was 50% for many combinations of catcher and count. To highlight the relative size of the catcher and count effects, we have subtracted a baseline 54%, the called strike probability when the catcher is Pena and the count is 0–0, from all of these probabilities.

Recall that the baseline called strike probability is 54% on such a pitch. According to our model, the effect of changing the count from 0–0 to 0–2 when Posey is receiving the pitch is about the same as changing the count from 0–0 to 1–2 with the baseline catcher receiving the pitch. We note that the called strike probability forecasts for Tomas Tellis are much lower than for Hank Conger, Posey, and the baseline catcher. For instance, our model estimates that umpires on average would call this pitch a strike only 15% of the time if it were thrown in a 0–2 count and Tellis was receiving, in contrast to 40% for Conger, 36% for Posey, and 28% for Pena.

References

Albert, J. 2010. "Using the Count to Measure Pitching Performance." *Journal of Quantitative Analysis in Sports* 6. https://doi.org/10.2202/1559-0410.1279.

Albert, J. 2015. "Beyond Run Expectancy." *Journal of Sports Analytics* 1:3–18.

Cameron, D. 2008. "Win Values Explained: Part Five." http://www.fangraphs.com/blogs/win-values-explained-part-five/. Accessed on August 2, 2016.

Cameron, D. 2014. "The Cost of a Win in the 2014 Off-Season." http://www.fangraphs.com/blogs/the-cost-of-a-win-in-the-2014-off-season/. Accessed on August 2, 2016.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. 2016. "Stan: A Probabilitistic Programming Language." *Journal of Statistical Software* 20:1–37.

Chen, D., T. J. Moskowitz, and K. Shue. 2016. "Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *Quarterly Journal of Economics* 131:1181–1241.

Chipman, H. A., E. I. George, and R. E. McCulloch. 2010. "Bart: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4:266–298.

Drellich, E. 2014. "What Conger Trade Means for Castro and Astros Rotation." http://blog.chron.com/ultimateastros/2014/11/05/astros-make-trade-with-angels-for-catcher-hank-conger/. Accessed on May 27, 2016.

Fast, M. 2010."What the Heck is PITCHf/x?." In: *The Hardball Times Annual, 2010*, Chicago, IL: ACTA Publications, 153–158.

Holt, R. 2014. "How Important is Pitch Framing?" http://www.beyondtheboxscore.com/2014/12/12/7375383/how-important-is-pitch-framing. Accessed on April 15, 2016.

Jensen, S. T., K. E. Shirley, and A. J. Wyner. 2009. "Bayesball: A Bayesian Hierarchical Model For Evaluating Fielding in Major League Baseball." *Annals of Applied Statistics* 3:491–520.

Judge, J., H. Pavlidis, and D. Brooks. 2015. "Moving Beyond WOWY: A Mixed Approach to Measuring Catcher Framing." http://www.baseballprospectus.com/article.php?articleid$=$25514. Accessed on February 8, 2015.

Kim, J. W. and B. G. King. 2014. "Seeing Stars: Matthew Effects and Status Bias in Major League Baseball Umpiring." *Management Science* 60:2619–2644.

Lindbergh, B. 2013. "The Art of Pitch Framing." http://grantland.com/features/studying-art-pitch-framing-catchers-such-francisco-cervelli-chris-stewart-jose-molina-others/. Accessed on February 8, 2015.

Lindsey, G. R. 1963. "An Investigation of Strategies in Baseball." *Operations Research* 11:477–501.

Marchi, M. 2011. "Evaluating Catchers: Quantifying the Framing Pitches Skill." http://www.hardballtimes.com/evaluating-catchers-quantifying-the-framing-pitches-skill/. Accessed on April 23, 2016.

Mills, B. M. 2014. "Social Pressure at the Plate: Inequality Aversion, Status, and Mere Exposure." *Managerial and Decision Economics* 35:387–403.

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

Mills, B. M. 2017b. "Technological Innovations in Monitoring and Evaluation: Evidence of Performance Impacts Among Major League Baseball Umpires." *Labour Economics* 46:189–199.

Parsons, C. A., J. Sulaeman, M. C. Yates, and D. S. Hamermesh. 2011. "Strike Three: Discrimination, Incentives, and Evaluation." *American Economic Review* 101:1410–1435.

Pavlidis, H. 2014. "You Got Framed." http://espn.go.com/espn/feature/story/_/id/11127248/how-catcher-framing-becoming-great-skill-smart-teams-new-york-yankees-espn-magazine. Accessed on August 15, 2015.

Pollis, L. 2013. "How Much Does a Win Really Cost?" http://www.beyondtheboxscore.com/2013/10/15/4818740/how-much-does-a-win-really-cost.

Roegelle, J. 2014. "The Strike Zone Expansion Is Out of Control." http://www.hardballtimes.com/the-strike-zone-expansion-is-out-of-control/.

Rosales, J. and S. Spratt. 2015. "Who Is Responsible For a Called Strike?" in *Sloan Sports Analytics Conference 2015*. Boston, MA.

Sanchez, R. 2015. "Jonathan Lucroy Needs a Raise." http://espn.go.com/mlb/story/_/id/12492794/jonathan-lucroy-needs-raise. Accessed on January 29, 2016.

Sidhu, G. and B. Caffo. 2014. "MONEYBaRL: Exploiting Pitcher Decision-Making Using Reinforcement Learning." *Annals of Applied Statistics* 8:926–955.

Sullivan, J. 2015. "How the Astros Wound Up With a Bigger Zone." http://www.fangraphs.com/blogs/how-the-astros-wound-up-with-a-bigger-zone/. Accessed on October 10, 2015.

Sullivan, J. 2016. "The Beginning of the End for Pitch-Framing?" http://www.fangraphs.com/blogs/the-beginning-of-the-end-for-pitch-framing/. Accessed on September 9, 2016.

Tainsky, S., B. M. Mills, and J. A. Winfree. 2015. "Further Examination of potential discrimination among MLB Umpires." *Journal of Sports Economics* 16:353–374.

Thorn, J. and P. Palmer. 1985. *The hidden game of baseball*. New York, NY: Doubleday.

Turkenkopf, D. 2008. "Framing the Debate." http://www.beyondtheboxscore.com/2008/4/5/389840/framing-the-debate. Accessed on April 23, 2016.

Walsh, J. 2007. "Strike Zone: Fact vs. Fiction." http://www.hardballtimes.com/strike-zone-fact-vs-fiction/. Accessed on April 23, 2016.

Woodrum, B. 2014. "The State and Future of Pitch-Framing Research." http://www.hardballtimes.com/the-state-and-future-of-pitch-framing-research/. Accessed on February 10, 2015.