DRAFT

Talkin' Baseball: Peer Effects in Expert Decision Making

Jed M. S. Armstrong[†]

September 27, 2022

Abstract

This paper uses data on Major League Baseball (MLB) umpire crews to estimate the effect of an expert's peer network on decision-making. I quantify decision-making by using the quality of strike / ball calls made by home-plate umpires. I then empirically analyze the extent to which an umpire paired to work with high-quality peers experiences an improvement in decision-quality. MLB umpire data has two advantages that make this analysis possible: 1) umpires are assigned to a crew each season / game in a way that is independent of decision-making quality, and have significant churn within these crews; and 2) decisions can be uniquely attributed to a single umpire, and are able to be quantified in a detailed way. Using these data, I show that a one-standard deviation improvement in the decision-making quality of an umpire's peer network raises their own-call quality by around 0.03-0.1 standard deviations. Experience is a key driver of this effect, with more-experienced umpires having smaller peer effects than less-experienced umpires. I also show that there is some persistence to these peer effects. I consider expert decisions in a Bayesian framework, and show that decision-making spill-over can be characterized in terms of two types of accuracy: bias and precision. I then demonstrate that both types of accuracy transmit through peer networks.

[†]New York University: jedmsarmstrong@nyu.edu.

1 Introduction

Economists have a considerable interest in determining how experts make decisions. Given the role that expert decisions have on a number of spheres of life, including politics, finance, and public safety, understanding the determinants of expert decisions has extremely broad implications. A number of frameworks for expert decision-making have been proposed, incorporating both rational (Baysian) and behavioral components. In this paper, I study one particular influence channel on expert decision-making: the decisions made by the expert's peer network.

Peer spillovers are a popular mechanism for study by economists. There is a strong literature focused on identifying and estimating the effect of an individual's peer network on productivity and ability. Quantifying the productivity spill-overs from a group to an individual (and *vice versa*) has implications for the optimal allocation of people to teams, such as workers across and within firms, and students across and within schools (Jackson & Bruegmann, 2009). There is also a literature focused on measuring the direct effect that peers have on decision-making. Most of the research in this field has looked at non-expert decision-making, instead quantifying peer effects on decisions such as consumers' decisions to purchase new goods (Bailey, Johnston, Kuchler, Stroebel, & Wong, 2022). In this paper, I investigate how peer networks can affect even highly-trained experts, whose decisions are intensely scrutinized, and who are strongly motivated to make good decisions.

Empirical estimates of the peer effect on decision-making are limited in the literature. To a large extent, this is due to data requirements. Quantifying peer effects on decision-making requires considerable amounts of data, in terms of both depth and breadth.

From the perspective of data depth, decisions need to be both quantifiable and observed at a highly granular level (ideally at an individual level). Quantifying decisions can be very difficult. In many cases decisions are made privately, and only the outcomes of those decisions are observable. These observed outcomes are certainly determined by expert decisions, but are typically also impacted by external factors, such as constraints or stochasticity. For instance, the performance of an expert stock broker is determined by their investing decisions, but also by a range of macroeconomic factors that are outside their control. Extracting underlying decisions from outcomes is a difficult and noisy process.

Similarly, trying to assign individual decisions to individual experts can be extremely difficult. Decisions are often made in by a group of experts working together, and so it difficult to allocate outcomes to an individual and often requires extremely detailed data. Many previous papers in this literature have have taken one of two approaches. The first approach is to use fixed-effect models to extract individual elements from firm data, which is a noisy and difficult process. The second approach is to focus on specific types of clearly observable non-expert decisions, such as which college major to declare. Neither of these approaches provides a clean method for identifying expert decision-making.

From the perspective of data breadth, individuals must be observed across a range of different groups within a network, in order to identify peer effects from individual data. This requires a long time-series, as well as rich data on a very broad range of groups. Most prior studies have tended to look at particular institutional settings (such as schools or residence halls) where individuals and their interactions can be tracked with some degree of precision, or have relied on extremely large administrative datasets covering all workers in an economy. One potential flaw of these empirical settings is that it can be difficult to fully isolate peer effects due to the influence of outsiders who don't appear in the data (for example a student's peers who attend a different school).

In this paper, I make use of an empirical setting that satisfies both depth and breadth requirements on data: the 'crews' of umpires in charge of Major League Baseball (MLB) games. MLB umpire crews are typically made up of four umpires, each of which is assigned to work one of the four base each game. The umpire assigned to the home plate for a game is tasked with making calls on whether a non-swung pitch is a strike or a ball, and is the *only* one of the umpires responsible for making these calls in a particular game. Since 2008, technological advances in ball-tracking software mean that these decisions are able to be precisely quantified, meaning that the home plate umpire's calls can be determined to be incorrect calls (true strikes called as balls or *vice versa*) or correct calls. Given that the umpire's main duty is to officiate the game fairly,[1] making correct decisions is a first-order priority for umpires.

I use ball-tracking software to quantify the decision-quality of all MLB umpires, and then assess the impact of an umpires crew on their quality. I find that working with a good crew raises call quality, and *vice versa*, indicating that there are sizable peer effects in decision-making among experts. Given

[1] The MLB's Official Baseball Rules state that the umpires' "first requisite is to get decisions correct. Umpire dignity is important but never as important as being right." (Office of the Commissioner of Baseball, 2019)

that MLB umpires are selected and trained to be the best umpires, it is perhaps surprising that they are influenced by peer networks. I outline some potential channels for this peer influence, including a learning channel and an effort channel drawing on the literature on rational inattention.

The rich data allow me to quantify the peer effects on expert decision making in extremely detailed ways. For example, I'm able to use historical crew data to examine persistence in peer effects. I show that an umpires last-season crew has the largest peer effect, with previous-season crews having smaller effects (but still positive for up to two seasons). This type of analysis helps to provide insights into how dynamic peer effects can manifest (Johnson & Jackson, 2019). I am also able to examine how experience plays a role in the peer effects on expert decision-making. Using the number of seasons each umpire has worked in the MLB as a measure of experience, I show that less-experienced experts have much stronger peer effects than more-experienced experts.

I then combine my empirical results on peer effects with theoretical models of rational decision making. Experts, including MLB umpires, have been modeled as Bayesian decision makers (Green & Daniels, 2018). As such, it is possible to decompose the bias and imprecision in their decisions. I construct an 'optimal strike zone' for each game, which captures a measure of what the umpire was making their decisions based on. I use this optimal strike zone to extract bias and imprecision scores for each umpire, and show that there is tentative evidence that both aspects to decision-making are impacted by peer networks.

Related literature

Building on insights from prior work, this paper aims to make contributions to three broad strands of the literature.

The first literature I contribute attempts to quantify peer and network effects on decision-making, productivity, and work performance. Spill-over effects of peer quality and behavior have been acknowledged and studied by economic researchers for many decades. Early studies identified links between school peer characteristics and grade point average (GPA) (Betts & Morell, 1999), criminal behavior and drug use (Case & Katz, 1991), and life decisions in college (Sacerdote, 2001), among others. More recently, the rise of large administrative datasets has allowed for studies into labor market and net-

work effects. Due to the data burden, these studies have tended to focus on subsets of the population for which detailed data are available and group movements are common. For instance, Jackson and Bruegmann (2009) look at the effect of a teacher's peer network on value added.

The second literature I contribute to is focused on using sports data to analyze labor market outcomes and phenomenon. Sports provides a useful setting for considering labor market outcomes, because the rich and often publicly-available dataset allow for useful and replicable empirical analyses Kahn (2000). The paper most closely linked to mine is Guryan, Kroft, and Notowidigdo (2009), which combines sports data and peer networks, using random player matching in golf tournaments to identify peer effects on performance. They find little peer effects.

Baseball is a particularly popular sport to use for empirical analysis, due to the large amounts of data available, and the structured nature of the game. One of the first papers to use MLB data to answer empirical labor market questions is Parsons, Sulaeman, Yates, and Hamermesh (2011), which uses pitch and umpire data to quantify racial discrimination, looking at the effect of the umpire's and player's races. Further work has used MLB as a testing ground to investigate the effect of technology and monitoring tools on labor performance (Mills, 2017), the effect of temperature and weather on productivity Fesselmeyer (2019), and the role of on-the-job training in improving work standards (Mills, 2014).

The third literature I contribute to is on quantifying expert's decision-making processes in terms of Bayesian updating. There is a considerable literature showing that the behaviour of many decision-makers can be reasonably modeled using a Bayesian approach (for surveys, see Baley and Veldkamp (2021), DeGroot (2005), andParmigiani (2001)). There is also a broad literature finding *deviations* from a rational Bayesian decision-making framework (Kahneman and Tversky (1977) and Tversky and Kahneman (1974) provide foundational examples of these biases, and see Camerer, Loewenstein, and Rabin (2004) for a more-recent survey of the research on biases and departures from rationality in decision-making). In a baseball setting, Green and Daniels (2018) show that MLB umpiring decisions can be interpreted using a Bayesian framework, which I build on by examining how peer effects manifest through bias and precision.

Layout

The rest of this paper is organized as follows. Section 2 provides an overview of the setting for this paper, discussing MLB umpiring and crew assignment. Section 3 discusses the data used, and section 5 outlines the empirical strategy to estimate the peer effects. Section 4 discusses the identification of these effects, and how plausible it is to interpret the empirical effects causally. Section 6 presents empirical results. Section 7 discusses a decomposition of decision-making into *bias* and *precision* elements, and section 8 discusses some possible mechanisms for the effects, as well as outlining some evidence for these mechanisms. Finally, section 9 concludes.

2 Background on MLB umpiring

Umpires play a crucial role in any professional sport, ensuring that rules are adhered to and a fair playing field is maintained. Typically, umpires must exercise judgment and make decisions about whether a particular rule is violated, as in many situations plays happen quickly and decision margins are very small. In baseball games (and in particular games in the MLB), umpires make a variety of these judgment calls, including whether a base runner is 'safe' or out, or whether a batted ball is a foul or fair.

2.1 Strike and Ball calls

One particularly evident judgment call that MLB umpires make is whether a non-swung pitch is a 'strike' or a 'ball'. To make this decision, umpires consider a 'strike zone' which is a three-dimensional area whose top and bottom face is defined by the home plate and whose height is defined by the batters torso. A pitch that is not swung at by the batter[2] is judged by the home-plate umpire (who is positioned immediately behind the catcher, and hence behind home plate) to be a 'strike' if it passes through the strike zone, and a 'ball' if it does not. Appendix A provides more information on the strike zone and how it is used to determine strikes and balls. Accumulating strikes is a key way that the fielding team gets a batter out, accounting for almost a quarter of outs, and hence correct calls

[2]If the batter swings at a pitch and does not hit the ball, the result is *always* a strike, irrespective of the position of the ball.

by the umpire are crucial for a fair contest between the teams.

2.2 Who calls the pitch: Assignment of umpires to pitches

The home-plate umpire plays a key role in baseball officiating, and the assignment of a particular umpire to calling a given pitch is central to identifying peer effects in this paper. In this section I outline the process by which a given umpire is responsible for making the strike or ball call on a particular pitch. In general, this assignment process can be thought of as being the result of three sub-processes, which are discussed in further detail:

1. Assignment of umpires to crews, either at the start of the season or during the season due to injury, illness, umpire vacation, etc.
2. Assignment of crews to games.
3. Assignment of a particular umpire within the crew to be the home plate umpire for a particular call.

2.2.1 Assignment of umpires to crews

At the start of each season, MLB umpires are allocated to crews of (typically) four.[3] In my sample, the 2008-2013 seasons each had 17 umpire crews allocated, and the 2014-2019 seasons had 19 crews allocated, with the extra crews allowing for more vacation time and for crews to spend time in MLB's new video replay center. This process starts by selecting a a 'Crew Chief'. The Crew Chiefs for a given season are known ahead of time, and are typically the most senior umpires. The remaining pool of umpires are assigned to crews by the MLB Umpiring Department, following conditions outlined in the collective bargaining agreement between the MLB and the MLB and the umpire's union.

The MLB doesn't publish definitive guidelines for how this assignment is made. However, communication with MLB officials makes it clear that past performance is not taken into account when assigning umpires to crews[4], and hence it appears that the assignment is *as-good-as random* with respect to

[3] All information on crews is obtained from http://www.stevetheump.com/Proumpires.htm.

[4] For instance, officials note that assignment is made to "construct crews that will work well as a cohesive unit", and that experience is considered.

call quality. The randomness of allocation with respect to call quality can be statistically tested, and is done so in section 4.

In theory, the crew remain together for the season.[5]. However, in practice, there are a number of 'shuffles' throughout the season. These are due to exogenous factors such as injury and illness, as well as pre-announced factors such as vacations.[6] When an umpire is away from a crew for a period of time, they are either replaced by another MLB umpire who comes across from another crew, or by one of several Class-AAA (Minor League Baseball) umpires. The replacement decision is based entirely on scheduling and location issues, and is hence exogenous to call quality.

2.2.2 Assignment of crews to games

Assignment of umpires to games happens at a crew level (rather than at an umpire-specific level). Formally, the MLB regular season is divided into a number of 'series' between teams (typically comprising three or four games in a single city), and an umpire crew assigned to a officiate all of the games in a particular series. MLB umpires don't have a 'home base' where they do most of their umpiring – instead they are often on the road for significant periods of time (up to three weeks), officiating games for many teams in many locations across the country. Assignment of umpire crews to series is a particularly complicated combinatorial problem, and is studied in great detail by mathematicians and operations research analysts as the Traveling Umpire Problem.

The process is typically solved algorithmically. The particular assignment process used by the MLB is modified from that outlined in Trick, Yildiz, and Yunes (2012). The assignment of crews to games must satisfy a number of constraints (such as that each umpire crew sees each team at home and on the road, and travels to each of MLB's 27 cities at least once). The algorithmic approach means that selection of crews to series and games can be thought of as random with respect to any salient variables in this analysis.

[5] In my analysis, I consider only the regular season, in which each team plays 162 games in a round-robin format to determine seedings for the post-season and World Series. Umpire allocations for the post-season are made independently of regular season crews, and typically are used to reward high-performing and long-tenured umpires. I also omit All-Star Games, for similar reasons.

[6] The MLB Umpire's CBA entitles umpires to 4 weeks vacation time during the regular season.

2.2.3 Assignment of umpire within crew to call a particular pitch

Each game, the four members of the assigned crew are each assigned to a base.[7] Umpires on first, second, and third base are primarily responsible for determining whether runners to their base are safe or out.[8] The umpire assigned to home plate is responsible for making all strike and ball calls. The assignment of umpires to bases is exogenous and follows a prescribed pattern – in the first game of the season the Crew Chief is assigned to home plate. Each subsequent game, the umpires in the crew rotate around the diamond clockwise, so that, for example the home plate umpire rotates to third base in the next game. In the dataset I use, 94.5 percent of games followed this rule, with the home-plate umpire having been the first-base umpire in the previous game the crew officiated. The results are robust to removing the 5.5 percent of games where this was not the case.

3 Data

There are two key types of data required for this analysis: data on pitches to determine whether calls are correct or incorrect, and data on umpire crews and assignments to identify peer networks.

3.1 PITCHf/x

Starting in the 2008 season, MLB partnered with PITCHf/x to collect information on the speeds and trajectories of pitched baseballs. The technology uses multiple mounted cameras in ballparks[9] to track a baseball in flight, and calculate its co-ordinates in 3-dimensional space. PITCHf/x is used by sabermetricians and other analysts to analyze pitch patterns and player behavior, and provides a useful cross-check against the strike and ball calls made by umpires.

In particular, PITCHf/x data are collected for the entire flight of the baseball, even as the baseball travels past the batter. PITCHf/x can thus provide data on the x and y position of the baseball in the moment it passes the batter, allowing for *post-hoc* analysis of whether a pitch was in or out of the strike zone. PITCHf/x also take account of the height, proportions, and stance of the batter to

[7] In cases where there are only three umpires allocated to a game, or when one of crew leaves the game, second base is left un-manned.

[8] Other responsibilities include determining whether a hit ball is foul or fair, and ruling on checked swings.

[9] PITCHf/x is not available for many games played outside of usual ballparks, such as international fixtures.

generate a 'true strike zone' which takes into account that the vertical position and size of the strike zone depends on the batter. These PITCHf/x data can be used to determine whether a non-swung pitch was true strike or a true ball, which can be cross-checked against the umpire's call to create a measure of decision-making quality made for the home-plate umpire.[10]

Figure 1 shows an example of the PITCHf/x data used in this analysis, for the first US-based game of the 2019 MLB season.[11] I use data on every pitch thrown in a MLB game from the 2008 season to the 2019 season.[12] In total, there were 8,685,757 pitches thrown over the 12 seasons in the data.

The main PITCHf/x variables of interest are $pitch_x$ and $pitch_y$ which give the x and y co-ordinates of the ball as it passes the strike zone, $strike\_zone_y$ which provides the height and position of the true strike zone, and the outcome of the pitch: for example a called strike, a called ball, a swung strike, or a hit. I keep only those pitches where the home-plate umpire was required to make a judgment call (i.e. called strikes and called balls)[13]. Using the PITCHf/x position data, I construct an indicator for whether the ball was in the strike zone as it passed the batter (a 'true strike'), or outside the strike zone (a 'true ball'). I then produce an variable q for each pitch p designating the quality of the call (i.e. a correct call or on incorrect call).

$$q_p = \begin{cases} \text{Correct} & \text{(called strike AND true strike) OR (called ball AND true ball)} \\ \text{Incorrect} & \text{otherwise} \end{cases}$$

3.2 Umpire Crews

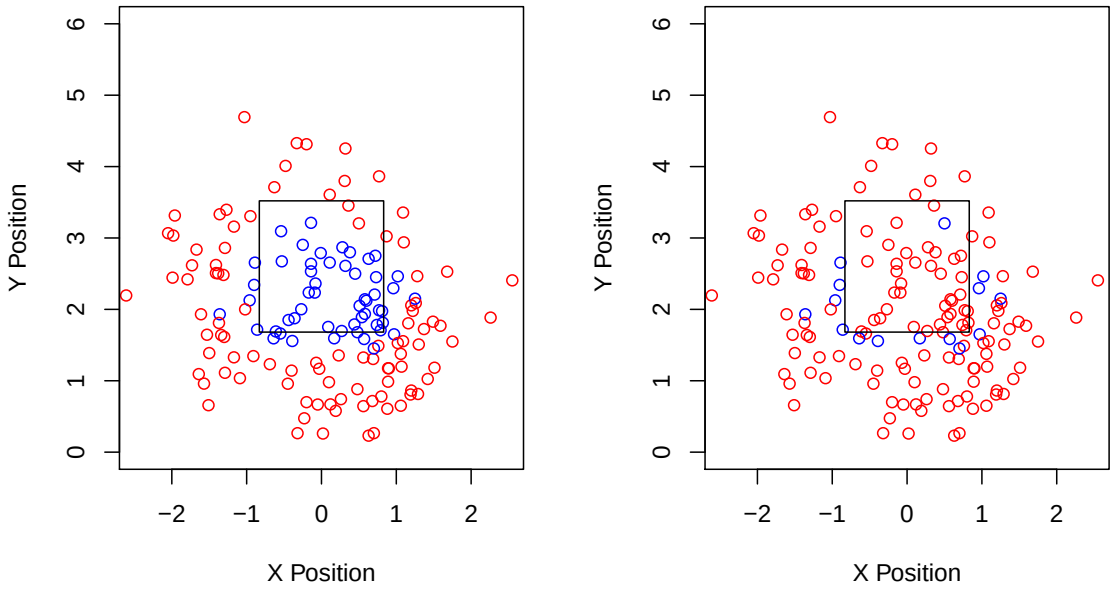
As well as pitch-level ball-tracking information, I use a number of game-level characteristics. Most important is the crew of umpires who worked each game, and their base assignment. This is used to determine who the home-plate umpire was for a particular game, which allows me to attribute game calls to a particular umpire. I also obtain information such as the player lists, ballpark, attendance,

[10] According to the official rules of baseball, the umpire's strike / ball call is final, and can not be challenged, disputed, or overturned by either team. Thus, PITCHf/x data are only used in *post-hoc* analysis, not during the game itself.

[11] There were two earlier games in the 2019 season, between the Seattle Mariners and the Oakland Athletics on March 20 and March 21. These games were both played at the Tokyo Dome, in Japan, and hence PITCHf/x data are not available for these games.

[12] The 2020 and 2021 MLB season were affected by the Covid-19 pandemic, which introduced a number of irregularities.

[13] Around 1 percent of balls are classified as 'intentional balls' in PITCHf/x. These refer to situations in which the pitcher throws the pitch far outside the strike zone in order to intentionally walk the batter. I drop these pitches as well, as they are unlikely to be true judgment calls by the umpire.



(a) Game Calls

(b) Correct and Incorrect calls

Figure 1: Called pitches for a game between the Milwaukee Brewers and the St. Louis Cardinals on Marc 28, 2019.

Both figures show all called pitches (i.e. pitches that the home-plate umpire was required to make a judgment call on whether it was a strike or a ball) for the game. The black rectangle is the normalized strike zone, which takes into account the height and stance of the batter. All Y-positions are scaled to this normalized strike zone. In the left panel, pitches in blue were called strikes, while pitches in red were called balls. In the right panel, pitches in blue are incorrect calls, while pitches in red are correct calls. There were 149 called pitches in this game, of which 134 were correct. Hence, the home-plate umpire for this game was assigned a score of $q_{i,cgt} = 134/149 = 0.899$.

and game time. All game-level data for this analysis are sourced from RetroSheet.[14]

3.3 Analysis variables

I combine pitch and umpire data to construct a measure of call-quality for each umpire i and game g (in season t and when working with crew c) by the proportion of correct calls made by an umpire in that particular game:

$$q_{i,cgt} = \frac{1}{N_p(g)} \sum_{q_p} \mathbb{1}_{q_p=\text{Correct}}$$

where $N_p(g)$ is the number of pitches in game g for which the umpire was required to make a strike / ball call. I also aggregate this to a season average for each umpire to construct a season-level quality measure.

$$Q_{i,\overrightarrow{c}\,t} = \frac{1}{N_g(s)} \sum_{g} q_{i,cgt}$$

where $\overrightarrow{c}$ is the vector of crews that umpire i worked with, and $N_g(s)$ is the number of games in season s for which umpire i was the home-plate umpire.

My main measure of network call-quality involves taking all games in season s for which umpire i was in the crew but *not* the home-plate umpire, and taking the average game call-quality over those games. This works out to be a weighted average of call-quality scores for the network of umpires that i worked with, based only on games in which i was present, and weighted by the number of games for which i was in the crew. I typically focus on last season's $(t-1)$ network quality as a driver for own quality in the current season.

$$\overline{Q}_{-i,\overrightarrow{c}\,t-1} = \frac{1}{N_g(s)} \sum_{g} q_{-i,cgt-1}$$

[14] https://www.retrosheet.org/gamelogs

4 Identification

If assignment to crews were fully random, the effect of an umpire's network on decision-making could be estimated using OLS. In this section, I show statistically that there is no selection on umpire quality on assignment, and hence argue that crews are *as-good-as-randomly* assigned.

The main approach to demonstrating randomness in peer matching in the literature typically involves regressing the own-characteristic on the average peer characteristic. In the case of umpire networks, the regression would be of the following form.

$$Q_{i,ct} = \pi_0 + \pi_1 \overline{Q}_{-i,ct} + \delta_t + \varepsilon_{ict} \tag{1}$$

In this regression, we would expect $\pi_1 = 0$ in cases where there is no selection to groups on quality. Guryan et al. (2009) show that the coefficient π_1 is biased, due the fact that the peer characteristic $\overline{Q}_{-i,ct}$ necessarily excludes the own characteristic. To correct for this, they propose the following regression model

$$Q_{i,ct} = \pi_0 + \pi_1 \overline{Q}_{-i,ct} + \varphi\pi_1 \overline{Q}_{-i,t} + \delta_t + \varepsilon_{ict} \tag{2}$$

where $\overline{Q}_{-i,t}$ is the average characteristic of *all* other individuals in time t, not just those in the peer-network of i. The test of randomness associated with Guryan et al. (2009) is still that $\pi_1 = 0$.

Table 1 shows estimated coefficients for the above regression models. In each case, I run the regressions using the crew to which umpire i was assigned at the start of the season, as well as all crews with which umpire i worked during the season. The estimated coefficients of interest are all statistically indistinguishable from zero at conventional levels of significance. Thus, I conclude that assignment to crews is as good as random with respect to quality, and hence each new crew assignment can be thought of as a random draw in terms of network quality.

Table 1

	Dependent variable:			
	Call Quality			
	(1)	(2)	(3)	(4)
Crew	All		Assigned	
Past Crew Qual.	-0.00468	-0.000789	0.05330	0.00068
	(0.01377)	(0.000589)	(0.04373)	(0.00198)
Leave-out Qual.		-88.19***		-85.83***
		(0.0301)		(0.176)
Controls	Season	Season	Season	Season
Observations	843	843	1,047	1,047
R^2	0.771	1.000	0.756	1.000
Adjusted R^2	0.768	1.000	0.753	1.000
Note:				*p<0.1; **p<0.05; ***p<0.01

5 Empirical strategy

There are two key channels of interest when considering peer effects on performance: 1) the effect of a high- or low-performance individual on a given group; and 2) the effect of a high- or low-performance group on an individual. The type of channel identifiable in a particular setting usually depends on the type of team allocation that is seen – when groups remain together with few switchers entering it is possible to identify the individual-to-group effect, while when groups are largely broken up and re-allocated the group-to-individual effect is more clear. In the setting of MLB umpires, most umpires are allocated to an entirely new crew each season, meaning that the effect of crews on individual umpires is more evident.

The first way of identifying a peer effect is to regress an umpire's season-level call quality on the average call quality of their peer network last season (as defined in section 3). I include time fixed effects to account for drift in call quality. In the following regression, the coefficient of interest λ provides an indication of how umpire call quality depends on past network quality.

$$Q_{i,ct} = \alpha + \lambda \overline{Q}_{-i,c(t-1)} + \delta_t + \varepsilon_{i,ct} \tag{3}$$

A more comprehensive model for capturing the peer effect takes into account own quality and peer quality. The main empirical specification of the paper is the following regression, which is common in

the peer-effects literature (e.g. Jackson and Bruegmann (2009) and Guryan et al. (2009)). It regresses game-level call-quality of umpire i on a quality measure for umpire i (proxied by last season's call quality) and the network quality measure (last season's umpire network call quality). I include time fixed effects to capture the seasonal effect described above.

$$q_{i,cgt} = \alpha + \varphi \overline{Q}_{i,c(t-1)} + \beta \overline{Q}_{-i,c(t-1)} + \delta_t + \varepsilon_{i,cgt} \tag{4}$$

The parameter of interest is β. Given the as-good-as-random assignment of umpires to crews and pitches, β can be consistently estimated from 4 using OLS.

6 Results

As an indication of the results, figure 2 shows the raw correlation between the call quality of an umpire in o given season, and the call quality of thei umpire's previous network. Although some of this is likely to be mechanical if there is a persistent trend in call quality (such as umpires generally becoming more accurate in their calls over time), it is instructive of the type of relationship between peer decision-making and individual decision-making that this paper seeks to uncover.

The upward sloping association in figure illustrates that individuals with a higher-quality past network are associated with higher performance (that is, lower mistake making). The remainder of the results in this section explore this association in further detail.

Table 2 presents some of the results from these regressions. The coefficient on past network quality is positive and typically statistically significant. The effect of increasing the average quality of an umpire's peer network in the previous season raises own call quality by between 3 and 10.7 percentage points.

6.1 Role of experience

One further consideration is the role that umpire experience plays in network effects. There is considerable heterogeneity in the level of Major League umpiring experience among the umpires in the

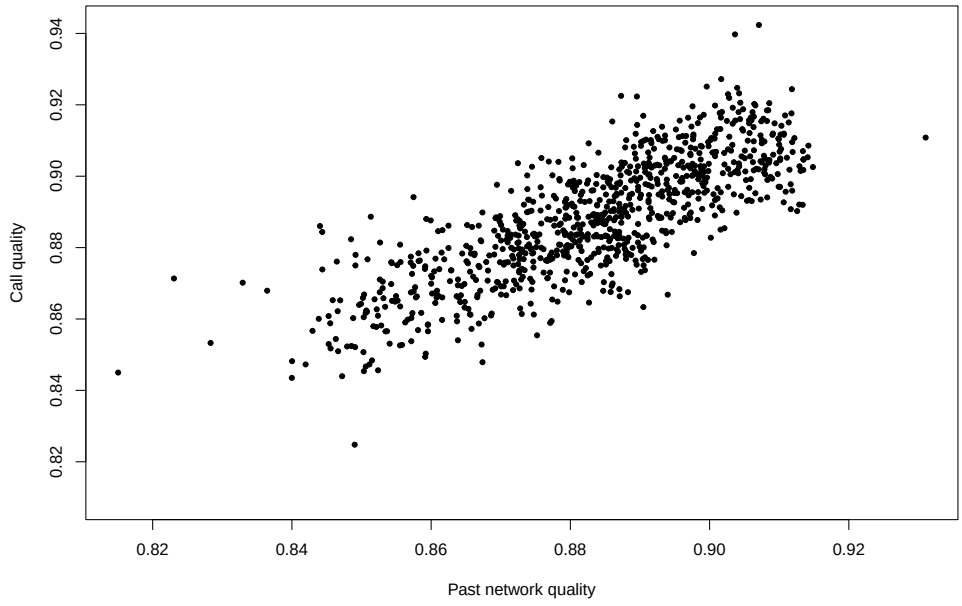


Figure 2: Raw correlation between Past Crew Error and Current Error Rate at a season level

Note: Each point is a umpire-season. I was only able to plot umpire-seasons for seasons in which that umpire also umpired the previous season.

Table 2

	Dependent variable:		
	Call Quality: Season Level		
	(1)	(2)	(3))
Past Crew Quality	0.1588***	0.09572**	0.08101*
	(0.0531)	(0.04686)	(0.04185)
Past Self Quality		0.4999***	
		(0.02732)	
Controls	Season	Season	Season, Ump.
Observations	911	909	911
R^2	0.708	0.788	0.869
Adjusted R^2	0.704	0.785	0.847

sample. In order to test how spillovers depend on experience, augment regression 4 to include interaction terms on network quality for experience (defined as years since an umpire made their MLB umpiring debut).

$$q_{i,gct} = \pi_0 + \pi_1 \overline{Q}_{-i,c(t-1)} + \varphi_1 X_{i,t} + \varphi_2 \overline{Q}_{-i,c(t-1)} \times EXP_{i,t} + \pi_2 \overline{Q}_{i,c(t-1)} + \delta_t + \varepsilon_{ict} \tag{5}$$

In this regression, the hypothesized direction of the experience-peer-effect interaction term φ_2 is *a priori* ambiguous. It could be that more-experienced umpires learn more from their peers, in which case we would expect $\varphi_2 > 0$. Alternatively, it could be that more-experienced umpires are more 'set in their ways' (or, equivalently, less-experienced umpires are more open to peer suggestions), in which case we would expect $\varphi_2 < 0$.

The estimated coefficients of regression 6.1 are shown in table 3. I find that experience increases call quality, with each extra year of MLB umpiring experience leading to a 1.5 percent increase in average call quality. Moreover, there are strong interaction effects between experience and network quality. Taking account of experience, an increase in network quality raises average umpire call quality by 8.6 percentage points. This effect is smaller for more-experience umpires – each additional year of experience reduces the peer effect by around 0.3 percentage points.

6.2 Persistence of results

An important question in the peer-effects literature is the persistence of peer effects. In my main specification, I identify peer effects with some persistence – my measure of network quality is *last-season's* quality. I re-run the main specification, but using the crew-quality from two to five seasons ago as my explanatory variable, rather than one season ago. The results are summarized in figure 3. I find that the

Table 3

	Dependent variable:			
	Call Quality: Game Level			
	(1)	(2)	(3)	(4)
Past Crew Quality	0.1051***		0.09496**	0.09238**
	(0.0354)		(0.0414)	(0.0459)
Exp.		0.01505***	0.00208*	0.01843***
		(0.00123)	(0.00114)	(0.00195)
Exp × Past Crew Qual.			-0.00259**	-0.00282*
			(0.00128)	(0.00167)
Past Self Quality			0.4628***	
			(0.02105)	
Controls	Season	Season, Ump.	Season	Season, Ump.
Observations	25,978	29,091	25,976	25,978
R^2	0.179	0.268	0.203	0.224
Adjusted R^2	0.179	0.264	0.202	0.220

Table 4

	Dependent variable:			
	Call Quality: Game Level (Close Calls)			
	(1)	(2)	(3)	(4)
Past Crew Quality	0.168**		0.299***	0.248**
	(0.086)		(0.101)	(0.113)
Exp.		0.022***	0.011***	0.033***
		(0.003)	(0.003)	(0.005)
Exp × Past Crew Qual.			−0.013***	−0.011***
			(0.003)	(0.004)
Past Self Quality			0.709***	
			(0.051)	
Controls	Season	Season, Ump.	Season	Season, Ump.
Observations	25,978	29,091	25,976	25,978
R^2	0.074	0.127	0.085	0.098
Adjusted R^2	0.073	0.122	0.084	0.093

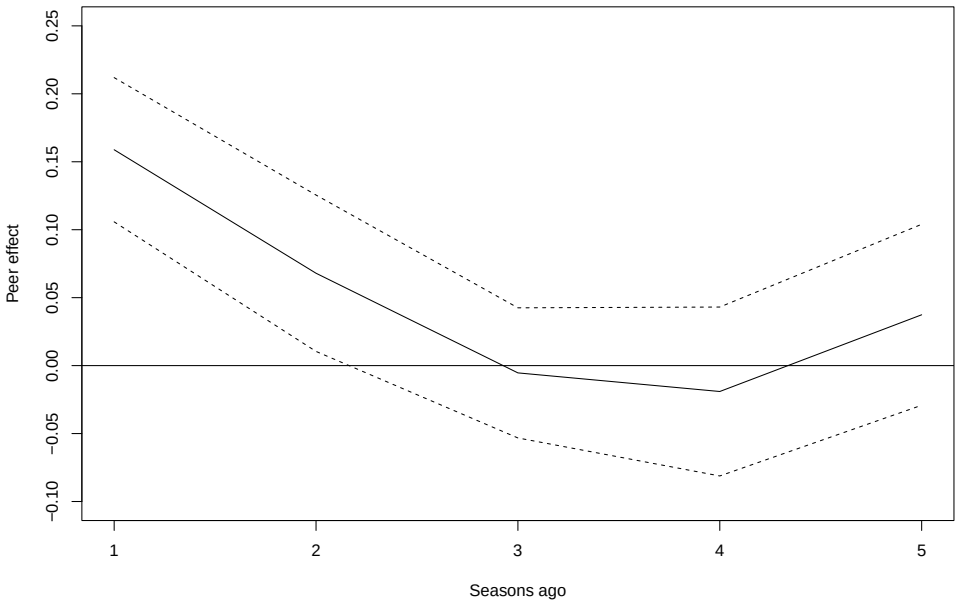


Figure 3: Peer-effect coefficient over time

Note: Each coefficient for x seasons ago is calculated from a regression of the form 4, with x season-lagged crew quality as the explanatory variable: $\overline{Q}_{-i,c(t-x)}$. The dashed lines show ± 1 s.d. error bars.

6.3 Robustness

There are a number of checks and falsification exercises that I run to ensure the robustness of the results.

6.3.1 Falsification exercises

One possible concern with the findings is that spurious results may arise in the absence of peer learning if umpires experience season-to-season mean-reversion in quality, and umpires are assigned to crews based on post performance. In such a setting, it would be possible for an relationship between past network quality and current quality to arise spuriously: an umpire who has a bad year will be matched with umpires who had good years, and if there is mean reversion, we would expect that in the following season the umpire who did poorly will improve, and hence we will see a (spurious) relationship between good umpires in the past and good umpires today.

Partly this is dealt with by demonstrating that there is no quality-based assignment, as shown in

section 4. I can also test for this spurious effect using a falsification exercise in which the call-quality of an umpire is regressed on the average call-quality of umpires they worked with in the *following* season (instead of the *previous* season). If there is season-to-season mean reversion and quality-based crew assignment, then an umpire who does well one year will tend to be assigned to a crew with umpires who did poorly. If there is mean reversion, then we'd expect those who did poorly to improve, and so there will be a positive relationship between doing well one year and having a better crew next year. If the result is driven by peer learning instead, then we'd expect there to be no relationship – an umpire can not learn from people they have yet to work with.

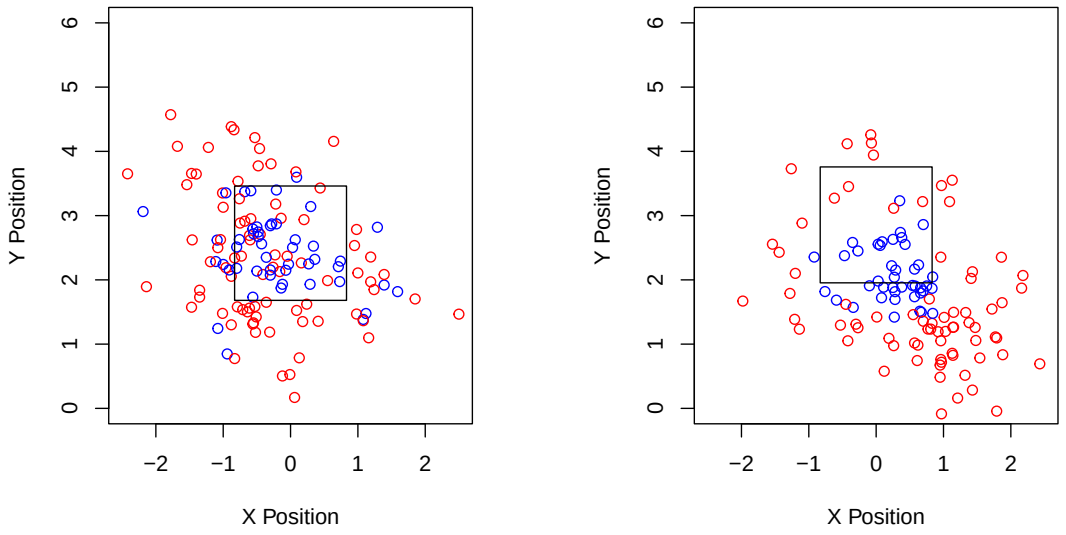
Table 5 re-runs the main results above but using *next-season's* network quality $\overline{Q}_{-i,c(t+1)}$ as the explanatory variable. The non-significant regression coefficients are consistent with the hypothesis that the effect is driven by peer learning, rather than mean-reversion.

Table 5

	Dependent variable:			
	Call Qual (Seas.)	Call Qual. (Game Level)		
	(1)	(2)	(3)	(4)
Next Crew Quality	−0.014	−0.013	−0.036	0.050
	(0.061)	(0.041)	(0.042)	(0.051)
Experience				0.001
				(0.002)
Exp. × Next Crew Quality				−0.002
				(0.002)
Past Self Quality		0.529***		0.474***
		(0.022)		(0.023)
Observations	915	22,876	25,864	22,876
R^2	0.742	0.183	0.250	0.186
Adjusted R^2	0.739	0.183	0.247	0.186
Note:				*p<0.1; **p<0.05; ***p<0.01

7 Bias and precision

Green and Daniels (2018) use similar ball-tracking data on MLB umpire pitch / ball calls to conclude



(a) Low bias, high prec.

(b) Low prec., high bias

Figure 4: Two games with decision making quality of 0.75

Note:

that home-plate umpires make decisions in a Bayesian way. As such, it is useful to consider two elements of decision making: bias and precision.

To the extent that umpires make decision in a Bayesian fashion, it may be the case that bias reflects mis-specified priors about the position of the true strike zone, while imprecision may reflect either a loosely-held prior or poor performance by the home plate umpire.

In order to demonstrate these differences, figure 4 shows the ball-strike call plot for two games in the data for which the percentage of correct calls was around three fourths.

I estimate bias by considering the 'optimal strike zone' associated with a particular game. Although the x-limits of the strike zone are well defined (by the position and width of the home plate, which is immutable pitch-to-pitch and even game-to-game), the y-limits depend on the build, stance, and position of the batter. This affords the home plate umpire some discretion in terms of where they may believe the correct strike zone is. For each game, I calculate the position of the optimal strike zone. I then calculate the bias of the game calls in terms of how far the optimal strike zone is from the official strike zone, and the precision of the game calls in terms of the residual variation after accounting

after accounting for bias in strike zone location. See Appendix B for details of the construction of this index.

I then repeat the analysis in the main section, but studying the peer effect of decision-making quality through both bias and precision. In particular, I estimate a version of 4, but with bias and precision (season-individual level and season-crew level) in place of aggregate decision-making quality. The results are shown in table 6. I find strong evidence that the precision of decisions is transmitted across networks, and some weaker evidence that bias transmits across networks.

Table 6

	Dependent variable:			
	Bias		Precision	
	(1)	(2)	(3)	(4)
Past Crew Value	0.0860**	0.0567	0.122***	0.0910**
	(0.04299)	(0.0501)	(0.0552)	(0.0439)
Controls	Seas.	Seas., Ump.	Seas.	Seas., Ump.

8 Possible mechanisms

MLB umpiring may be a surprising setting in which to find spill-overs: umpires' decisions are largely solitary, and hence it's not clear that there are externalities of network performance to compensation. This setting doesn't preclude spill-overs however, with Bandiera, Barankay, and Rasul (2010) and Mas and Moretti (2009) for example finding that even in groups with no production externalities there may be network effects on decision-making due to social pressures and monitoring. In a highly scrutinized environment such as MLB umpiring, such explanations for spill-overs might also be important.

In this section, I discuss some possible mechanisms for the network effects outlined above. The study of peer networks tends to ascribe spillovers to two main mechanisms: a direct mechanism like learning or assistance, and an indirect mechanism such as effort or social pressure. Both are potential explanations in the setting of umpires. The rise of monitoring technology for umpire decisions (including the PITCHf/x data used in this paper) means that umpire calls are more scrutinized than ever. To the extent that umpires work together to review footage and decisions, it's possible that an umpire working with a high-quality network could learn to make better calls, and hence have better call quality in

future seasons. In terms of indirect mechanisms, the effort required to make accurate umpire calls means it is likely that umpires employ heuristics and other approaches. Peer pressure on effort could thus manifest in a positive relationship between network quality and own call quality.

Learning

A common explanation for peer effects in productivity in the literature is a learning channel. In situations which require expert or institutional knowledge, working closely with knowledgeable people is likely to have some spill-over through teaching and learning channels. Although Major League umpires often have many years of experiencing umpiring at a range of lower levels (such as high-school, college, or Minor League games) before being selected for MLB duties, it's possible that there is additional knowledge required to be an MLB umpire. Thus, umpires assigned to work with knowledgeable crews (who are likely to make fewer mistakes in their strike and ball calls on average) will gain knowledge and have higher average call-quality themselves.

A particular example of this type of institutional knowledge in Major League Baseball is pitch framing. Pitch framing refers to the way in which the catcher positions themselves after catching a pitch, in order to manipulate the home-plate umpire into calling a strike. Pitch framing is a skill that a catcher can develop, and can provide an advantage to a team, even though it is technically within the rules of the game. Pitch framing was particularly evident in the late 2000s and early 2010s. However, over time, umpires developed more knowledge about pitch framing (in particular taking advantage of video replay and monitoring tools), and it fell out of fashion. During this time, if a given umpire was assigned to a crew who were highly knowledgeable about pitch framing, they might have learned from their peer about how to identify and correct for it. In such a way, their call-quality would be expected to increase, as a result of spill-overs from their peer network.

One way of testing for this is to think about the length of time that umpire crews spend together in a season. The frequent in-season umpire shuffles mean that some crews may work only a few games together before one or multiple umpires are reassigned to new crews. If there is a learning aspect, we might expect that crews that work together many games in a season should have stronger peer spillovers than crews that work few games.

Peer effects via learning are likely to take time to manifest, with the possible mechanisms discussed in section 8 all requiring repeated interaction. In my sample, most crews don't umpire a large number of games together – the mean is seven games worked together as a crew. In order to test the effect of time worked together on peer spill-overs, I split the sample into those crews who fewer than seven games together, and those who worked seven or more games together, and ran the main regression for each group. The results are shown in table 7. I find that the sample-wide results are driven entirely by the crews who worked at least seven games together.

Table 7

	Dependent variable:					
	Call Quality					
		≤6			7+	
	(1)	(2)	(3)	(4)	(5)	(6)
Past Crew Qual.	0.00324	0.0264	-0.00294	0.107***	0.0724*	0.0885**
	(0.0252)	(0.0322)	(0.00157)	(0.0307)	(0.0372)	(0.0423)
Seas.	✓	✓	✓	✓	✓	✓
Exp.		✓	✓		✓	✓
Past Qual.		✓			✓	
Ump.			✓			✓

Rational inattention and social norms

MLB umpiring requires considerable amounts of judgment calls, and hence it's likely that there is some degree of heuristic application and social norms. (Archsmith, Heyes, Neidell, & Sampat, 2021) demonstrate that MLB umpires make fewer mistakes when considering 'higher-stakes' calls, and contend that umpires have a depletable 'budget' of attention for a particular game. This provides a mechanism for spill-overs, via peer effects on effort. If umpires are able to improve their call quality by exerting more effort, we would expect to see a decision-making spill-over if peer effort (which is proportional to peer call quality) raises own effort and hence own quality. The peer effects in effort have been well studied (see, for example, (Cornelissen, Dustmann, & Schönberg, 2017)).

One test for this is to think about the 'stakes' of a particular call. Some calls have more bearing on game outcomes than others – for instance, when the (ball-strike) count is 1-1 the particular call won't lead to a strikeout or a walk, and so there is less at stake than if the count is 3-2 and then next ball

or strike will be decisive. Similarly, some decisions may have more impact on season outcomes, with late-in-season games for playoff contenders having more direct impact than early-in-season games or 'dead rubber' games between non-playoff teams at the end of the season.

9 Conclusion

This paper presents evidence of network spillovers in a novel setting – among Major League Baseball umpires. Umpires appear to learn about decision-making from their peers, with those umpires assigned to work in high-quality crews having higher call quality in subsequent games.

This paper contributes more micro-evidence to the study of network effects on workplace performance. In particular, the highly detailed dataset allows for clear quantification of decisions along multiple dimensions.

References

Archsmith, J. E., Heyes, A., Neidell, M. J., & Sampat, B. N. (2021). *The dynamics of inattention in the (baseball) field* (Tech. Rep.). National Bureau of Economic Research.

Bailey, M., Johnston, D., Kuchler, T., Stroebel, J., & Wong, A. (2022). Peer effects in product adoption. *American Economic Journal: Applied Economics*, *14*(3), 488–526.

Baley, I., & Veldkamp, L. (2021). *Bayesian learning* (Tech. Rep.). National Bureau of Economic Research.

Bandiera, O., Barankay, I., & Rasul, I. (2010). Social incentives in the workplace. *The Review of Economic Studies*, *77*(2), 417–458.

Betts, J. R., & Morell, D. (1999). The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *Journal of Human Resources*, 268–293.

Camerer, C. F., Loewenstein, G., & Rabin, M. (2004). *Advances in behavioral economics*. Princeton University Press.

Case, A. C., & Katz, L. F. (1991). *The company you keep: The effects of family and neighborhood on disadvantaged youths* (Tech. Rep.). National Bureau of Economic Research.

Cornelissen, T., Dustmann, C., & Schönberg, U. (2017). Peer effects in the workplace. *American Economic Review*, *107*(2), 425–56.

DeGroot, M. H. (2005). *Optimal statistical decisions*. John Wiley & Sons.

Fesselmeyer, E. (2019). *The impact of temperature on labor quality: Umpire accuracy in Major League Baseball.* (Working Paper: Available at SSRN: 3421241)

Green, E., & Daniels, D. (2018). Bayesian instinct. *Available at SSRN 2916929*.

Guryan, J., Kroft, K., & Notowidigdo, M. J. (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, *1*(4), 34–68.

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4), 85–108.

Johnson, R. C., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from head start and public school spending. *American Economic Journal: Economic Policy*, *11*(4), 310–49.

Kahn, L. M. (2000). The sports business as a labor market laboratory. *Journal of Economic Perspec-*

tives, *14*(3), 75–94.

Kahneman, D., & Tversky, A. (1977). *Intuitive prediction: Biases and corrective procedures* (Tech. Rep.). Decisions and Designs Inc Mclean Va.

Mas, A., & Moretti, E. (2009). Peers at work. *American Economic Review*, *99*(1), 112–45.

Mills, B. (2014). *Expert workers, performance standards, and on-the-job training: Evaluating Major League Baseball umpires.* (Working Paper: Available at SSRN: 2478447)

Mills, B. (2017). Technological innovations in monitoring and evaluation: Evidence of performance impacts among major league baseball umpires. *Labour Economics*, *46*, 189–199.

Office of the Commissioner of Baseball. (2019). *Official Baseball Rules: 2019 Edition* (Tech. Rep.). New York, USA.

Parmigiani, G. (2001). Decision theory: Bayesian. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (p. 3327-3334). Oxford: Pergamon.

Parsons, C. A., Sulaeman, J., Yates, M. C., & Hamermesh, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *American Economic Review*, *101*(4), 1410–35.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *Quarterly Journal of Economics*, *116*(2), 681–704.

Trick, M. A., Yildiz, H., & Yunes, T. (2012). Scheduling Major League Baseball umpires and the Traveling Umpire Problem. *Interfaces*, *42*(3), 232–244.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

Appendix A The strike zone and MLB umpiring

The strike zone in baseball is defined in the laws of the game: "that area over home plate the upper limit of which is a horizontal line at the midpoint between the top of the shoulders and the top of the uniform pants, and the lower level is a line at the hollow beneath the kneecap" (Office of the Commissioner of Baseball, 2019, pp. 153).

A pitch is defined to be within the strike zone if it's central point is within the strike zone. This pins down the exact width of the strike zone. Formally, the strike zone extends from the middle of home plate 19.94 / 2 inches in either direction. Home plate is 17 inches wide, and a regulation baseball is 2.94 inches wide. Thus, if the center of the ball is more than 17 + 2.94 / inches away from the center of home plate, it is outside the strike zone. The y position and height are more subjective, but are tracked using video technology.

In many instances, the strike zone is converted to a 'universal strike zone'. This scales the height and vertical position of the strike zone to be accurate for a typical 73-inch baseball player (near the average height of MLB batters). The y co-ordinate of pitches is also scaled. This allows for comparisons and clear presentation of ball-tracking data, in situations in which the identity of the batter is un-known or irrelevant. For all pitch-position charts shown in this paper, I use a universal strike zone and scaled y co-ordinates.

Figure 5 provides an illustration of the strike zone relative to home plate and a batter.

Appendix B The optimal strike zone

It is possible that the home-plate umpire has a different idea of the location of the strike zone, and makes his ball / strike calls based on his internal model of the strike zone. While doing this consistently over time would result in poor work performance and hence possible repercussions, it is possible that for a short stretch of time (e.g. a game), an internal model is used. I attempt to quantify this internal model (which I call an 'optimal strike zone') using a maximum-likelihood procedure outlined below.

I assume that the x-position of the optimal strike zone is exactly the same as the actual strike zone,

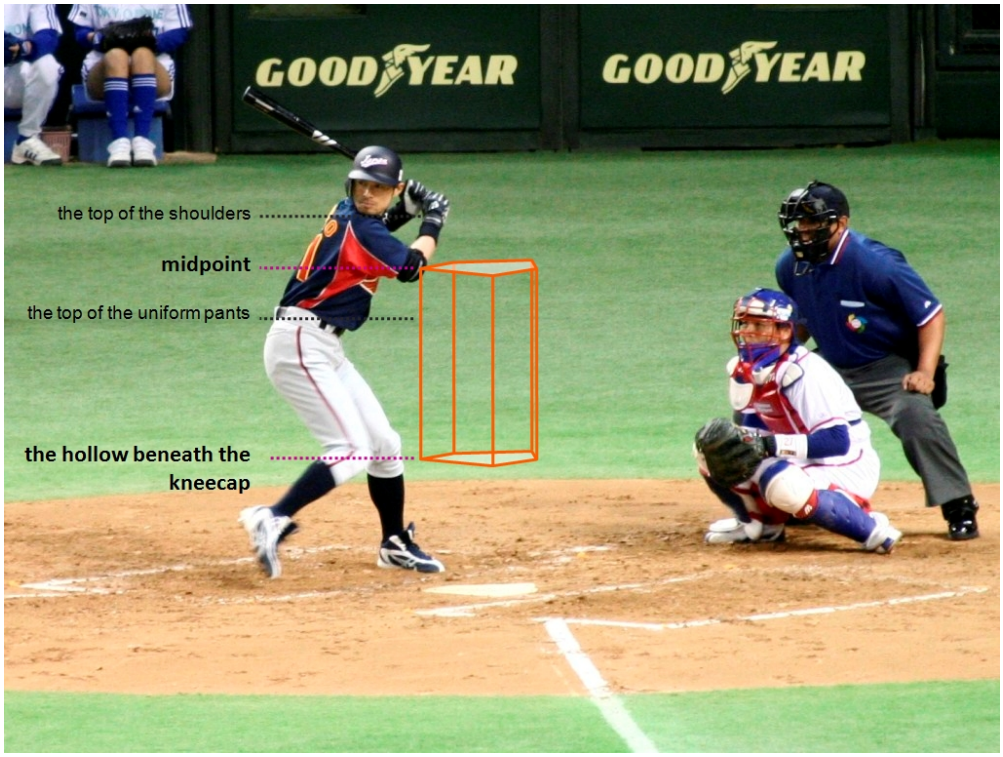
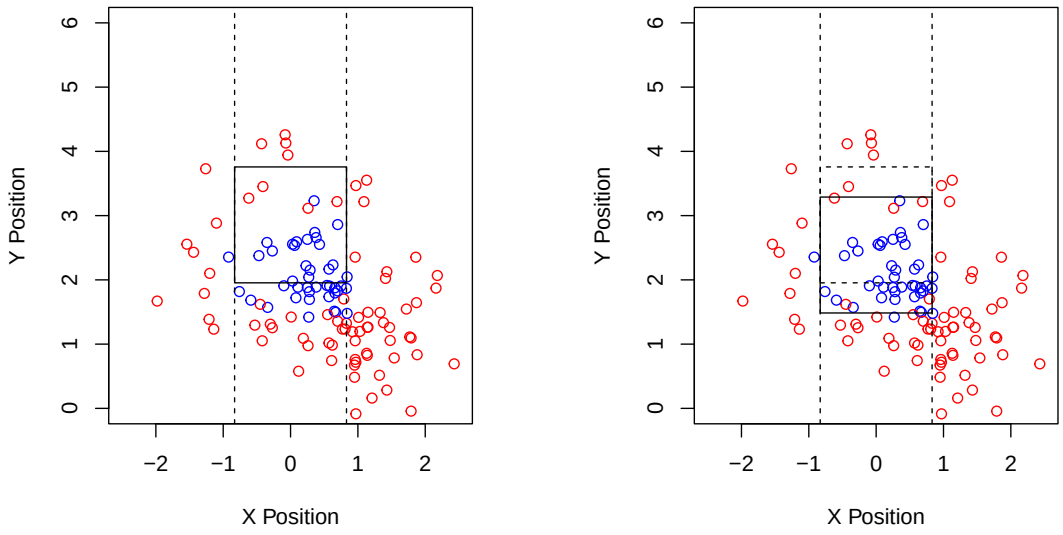


Figure 5: Strike zone super-imposed onto an image from a game

Note: Image retrieved from https://commons.wikimedia.org/wiki/File:Strike_zone_en.JPG, used under CC BY 2.0.



(a) Strike / Ball calls and fixed x position

(b) Actual and optimal strike zone

Figure 6: Choosing the optima y-position for the optimal strike zone

Note: In both charts, red dots are the position of pitches that were called balls, and blue dots are the positions of pitches that were called strikes. The dashed vertical lines are the (fixed) horizontal edges of the strike zone. In (b) the dashed line shows the actual strike zone, and the solid line shows the optimal strike zone.

as it is pinned down exactly by the position of home plate, which is constant for every pitch. I also assume that the dimensions (height and width) of the strike-zone itself are constant. The procedure then involves choosing the y-position of the actual strike zone in order to maximize the likelihood that the implied strike zone was the strike zone that the umpire had in mind. I do this by choosing the y-position to minimize the proportion of incorrect calls.

The bias is defined as the difference between the y-position of the optimal strike zone and the actual strike zone.

For the precision variable, I calculate whether each pitch call (called strike or called ball) would have been correct if the optimal strike zone was in use. I then define precision as the number of 'correct' calls under this optimal strike zone.

This procedure is demonstrated in figure 6. I also show the histogram of biases and a scatter plot of raw accuracy and my precision variable in figures 7 and 8.

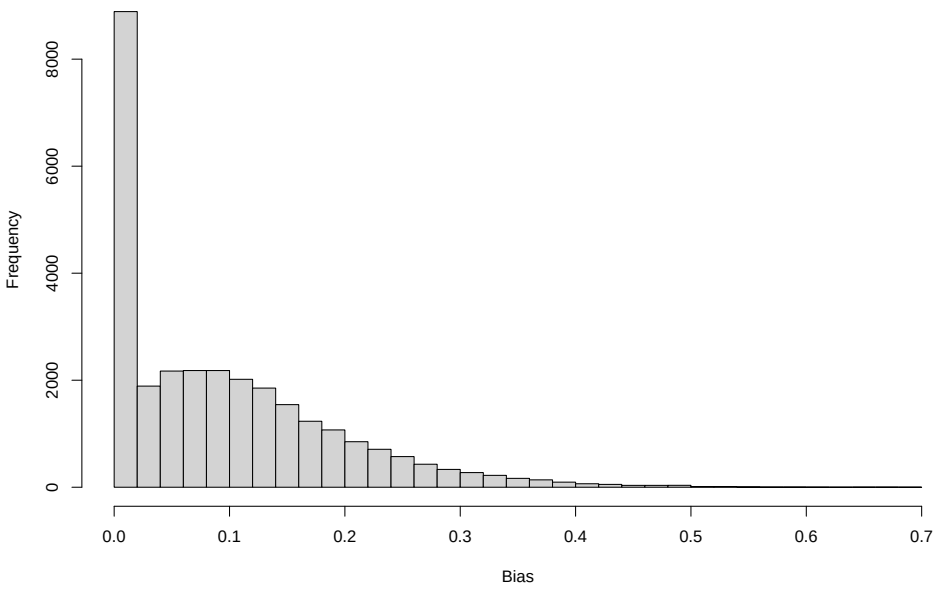


Figure 7: Umpire strike-zone biases calculated from optimal strike zones

Note:

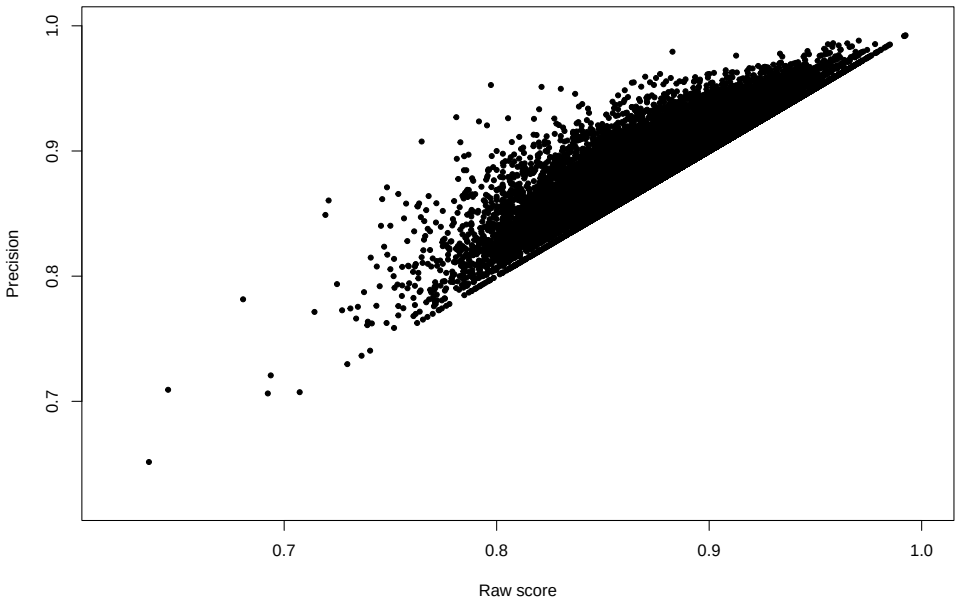


Figure 8: Raw accuracy and 'precision' calculated from optimal strike zones

Note: