DRAFT: Network Effects and Productivity in Major League Baseball Umpiring

Jed M. S. Armstrong[†]

April 22, 2021

Abstract

This paper uses data on Major League Baseball (MLB) umpire crews to estimate the effect of an individual's peer network on productivity, measured by the quality of strike / ball calls. MLB umpire data has two advantages that make this possible: 1) umpires are assigned to a crew each season / game in a way that is independent of call quality, and have significant churn within these crews; and 2) decisions can be uniquely attributed to a single umpire, and every call is able to be reviewed as either correct incorrect. Using these data, I show that a one-standard deviation improvement in the call quality of an umpire's peer network in the previous season 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.

[†]New York University: jedmsarmstrong@nyu.edu. I thank Mike Gilraine, Paul Scott, and Daniel Waldinger for useful feedback. I posted some of the early correlations presented in this paper to FanGraphs Community Research: https://community.fangraphs.com/. First draft posted: April 22, 2021.

1 Introduction

Economists have considerable interest in identifying and estimating the effect of networks on productivity. Quantifying the productivity spill-overs from a group to an individual and vice versa has implications for the optimal allocation of workers across and within firms (Jackson & Bruegmann, 2009). Despite this interest, empirical estimates of spill-over productivity effects in the literature are limited. To a large extent, this is due to the fact that quantifying peer effects on productivity requires considerable amounts of data, in terms of both breadth and depth. Firstly, productivity needs to be observed at a highly granular level (such as at a team / individual level), which requires extremely detailed data. Secondly, individuals must be observed across a range of different groups, which requires a long time series, as well as data on a very broad range of workplaces. As such, most spill-over papers have tended to focus on subsets of the population, such as teachers.

In this paper, I make use of an empirical setting that displays both of these data features: 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 a 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 mean that these calls are able to be verified, 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 calls can be thought of as a measure of productivity in this setting.

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, and section 7 discusses some possible mechanisms for the effects, as well as outlining some evidence for these mechanisms. Finally, section 8 concludes.

[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)

Related literature

This paper contributes to two main literatures: 1) the literature on peer and network effects on productivity and work performance; and 2) the literature using sports (and particularly baseball) data to identify labor market results.

Peer and network effects on productivity

The 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 productivity 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.

Despite the rise in interest among economists, establishing credible estimates of peer effects remains somewhat elusive in the literature.

Sports data and labor markets

The second literature this paper contributes to uses 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.

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).

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 in 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.

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 two-dimensional area whose width is defined by the home plate and whose height is defined by the batters torso.[2] A pitch that is not swung at by the batter[3] 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. Accumulating strikes is a key way to get a batter out, accounting for almost a quarter of outs, and hence correct calls by the umpire are crucial for a fair contest between the teams.

[2]Formally, MLB defines the 'official strike zone' as "the area over home plate from the midpoint between a batter's shoulders and the top of the uniform pants – when the batter is in his stance and prepared to swing at a pitched ball – and a point just below the kneecap."

[3]If the batter swings at a pitch, the result is always a strike, irrespective of the position of the ball.

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.

Assignment of umpires to crews

At the start of each season, MLB umpires are allocated to crews of (typically) four.[4] 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[5], and hence it appears that the assignment is *as-good-as random* with respect to call quality.

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

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

In theory, the crew remain together for the season.[6]. 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.[7] 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.

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

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 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.

[6]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.

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

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.[8] Umpires on base one, base two, and base three are responsible for determining whether runners to their base are safe or out. 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.

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

[8] 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.

[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 'mistakes' made by 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]

[Figure 1 about here.]

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 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).

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

I merge information on the umpire allocations (including the home-plate umpire for the game), and game characteristics (e.g. ballpark, attendance, and game time) using game logs compiled by Ret-

[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.

roSheet.[14]. By combining pitch and umpire data, I can construct a measure of call-quality for each umpire i and game g (in season t and when working with crew c):

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

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.

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

4 Identification

If assignment to crews were fully random, the effect of an umpire's network on productivity could be estimated from 4 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.

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

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

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

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 about here.]

5 Empirical strategy

There are two key channels of interest when considering peer effects on productivity: 1) the effect of a high or low productivity individual on a given group; and 2) the effect of a high or low-productivity 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 past-crew mistake index and the umpire error rate at a season level. 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 productivity and individual productivity that this paper seeks to uncover.

[Figure 2 about here.]

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.

[Table 2 about here.]

The effect of experience on peer influence

There is considerable heterogeneity in the level of Major League umpiring experience among the umpires in the sample. In order to test how spillovers depend on experience, I take the game-level regression model 4, and include interaction terms on network quality for experience (defined as years since an umpire made their MLB umpiring debut). The results are shown in table 3.

[Table 3 about here.]

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.

Robustness

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

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 4 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 4 about here.]

7 Possible mechanisms

In this section, I discuss some possible mechanisms for the network effects outlined above. 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 productivity to compensation. This setting doesn't preclude spill-overs however, with Bandiera, Barankay, and Rasul (2010) and (Mas & Moretti, 2009) for example finding that even in groups with no production externalities there may be network effects on productivity due to social pressures and monitoring. In a highly scrutinized environment such as MLB umpiring, such explanations for spill-overs might also be important.

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.

8 Conclusion

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

This paper contributes more micro-evidence to the study of network effects on productivity. Productivity spill-overs remain an important aspect of many theoretical model of production and organization structure, and yet quantifications of network effects on productivity are limited.

References

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.

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.

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

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.

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)

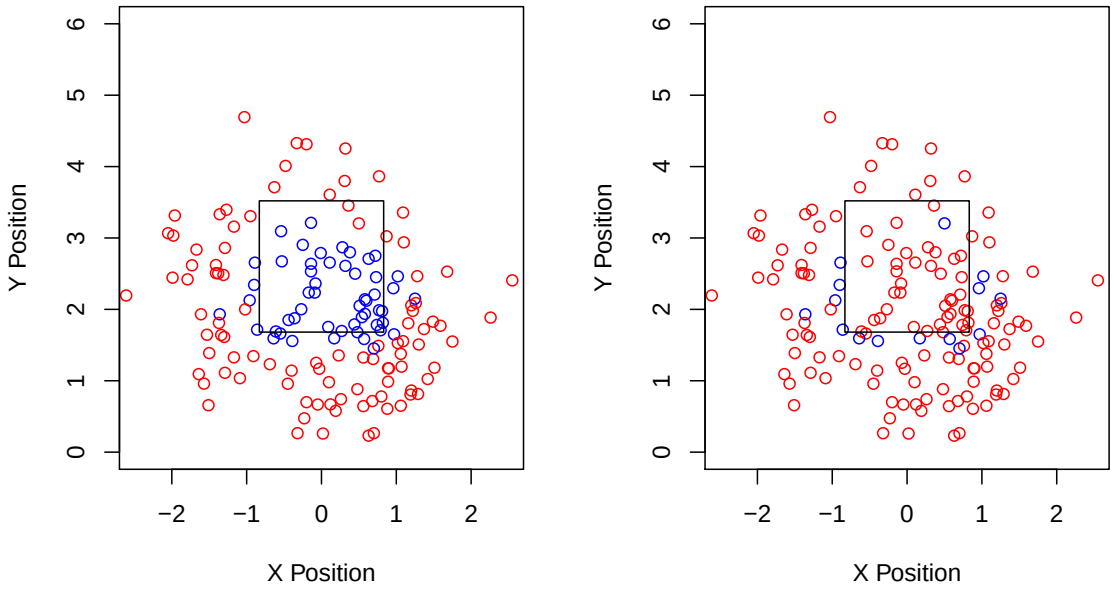
Mills, B. (2017). Technological innovations in monitoring and evaluation: Evidence of performance impacts among major league baseball umpires. *Labour Economics*, *46*, 189–199.

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

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. *The 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.



(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 $q_{i,cgt} = 134/149 = 0.899$.

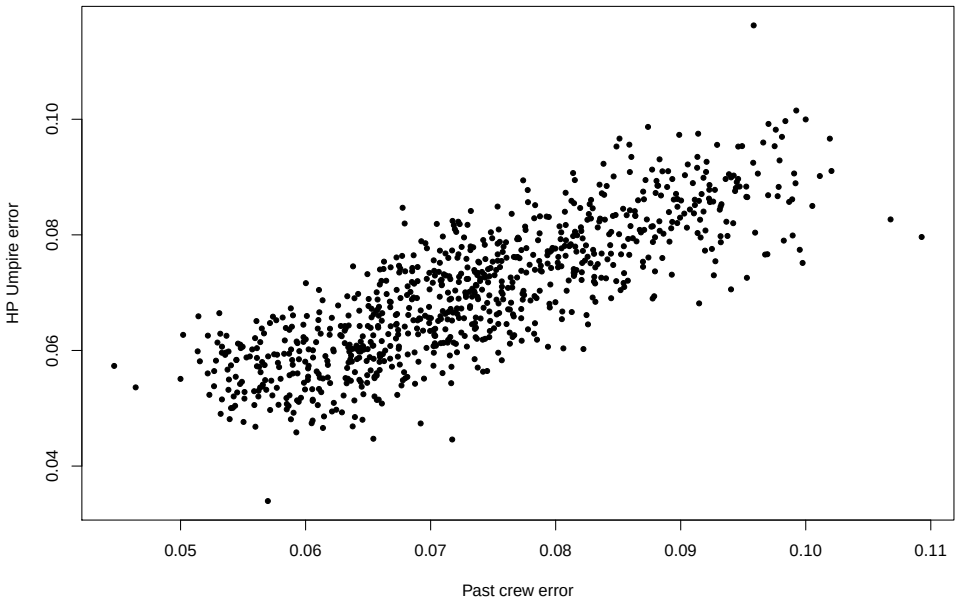


Figure 2: Raw correlation between Past Crew Error and Current Error Rate at a season level

Table 1

	Dependent variable:			
	is_good			
	(1)	(2)	(3)	(4)
crew_quality	0.088	0.003	0.102*	0.003
	(0.057)	(0.002)	(0.052)	(0.002)
all_other_quality		−77.324***		−85.831***
		(0.090)		(0.119)
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

Table 2

	Dependent variable:		
	is_good	is_good.x	
	(1)	(2)	(3)
past_self_quality		0.567***	
		(0.022)	
past_network_quality	0.107**	0.030	0.084***
	(0.049)	(0.030)	(0.032)
Observations	803	23,688	23,688
R^2	0.708	0.200	0.218
Adjusted R^2	0.704	0.199	0.214
Note:		*p<0.1; **p<0.05; ***p<0.01	

Table 3

	Dependent variable:	
	is_good.x	
	(1)	(2)
past_self_quality		0.512^{***}
		(0.023)
past_network_quality		0.086^{**}
		(0.039)
experience	0.015^{***}	0.002^{*}
	(0.001)	(0.001)
past_network_quality:experience		-0.003^{*}
		(0.001)
Observations	29,091	23,688
R^2	0.268	0.201
Adjusted R^2	0.264	0.201
Residual Std. Error	0.030 (df = 28949)	0.030 (df = 23673)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 4

	Dependent variable:	
	is_good.x	
	(1)	(2)
past_self_quality	0.529***	
	(0.022)	
past_network_quality	−0.013	−0.036
	(0.041)	(0.042)
Observations	22,876	25,864
R^2	0.183	0.250
Adjusted R^2	0.183	0.247
Residual Std. Error	0.030 (df = 22864)	0.031 (df = 25728)
Note:		*p<0.1; **p<0.05; ***p<0.01