Technological Innovations in Monitoring and Evaluation: Evidence of Performance Impacts among Major League Baseball Umpires[1]

Brian M. Mills*

*University of Florida
P.O. Box 118208
Gainesville, FL 32611
Tel: 352-294-1664
E-mail: bmmillsy@hhp.ufl.edu

***Abstract**: This paper examines the role of changes in monitoring, evaluation, and training technology as it relates to the performance of Major League Baseball umpires. I find substantial improvements in performance of umpire ball-strike calls consistent with innovations in monitoring technology and later evaluation and feedback enhancements across the league. These improvements are more pronounced for younger umpires that enter the league with similar skill sets to their more experienced peers. Implications for the differential effects of monitoring relative to evaluation and feedback are discussed.*

JEL Codes: Z22, M50, J42, J59

[1] ***Acknowledgements***: I would like to thank Charlie Brown, Rodney Fort, Scott Tainsky, Adrian Burgos, Alan Nathan, the participants at the 2014 Western Economic Association Conference in Denver, and the participants at the 2014 European Conference on Sports Economics in Antwerp for helpful comments on earlier versions of this manuscript in written and presentation form. Finally, I would like to thank Mike Fast, Jon Roegele, and Aaron Baggett for helpful discussions regarding the data and strike zone definitions in this paper.

Technological Innovations in Monitoring and Evaluation:
Evidence of Performance Impacts among Major League Baseball Umpires

"The owners basically see them like bases. They say, 'We need a base, we need an umpire, same thing. We've got to pay them, they're human beings, but they're basically bases.'"

-Fay Vincent (pp. 10, *As They See 'Em*)

1 Introduction

Implementing incentive systems within the workplace to increase employee performance and induce sorting of the labor pool is a central issue in personnel economics (Lazear & Rosen, 1981; O'Keeffe et al., 1984; Lazear 1986; Ehrenberg & Bognanno, 1990; Baker, 1992; Paarsch & Shearer, 1999; Prendergast, 1999; Lazear, 2000a; Lazear, 2000b; Lazear, 2000c; Dohmen & Falk, 2011). There has been substantial effort put forth in designing and evaluating performance pay systems, with empirical evaluation in the economics literature taking place in a number of contexts including, among others, CEOs and firm performance (Jensen & Murphy, 1990; Barro & Barro, 1990; Bulan, Sanyal, & Yan, 2010), education (Fryer, 2013; Goodman & Turner, 2013; Podgursky & Springer, 2007; Woessmann, 2011), healthcare (Lindenauer et al., 2007; Campbell et al., 2009), and professional sports (Simmons & Berri, 2011). The various strategies used to align incentives between principals and agents are often implemented where strict monitoring of employee effort is cost prohibitive, and often depend on limited monitoring paired with pay for performance or termination decisions within the workplace. However, in the event that technological shocks substantially reduce the cost of precise monitoring, employee performance may be positively impacted even without significant adjustments to incentive-based pay or prospective punitive outcomes.

Accordingly, there could be various ways in which precise evaluation affects performance of agents, depending on how these technological innovations are used. For example, recent research has suggested that agents may simply feel pressure to perform at a higher level due to increased scrutiny (price) of poor performance stemming from monitoring alone (Parsons, Sulaeman, Yates, & Hamermesh, 2011). This would imply a sudden change in effort in the face of monitoring from cost increase associated with the monitoring itself, particularly when tied directly to pay or termination decisions. However, while the tying of monitoring and performance measurement with punitive outcomes may provide short-term improvement (Lucifora & Origo, 2015), the human resources literature has suggested that this practice could undermine the ultimate intention of the system—improving employee and firm performance in the long run—due to psychological stress on employees, or other related factors (Stanton, 2009; Chalykoff & Kochan, 1989). This may be particularly true when these systems are not jointly developed and agreed upon with the employees themselves (Alder, 2001).

Alternatively, Taylor and Tyler (2012) note that performance evaluation can also be used as part of a long-run developmental process. For example, when objective measurement can be achieved at relatively precise levels, feedback from a trusted supervisor could in turn play a key role in improving employee productivity and encourage further human capital investment or effort over time by employees (Taylor & Tyler, 2012; Rockoff, Staiger, Kane, & Taylor, 2012; Jackson, Rockoff, & Staiger, 2014). Therefore, improvements in technology that increase both the precision and continuity of evaluation and feedback may result in performance improvements beyond that of a sudden structural shift at the initial introduction of monitoring. Alternatively subsequent to such a shift, continual human capital development among employees could take place at multiple time points of reflection throughout the career life cycle. Yet, beyond this small

body of work, there is little empirical evaluation of the long-run developmental effects of evaluation itself on employee performance, or whether these systems have differential impacts in terms of discrete or gradual improvements.

With this in mind, this paper uses Major League Baseball (MLB) umpire ball and strike calls—a measure of performance for one portion of an official's duties—as its setting to identify performance changes associated with implementation of, and innovation in, monitoring and evaluation in a novel context of expert workers with low turnover rates. Umpires are well-monitored agents with clear changes in the use of monitoring and evaluation throughout their labor history with MLB.

Using this labor context, I find evidence that changes in the use of technology in monitoring and evaluation have improved performance of umpires at different time points in their recent labor history. These changes are consistent with publicly mandated expectations with respect to the strike zone size and shape: umpires call balls and strikes in a way that is more consistent with the stated MLB Rulebook strike zone when monitored and evaluated, and call more strikes when directed to do so under monitoring.

Results point to evidence of a sudden response to the implementation of monitoring from MLB—in the form of a level shift in strike calling rates. However, using locational data to measure umpire accuracy rates, a more gradual continuous improvement is revealed after on-the-job evaluation and precise feedback began in 2009. These improvements are steeper for less experienced umpires—who tend to enter the league with similar base levels of performance—indicating that newer umpires benefit more from the training and evaluation than mid-career and senior umpires.

2 Contextual Setting

2.1 Major League Baseball as a Labor Laboratory

As with most team sports, the legitimacy of competition in MLB requires that its officials—called umpires—fairly and accurately judge a range of outcomes throughout the game. These plays include calling balls and strikes when working behind the plate, calling runners out or safe at each base, and judging whether balls land in fair or foul territory. They are further considered arbiters of the game in that they have the authority to eject players for unsportsmanlike play, or other violations of the rules. However, for this work, I focus specifically on strike zones of MLB umpires when they work behind the plate in a regular season game. It is important to note that evaluation of the strike zone addresses only a portion of the duties of MLB umpires. In fact, umpires serve behind the plate in only about one quarter of the games in which they work, rotating around the field at first base, second base, and third base approximately evenly across the season. Therefore, job performance improvements among this group of employees do not necessarily mean generalized effects across all portions of the duties of those employees. However, given limitations on data regarding other duties of umpires within MLB games—and the specificity with which monitoring and evaluation changes have taken place over the time period of interest—the strike zone is the most accessible performance outcome to address in this labor context.

As noted in Price and Wolfers (2010) and Parsons et al. (2011), the nature of the required expertise in the labor market for professional sports officials describes a group of employees that are at the top of their profession, yet are still subject to biases or mistakes (Dohmen & Sauermann, 2015; Garicano, Palacios-Huerta, & Prendergast, 2005; Green & Daniels, 2014; Kim & King, 2014; Lopez & Snyder, 2013; Matthewson, 2010; Mills, 2013; Moskowitz & Wertheim,

2011; Nevill, Balmer, & Williams, 2002; Price, Remer, & Stone, 2012; Sutter & Kocher, 2004; Tainsky, Mills, & Winfree, 2015). However, there are a number of additional characteristics of the umpires' labor market beyond social biases that make this context valuable for study of other economic phenomena.

First, the wealth of publicly available employee performance data is unlike many other industries where data are often proprietary and unavailable (Kahn, 2000), allowing for evaluation of economic phenomena through a forensic economics approach (Zitzewitz, 2012). Second, umpires operate under a monopsonistic buyer of labor like that of many professional athletes, but do not have the benefit of teams bidding for their services. Instead, contracts for umpires are secured at the league level, with no truly comparable alternative to MLB.[2] Therefore, these employees are subject to the league's demands, with little avenue for alternative employment. Third, umpires in MLB are effectively superstars within their profession, but with ability levels that are somewhat heterogeneous (Rosen, 1981; O'Keeffe, 1984; Franceschelli, Galiani, & Gulmez, 2010). While this is also a characteristic of baseball players, player performance is zero-sum in the context of league-level changes to compensation, training, or monitoring. At the umpire level, performance is standalone, and therefore changes in this performance due to changes in incentives, monitoring, or evaluation at the league level may be measured independent of the performance of the labor market competition (Fernie & Metcalf, 1996). Fourth, there have been improvements to the monitoring (and training) system in MLB, with recent monitoring implementation that is both cheaper and more precise than the previously used

[2] MLB Umpires with significant experience can make as much as $400,000, while new umpires can expect around $120,000, including per diems between $300 and $400. Minor league umpires, on the other hand, make between $1,900 and $3,500 per month for the five or six month season, and most work odd jobs—such as delivering pizzas—in the off-season for additional income.

manual methods. These systems were put in place alongside an expected minimum performance level to avoid further developmental and training requirements, discipline, or termination both prior to, and after, collective bargaining with umpires. [3] Lastly, the low turnover rates of umpires once they reach the MLB level provides a setting in which the threat of termination is relatively low. This seems to suggest that, as in Taylor and Tyler (2012), MLB umpires make up a group of professionals that could be considered motivated agents who are willing to act positively on information about their performance.

2.2 The Baseball Umpires' Labor Market: A Short History

The most recent changes to monitoring and evaluation relevant to this work include an implementation of limited technological monitoring in 2001 alongside a directive for umpires to call more strikes and limit the total pitches in games, and an agreement to improve the monitoring system in 2009 with new technology, which would further provide regular reports directly to umpires after every game they served behind the plate. However, labor relations between MLB and its officials have had a long and eventful history, beginning in the 19th Century and continuing through the early 21st Century (Alcaro, 2002; SDABU, 2014).

This section briefly reprises the more recent tensions in order to set the stage for the current analysis. While umpires are trained in the minor leagues by unaffiliated schools and organizations, MLB has shown interest in having a heavier hand in the umpire development process. Therefore, I focus on the MLB level of performance and monitoring, rather than the

[3] Performance levels included ensuring a certain percentage of correct calls were achieved in a given game called by the umpire. This information was provided through personal communication with an MLB umpire.

specifics of training throughout the minor leagues, but note this development is just as interesting.

I begin with more recent work stoppages to contextualize the eventual implementation of strike zone monitoring. In 1991, there was a short strike by umpires, which resulted in increased pay and a merit system related to postseason assignment and its resultant additional pay, though umpires were generally resistant to the idea of performance standards (O'Neill, 1990). During this time, umpires were handled separately by the American League (AL) and National League (NL). Despite the performance standards, the league's leaders were unhappy with the performance of umpires even after 1991. In 1992, the commissioner's office discussed centralizing umpire activities in order to combat any variation in the strike zone across the AL and NL umpires. Eventually, in 1995, a lockout took place as MLB players returned from their own strike. The lockout was short-lived, and a new 5-year collective bargaining agreement was reached. However, umpires agreed that they were not allowed to strike prior to the next collective bargaining meetings in 2000.

The following season, in 1996, the league changed the official definition of a strike for umpires to follow that has remained through 2014, and included expansion of the zone relative to its previous definition. This begins our interest in changes in the strike zone, as its definition has been constant since this time. Commissioner Bud Selig further publicly acknowledged his interest in obtaining direct control over all umpires and had Sandy Alderson issue a memo calling for uniform enforcement of the strike zone. Perhaps most importantly, Alderson noted that the strike zone should be called from two inches above the belt to the bottom of the knees, and asked team officials to manually chart pitches for umpires working their games, with publicly noted resistance from umpire leaders. However, while ensuring an accurately called

strike zone certainly requires some effort, it is unclear why umpires would be resistant to normalizing the boundaries across the league. The autonomy with which umpires enforced the strike zone—while presumably empowering to each individual umpire in a context where they were almost universally reviled—does not seem to be a worthwhile fight. Nevertheless, umpires made clear their discontent with the demands of Selig and Alderson, as the NL's senior umpire, Bruce Froemming, stated in 1999, "The only talk among National League umpires is the total disrespect that has been shown," (Chass, 2001).

Wary of another lockout, the leader of the union, Richie Phillips, orchestrated a mass resignation of umpires—57 of the league's 68 did so—in 1999 with the belief that this would bring MLB to the negotiating table (Callan, 2012). Instead, the league accepted all of the resignations and began hiring new umpires to replace them. Ultimately, a number of umpires rescinded their resignation, with 22 officially accepted by MLB. This strategy led to the disbandment of the union at that time, in favor of the newly formed World Umpires Association (WUA). The WUA reached a 5-year agreement, and umpires were brought under control of MLB, rather than separately under the AL and NL.

In 2001, MLB installed new pitch tracking technology called QuesTec in four of its parks to evaluate strike zones of umpires. This implementation marks the first innovation in monitoring relevant to this work, and was accompanied by instructions from the league to increase the rate of strike-calling to reduce total pitch counts and length of games.[4] This technology was later expanded to about half of MLB parks after a grievance from the WUA was settled in 2004 (Schwarz, 2009). The implementation of QuesTec in only some parks marks an event after which

[4] One umpire supervisor resigned over the directive from the league (Chass, 2001).

differences in monitoring- and non-monitoring-specific strike rate changes can be estimated from 2001 through 2006.

Finally, in 2007, MLB installed new technology in its parks halfway through the season, known colloquially as Pitch f/x. The system was installed in all parks by the start of the 2008 season. This system monitored every pitch as it cross the plate, and subsequent call by the umpire in the event that the batter did not swing. The data were publicly available for half of 2007 and all of 2008, despite the league not using this explicitly to evaluate umpires. This changed in 2009, as the WUA and MLB agreed on new terms in collective bargaining, a new system took hold in the evaluation and training of umpires called Zone Evaluation (ZE), based on the Pitch f/x system. Under this new system, umpires receive reports of their performance after every game, and the ZE system is claimed to have been more accurate (Schwarz, 2009). The direct nature of the umpire reporting and monitoring here provides a second implementation of improved monitoring at low cost. But most importantly, umpires are provided with immediate feedback after each game they umpire under the ZE system, allowing for evaluation and improvement through direct review of an umpire's performance in that day's game. As noted in previous literature, there could be further effects of human capital development with relatively precise feedback presented to employees. Given existing presence of monitoring across the league at this time, it seems more likely that this effect would introduce gradual—rather than sudden upward shifting—improvement in umpire performance during this era.

This short history provides us with key policy changes for analysis in this work, and clear directives at the league level regarding expectations of employee performance, and feedback through evaluation and training. I use two different data sets to estimate the ball-strike calling effects of technological monitoring and innovations related to employee evaluation, at two key

time points. Effects of interest include changes in the rate of total and called strikes, as directed by the league in 2001 with the implementation of Questec, and accuracy changes after the use of evaluation and feedback for umpires. The following section describes the data and estimation strategies used, and the validity of using these as performance measures in the context of expected changes by MLB officials over the years included in this study.

3 Data and Estimation Strategy

3.1 Estimation I: QuesTec Piecewise Difference-in-Difference Model

3.1.1 Data and Measurement

The first estimation addresses the monitoring change in 2001 using QuesTec, modeling strike rate changes concurrent with increases in technological monitoring paired with the directive to call more strikes during this time period. These data begin in 1997, just after the redefinition of the strike zone by MLB. This series of data on yearly-level umpire strike rates allows quasi-experimental estimation of monitoring impacts, leveraging information on which stadiums are equipped with the QuesTec monitoring system, as in Parsons et al. (2011) and Tainsky et al. (2015). I use these data, aggregated from Tainsky et al. (2015) to estimate discrete shifts in strike calling rates after the implementation of new technology for monitoring of umpires.

The measures developed for this portion of the analysis are proxies for umpire performance. Because locational data is not available until a later date—QuesTec data are and have always been kept privately—I assume increases in the rate of strikes called can proxy umpires' willingness to conform with the requests of MLB to call more strikes and reduce the number of pitches throughout a game. I use two separate measures as dependent variables:

$TotalStrikeRate_{it}$ and $CalledStrikeRate_{it}$. The first measure is calculated by dividing the total number of strikes for umpire *i* by the total number of pitches during the season *t*. This calculation is performed separately for stadiums with and without the QuesTec system installed. Note that this measure can be affected by behaviors of the batters and pitchers in the game. If batters swing more, or pitchers throw pitches more likely to be missed by the batter, then the strike rate can be affected by these outcomes. Therefore, the second measure, $CalledStrikeRate_{it}$, uses only pitches which require judgment by the umpire. If a batter does not swing, and the pitch crosses through the rulebook strike zone, then the umpire's duty is to call this a strike. If the pitch crosses the plate outside the rulebook strike zone, then the umpire should call the pitch a ball. Accordingly, $CalledStrikeRate_{it}$ is calculated by dividing the total number of pitches that were called strikes by the umpire, divided by the total number of pitches at which batters did not swing. This reduces the impact that batter swings could have on the performance measure, and evaluates umpires' willingness to call more pitches strikes during this time.

The data span the years 1997—shortly after MLB's redefinition of the strike zone—through 2006. I avoid using the 2007 and 2008 seasons in the estimation here due to the installation of the Pitch f/x system in all or most parks beginning in mid-2007. While the league did not explicitly use the newer system to monitor umpires in lieu of QuesTec, the public data was available to anyone during this time, leading to ambiguity over the differentiation between monitoring conditions in these years. Monitoring conditions through 2006, however, are clearly distinguished by the presence of QuesTec in the given stadium.

3.1.2 Estimation Strategy

I use piecewise regression in a difference-in-differences (DiD) framework to estimate changes in umpire strike rates attributable specifically to QuesTec monitoring after the 2001 season. The piecewise estimation requires transforming a dummy variable indicating the use of QuesTec in the stadium for the observation, $q \in \{0,1\}$, and a yearly trend variable, $t \in [1997, 2006]$, to two new variables, $Q_j \; and \; Z_j, \; (j = 1, 2, 3)$, respectively, such that:

$$Z_1 = \begin{cases} 0, & t < 2001 \\ t - 2001, & t \geq 2001 \end{cases}$$

$$Z_2 = \begin{cases} 0, & t < 2001 \\ 0, & t \geq 2001 \; and \; q = 0 \\ t - 2001, & t \geq 2001 \; and \; q = 1 \end{cases}$$

$$Z_3 = \begin{cases} 0, & t < 2001 \\ t - 2001, & t \geq 2001 \; and \; q = 0 \\ 0, & t \geq 2001 \; and \; q = 1 \end{cases}$$

$$Q_1 = \begin{cases} 0, & t < 2001 \\ 1, & t \geq 2001 \end{cases}$$

$$Q_2 = \begin{cases} 0, & t < 2001 \\ 0, & t \geq 2001 \; and \; q = 0 \\ 1, & t \geq 2001 \; and \; q = 1 \end{cases}$$

$$Q_3 = \begin{cases} 0, & t < 2001 \\ 1, & t \geq 2001 \; and \; q = 0 \\ 0, & t \geq 2001 \; and \; q = 1. \end{cases}$$

In this representation, Z_1 is the time trend prior to the use of QuesTec, Z_2 is the time trend in stadiums that have QuesTec installed, and Z_3 is the time trend in stadiums that are not equipped with QuesTec after the start of its use in 2001. The regression model is estimated without a

constant, and therefore the Q_n indicators serve as regime-specific intercepts for the pre-QuesTec period, Q_1, the stadiums equipped with QuesTec after its initial implementation, Q_2, and parks not equipped with QuesTec after its implementation, Q_3.

These transformations are used to estimate a model with either $CalledStrikeRate_{it}$ or $TotalStrikeRate_{it}$ as dependent variables, y_{it}. Both DiD models are fit with and without umpire fixed effects, and take the following form:

$$y_{it} = \beta_1 Q_{1it} + \beta_2 Q_{2it} + \beta_3 Q_{3it} + \alpha_1 Z_{1it} + \alpha_2 Z_{2it} + \alpha_3 Z_{3it} + \delta_i + \varepsilon_{it}.$$

Here, *i* and *t* index the umpire and year, respectively. Intercepts and slopes for the regime- and QuesTec- presence specific samples are represented by β_k and α_k, respectively, with individual umpire fixed effects represented as δ_i (where appropriate). The umpire-year specific error term is represented as ε_{it}. Standard errors are clustered by umpire with observations weighted by $\sqrt{n_{it}}$, where n_{it} is the number of pitches called by umpire *i* in year *t*, and specific to the QuesTec monitoring condition for those years after the system's initial implementation. Each of the coefficient estimates can be directly interpreted as the regime-condition-specific slope and intercept, and therefore the Wald test is applied to each linear combination of β_k and α_k to evaluate the statistical significance of the *differences* in intercepts and slopes across regimes, respectively.

3.2 Estimation II: Accuracy Rate Improvements and Segmented Panel Regression

3.2.1 Data and Measurement

To address the effects of evaluation and feedback on umpire performance, I use more detailed locational data on umpire ball-strike call accuracy just prior to, and after, the introduction of the Zone Evaluation System in 2009. Here, rather than aggregate strike rates, I leverage information on pitch location available from the Pitch f/x system for every pitch in the regular season from 2008 through 2014. The availability of these data allows a direct measurement of performance unlike the implied proxies calculated from earlier data samples. I further use these data to identify seniority effects in performance and accuracy improvement rates over the course of the data set to test for heterogeneous across early, mid, and late career umpires.

The availability of data specific to the accuracy of umpire ball-strike calls through Sportvision's regular season Pitch f/x data allows identification of whether a call by the umpire was correct based on the rulebook strike zone, and how these have changed over its use from 2008 through 2014.[5] This analysis directly measures performance improvements among umpires, rather than strike rate measurements that could be influenced by pitchers changing the location of their pitches over time. To measure umpire performance changes, I reduce the data only to pitches subject to judgment by the umpire (called balls and called strikes), and removed pitches labeled as pitchouts, balls in the dirt, or intentional balls, as these require little subjective evaluation by the umpire. The sample size of this subset of data was just over 2.47 million observations.

[5]Discussion with experts in the Pitch f/x system led to the removal of locational data from 2007, as improvements and adjustments to the system were made during this time that preclude the ability to differentiate improvements in umpire accuracy versus improvements in the measurement system. Therefore, all locational analysis takes place on data beginning in 2008.

Accuracy of strike calls requires the identification of the strike zone indicated within the MLB Official Rulebook.[6] I use measurements of anthropometric knee, waist, and shoulder height of males from NASA's Human Integration Design Handbook (NASA, 2000), applied in the context of the average height of MLB batters during the 2008 through 2014 seasons (approximately 73.5 inches).[7] This places the lower and upper boundary of the strike zone at approximately 18.2 and 41 inches, respectively. The width of the strike zone is measured as the 17 inch plate width as noted in the official rules. However, because a pitch is considered a strike even if a portion of the ball crosses over the plate—and Pitch f/x measurements are associated with the center of the baseball—I add the radius of the ball to both the inside and outside edges of the plate width. The final result is a strike zone width of approximately 19.92 inches.

Four pitch types were categorized by the respective accuracy of the umpire call for the 2008 through 2014 regular seasons. The first of these pitch types is a correctly called strike defined as a pitch that crosses the front of the plate within the strike zone plane and is subsequently called a strike by the umpire (Column 1, Figure 2). This can be thought of as a true positive in the statistical sense. The second pitch type is a correctly called ball (Column 3, Figure 2). This occurs when a ball located outside of the strike zone is called a ball by the umpire, or a true negative. The last two classifications include incorrectly called strikes (a false positive;

[6] The STRIKE ZONE is 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 knee cap. The Strike Zone shall be determined from the batter's stance as the batter is prepared to swing at a pitched ball," (MLB, 2010). I use the "2 inches above the waist" directive from Sandy Alderson noted earlier in this paper as the top of the strike zone.

[7] Individual batter height was not included in the data. However, umpires generally see a similar sample of batter heights within and across seasons, which would leave the estimates of changes across years unaffected in any substantive way. In the case that this does not portray true accuracy rates, all rates should be biased downward in a similar fashion across both years and umpires.

Column 2, Figure 2) and incorrectly called balls (a false negative; Column 4, Figure 2). Together, these identify strike calls on pitches outside the strike zone, and ball calls on pitches crossing through the strike zone plane, respectively. The proportion of correct strikes to true strikes is a measure of *sensitivity* in the statistical sense, while the proportion of correct balls to true balls is a *specificity* measure. The changes in correct strike and correct ball rates, as well as overall correct call rates, are reported in Table1. As you can see from this table, correct strike rates (*CorStrRate*) have increased by much more than correct ball rates (*CorBallRate*) since the implementation of the Zone Evaluation system.

3.2.2 Segmented Panel Regression Estimation Strategy

The models using the accuracy rate and regime changes—before and after the use of the Zone Evaluation system—are estimated using a panel regression in the segmented regression context, with fixed effects at the umpire level.[8] I again weight observations by $\sqrt{n_{it}}$ and cluster standard errors at the individual umpire level. Further, I remove any umpires that only worked one season in the time span of the data, and restrict umpire-year observations to at least 1,500 called pitches in a given season (about 10 games). Here, there are two regimes of interest, as there is only the single change in evaluation conditions during the time period from 2008 to 2014, and the ZE System is present in all stadiums.

I control for experience in these estimations and test for heterogeneous effects of accuracy improvement for umpires across the spectrum of experience in the league with an interaction of the experience and post-ZE System time trend variable. Umpire experience is measured using the number of years umpired prior to the current season. Therefore, if an umpire

[8] Estimations without umpire fixed effects are also presented.

is in his rookie year, the experience variable takes the value of zero, and if the umpire is in his 30th year, then the experience variable takes a value of 29. This tests predictions of Lazear (2000a) as to whether newer employees begin at a lower competency level, but increase their performance more quickly than older employees. Further, as noted in Taylor and Tyler (2012), prior work largely assumed that mid-career evaluation would have little effect on performance of these employees. However, the authors find some evidence of benefit across the tenure spectrum. I directly test for this by comparing the base improvement trend during the Zone Evaluation era across experience levels of umpires in MLB.

Regime indicator and yearly trend variables are included as regressors to indicate the era in which each observation takes place. I allow for both shifts in the intercepts and changes in the time trend. The form of the regression model is as follows:

In this specification, I identify the ZE era as beginning in 2010. This allows for a one-year delay in the expected improvement from evaluation, feedback, and learning with Zone Evaluation. For example, if training first takes place in time t, then we would expect to see changes in performance in time $t + 1$. Therefore, the pre-ZE intercept and slope are estimated using 2008 and 2009, while the post-ZE intercept and slope are estimated using data from 2010 through 2014. With the subscript *ZE* identifying the Zone Evaluation regime, the model takes the following form:

$$AccuracyRate_{it} = \beta_0 + \beta_1 T_t + \beta_2 D_{ZE} + \beta_3 Exp_{it} + \beta_4 D_{ZE}\tau_{ZE} + \beta_5 D_{ZE}\tau_{ZE} Exp_{it} + \delta_i + \varepsilon_{it}.$$

Here $AccuracyRate_{it}$ is the rate of correct calls for umpire i in year *t*, β_0 estimates the intercept at the beginning of the data set (2008), and T_t identifies the year. Further, D_{ZE} is a dummy

variable indicating the that the observations takes place after ZE implementation, and τ_{ZE} identifies the ZE era-specific time trend change relative to T_t, beginning at zero for each regime and taking the value of zero for years in any other regime. Coefficients for both intercept and trend after the implementation of the ZE System in this estimation are interpreted as the *change* relative to the previous period. ε_{it} is the individual error for each observation in the panel.

4 Results and Discussion

4.1 QuesTec Piecewise DiD Estimation

The results of the DiD estimation for QuesTec and non-QuesTec stadiums can be found in Table 2. Findings are largely consistent both with and without umpire fixed effects. Note that the coefficients in this table are read directly as the intercept and trend during the given regime-QuesTec specific samples.

To begin, at the implementation of QuesTec, *TotalStrikeRate* and *CalledStrikeRate* shifted dramatically upward relative to the year before, as represented by the Wald Test for $(\beta_2 - \beta_1)$ and $(\beta_3 - \beta_1)$. Further, there is clear evidence of a slope change from the pre-QuesTec condition to the post-QuesTec condition across the league. The visualization Figure 1 shows the change in slope, as well as the substantial intercept shift upward in 2001. After the call for increased strike calls, the increase in *TotalStrikeRate* was as much as 2.5 percent relative to the year before. For *CalledStrikeRate*, the increase was as much as 6 to 7 percent relative to the prior season. In both cases, the negative trend in strike calling was no longer apparent after the beginning of the QuesTec era.

Of most importance in these estimations is the relative change in intercept for the QuesTec and non-QuesTec MLB stadiums after its initial implementation. The Wald Test for

$(\beta_3 - \beta_2)$—explicitly testing the difference in the change in shifts after the use of QuesTec in parsk that are and are not equipped with the technology—indicates statistically significant differences in the intercept shifts for the two QuesTec conditions for *CalledStrikeRate,* but not *TotalStrikeRate*. I focus on *CalledStrikeRate* given that it more directly measures the change in behavior among umpires in their ball-strike calling. Specifically for *CalledStrikeRate,* those parks without QuesTec experienced a smaller upward shift than those in the Questec-monitored condition. Further, while the slopes in the two QuesTec conditions are not statistically different according to the Wald Test of $(\alpha_3 - \alpha_2)$—an explicit test of trend differences across conditions after the general QuesTec introduction—there seems to be some convergence of *CalledStrikeRate* across the conditions as represented by the statistically significant (from zero) trend in parks that were not equipped with QuesTec.. This apparent convergence could indicate acclimatization to monitoring and the new strike zone over time, and a subsequent habituation to the newer strike zone that would be difficult to change drastically across conditions, particularly after QuesTec was used more often.

Ultimately, the difference in sudden shifts across the conditions identifies the relative impacts of generalized changes likely due to a directive from the league to call more strikes, and the effect of technological installation in some parks in the league. Without the DiD estimation, it would be somewhat disingenuous to attribute the entire shift in strike rates in 2001 to the installation of monitoring technology. However, it is possible that the presence of the technology had some generalized impact on strike rates that is larger than the difference in conditions. For example, under the acclimatization scenario, changing the split-second judgment substantially across conditions may not be possible, and therefore some of the generalized shift could be attributed to changing behavior because of the technology. However, this effect is not estimable

from the data without a comparable counterfactual league where monitoring was not implemented at all.

4.2 Zone Accuracy Estimation

Lastly, turning to the direct measurements of accuracy after the ZE system implementation, Table 1 shows the yearly changes in accuracy rates of umpire ball-strike calls in aggregate for the league, as well as separated by correct strike calls (*CorStrRate*) and correct ball calls (*CorBallRate*). Throughout this time—and especially so after the 2009 season—umpires both increased the percentage of correct strikes and correct balls in each year. Prior to the ZE system implementation, umpires were calling strikes on only 76.7 percent of pitches inside the zone. By 2013, the rate of pitches within the zone that were correctly called strikes was 85.7 percent. While pitches outside the zone were more accurate to begin with—likely due to many pitches being clearly balls—the accuracy rate of ball calls on pitches outside the strike zone also increased, eclipsing 90 percent in the data use here.[9] Overall accuracy improved from 85.16 percent in 2008 to 88.99 percent in 2014.

Moving onto the regression estimations, Table 3 presents results for the analysis of accuracy rate improvements and improvements interacted with umpire experience after the

[9] While the Pitch f/x data are the same as used in the ZE system, the strike zone is measured differently here than by the ZE system. The reason for the discrepancy is the two-dimensional representation of the strike zone presented here, while the rulebook zone and ZE measurement is three-dimensional. Umpires are generally required to make a certain percentage of their calls correctly in any given game or be subject to discipline or further developmental requirements under the ZE measured zone, with a two-inch leeway on either side of the plate for a certain percentage of calls (Moore, 2013). Therefore, any percentage of correct calls based on measurements from the two-dimensional zone data is likely to underestimate the absolute performance of umpires as determined by MLB. However, since the Pitch f/x system has been largely unchanged from 2008 through 2014, the *relative* changes in accuracy rates are most useful in the context presented here.

implementation of the ZE system. Models both with and without fixed effects estimate a clear increase in accuracy rates for umpires as a whole during the time that the Zone Evaluation system has been in place, with an accuracy increase of nearly 0.8 percentage points per year, supporting the proposition that evaluation and reflection on relatively precise reports of performance can increase performance over time, and do so across the career lifespan. As there was an already existing monitoring presence, it seems likely that the resultant performance improvements stemmed from the human capital development process predicted from an increase in evaluation and feedback. The steep change in trend after evaluation is robust across model specifications with and without umpire fixed effects, and when including experience as a control variable in the estimation.

For robustness, a quadratic term was used for the time trend effects to test for the possibility of diminishing human capital returns to the evaluation over time, as it is seems feasible that there is a limit on the perceptive performance of umpires when judging the location of a pitch traveling at high speed. However, there was little evidence that—at least during the time period of this data set—there were significant slope changes later on after the implementation of the evaluation system. This does not preclude the possibility of reductions in performance improvements into the future, and it may be worth revisiting at that point. This possibility is particularly important in the context of using these systems at lower levels for training umpires before they arrive at the MLB level, where much of the performance improvements would take place before observing them for the higher level league.

Further, while the generalized existence of an improvement in accuracy after evaluation is robust across umpire experience levels, there are significant differences in the *change* in rate of improvement depending on umpire experience. Specifically, more experienced umpires do not

improve their performance at the same rate as younger umpires. Perhaps most interestingly, this is not due to the newer umpires entering the league with an inferior skillset and more room for improvement than their more experience peers. Rather, less experienced umpires are already superior performers.

This relationship is clearly visualized in Figure 3, with the exhibited diverging slope estimates pooled among umpires with less than 15 years of experience, and umpires with more than 15 years of experience, respectively. As you can see in the figure, less experienced umpires both begin with slightly higher accuracy rates than longer tenured umpires, and also increase their performance during the use of Zone Evaluation more quickly. This is consistent with past literature on learning and performance of newer employees (Lazear, 2000a; Kostiuk & Follmann, 1989), and provides further evidence of heterogeneous effects of evaluation and feedback across employee tenure (Coffey & Maloney, 2010). However, these results also show that despite the heterogeneity in experience, evaluation can spark learning and improvement among employees (umpires) considered mid- or late-career.

The implementation of this feedback loop has seemingly provided clear advantages to MLB as it looked to improve the performance of its umpires, who are already pre-selected as employees that are at the very top of their field. Given the effect that the precise evaluation system has had on these expert employees, it seems reasonable to believe that performance impacts could be even higher for less skilled workers. These workers would be further from the bounds of possible performance levels, which may imply that room for improvement could be rather large.

The differential findings between QuesTec and Zone Evaluation contrast in an interesting way as it relates to the literature on monitoring and evaluation. Specifically, the shift in strike

rate during the QuesTec era, rather than a gradual change, indicates immediate shifts in behavior or effort of umpires, rather than skill development over time. This contrasts with that found when estimating accuracy rate changes during evaluation, where there is evidence of a gradual increase in the skillset of umpires, rather than a pure, one-time behavioral or effort change. These findings reveal an important dichotomy in the context of firms implementing systems to increase employee performance. If the goal is to simply ensure full effort is being given, monitoring may be sufficient in doing so. However, if firms would prefer to increase the ability of their workers, then providing evaluation and feedback could improve outcomes by even more.

5 Summary and Conclusions

This paper adds to the literature by empirically evaluating performance effects among an elite group of professionals—MLB umpires—as they related to technological improvements in monitoring, evaluation, and training. I find that employees (umpires) have substantially changed their ball-strike calling behavior concurrent with various monitoring and training improvements in expected ways. These results are apparent both in terms of increased strike calling and increased accuracy on ball-strike calls. The effects of monitoring and evaluation, while largely described as ways to increase effort, diverge in their impact on employee performance in important ways. Specifically, the evidence presented here identifies three main effects of interest to improving labor productivity: 1) technological innovations in monitoring, paired with additional directives to change behavior, resulted in sudden level shifts in performance and/or effort, 2) implementation of evaluation and training using further technological innovations increased umpire performance gradually over time, presumably through skill development, and 3) there are heterogeneous returns to skill development from evaluation depending on experience

of the employee. In the context of MLB, the league has experienced improvements in the ability of its labor pool of umpires, with newer umpires arriving with near or better performance than their more experienced peers, and improving with evaluation at a faster rate. Each of these findings have important lessons for developing monitoring and evaluation systems in the workplace, particularly when technological shocks decrease the cost and increase the precision of such systems.

It is important to note that this paper investigates only one portion of an umpire's duties on the field. Most recently, in the 2015 season, the league implemented replay and manager challenges within games to assist umpires in making calls at bases and on home runs. Future research would be well served to make use of outcomes with this technology to track rates at which calls are overturned for specific umpires over time to shed light on the heterogeneity in these workers in other areas of their job not addressed here. This could extend the literature beyond the myopic scope here that focuses only on ball-strike calls. The implementation of replay in 2014 has further required the hiring of additional umpires to man the replay booth. It would be worthwhile to continue to track umpire performance to evaluate whether the requirement of a larger pool of labor—with respect to reviewing replay—has decreased average performance among all umpires in the league.

The findings here have important implications beyond the setting of elite professional sports. As many firms attempt to introduce monitoring and evaluation thanks to reduced costs attributable to technological innovations, understanding the impact on employee performance across the lifespan will be an important step in evaluating the benefit of such systems. Further, if precise monitoring and evaluation are relatively low cost, and younger employees increase their abilities more quickly, then the tradeoff for firms in training employees in transferrable skills

could be improved to the point that these systems provide an alternative to otherwise expensive formal training within the workplace. However, given the previously cited psychological and privacy issues related to close monitoring and evaluation of employees, care should be taken in implementation in different settings. Further work would be well served to investigate the relationship between collectively bargaining the use of technology in monitoring and training, and the effectiveness these systems have on the performance of employees relative to when firms force such systems on its workers without mutual agreement.

References

Alcaro, Frederick. 2002. When in doubt, get locked out: A comparison of the 2001 lockout of the National Football League Referees' Association and the failed 1999 resignation scheme of the Major League Baseball Umpires' Association. *University of Pennsylvania Journal of Labor & Employment Law* 5:335-361.

Alder, G. Stoney 2001. Employee reactions to electronic performance monitoring: A consequence of organizational culture. *The Journal of High Technology Management Research* 12:323-342.

Baker, George P. 1992. Incentive contracts and performance measurement. *Journal of Political Economy* 100:598-614.

Barro, Jason R. & Barro, Robert J. (1990). Pay, performance, and turnover of bank CEOs. *Journal of Labor Economics, 8*, 448-481.

Baseball Prospectus. 2014. Custom statistic report: Umpire yearly. http://www.baseballprospectus.com/sortable/index.php?cid=1316050 (accessed January 30, 2014).

Baseball Reference. 2014. Major League Baseball pitches batting. http://www.baseball-reference.com/leagues/MLB/2014-pitches-batting.shtml (Accessed June 20, 2015).

Bulan, Laarni, Sanyal, Paroma & Yan, Zhipeng. 2010. A few bad apples: An analysis of CEO performance pay and firm productivity. *Journal of Economics and Business* 62:273-306.

Callan, Matthew. 2012. Called out: The forgotten baseball umpires strike of 1999. *The Classical*. http://theclassical.org/articles/called-out-the-forgotten-baseball-umpires-strike-of-1999 (accessed April 1, 2014).

Campbell, Stephen M., Reeves, David, Kontopantelis, Evangelos, Sibbald, Bonnie & Roland, Martin. 2009. Effects of pay for performance on the quality of primary care in England. *The New England Journal of Medicine* 361:368-378.

Chass, Murray. 1995. Baseball; Umpires hope a law might be on their side. *The New York Times*. http://www.nytimes.com/1995/04/24/sports/baseball-umpires-hope-a-law-might-be-on-their-side.html (accessed March 25, 2014).

Chass, Murray. 2001. Baseball: Now even umpires are arguing over number of balls and strikes. *The New York Times*. http://www.nytimes.com/2001/07/16/sports/baseball-now-even-umpires-are-arguing-over-number-of-balls-and-strikes.html (accessed July 1, 2015).

Coffey, Bentley & Maloney, M.T. 2010. The thrill of victory: Measuring the incentive to win. *Journal of Labor Economics* 28:87-112.

Dohmen, Thomas & Falk, Armin. 2011. Performance pay and multidimensional sorting: Productivity, preferences, and gender. *The American Economic Review* 101:556-590.

Dohmen, Thomas & Sauermann, Jan. 2015. Referee bias. *Journal of Economic Surveys*. DOI: 10.1111/joes.12106.

Ehrenberg, Ronald G. & Bognanno, Michael L. 1990. The incentive effects of tournaments revisited: Evidence from the European PGA Tour. *Industrial and Labor Relations Review* 43:74S-88S.

Fernie, Sue & Metcalf, David. 1999. It's not what you pay it's the way that you pay it and that's what gets results: Jockeys' pay and performance. *Labour* 13:385-411.

Franceschelli, Ignacio, Galiani, Sebastian & Gulmez, Eduardo. 2010. Performance pay and productivity of low-and-high-ability workers. *Labour Economics* 17:317-322.

Fryer, Ronald G. 2013. Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics* 31:373-407.

Garicano, Luis, Palacios-Huerta, Ignacio & Prendergast, Candice. 2005. Favoritism under social pressure. *The Review of Economics and Statistics* 87:208-216.

Goodman, Sarena F. & Turner, Lesley J. 2013. The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program. *Journal of Labor Economics* 31:409-420.

Green, Etan & Daniels, David P. 2014. Impact aversion: Agency failure and decision bias at high stakes. *SSRN Working Paper*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id =2391558.

Jackson, C. Kirabo, Rockoff, Jonah E., & Staiger, Douglas O. 2014. Teacher effects and teacher-related policies. *Annual Review of Economics* 6:801-825.

Jensen, Michael C. & Murphy, Kevin J. 1990. Performance pay and top-management incentives. *Journal of Political Economy* 98:225-264.

Kahn, Lawrence M. 2000. The sports business as a labor market laboratory. *The Journal of Economic Perspectives* 14:75-94.

Keh, Andrew. 2012. For umpiring school, a staff party proves costly. *The New York Times*. http://www.nytimes.com/2012/02/10/sports/baseball/umpiring-school-loses-baseball-relationship-over-behavior-at-party.html (accessed June 10, 2014).

Kim, Jerry W. & King, Brayden G. 2014. Seeing stars: Matthew effects and status bias in Major League Baseball umpiring. *Management Science*. DOI: http://dx.doi.org/10.1287/mnsc.2014.1967.

Kostiuk, Peter F. & Follmann, Dean A. 1989. Learning curves, personal characteristics, and job performance. *Journal of Labor Economics* 7:129-146.

Lazear, Edward P. 2000a. Performance pay and productivity. *American Economic Review* 90:1346-1361.

Lazear, Edward P. 2000b. The power of incentives. *American Economic Review Papers and Proceedings* 90:410-414.

Lazear, Edward P. 2000c. The future of personnel economics. *The Economic Journal* 110:F611-F639.

Lindenauer, Peter K., Remus, Denise, Roman, Sheila, Rothberg, Michael B., Benjamin, Evan M., Ma, Allen, & Bratzler, Dale W. 2007. Public reporting and pay for performance in hospital quality improvement. *The New England Journal of Medicine* 356:486-496.

Lopez, Michael J. & Snyder, Kevin. 2013. Biased impartiality among national hockey league referees. *International Journal of Sport Finance* 8:208-223.

Lucifora, Claudio & Origo, Federica. 2015. Performance-related pay and firm productivity: Evidence from a reform in the structure of collective bargaining. *Industrial and Labor Relations Review* 68:606-632.

Mills, Brian M. 2013. Social pressure at the plate: Inequality aversion, status, and mere exposure. *Managerial and Decision Economics* 35:387-403.

MLB. 2010. Major League Baseball: Official baseball rules.

Moore, Matt. 2013. *Baseball balls & Strikes: Every pitch counts*. Referee Enterprises, Inc.: Franksville, WI.

Moskowitz, Tobias J. & Wertheim, L. Jon. 2011. *Scorecasting: The Hidden Influences behind How Sports are Played and Games are Won*. New York: Crown Archetype.

NASA. 2000. Human Integration Design Handbook. http://msis.jsc.nasa.gov/sections /section03.htm (accessed February 4, 2014).

Nevill, Alan M., Balmer, Nigel J., & Williams, A. Mark. 2002. The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise* 3:261-272.

O'Neill, Dan. 1990. Umpires are victimized by lockout, too. *Chicago Tribune*. http://articles. chicagotribune.com/1990-03-18/sports/9001230580_1_umpires-spring-training-lockout-dave-phillips (accessed March 20, 2014).

Paarsch, Harry J. & Shearer, Bruce S. 1999. The response of worker effort to piece rates: Evidence from the British Columbia tree-planting industry. *The Journal of Human Resources* 34:643-667.

Parsons, Christopher A., Sulaeman, Johan, Yates, Michael C., & Hammermesh, Daniel S. 2011. Strike three: Discrimination, incentives, and evaluation. *The American Economic Review* 101:1410-1435.

Podgursky, Michael J. & Springer, Matthew G. 2007. Teacher performance pay: A review. *Journal of Policy Analysis and Management* 26:909-949.

Prendergast, Candice. 1999. The provision of incentives in firms. *Journal of Economic Literature* 37:7-63.

Price, Joseph. & Wolfers, Justin. 2010. Racial discrimination among NBA referees. *Quarterly Journal of Economics* 125:1859-1887.

Price, Joseph., Remer, Marc, & Stone, Daniel F. 2012. Subperfect game: Profitable biases of NBA referees. *Journal of Economics and Management Strategy* 21:271-300.

Rockoff, Jonah E., Staiger, Douglas O., Kane, Thomas J., & Taylor, Eric S. 2012. Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review* 102:3184-3213.

Schwarz, Alan. 2009. Ball-strike monitor may reopen wounds. *The New York Times*. http://www.nytimes.com/2009/04/01/sports/baseball/01umpires. html?_r=1& (accessed April 1, 2014).

Simmons, Rob & Berri, David J. 2011. Mixing the princes and the paupers: Pay and performance in the National Basketball Association. *Labour Economics* 18:381-388.

Stanton, Jeffrey M. 2009. Reactions to employee performance monitoring: Framework, review, and research directions. *Human Performance* 13:85-113.

Taylor, Eric S. & Tyler, John H. 2012. The effect of evaluation on teacher performance. *American Economic Review* 102:3628-3651.

Tainsky, Scott, Mills, Brian M., & Winfree, Jason A. 2015. An examination of potential discrimination among MLB umpires. *Journal of Sports Economics*, *16*, 353-374.

Woessmann, Ludger. 2011. Cross-country evidence on teacher performance pay. *Economics of Education Review* 30:404-418.

Zitzewitz, E. 2012. Forensic economics. *Journal of Economic Literature* 50:731-769.

Table 1: Summary of Dependent Variables

Year	Umps	Pitches	Called Pitches	*TotalStrikeRate*	*wCV*	*CalledStrikeRate*	*CorStrRate*	*CorBallRate*	*AccuracyRate*
1997	88	621,672	331,559	61.134	1.559	29.541	----------	----------	----------
1998	87	668,253	380,912	61.351	1.436	29.746	----------	----------	----------
1999	103	712,976	390,325	60.778	1.554	29.236	----------	----------	----------
2000	84	729,814	394,113	60.764	1.499	29.402	----------	----------	----------
2001	88	711,322	377,353	62.234	1.475	30.976	----------	----------	----------
2002	84	711,002	378,061	62.033	1.456	30.612	----------	----------	----------
2003	84	714,789	382,123	62.232	1.438	31.250	----------	----------	----------
2004	84	723,273	386,946	62.082	1.344	31.046	----------	----------	----------
2005	85	708,797	377,011	62.652	1.279	31.629	----------	----------	----------
2006	83	720,657	383,780	62.330	1.280	31.195	----------	----------	----------
2007	85	725,418	387,388	62.273	1.290	31.307	----------	----------	----------
2008	83	727,729	389,376	62.206	1.299	31.313	78.691	88.445	85.353
2009	85	731,396	394,793	62.077	1.232	31.743	78.721	88.818	85.504
2010	85	725,140	390,651	62.427	1.183	32.307	79.994	88.923	85.965
2011	83	718,796	384,339	62.619	1.195	32.250	81.421	89.123	86.606
2012	82	715,687	382,647	62.820	1.171	32.632	83.688	89.394	87.538
2013	82	718,733	383,217	62.948	1.014	32.500	85.740	90.021	88.651
2014	90	713,047	378,124	63.318	1.132	32.788	87.177	89.850	88.998

TABLE 2: QuesTec Panel DiD Estimation

	TotalStrikeRate	***TotalStrikeRate***	***CalledStrikeRate***	***CalledStrikeRate***
Umpire F.E.	No	Yes	No	Yes
Pre-QuesTec				
Q_1	60.897***	60.931***	29.482***	28.625***
	(0.116)	*(0.115)*	*(0.183)*	*(0.178)*
Z_1	-0.213***	-0.221***	-0.132**	-0.125**
	(0.038)	*(0.039)*	*(0.060)*	*(0.060)*
Post-Questec, Installed				
Q_2	62.628***	62.659***	31.672***	30.803***
	(0.145)	*(0.148)*	*(0.191)*	*(0.185)*
Z_2	0.025	0.017	0.028	0.036
	(0.036)	*(0.039)*	*(0.046)*	*(0.049)*
Post-Questec, Not Installed				
Q_3	62.489***	62.520***	31.187***	30.333***
	(0.090)	*(0.088)*	*(0.139)*	*(0.114)*
Z_3	0.050**	0.042*	0.108***	0.113***
	(0.021)	*(0.023)*	*(0.029)*	*(0.030)*
Obs.	1,347	1,347	1,347	1,347
$\mathbf{R^2}$	0.233	0.473	0.225	0.574
Wald Test (*t*)				
$(\beta_2 - \beta_1)$	12.16***	11.53***	10.62***	2.08***
$(\beta_3 - \beta_1)$	14.10***	13.27***	10.86***	3.44***
$(\beta_3 - \beta_2)$	-1.07	-1.01	-3.14***	-2.86***
$(\alpha_2 - \alpha_1)$	4.20***	4.06***	2.04**	9.87***
$(\alpha_3 - \alpha_1)$	6.00***	5.82***	3.47***	10.04***
$(\alpha_3 - \alpha_2)$	0.65	0.60	1.74*	1.56

Notes: Estimation includes umpire fixed effects and standard errors clustered by umpire, and observations are weighted by $\sqrt{n_{i,t}}$.

FIGURE 1: QuesTec Panel DiD Visualization

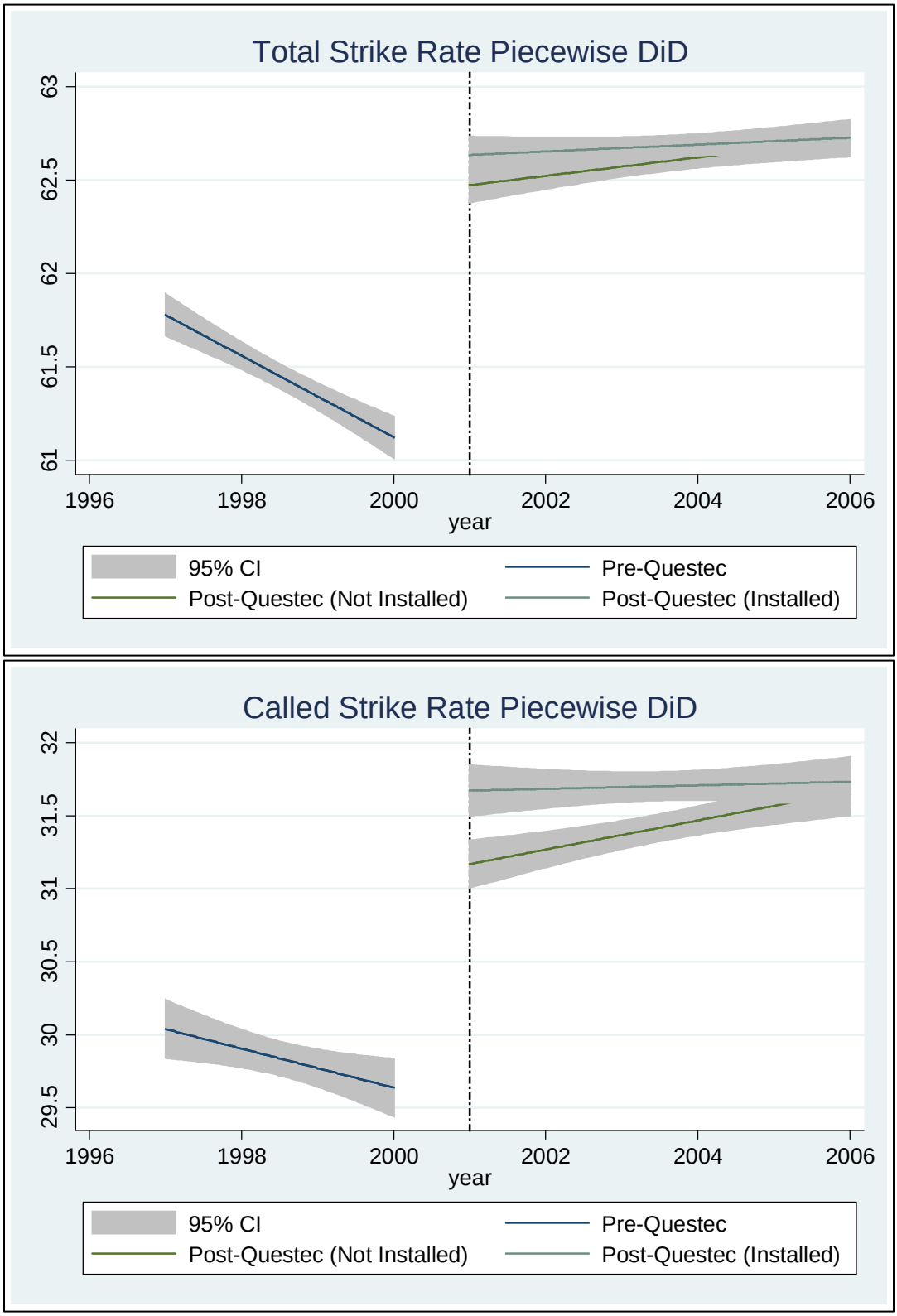

FIGURE 2: Exhibition of Correct Call and Incorrect Call Classifications

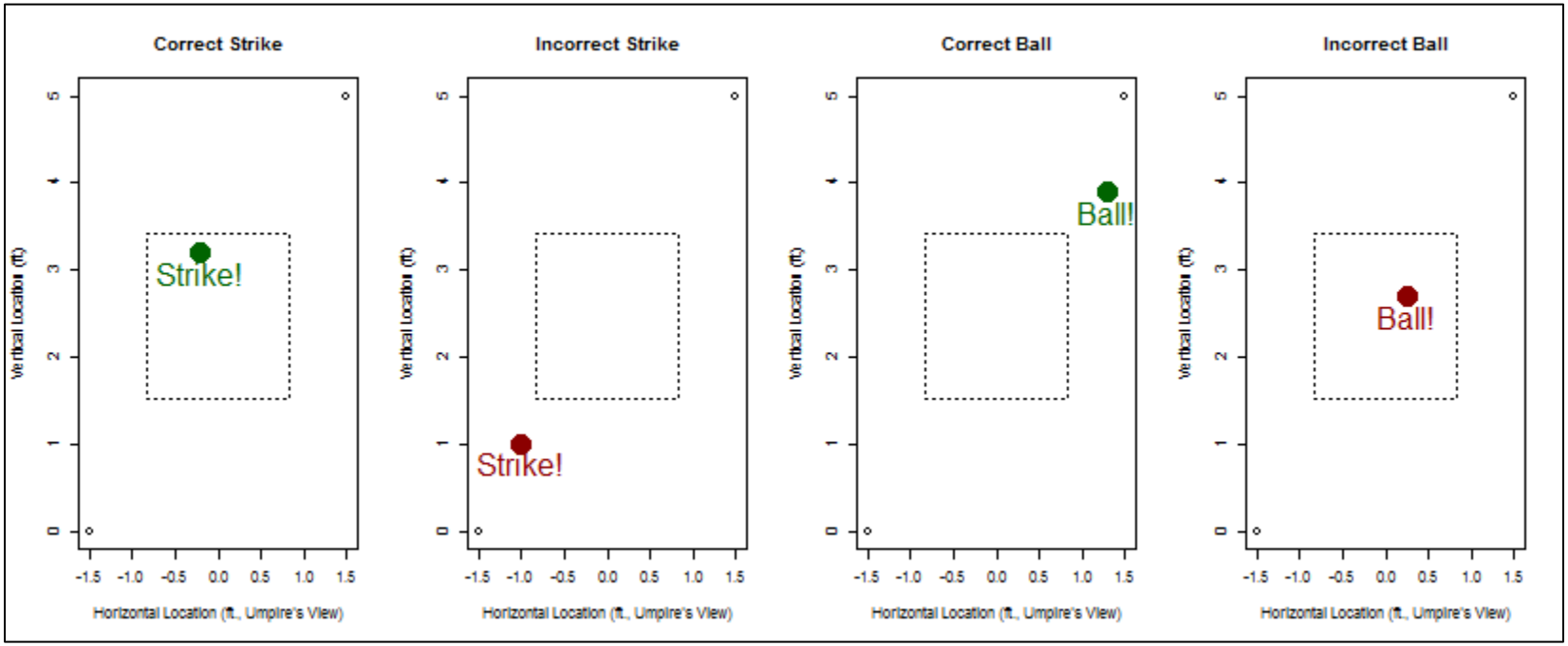

TABLE 3: Accuracy Rate Changes by Zone Evaluation Condition

	AccuracyRate	***AccuracyRate***	***AccuracyRate***	***AccuracyRate***	***AccuracyRate***	***AccuracyRate***
Umpire F.E.	No	Yes	No	Yes	No	Yes
Intercept Pre-ZE	85.344***	86.071***	85.734***	86.020***	85.546***	85.727***
	(0.111)	*(0.116)*	*(0.145)*	*(0.110)*	*(0.188)*	*(0.152)*
Trend Pre-Zone ZE	0.134	0.127	0.153	0.152	0.144	0.146
	(0.121)	*(0.135)*	*(0.122)*	*(0.133)*	*(0.123)*	*(0.133)*
Δ Intercept Post-ZE	0.317	0.294	0.298	0.294	0.309	0.297
	(0.205)	*(0.229)*	*(0.206)*	*(0.229)*	*(0.208)*	*(0.229)*
Δ Trend Post-ZE	0.673***	0.642***	0.651***	0.642***	0.787***	0.781***
	(0.125)	*(0.136)*	*(0.124)*	*(0.136)*	*(0.145)*	*(0.152)*
Experience	----------	----------	-0.026**	-0.025***	-0.014	-0.017***
	----------	----------	*(0.010)*	*(0.003)*	*(0.013)*	*(0.004)*
Exp*Post-ZE Trend	----------	----------	----------	----------	-0.008**	-0.008**
	----------	----------	----------	----------	*(0.003)*	*(0.003)*
Obs.	536	536	536	536	536	536
R^2	0.631	0.858	0.646	0.858	0.650	0.861

***, **, * refer to statistical significance at the 1%, 5%, and 10% levels, respectively. Panel regressions weighted by $\sqrt{n_{i,t}}$ and limited to observations where the umpire called at least 1,500 pitches in the given season (approximately 10 games). Umpires that worked 1 year or less during this sample were removed from the analysis.

FIGURE 3: Accuracy Rate Trend Change Visualization