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 performance-enhancing role of changes in monitoring, evaluation, and training technology in an experience-diverse expert workforce: Major League Baseball umpires. I find improvements in job performance across the league that is consistent with the timing of innovations in monitoring technology and later improvements in training and feedback. Performance improvements associated with increased training and immediate feedback are more pronounced for younger umpires. Implications for the differential effects of evaluation and feedback across experience levels of firm employees 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, the participants at the 2014 European Conference on Sports Economics in Antwerp, and the anonymous referees of this paper 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 used here.

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

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 an empirical 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 that experienced identifiable changes in the use of monitoring and evaluation throughout their labor history with MLB, providing a useful context for examining the effects of these activities on measureable performance output. Using pitch-level data on ball and strike calls, I find evidence that the use of monitoring and evaluation substantially improved umpire performance when implemented across the league.

Changes in umpire behavior took place at two separate points where new monitoring and evaluation was introduced. The observed changes are consistent with publicly mandated league expectations with respect to the size and shape of the MLB Rulebook strike zone when monitored and evaluated. Under the initial use of technology to monitor ball-strike calls in 2001, there was a one-time upward shift in performance, with similar impacts across the entire distribution of umpires. After the addition of more immediate evaluation and feedback from a more accurate monitoring system in 2009, performance improved continually in each year. However, this effect was heterogeneous depending on the experience level of the umpire.

Specifically, umpires with less experience benefited more from the implementation of the new evaluation system. These findings have important implications for the use and implementation of monitoring, evaluation, and training systems in the context of an experience-diverse workforce, and provide useful guidance related to how these systems may be implemented to maximize sustained agent improvement.

In the next section, I describe the general monitoring and evaluation setting in addition to the MLB umpire labor market context, introducing the points at which policy changes would be expected to impact umpire behavior. Section 3 presents the data, empirical specification, and results from analysis of the introduction of a new monitoring system, called QuesTec. Section 4 does the same for an improved evaluation and feedback system, known as Zone Evaluation. I contrast these outcomes in Section 5 and provide concluding comments and suggestions for future inquiry.

2 Literature and Contextual Setting

2.1 Monitoring and Evaluation in the Workplace

Implementing incentive systems within the workplace to increase employee performance and induce sorting of the labor pool is a central issue in personnel economics (Baker, 1992; Paarsch & Shearer, 1999; Prendergast, 1999; Lazear, 2000a; Lazear, 2000b). There has been substantial effort put forth in designing and evaluating performance pay systems, with empirical evaluation in the economics literature taking place across a number of industries and in education (Duflo, Hanna, & Ryan, 2012; Lavy, 2009; Podgursky & Springer, 2007). The various strategies used to align incentives between principals and agents are often implemented where strict monitoring of employee effort is cost prohibitive, and usually depend on limited monitoring paired with pay for performance or termination decisions.

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. 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), implying a sudden change in effort in the face of monitoring from cost increase associated with the monitoring itself. While tying monitoring and performance measurement to punitive outcomes may provide short-term improvement (Lucifora & Origo, 2015), the human resources literature has suggested that this practice could undermine long-run employee performance improvements 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). This contrasts with feedback simply for appraisal purposes that could have deleterious effects on employees continuing to improve performance (Silverman, Pogson, & Cober, 2005). Therefore, improvements in technology that increase both the precision and continuity of evaluation and feedback may result in performance improvements beyond a short term structural shift at the initial introduction of monitoring.

Specifically, 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 limited empirical evaluation of the long-run developmental effects of evaluation and feedback itself on employee performance, or whether these systems have gradual improvements in human capital development rather than a sudden, discontinuous change in performance that might be expected from monitoring alone. Further, while Lazear (2000a) finds evidence of impacts of experience on performance levels, it is possible that employees with different experience levels have heterogeneous human capital development responses to various monitoring and evaluation mechanisms. If younger workers are more (less) willing to develop skills expected by the employer, then this could result in a diverse response to incentive and feedback changes across the workforce that would be relevant to optimal implementation of new performance measurement systems.

2.2 Major League Baseball as a Labor Laboratory

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). And, 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 (for a full review, see Dohmen &

Sauermann [2015]). However, there are a number of characteristics of the umpires' labor market beyond social biases that make this context valuable for study of other economic phenomena.

First, umpires operate under a monopsonistic buyer of labor like 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.

Second, umpires in MLB are effectively superstars within their profession, but with ability levels that are somewhat heterogeneous (Rosen, 1981; 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 resulting from 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).

Third, 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 manual methods. These systems were put in place alongside an expected minimum

[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.

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. Any changes in performance would then suggest that MLB umpires could be considered motivated agents who are willing to act positively on information about their performance, an important characteristic noted in Taylor and Tyler (2012).

As with most team sports, the legitimacy of competition in MLB requires that its officials 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. 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. While umpires serve behind the plate in only one quarter of the games in which they work, ball-strike calls are the most salient decisions they make. These calls often lead to complaints and disagreements from players and coaches, and present the possibility

[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.

of strategic behavior on the part of these players. Therefore, these decisions are likely to have the most impact on game outcomes and are at the center of this inquiry.

2.3 The Baseball Umpires' Labor Market: A Short History

Although umpires are trained in the minor leagues by unaffiliated schools and organizations, I focus on performance and monitoring at the MLB level, rather than the specifics of earlier training practices. There are two recent changes to monitoring and evaluation relevant to this work. The first was an implementation of limited technological monitoring of the accuracy of ball-strike calls in 2001 alongside a directive for umpires to call more (higher) strikes and limit the total pitches in games. The second was an agreement to improve the monitoring system in 2009 with new ball-tracking technology, which would immediately provide visual accuracy reports directly to umpires after every game they served behind the plate.

In 1996, the league changed the official definition of a strike that has remained through 2014, which included expansion of the zone relative to its previous definition. The commissioner's office issued a memo shortly thereafter calling for uniform enforcement of the strike zone. Perhaps most importantly, the memo noted that the strike zone should be called from two inches above the belt to the bottom of the knees, and directed team officials to manually chart pitches for umpires working their games. There was publicly noted resistance from umpire union leaders and limited indication of any changes to umpire behavior at this time.

While the definition has remained since 1996, the league further mandated that umpires call more strikes on high pitches prior to the 2001 season. This coincided with the installation of a new pitch tracking technology called QuesTec in three of its parks to evaluate strike zones of umpires that would more precisely evaluate the accuracy of umpires than manual charting by teams. This technology was later expanded to about one third of MLB parks without the agreement of umpires for its use in termination or discipline decisions. However, its use to evaluate umpires was finally agreed upon after a grievance from the umpires' union was settled in 2004 (Schwarz, 2009). As shown in Rader and Wilson (2008), there was a sudden shift in offensive levels at the time of the mandate to expand the strike zone alongside the introduction of QuesTec monitoring in 2001. There is also evidence that this had much to do with an increase in the rate at which umpires called strikes (an expanded strike zone). I examine the changes to umpire behavior more closely in this work.

Subsequently, in 2007, MLB installed new technology in its parks halfway through the season, known colloquially as Pitch f/x. This system monitored every pitch as it cross the plate and recorded each call by the umpire in the event that the batter did not swing. The data produced by this system have been publicly available since 2007, and it was fully installed in all parks by the start of the 2008 season. However, in 2007 and 2008, these data were not used to explicitly evaluate umpires by the league.

This changed in 2009, as the umpires' union and MLB agreed on new terms in collective bargaining. The new agreement allowed this new system to replace QuesTec, and included a host of expectations for minimum performance levels and feedback for the continued training of umpires. The new evaluation and feedback system, called Zone Evaluation (ZE), used the Pitch f/x technology, which is claimed to be more accurate (Schwarz, 2009). Under this new system, umpires receive more detailed and accessible reports of their performance after every game. 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 detailed 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. This data was presented to umpires immediately after each game, with a minimum expected accuracy rate to maintain to avoid any disciplinary action by the league.

The remainder of this work focuses specifically on changes associated with the implementation of QuesTec and Zone Evaluation. I address how these monitoring and feedback systems differentially impacted umpires’ willingness to conform to league expectations regarding strike calls and zone accuracy.

3 QuesTec and Umpire Behavior

3.1 Data and Measurement

This section 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 game-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).

The measures developed for this portion of the analysis are proxies for umpire performance. Because locational data is not available until the use of the Pitch f/x system—QuesTec data are and have always been kept private—I assume increases in the rate of strikes can proxy umpires’ willingness to conform with the specific 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*$_{ijt}$ and *CalledStrikeRate*$_{ijt}$.[4] *TotalStrikeRate*$_{ijt}$ is calculated by dividing the total number of strikes by the total number of pitches during which umpire *i* worked behind the plate for game *j* in season *t*. 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. For robustness, I also calculate *CalledStrikeRate*$_{ijt}$, using 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*$_{ijt}$ 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 somewhat 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.

3.2 Estimation Procedure

[4] These tabulations come directly from data used in Tainsky, Mills, & Winfree (2015).

I use a difference-in-differences (DiD) estimation to identify changes in umpire strike rates attributable specifically to QuesTec monitoring starting in the 2001 season. The key indicators in this estimation include a dummy variable identifying the existence of QuesTec overall (equal to one for all observations in 2001 and later, *QuesTecExist*) and a dummy variable identifying whether the specific stadium, *j*, has QuesTec installed in year *t* (*QuesTecInstall*). I also test for the possibility of heterogeneous impacts of monitoring across experience levels with the *Experience* variable, equal to one in the first year of an umpire's MLB career. Estimates include yearly effects and fixed or random effects for umpire and stadium. Observations are at the umpire-stadium-season level. The regression estimation takes the following form:

$$\begin{aligned} Y_{ijt} = \beta_0 + \beta_1 QuesTecExist_t + \beta_2 QuesTecInstall_{jt} + \beta_3 \ln(Experience)_{it} \\ + \sum_{t=1998}^{2000} [\beta_{t+3-1997} \mathrm{I}(PreQuesTecYear_t = t)] \\ + \sum_{t=2002}^{2006} [\beta_{t+6-2001} \mathrm{I}(PostQuesTecYear_t = t)] \\ + \beta_{12}[\ln(Experience)_{it} \times QuesTecExist_t] + \delta_i + \gamma_j + \varepsilon_{ijt}. \end{aligned}$$

Where Y_{ijt} is one of *TotalStrikeRate* or *CalledStrikeRate* for umpire *i* in stadium *j* in season *t*. The *QuesTecInstall* indicator is equal to one if stadium *j* has the QuesTec monitoring system installed in the given season, *t*, and zero otherwise. δ_i and γ_j refer to umpire and stadium-specific effects, and ε_{ijt} is the umpire-stadium-year specific error term. Standard errors are robust to individual umpire clustering and heteroscedasticity. The natural log of *Experience* is used to account for any non-linear effects of umpire experience.

It is important to note that two of the yearly effects are dropped from this estimation. The 1997 effect is used as the base year in the standard fashion to avoid multicollinearity. But, because of the inclusion of the explicit *QuesTecExist* variable, the 2001 season effect is also excluded. Each yearly effect estimate after 2001 is therefore the effect size relative to the *QuesTecExist* coefficient estimate.

3.3 Results of QuesTec Estimation

The results of the DiD estimation for QuesTec and non-QuesTec stadiums can be found in Table 3. Beginning with *TotalStrikeRate* (Columns 1 and 2), there was a clear upward shift concurrent with the initial installation of QuesTec in MLB and demand for a larger strike zone, as represented by the coefficient for the *QuestecExist* variable. This is presented visually in Figure 1 (top panel), where the discontinuity can be seen clearly in 2001. Most importantly, there was an additional upward shift specific to stadiums that were installed with the QuesTec monitoring technology relative to those that were not. There was no evidence that more experienced umpires had lower (higher) *TotalStrikeRate*, nor was there evidence of heterogeneous effects of monitoring across the umpire experience distribution using this measure.

Interestingly, the *CalledStrikeRate* model results (Table 3, Columns 3 and 4) differ from *TotalStrikeRate* in two key ways. First, while there was a large shift at the time of QuesTec implementation, the level shift specific to parks fitted with the technology was actually *smaller* than those without. This is visualized in Figure 1 (bottom panel). Further, while was no apparent generalized impact of experience level on the rate of called strikes, there is some limited evidence that more experienced umpires changed their behavior by less than younger umpires after the implementation of monitoring.

The seemingly divergent results for *TotalStrikeRate* and *CalledStrikeRate* could point to additional complexities in the relationship between umpire *called* strikes and strategic responses by batters and pitchers. If batters know that umpires are more likely to call pitches strikes in parks where QuesTec is installed, then they may be more likely to swing at those pitches (and miss more often). While the umpires would therefore not be *calling* more strikes in these parks, the total number of strikes in these games could increase.

Despite this contrast, there is evidence that *TotalStrikeRate* and *CalledStrikeRate* in parks with and without QuesTec converged toward one another in the QuesTec era. Specifically, *TotakStrikeRates* were consistently lower in the pre-QuesTec era in parks that would later be installed with QuesTec (and *CalledStrikeRates* were consistently higher). After the use of monitoring, these rates were more similar across parks, implying higher levels of consistency in the strike zone. If umpires are acclimatizing across conditions, this observed increase in consistency in should be expected, and therefore some of these changes may be attributed to technology induced behavioral changes.

Ultimately, the difference in sudden shifts across the conditions identifies the relative impacts of generalized changes plausibly stemming from both the directive from the league to call more strikes and the effect of technological innovation in monitoring. Given the DiD estimation results, it would be aggressive to attribute the entire shift in strike rates in 2001 to the installation of monitoring technology, though there seems to be strong evidence of effects at its introduction with large generalized increases in both *TotalStrikeRate* and *CalledStrikeRate* in 2001.

4 Zone Evaluation and Umpire Performance

4.1 Data and Measurement

This portion of the empirical estimation uses detailed locational data on umpire ball-strike call accuracy just prior to, and after, the introduction of the Zone Evaluation System in 2009. I use information on pitch location available from the Pitch f/x system for every pitch in the regular season from 2008 through 2014 (Willman, 2015). The availability of these data allows a direct measurement of performance unlike the implied proxies calculated from earlier data samples. I also use these data to heterogeneous effects of feedback and evaluation 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 proxies that could be influenced by pitchers changing the location of their pitches over time, batters swinging more (less) often, or other factors. To measure changes specific to umpire behavior, 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 a definition 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. This can be thought of as a true positive. The second pitch type is a correctly called ball. 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—a called strike on a pitch that does not cross through the strike

[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. Models were estimated on a subset of the data that did have height available, and no substantive differences in results were found. These are available upon request from the author.

zone plane) and incorrectly called balls (a false negative—called balls that do cross through the strike zone plane). The proportion of correct strikes to true strikes can be thought of as a measure of sensitivity, while the proportion of correct balls to true balls can be thought of as a specificity measure. The changes in correct strike and correct ball rates, as well as overall correct call rates, are reported in Table 1 from 2008 through 2014. 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.

4.2 Estimation Procedure

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 with random or fixed effects at the umpire level.[8] These models are estimated with a time trend variable and ZE-presence specific time trend, as well as with a yearly indicator in lieu of a time trend to account for any non-linear effects after the implementation of ZE. 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 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).

In the linear time trend models, the ZE era is identified as beginning in 2010. This allows for a one-year delay in the expected improvement from evaluation, feedback, and learning with

[8] Heteroskedasticity related to the number of pitch-level observations was tested using the modified Breusch-Pagan test (suggested by Solon, Haider, & Wooldridge, 2013). This test indicated that there were no apparent heteroscedasticity issues from sample size differences across umpires. I therefore proceed with unweighted estimates, using standard errors robust to heteroscedasticity and clustering within umpire.

Zone Evaluation. For example, if training first takes place in time t, then we might 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:

$$AccRate_{it} = \beta_0 + \beta_1 T_t + \beta_2 D_{ZE} + \beta_3 D_{ZE}\tau_{ZE} + \delta_i + \varepsilon_{it}$$

And with experience included:

$$AccRate_{it} = \beta_0 + \beta_1 T_t + \beta_2 D_{ZE} + \beta_3 \ln(Exp_{it}) + \beta_4 D_{ZE}\tau_{ZE} + \beta_5 D_{ZE}\tau_{ZE} \ln(Exp_{it}) + \delta_i + \varepsilon_{it}$$

$AccuracyRate_{it}$ 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. ε_{it} is the individual error for each observation in the panel. δ_i refers to umpire effects, which are fixed effects when experience is excluded and random effects when experience is included.[9]

The yearly effects model, in lieu of the ZE-specific time trends, takes the form:

[9] Umpire fixed effects in the models that do not include experience establish that the performance improvement is not due to new entry of umpires into the league, and actual increases in accuracy for individual umpires. However, the inclusion of logged experience makes estimation difficult due to the collinearity with the yearly effects in the model. Therefore, random effects are preferred in models that include experience as a right hand side variable. These choices are confirmed by the use of a Hausman specification test, the results of which are available upon request from the author.

$$AccRate_{it} = \beta_0 + \sum_{t=2009}^{2014} [\beta_{t-2008}\mathrm{I}(T_t = t)] + \delta_i + \varepsilon_{it}$$

And with experience included:

$$AccRate_{it} = \beta_0 + \sum_{t=2009}^{2014} [\beta_{t-2008}\mathrm{I}(T_t = t)] + \beta_7 \ln(Exp_{it}) + \sum_{t=2009}^{2014} [\beta_{t+7-2008}\mathrm{I}(T_t = t) * \ln(Exp_{it})] + \delta_i + \varepsilon_{it}.$$

I include the number of years of experience for each umpire in each season as an added variable to test for heterogeneous effects of accuracy improvement for umpires across the distribution of experience in the league. The experience variable is calculated as in the previous QuesTec analysis, and the natural log of this variable is used in the estimation. This tests predictions of Lazear (2000a) as to whether newer employees begin at a lower (higher) 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.

4.3 Results of ZE Estimation

Table 1 presents a summary of 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*).[10] 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 2014, the rate of pitches within the zone that were correctly called strikes was more than 87 percent. While pitches outside the zone are generally 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 based on the strike zone definition used here. Overall accuracy improved from just over 85 percent in 2008 to 89 percent in 2014.

Table 4 presents results for the regression estimations of accuracy rate improvements on ZE presence and umpire experience. Both estimations show stark increases in accuracy rates for umpires as a whole during the time that the Zone Evaluation system has been in place. The increase in accuracy was estimated to be approximately 0.63 to 0.98 percentage points per year, supporting the proposition that evaluation and reflection on relatively precise reports of performance could have positive effects on performance. While the percentage point change sounds relatively small in absolute value, it is important to note that the 2008 to 2014 change is enough to put a 2008 umpire in the 5th percentile (83.63) above the 95th percentile umpire

[10] 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 used in this work, whereas 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 (Moore, 2013). Therefore, any percentage of correct calls based on measurements from the two-dimensional zone data may 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.

performance from 2008 (87.05). Essentially, this 5 year change resulted in nearly all umpires in 2014 being as accurate, or more accurate, than 95% of the 2008 umpires.

As there was an already existing monitoring presence (QuesTec) prior to ZE implementation, it seems likely that the resulting 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, and is rather consistent with a largely linear effect after 2009, with the exception of the most recent year (2014). Performance improvements could have some limitations due to human error, and umpires may be starting to reach peak performance on their task. However, more data is needed to evaluate whether there are clear diminishing returns to the evaluation system beyond 2014.

Interestingly, while the existence of an improvement in accuracy after evaluation is robust across umpire experience levels, there is experience-related heterogeneity in the *change* in rate of improvement. Specifically, younger umpires have experienced larger gains in accuracy than more experienced ones. Perhaps most interestingly, this is not due to the newer umpires entering the league with a substantially inferior skillset and more room for improvement than their more experience peers. There is evidence that they enter the league with very similar skill levels, or at least only slightly worse than their more experienced counterparts.

This contrasts with prior literature (Lazear, 2000a), finding less experienced workers arrive with much lower productivity. But it is important to note that rising to the MLB umpire labor market is already extremely competitive, and requires years of development at the minor league level. Ultimately, those chosen to advance to the highest level are more likely to be at the extremes of the skill distribution. Further, given that the ball-strike calling task requires rather specific visual acuity, younger umpires may have an advantage physiologically, or at least the opportunity to take advantage of any disparities with additional training through Zone Evaluation and learn more quickly. While this result could be specific to physically intensive jobs, the tenure-like structure of the league could simply be driving the disparity in results: younger umpires often must wait a few years before receiving a permanent contract, while older umpires are heavily protected by union agreements on contracts with MLB.

The experience-performance relationship is clearly visualized in Figure 2. Notice the diverging slope estimates for the groups of umpires pooled for those with less than 15 years of experience, and those with more than 15 years of experience, respectively. The top panel presents fitted results from the yearly trend model, while the bottom panel presents the yearly effects model overlaid with a fitted median spline for comparison of the experience related differences across time. Less experienced umpires increase their performance during the use of Zone Evaluation more steeply. The learning effect is consistent with past literature on learning and performance of newer employees (Lazear, 2000a; Kostiuk & Follmann, 1989), and provides further evidence of heterogeneous impacts of evaluation and feedback across the employee tenure distribution (Coffey & Maloney, 2010). However, these results also show that despite the heterogeneity in experience, evaluation can spark learning and improvement among even those employees (umpires) considered mid- or late-career.

The implementation of this feedback loop has seemingly provided 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.

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, but different, ways.

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. First, technological innovations in monitoring, paired with additional directives to change behavior, were associated with sudden level shifts in performance or effort. Second, the implementation of evaluation and training using further technological innovations increased umpire performance gradually over time, presumably through skill development. And third, there are heterogeneous returns to skill development from evaluation and training depending on experience of the employee. 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.

These findings reveal an important dichotomy in the context of firms implementing systems to increase employee performance beyond the setting of professional sports. If the goal is to simply ensure that employees provide higher effort, monitoring may be sufficient in doing so. However, if firms would prefer to increase the ability of their workers, then providing immediate evaluation and feedback could improve outcomes by even more.

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 career 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 (or external to) the workplace.

Improvements in measurement accuracy could be particularly relevant if especially strong incentive mechanisms in the workplace become de-motivating. The de-motivation may be mitigated by increased measurement accuracy, particularly if the legitimacy of the performance measure is accepted by agents themselves. Further inquiry into measurement accuracy and employee performance may shed light on this relationship.

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.

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).

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

Duflo, Esther, Hanna, Rema, & Ryan, Stephen P. 2012. Incentives work: getting teachers to come to school. *American Economic Review* 102:1241-1278.

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.

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

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).

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

Lavy, Victor. 2009. Performance pay and teachers effort, productivity, and grading ethics. *American Economic Review* 99:1979-2011.

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.

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).

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.

Rader, Benjamin G. & Winkle, Kenneth J. Baseball's great hitting barrage of the 1990s (and beyond) reexamined. *NINE: A Journal of Baseball History and Culture* 17:70-96.

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.

Rosen, Sherwin. 1981. The economics of superstars. *American Economic Review* 71:845-858.

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).

Silverman, Stanley B., Pogson, Corrie E., & Cober, Alana B. 2005. When employees at work don't get it: A model for enhancing individual employee change in response to performance feedback. *The Academy of Management Executive* 19:135-147.

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.

Solon, Gary, Haider, Steven J., & Wooldridge, Jeffrey M. 2013. What are we weighting for? *Journal of Human Resources* 50:301-316.

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

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

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

Willman, Daren. 2014. Pitchf/x Search. *Baseball Savant*. https://baseballsavant.com/ (accessed September 1, 2014).

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

TABLE 1: Timeline of Umpire Labor History

Year	Event	Description
1996	New Strike Zone Definition	MLB redefined the strike zone, moving lower end from top to bottom of knees
2001	Introduction of QuesTec	Introduction of technology to monitor umpire accuracy on ball-strike calls
2004	Ratification of QuesTec	Grievance by union settled; QuesTec used partially for evaluation; new CBA
2007	Introduction of PITCHf/x	Sportvision's PITCHf/x introduced for the Gameday website; data is publicly released
2009	Introduction of Zone Evaluation	Umpire union agrees to use of PITCHf/x data is for evaluation and training; new CBA

TABLE 2: Summary of Dependent Variables

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

TABLE 3: Empirical Estimation of Strike Rates by QuesTec Condition

	TotalStrikeRate	***TotalStrikeRate***	***CalledStrikeRate***	***CalledStrikeRate***
Umpire FE	No	Yes	No	Yes
Constant	62.080***	61.513***	30.920***	29.072***
	(0.224)	*(0.334)*	*(0.340)*	*(0.423)*
1998	0.175	0.198*	0.172	0.201
	(0.118)	*(0.117)*	*(0.160)*	*(0.162)*
1999	-0.446***	-0.398***	-0.411***	-0.372**
	(0.108)	*(0.117)*	*(0.144)*	*(0.153)*
2000	-0.530***	-0.350**	-0.246	0.005
	(0.132)	*(0.143)*	*(0.189)*	*(0.205)*
2001	----------	----------	----------	----------
	----------	----------	----------	----------
2002	-0.253**	-0.218**	-0.205	-0.173
	(0.105)	*(0.105)*	*(0.156)*	*(0.156)*
2003	-0.079	-0.014	0.404**	0.485***
	(0.121)	*(0.129)*	*(0.159)*	*(0.181)*
2004	-0.202*	-0.097	0.208	0.298
	(0.110)	*(0.134)*	*(0.168)*	*(0.207)*
2005	0.390***	0.483***	0.791***	0.863***
	(0.119)	*(0.153)*	*(0.169)*	*(0.220)*
2006	0.106	0.243	0.382**	0.527**
	(0.134)	*(0.192)*	*(0.184)*	*(0.262)*
QuesTecExist	1.103***	1.624***	1.836***	2.388***
	(0.228)	*(0.368)*	*(0.339)*	*(0.470)*
QuesTecInstall	0.217**	0.177**	-0.472***	-0.454***
	(0.089)	*(0.088)*	*(0.117)*	*(0.111)*
ln(*Experience*)	-0.108	-0.239	-0.142	-0.215
	(0.066)	*(0.188)*	*(0.115)*	*(0.271)*
Exist × ln(*Exp*)	-0.061	-0.225*	-0.142	-0.299*
	(0.083)	*(0.116)*	*(0.136)*	*(0.156)*
Obs.	23,819	23,819	23,819	23,819
R^2	0.063	0.099	0.082	0.153

***, **, * refer to statistical significance at the 1%, 5%, and 10% levels, respectively. Models include umpire and stadium dummy fixed effects.

FIGURE 1: *TotalStrikeRate* and *CalledStrikeRate* by QuesTec Condition

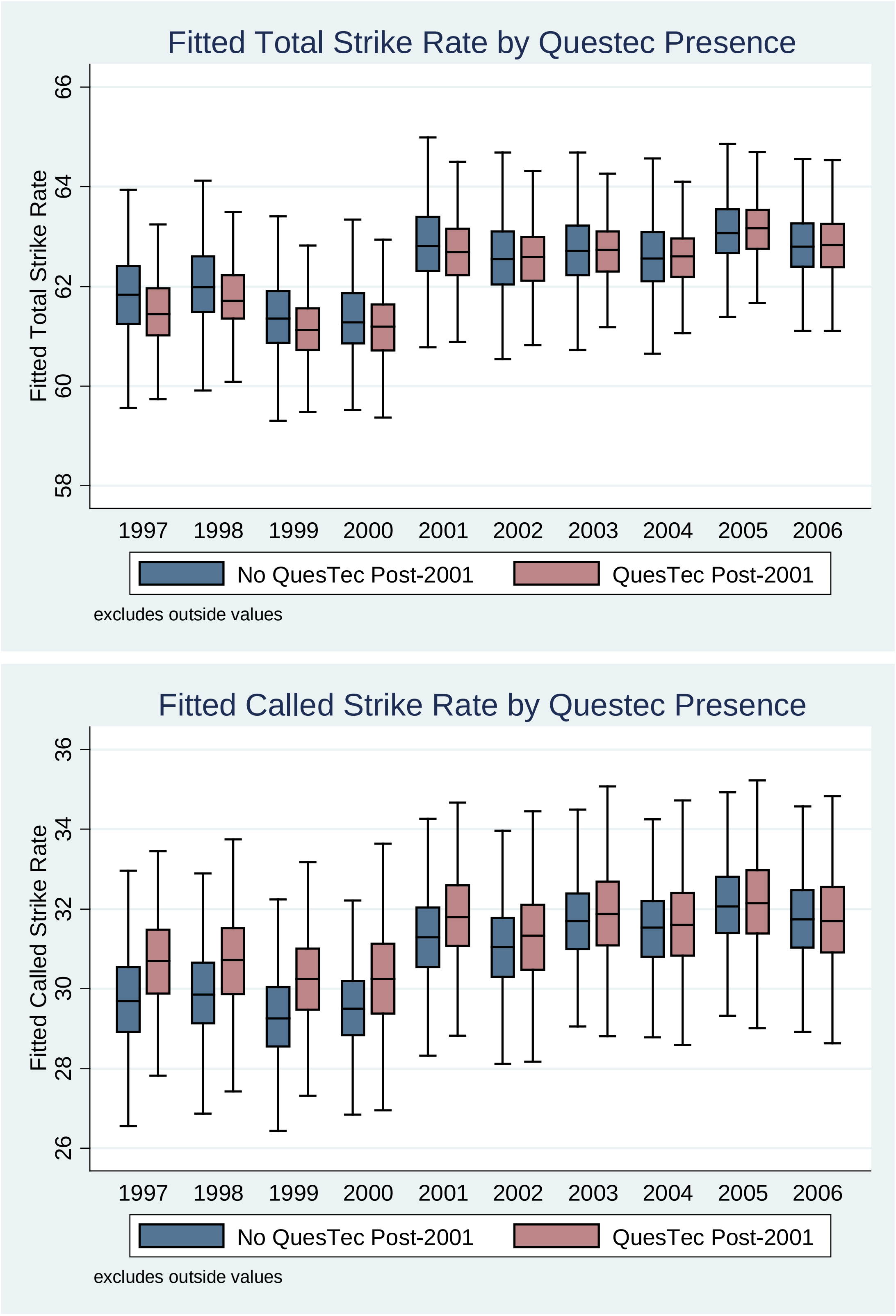
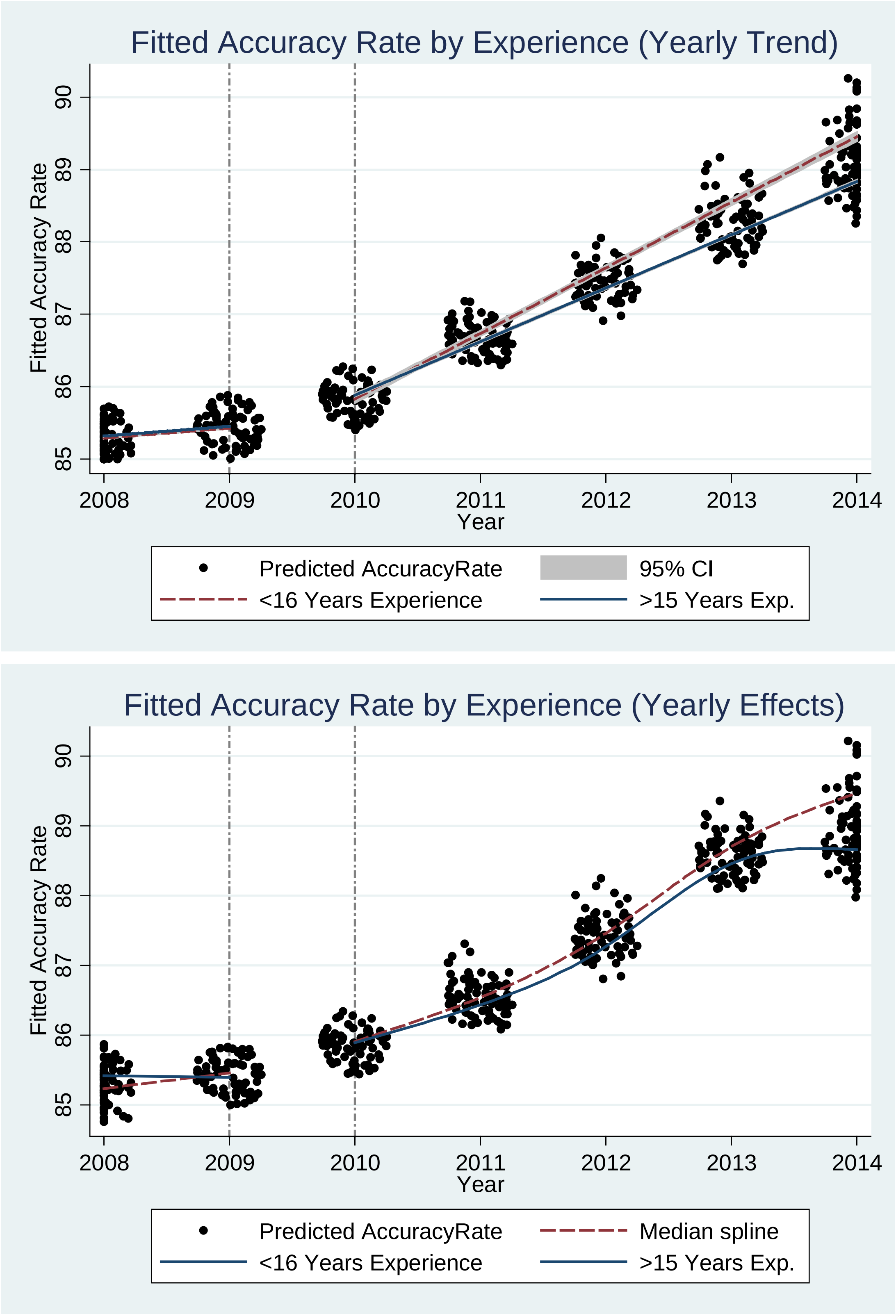


TABLE 4: Accuracy Rate Changes by Zone Evaluation and Experience Level

	Yearly Trend FE	**Yearly Trend RE**	**Yearly Effects FE**	**Yearly Effects RE**
Constant	85.411*** *(0.086)*	85.242*** *(0.213)*	85.412*** *(0.087)*	84.758*** *(0.198)*
Trend Pre-Zone ZE	0.134 *(0.122)*	0.133 *(0.122)*	---------- ----------	---------- ----------
Δ Intercept Post-ZE	0.285 *(0.206)*	0.307 *(0.206)*	---------- ----------	---------- ----------
Δ Trend Post-ZE	0.633*** *(0.124)*	0.978*** *(0.160)*	---------- ----------	---------- ----------
ln(*Experience*)	---------- ----------	0.025 *(0.092)*	---------- ----------	0.217** *(0.087)*
ln(*Experience*) × Trend Post-ZE	---------- ----------	-0.125*** *(0.030)*	---------- ----------	---------- ----------
2009	---------- ----------	---------- ----------	0.134 *(0.122)*	0.888* *(0.460)*
2010	---------- ----------	---------- ----------	0.586*** *(0.127)*	1.197** *(0.474)*
2011	---------- ----------	---------- ----------	1.184*** *(0.116)*	2.283*** *(0.337)*
2012	---------- ----------	---------- ----------	2.063*** *(0.130)*	3.551*** *(0.339)*
2013	---------- ----------	---------- ----------	3.168*** *(0.119)*	4.455*** *(0.339)*
2014	---------- ----------	---------- ----------	3.417*** *(0.133)*	5.460*** *(0.317)*
2009 × Experience	---------- ----------	---------- ----------	---------- ----------	-0.298* *(0.164)*
2010 × Experience	---------- ----------	---------- ----------	---------- ----------	-0.236 *(0.185)*
2011 × Experience	---------- ----------	---------- ----------	---------- ----------	-0.418*** *(0.126)*
2012 × Experience	---------- ----------	---------- ----------	---------- ----------	-0.558*** *(0.133)*
2013 × Experience	---------- ----------	---------- ----------	---------- ----------	-0.464*** *(0.128)*
2014 × Experience	---------- ----------	---------- ----------	---------- ----------	-0.755*** *(0.124)*
Obs.	536	532	536	532
R^2	0.775	0.647	0.787	0.657

***, **, * refer to statistical significance at the 1%, 5%, and 10% levels, respectively. Panel regressions limited to observations in which an 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. Models without experience include umpire fixed effects. Models including experience are estimated with umpire random effects, chosen from the output of a Hausman test.

FIGURE 2: Accuracy Rate by Year and Experience Level


Fitted Accuracy Rate by Experience (Yearly Trend)
Fitted Accuracy Rate
Year
● Predicted AccuracyRate
95% CI
<16 Years Experience
>15 Years Exp.
Fitted Accuracy Rate by Experience (Yearly Effects)
Fitted Accuracy Rate
Year
● Predicted AccuracyRate
Median spline
<16 Years Experience
>15 Years Exp.