Expert Workers, Performance Standards, and On-the-Job Training: Evaluating Major League Baseball Umpires[1]

Brian M. Mills*

August 27, 2014

*Assistant Professor
University of Florida
Department of Tourism, Recreation, and Sport Management
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 performance standards, monitoring, technology, and training as they relate to the performance of Major League Baseball umpires from 1988 through 2013. I find structural changes in performance concurrent with known bargaining struggles, and substantial improvements after implementation of incentive pay and new monitoring and training. However, variability in performance across umpires has decreased substantially during this time, contrary to past literature on performance pay. I detail these changes in the context of a reduction in offensive output that is often attributed to a crackdown on performance enhancing drug use in MLB.

JEL Codes: L83, J42, J5

[1] I would like to thank Charlie Brown and Rodney Fort for helpful comments on earlier versions of this manuscript in written and presentation form, as well as the participants at the 2014 Western Economic Association Conference in Denver. Finally, I would like to thank Mike Fast, Jon Roegele, Aaron Baggett, and Alan Nathan for helpful information and comments regarding the data and strike zone definitions in this paper.

Electronic copy available at: http://ssrn.com/abstract=2478447

Expert Workers, Performance Standards, and On-the-Job Training: Evaluating 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). Two general structures of performance based pay are commonly addressed, which include considerations of monitoring and contingent pay. The first, a piece rate system, consists of direct objective pay per output, while the second—a tournament structure—requires that performance relative to peers determines pay or promotion. In the piece rate context, Lazear (2000a) identifies the impacts of switching from hourly wages to piece rates using a unique data set from Safelite, exhibiting performance improvements as well as increases in the variability of performance among workers in the new pay scheme. Alternatively, Fernie and Metcalf (1996) evaluate a tournament structure—where performance pay is based on the relative performance in the context of horse jockeys—finding that pay and performance are associated, and that large retainer fees that do not depend on subsequent performance can results in deterioration of performance. Further work identifies performance pay impacts for CEOs and firm performance (Jensen & Murphy, 1990; Barro & Barro, 1990; Bulan, Sanyal, & Yan, 2010), and applies lessons specifically to 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).

Expanding upon this empirical work, this paper uses Major League Baseball (MLB) umpire ball and strike calls (performance data) to identify performance effects associated with known labor disputes and subsequent collective bargaining agreements (CBA) in a novel context of expert workers. As noted in Price and Wolfers (2010) and Parsons et al. (2011), the nature of the required expertise and monitoring involved in the umpire labor market describes a group of

Electronic copy available at: http://ssrn.com/abstract=2478447

employees that are at the top of their profession, yet are still subject to biases or mistakes (Garicano, Palacios-Huerta, & Predergast, 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, 2013; Walsh, 2010). 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, the wealth of publicly available employee performance data is unlike many other industries where data is often proprietary and unavailable (Kahn, 2000). Secondly, 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] 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, their 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 incentive effects at the league level may be measured independent of the performance of the labor market competition (Fernie & Metcalf, 1996). Fourth, the umpire labor market includes seniority considerations in pay—with all employees having a fixed component of their pay—identified in past literature as a way to obscure the likelihood of reaching a prize in a contest and/or reduce the risk involved in a pure piece rate pay system (O'Keeffe et al., 1984). Fifth, contests are repeated each year such that there is not a post-tournament period where incentives no longer remain (Leeds, 1988). Lastly, there have been changes to the monitoring (and training) system and strategy of MLB, with recent monitoring implementation that is both cheaper and more precise than manual methods. These systems were put in place alongside an expected minimum performance level[3] to avoid further developmental and training requirements, discipline, or termination.

[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.

[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.

There have been multiple recent time points in which MLB has been able to provide increased incentives for performance improvements through increased monitoring and changes in rewards such as additional payment for assignment to post-season play. The most recent labor disputes and/or agreements relevant to this work include a 1991 and a 1995 strike by the Major League Umpires' Union (MLUA), a 1999 failed mass resignation of umpires that led to the formation of the World Umpires' Association (WUA), limited technological monitoring in 2001, agreements for increased technological monitoring in 2004 and 2009, and a unique disciplinary action taken by MLB after the 2009 season.

Using this labor context, I find that changes in contingent pay and use of technology in training and monitoring have improved performance and led to a reduction in performance variability across umpires. These changes are consistent with a publicly specified requirement 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 given the incentive to do so. There are structural *shifts* in performance measures—concurrent with labor disputes and/or monitoring introduction—in addition to structural *trend* changes that indicate more gradual performance improvements after labor agreements, consistent with increased training and gradual improvements over time.

More specifically, after a failed 1999 mass resignation strategy by the MLUA and a new CBA the average strike rate in MLB increased from approximately 60.75% to approximately 62.2% and continued an increasing trend to 62.95% in 2013. When the sample is restricted specifically to pitches *called* by umpires, the average strike rate from 2000 through 2013 is 31.44% compared with 28.84% from 1988 through 1999. This increase in strike rate began just after the 1999 labor dispute, and has peaked in 2013 at 32.5%—an increase of 9% overall. There has also been a 30% to 40% *decrease* in the variability of strike rates across individual umpires, again concurrent with pivotal labor disputes in the 1990s. These results point to evidence of sudden response to incentive pay and monitoring from MLB—in the form of a level shift in strike calling rates—and evidence of impacts of on-the-job training specific to umpiring alongside improvements in technology.

More recent data (2007-2013; Sportvision, 2013) on pitch location allowed analysis of the spatial characteristics of the strike zone called by MLB umpires. Using a Generalized Additive Model (GAM), I find that the strike zone has expanded downward and was reined in on

the outside corner since a new evaluation system was put in place after the 2009 season.[4] Additionally, umpires have shown substantial improvement in their ball-strike call *accuracy*. In 2007, umpires correctly called only 76.8% of pitches within the rulebook zone and 87.4% of pitches outside the rulebook zone based on the Sportvision data. However, by 2013, umpires were correctly calling 85.7% and 90.0% of pitches within and outside the zone, respectively. An important result from this is a large increase in the strike rate called by umpires that has likely resulted in a decrease in offense often attributed to the league's crackdown on the use of performance enhancing drugs.

This paper proceeds as follows. In Section 2 I briefly describe the history of changes in MLB umpire labor agreements and note the relevant theory as it pertains to incentive pay. Section 3 details the long-term data used here, along with measurement strategies, and econometric methods used to identify performance effects over the period from 1988 through 2013. Section 4 describes the more detailed performance data, measurements of accuracy rates, and results over time concurrent with changes in monitoring and performance pay conditions. This section also presents estimates of the impact of age and experience on performance. Section 5 presents non-parametric methods used to estimate accuracy and changes in a two-dimensional strike zone plane called by umpires. Section 6 estimates the proportion of changes in offense in MLB over the past ten years attributed to umpire ball-strike calling behavior and subsequent pitching strategy changes. Finally, Section 7 rounds out the analysis with conclusions and suggestions for future inquiry.

2. Background and Literature

2.1 The Baseball Umpires' Labor Market

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. This section briefly reprises these tensions in order to set the stage for the current analysis. Much of this history is pieced together from the San Diego Association of Baseball Umpires (2014), Armour (2009), and Alcaro (2002). 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

[4] A popular press article released during the construction of this work found similar results using a similar data set using a different modeling technique (Roegle, 2014).

process. Most recently, the league and one of the more popular umpiring schools—the Jim Evans Acadamy for Professional Umpiring—ended their relationship in 2012.[5] Therefore, I focus on the MLB level of performance standards, rather than the specifics of training throughout the minor leagues, but note this development is just as interesting.

The main events of interest for this work are noted in Table 1. The 1990s mark the beginning of our interest in the umpires' labor market in this work, with the data series beginning in 1988. 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 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 during this time.

In 1998 and 1999, Commissioner Bud Selig publicly acknowledged his interest in obtaining direct control over all umpires (Callan, 2012) and had Sandy Alderson issue a memo calling for uniform enforcement of the strike zone.[6] While there certainly exists effort in ensuring an accurate strike zone, 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, 1999).

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

[5] While the relationship was officially ended due to a scandal related to racist costumes at a yearly academy bowling party, Evans and others have voiced concerns over the event being used as a way for the league to distance itself from the school (Keh, 2012).

[6] 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 (Callan, 2012).

bring MLB to the negotiating table. Instead, the league accepted all of the resignations and began hiring new umpires to replace them. Ultimately, a number of umpires rescinded their resignation, though 22 were officially accepted. 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 by 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 official technological monitoring of umpires in MLB (leaving aside Alderson's directives for teams to chart pitches in 1998). The WUA filed a grievance over the technology, which was resolved in 2004 with a new labor agreement (Schwarz, 2009). This agreement noted that umpire employment decisions were not subject *solely* to QuesTec monitoring, and that any poor performance based on the technology would require further monitoring using other means. At this point, QuesTec was installed in approximately half of MLB parks.

In 2009—as the WUA and MLB agreed on terms in collective bargaining—a new system took hold in the evaluation and training of umpires simply called Zone Evaluation (ZE). Umpires receive reports of their performance after every game, and the ZE system is claimed to be more accurate (Schwarz, 2009). Shortly after implementation of the ZE system, MLB fired three umpires due to highly publicized poor calls in the 2009 postseason, leaving another point of interest for analysis here (Nightengale, 2010).[7]

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. The implementation of postseason assignment based on merit—and subsequent adjustment to pay in 1995—identifies the first two points of interest. The 1999 resignation strategy and the technological monitoring through QuesTec that followed was a clear point at which the league gave a directive for improved performance, and followed through with what equated to forced retirement for some umpires. The 2004 season marked the first year in which umpires accepted monitoring technology as part of collective bargaining, with further adjustments to this in 2009. Finally, the league's release of three umpires after the 2009 season marks the first time this sort of disciplinary action had been taken in many years. Each of these time points identifies a point

[7] It is important to note that these calls were not ball-strike calls, but that this marks one of the few times that MLB used its power to remove umpires from their positions.

at which umpire strike calling behavior may be expected to change. Therefore, the following sections detail these changes, and relate them back to the expectations from theoretical foundations in the performance pay and monitoring literature.

3. Long-Term Trends in Strike Rate and Strike Rate Variability

3.1 Data and Estimation Strategy

Longer-term trends in umpire performance were tracked using multiple data sources, providing a base for three primary measures of umpire behavioral changes of interest (1988-2013). The time series allow nearly equal length before and after the 1999 labor dispute of most interest in this work, and includes other relevant changes to umpire performance pay and monitoring conditions. First, Baseball Reference (2014) was used to calculate, 1) the overall league strike rates as a percentage of the total number of pitches seen in a given season and, 2) league aggregated called strike rates as a percentage of called pitches only. This first measure includes both called and swinging strike changes aggregated yearly (henceforth referred to as *TotalStrikeRate*). It is calculated by dividing the number of total strikes by the number of total pitches in any given season. The second more closely measures the behavior of the umpire: whether or not more strikes are being called on pitches at which the batter did not swing (henceforth, *CalledStrikeRate*). This measure divides the number of strikes called by the umpire by the total number of pitches in a given season that are subject to an umpire's judgment (pitches at which the batter did not swing). While these measures may change over time due to changes in strategy of pitchers or performance of hitters, I search for substantial structural changes in the data near policy points of interest. Sudden structural changes in these measures would seem more likely attributed to umpires than a sudden evolution in the game itself.

Though these first two measures provide some information regarding changes in umpire propensity to call strikes, there is the possibility that either batters are swinging more often (*TotalStrikeRate*) or that pitchers are throwing closer to the middle of the strike zone more often (*CalledStrikeRate)*. Therefore, I also measure conformity in strike calling among umpires in MLB. In this analysis, individual umpire strike calling data (Baseball Prospectus, 2014) were used to calculate the weighted coefficient of variation of *TotalStrikeRate* across the entire league. Prior to 2007, data to calculate *CalledStrikeRate* at the individual level were unavailable. This measure is henceforth referred to as *wRateCV*. However, it is assumed that the distribution of

pitchers' pitch location is similar across umpires throughout a given season. Therefore, changes to the variability of the strike rates across umpires should be largely attributed to changes in the umpire's judgment calls.

This measure is calculated as the ratio of the weighted standard deviation and the weighted mean of umpire strike calls:

$$wRateCV_t = \frac{wS_t}{\overline{wX}_t}$$

Where wS_t is the standard deviation of strike rates across all umpires weighted by the total number of pitches that umpire saw in season t, and $\overline{wX}_t$ is the weighted average of strike rate across all umpires in season t (equivalent to *TotalStrikeRate*), again using total number of pitches seen as the weights. These are defined as:

$$\overline{wX}_t = \frac{\sum n_i x_i}{N}$$

And,

$$wS_t = \sqrt{\frac{N \sum n_i (x_i - \overline{wX}_t)^2}{N^2 - \sum n_i^2}}$$

Where i indexes each umpire, x_i and n_i are the strike rate and number of pitches seen by umpire i, respectively, and N is the total number of pitches in the season level sample. A larger *wRateCV* implies that umpires are more dissimilar in their propensity to call strikes. As the measure decreases, it is assumed that the strike zone in MLB is becoming more uniform. This measure is useful in that it does not depend on changing rates of batters swinging or pitchers throwing closer to the center of the strike zone, as with the previous two measures.

For robustness, an additional version of *wRateCV* and *TotalStrikeRate* was calculated using a subsample of the data. The subsample is taken only from umpires who were active in MLB before and after work stoppages in 1994-95 and/or 1999 to evaluate changes to umpire

behavior, rather than changes caused by replacement officials during this time. The initial measure includes all umpires in the data, including replacement umpires during strikes and lockouts, as well as replacement umpires hired permanently after the 1999 failed resignation strategy. These are used to differentiate any effects that took place among individual umpires due to performance incentive changes from those effects occurring due to the hiring of new umpires. Again, individual *CalledStrikeRate* data at the umpire level were unavailable, so this subsample strategy is not performed with this variable. The full time series of these variables can be found in Table 2.

The long-term time series data were used to estimate structural changes in strike rate and variability. Stationarity of each series was assessed using the step-by-step approach laid out in Fort and Lee (2006). I begin by testing each series using the Augmented Dickey-Fuller and Phillips-Perron tests, followed by Lagrange Multiplier tests for a unit root with two breaks and one break, respectively (Lee & Strazicich, 2001; 2004). Those series found to be stationary were then subjected to the structural change estimation method developed by Bai and Perron (1998; 2003; henceforth the BP Method). This method identifies structural changes within time series data at previously unspecified time points. These regressions contain parameters only for level and trend of the data. I consider the following regression model with m breaks (and $m + 1$ regimes):

$$y_t = x_t'\alpha_j + z_t'\beta_j + u_t, \qquad j = 1, \dots, m + 1.$$

Here, y_t is the dependent variable—*TotalStrikeRate*, *CalledStrikeRate*, or *wRateCV*—at time t, where x_t and z_t are the time variable and a constant, respectively, and α_j and β_j are the corresponding unknown trend and intercept coefficients. Indices $(T_1, \dots, T_m)$ are treated as unknown breakpoints to be estimated through the Bai and Perron approach that separate *j* regimes. Coefficients for both x_t and z_t—in this case the trend and level of the data—are allowed to vary across regimes, with no further covariates entered into these models. This is therefore referred to as a pure structural change model (Hansen, 2001; Perron, 1988; Perron, 1989). The model with varying trend and level is estimated using the sequential method with the repartition procedure and assumed homogenous variance across regimes with a trimming parameter of $\varepsilon = 0.15$, ensuring that regime lengths at any point of the time series are made up

of at least four years of data (Bai & Perron, 2006). For further details on the structural change estimation procedure, the reader is referred to Bai and Perron (1998; 2003; 2006).

In this section, I also analyze differences in called strike rates across monitoring conditions with QuesTec, and how these changed from 2001 through 2008. This data comes directly from Tainksy et al. (2013). This portion of the analysis identifies strike rates at the umpire-pitcher-game level—clustering these observations as in Tainsky et al. (2013)—and weights the least squares regression by the number of called pitches for each umpire-pitcher-game level observation. Theory predicts that—if MLB is interested in increasing strike rates overall—strike rates should be higher with QuesTec monitoring than without. This is used to supplement the findings of the longer-term structural change regressions.

3.2 Results of Strike Rate Models

All series were found to be stationary with at least one break, and were subjected to the BP Method with a maximum of two breaks. All structural changes presented here were found to be statistically significant. The testing procedure, break dates with associated confidence intervals, and regression results can be found in Tables 3 through 6. Beginning with the long-term analysis of umpire strike-calling behavior, Figure 1 exhibits the structural change model for the *CalledStrikeRate* time series from 1988 through 2013. Figures 2 and 3 include the plotted structural change models for *TotalStrikeRate* of all non-scab umpires, and those umpires working both before and after the 1999 resignation strategy, respectively. As you can see from these figures, there was a large structural shift in both *CalledStrikeRate* and *TotalStrikeRate* just after the 1999 dispute—with confidence intervals from 1999 through 2001—and near the implementation of QuesTec, with a continuing upward trend thereafter for both measures. The model including only umpires that worked both before and after the 1999 season identifies a second change in trend following the 2009 implementation of the ZE system. As will be shown later, this pattern is also found for the more nuanced locational data in a shorter term data series.

The discontinuity in *CalledStrikeRate* and *TotalStrikeRate* identifies a point at which the league experienced substantial changes in umpire strike calling rates and overall strike rates as they had wanted since the early 1990s. While it is unlikely that the entire trend in increased strikes and strike calls is due to umpire strike zone changes, it is unlikely that the rather large structural shift would be attributed to a sudden change in the style of play by batters and pitchers.

Shortly after MLB made clear its intentions to remove umpires that did not comply with their expectations of an expanded strike zone, umpires changed their behavior drastically by calling substantially more strikes. Taking advantage of this event, I also estimated the difference among umpires that were allowed to remain in MLB, and those who ultimately had resignations accepted by the league. Figures 4 and 5 exhibit strike rates among umpires that were allowed to continue employment with MLB after the failed 1999 resignation strategy, and those that had resignations accepted even after litigation from the union.[8] As you can see, prior to the 1995 agreement, the group of umpires that did not continue employment with MLB after 1999 called significantly more strikes than those that continued employment. However, this effect was no longer apparent after the 1995 agreement, and is in fact reversed. After the more recent 1995 labor agreement, umpires that continued on in the league were the ones that had increased their strike calling rates by the most when directed to do so, providing evidence that the decisions made by MLB were at least tangentially related to whether or not these umpires conformed with its requests in expanding the strike zone during this time. This effect is estimated by the regression model Table 7. While there is a positive coefficient on the rate of strike calls for this group over the time period, the relationship between strike rate and resignation acceptance reverses near the time of the 1995 CBA. The figures make clear that they were surpassed by their peers during the latter part of the decade, possibly impacting the choice of which umpires would remain in the league after 1999.

Lastly, Figures 6 and 7 present the structural change models for *wCV*. Between 1991 and 1995, there were large increases in the variability in strike rates. Shortly after the 1995 season, variability in strike rates experienced a large downward structural shift, followed by a notable decreasing trend through 2013. The structural changes in the variability in strike rates across umpires identify a curious pattern. There is some evidence suggesting that the league may have set minimum standards too low—or did not provide enough incentive pay—in their initial introduction of merit-based postseason assignments in1991. During the time between 1991 and 1995, some umpires may have been performing at the minimum expected performance level, while higher ability umpires called strikes at a rate above this minimum level with the same low effort level. However, after the newly negotiated collective bargaining agreement in 1995, variability in strike rates decreased suddenly and continued to do so through the end of the time

[8] 1991 and 1995 scab umpires were not included in this analysis.

series. If minimum standards were increased, then the reduction in variability in strike calls could be due to a larger improvement in performance by lower-ability umpires relative to their higher-ability peers, possibly indicating that the new agreement improved performance overall, when paired alongside the increased in *TotalStrikeRate* and *CalledStrikeRate*. Alternatively, the increased protections of collective bargaining and implementation of performance pay could have increased the labor pool of umpires looking to work at the MLB level. In this scenario, the increase in talent among the labor supply could have left MLB with more options when promoting new umpires, and allowing them to choose only those at the highest level. As this sorting process takes place, the average umpire at the highest level should increase, while the variability among these umpires decreases. As a whole, it is reasonable to conclude that umpires with the ability to make calls in line with MLB expectations did so with the correct incentives in place, and all umpires gradually improved their ability to comply through increases in technology that decreased the costs and precision of monitoring and training feedback.

3.3 QuesTec Effects

Lastly, Table 8 presents the results from the weighted regressions from the Tainsky et al. (2013) data, with the effect visualized in Figure 8. This model estimates a statistically significant difference in strike rate under monitoring than when there is no QuesTec present. However, this dynamic changes over time. After umpires filed their grievance regarding the use of QuesTec in performance evaluation and included it in the 2004 collective bargaining agreement, the difference in strike rates between QuesTec and non-QuesTec parks decreased and remained relatively small. But it is clear that after 2000, strike rates had increased dramatically across the league irrespective of the presence of QuesTec. This seems to indicate that while the monitoring impacted behavior among umpires—especially when being monitored—there was an adjustment period in maintaining a similar the zone across monitoring conditions between 2001 and 2004.

4 Accuracy Rates

4.1 Data, Measurement and Modeling Strategy

For this portion of the analysis, I evaluate accuracy rates of umpire ball and strike calls using Sportvision's regular season Pitch f/x data, and how these have changed over its use from

2007 through 2013. This analysis allows the measurement of true performance improvements among umpires, rather than strike rate measurements that could be influenced by pitchers changing the location of their pitches over time, on a shorter time period of available data. This data set consists of a large amount of information about each pitch thrown in the MLB regular season, including the location of the pitch when it crosses the front of the plate, the pitch velocity, a pitch type classification, whether the batter and pitcher are right or left handed, the ball-strike count when the pitch is thrown, the game outcome from the pitch (hit, out, hit type, groundout, error, fly out, etc.), and a number of other game-level variables. To measure umpire behavior 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 from the data set, as these require no real judgment by the umpire. The sample size of this subset of data was just over 2.47 million observations.

Accuracy of strike calls requires the identification of the strike zone indicated within the MLB Official Rulebook.[9] 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 2007 through 2013 seasons (approximately 73.5 inches).[10] 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 the 17 inch plate width as noted in the official rules. Finally, 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 2007 through 2013 regular seasons (Figure 9). The first of these pitch types is a correctly called strike (*CorStr)* described as a pitch that crosses the plate inside the strike zone plane and is called

[9] 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.

[10] Individual batter height was not included in the data. However, umpires generally see a similar random 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.

a strike by the umpire (Column 1, Figure 9). This can be thought of as a true positive in the statistical sense. The second pitch type is a correctly called ball (*CorBall*). This occurs when a ball located outside of the strike zone is called a ball by the umpire, or a true negative (Column 3, Figure 9). The last two classifications include incorrectly called strikes (*IncStr*—a false positive, Column 2, Figure 9) and incorrectly called balls (*IncBall*—a false negative, Column 4, Figure 9). Together, these identify strike calls on pitches outside the strike zone, and ball calls on pitches inside the strike zone, 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 proportions of correct calls on strikes, balls, and all pitches are included in their own respective umpire fixed effects regressions; however, because of the shorter span of data, no structural change model is estimated for strike zone accuracy. The variable of interest in this regression is the time trend, with a positive trend indicating increased performance among umpires. Additionally, I include umpire experience and debut date to gauge heterogeneous effects of new and existing employees. These are included in a panel model to test the impact of experience, and an umpire debuting before and after the implementation of the failed resignation strategy, implementation of the QuesTec system, and implementation of the ZE System individually. 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, and could indicate that more recent hires have higher levels of intrinsic ability than those hired before significant changes in performance pay or monitoring.

4.2 Zone Accuracy Results

Figure 10 shows the yearly changes in sensitivity and specificity of umpire ball-strike calls in aggregate for the league. Throughout this time—and especially so after the 2009 season—umpires both increased the percentage of correct strikes (sensitivity) and correct balls (specificity) in each year. Prior to the ZE system implementation, umpires were calling strikes on only 76.7% of pitches inside the zone. By 2013, the rate of pitches within the zone that were correctly called strikes was 85.7%. 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% in the data use here.[11] Overall accuracy improved from 84.0% in 2007 to 88.6% in 2013. The accompanying regression models showing statistical significance of the time trend and year fixed effects are presented in Tables 9 and 10.

4.3 Accuracy Results by Age and Experience

Tables 9 and 10, along with Figure 11, also present results for the analysis of experience and debut date on accuracy rates and improvements in accuracy rates. Beginning with Table 9, weighted ordinary least squares was used to estimate the effect of year and umpire experience variables on accuracy rates. These regressions were fit both with (Table 9, Columns 1 and 3) and without (Table 9, Columns 2 and 4) umpire dummy fixed effects, and with (Table 9, Columns 2 and 4) and without an interaction between year and experience (Table 9, Columns 1 and 2). For brevity, I discuss only Column 4 here, which includes both fixed effects and an interaction of the year and experience level of the umpire making the calls.

The model estimates a clear increase in accuracy rates for umpires as a whole during the time that the Pitch f/x system has been in place, with an accuracy increase of approximately one percent per year. Additionally, more experienced umpires tend to be better at their jobs than younger ones. However, as noted by the interaction between year and experience, less experienced umpires increase their performance more quickly than that of more experienced umpires. This relationship is shown in Figure 11 at a number of selected experience levels. As you can see in the figure, less experienced umpires begin their careers lagging behind in performance, but quickly catch up during the Pitch f/x era, with some evidence that they eclipse their more experienced peers. This is consistent with past literature on learning and performance of newer employees (Lazear, 2000a; Kostiuk & Follmann, 1989), and parses the data to some extent between apparent selection of new umpires, and increases in existing umpire ability (Coffey & Maloney, 2010). While some of this effect could be due to the attraction of a more

[11] It is important to note that while the Pitch f/x data is 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 over the time period, the *relative* changes in accuracy rates are still useful in the context presented here.

skilled labor pool, the steeper increase for younger umpires could be due to a background in training with more advanced technology than their more experienced counterparts that arrived before implementation of QuesTec, Pitch f/x, and the ZE system.

Table 10 shows similar results to these previous models, but with specific dummy variables for time of debut. Each of the regressions in Table 10 identifies umpires that debuted before and after significant events identified in the history of the umpires' labor market: the 1999 resignation strategy (Columns 1, 2 and 3), the 2004 implementation of QuesTec (Columns 4, 5 and 6), and the post-2009 implementation of the ZE system (Columns 7, 8 and 9). The Pre-2000 variable is statistically significant on its own, but not when interacted with the year variable. However, the opposite is true for the Pre-2005 and Pre-2010 Debut variables, estimating both a difference in the performance level by debut, as well as a slope difference between the two groups. In Columns 8 and 10, umpires debuting before 2005 and 2010, respectively, performed better than their counterparts debuting after this time. However, consistent with the models in Table 9, umpires debuting more recently increased their performance more quickly.

5 The Shape and Size of the Strike Zone

5.1 Data and Semi-Parametric Estimation Strategy

Due to the multidimensional, non-monotonic nature of the strike zone, I also estimate a generalized additive model (GAM) using the locational data from Sportvision to measure its *size* and *shape* during the time. This method has been used in past work investigating bias in umpire ball-strike calls (Mills, 2014; Tainsky et al., 2013). While a sensitivity/specificity analysis is useful in the context of correct call rates, it is important to note that not all correct calls are created equal. For example, a pitch at the very edge of the strike zone is much more difficult to judge than one thrown directly down the middle. Therefore, the generalized additive model allows for a non-parametric estimation of the strike zone surface and associated changes in strike probability and uncertainty over the call as it relates to pitch location. GAMs have the advantage of identifying the more rounded edges of the zone, and fit separate surfaces for batter handedness, without over-fitting the data.[12] A full discussion of GAMs can be found in Wood, (2000; 2003; 2004; 2006; 2011) and Gu and Wahba (1993), but I reprise the basic method developed in Mills (2014) for pitch location data here. The flexible model allows measurement

[12] Mills (2014) finds substantial differences in the strike zone for right handed and left handed batters.

of changes in the likelihood of a strike call conditional on the location of the pitch when it crosses the plate. This is a binomial logistic model, with an additional non-parametric component estimated through generalized cross-validation that allows more flexibility than a parametric polynomial representation of the edges of the strike zone. The GAM additively combines this non-parametric estimation with parametric estimates of parameters for other covariates as follows:

$$g(\mu_{i,t}) = \boldsymbol{X_{i,t}\beta} + f_b(Z_{h,i,t}, Z_{v,i,t}) + \varepsilon_{i,t}$$

Where the response variable, y_i—a variable indicating whether the pitch i thrown in year t was called a strike (strike=1 and ball=0)—has mean μ_i, where $g(\cdot)$ represents the logit link function for the binomial response. $\boldsymbol{X_i}$ and $\boldsymbol{\beta}$ are vectors of predictor variables and their unknown coefficients, respectively, which can be estimated parametrically. Here, $\boldsymbol{X_i}$ simply refers to the season in which the pitch is thrown. The unknown function $f_b(Z_{h,i,t}, Z_{v,i,t})$ is estimated jointly for vertical and horizontal location—indexed by h and v, respectively—using generalized cross-validation, with the right and left handed batter surfaces indexed by b. Predictor variables in this model include year, umpire fixed effects, pitch type, and ball-strike count.

While the above version of the GAM is a semi-parametric model, I also use a fully non-parametric model to measure the surface area and shape of the strike zone separately for each year. This estimation procedure simplifies to the form below, with h and t indexing the handedness of the batter and the season that the pitch is thrown, respectively:

$$g(\mu_{i,t}) = f_{b,t}(Z_{h,i,t}, Z_{v,i,t}) + \varepsilon_{i,t}$$

Using this model, I calculate a number of measures of umpire performance in the calling of balls and strikes. First, I empirically identify the edge of the strike zone as the spatial boundary at which the probability of an umpire calling a ball or a strike is estimated to be equivalent (Mills, 2014; Tainsky et al., 2013), and how this has changed over time.[13] The

[13] This strike zone definition can be adjusted to include a higher or lower probability of being called a strike. A number of strike zone definitions were analyzed, and are available from the author upon request.

measureable surface area within this empirically estimated boundary is considered the strike zone. I also adjust the strike zone definition using different strike probability contours at ten percent intervals, and calculate the difference in surface area between these. This identifies the area of uncertainty between these two strike zone boundaries, as exhibited in Figure 12. As the area between strike zone boundaries decreases, I assume umpire uncertainty is decreasing, and therefore both sensitivity and specificity are improving across time.

Secondly, I measure the proportion of surface area within the MLB Rulebook strike zone that is *excluded* from the empirically derived strike zone, and the proportion surface area of the empirically derived strike zone that lies *within* the rulebook zone. As an example, Figure 13 visualizes these measures in the context of the empirically derived zone with the 2013 right handed batter 50% probability boundary. The red contour identifies the strike zone derived from the non-parametric model. The blue area is the portion of the empirically derived strike zone that lies within the rulebook strike zone. The white area inside the dashed lines is considered part of the rulebook strike zone, but is not included in the empirically derived strike zone. The white area within the red contour, but outside the dashed lines, is the area of the empirically derived strike zone that is not included within the rulebook definition.

These measures implicitly identify a more detailed spatial representation the sensitivity and specificity of umpire strike calls across years. As the area excluded within the MLB Rulebook strike zone decreases, I assume that umpires are becoming more sensitive to strike calls. As the area included in the empirically derived strike zone lying outside the MLB Rulebook strike zone is reduced, I assume umpire specificity is improving. Therefore, rather than measuring a simple increase or decrease in the rate of strike calls—which could simply mean that umpires are calling more strikes outside the zone, or fewer strikes inside the zone, respectively—these analyses identify conformity to a standardized strike zone size and shape as defined with the MLB Rulebook.

5.2 Results of Strike Zone Shape and Size Estimations

Figure 14 presents the empirically derived strike zones in 2007 and 2013 for right and left handed batters, estimated from the fully non-parametric GAM noted earlier. As you can see, the zone has changed in shape, morphing more closely to the rulebook zone shown by the dotted box. For both right and left handed batters, the outside edge of the umpires' strike zone has been

reined in toward the rulebook boundary, while the lower half of the zone has extended downward to meet the rulebook strike zone floor. The zone—defined as the 50% boundary noted earlier—has also increased in size. While the left-handed strike zone still stretches well beyond the outside half of the plate, it has been reined in to some extent since 2007. The right-handed batter strike zone—if including the two inch leeway allowed for umpires—almost perfectly aligns to both the inside and outside edges of the rulebook zone. However, the top and bottom corners of the rulebook zone continue to be places where umpires are reluctant to call strikes, despite the considerable monitoring and training they receive from the ZE system.

The changes in the size of the strike zone surface during the Pitch f/x era are presented in Figure 15. Note that most of the increase in the size of the strike zone took place after the implementation of the ZE system. During this time, the 50% strike zone boundary has increased in size by about 30 square inches for left handed batters, and 35 square inches for right handed batters. This is approximately the surface area of 3 to 4 baseballs. Taking into account the shape changes in Figure 14, the additional area to the bottom of the strike zone—with some surface area taken from the outside edge of the previously called zone—is approximately equivalent to a row of baseballs below the previous bottom boundary of the zone.

Figure 16 further breaks down the increase in the size of the strike zone across varying definitions of the strike zone using the strike call likelihood boundary.[14] While the zone size has increased at all boundaries, it is particularly apparent for those pitches closer to the center of the strike zone, with the boundary for pitches almost surely to be called strikes—90% or higher in the sample for the given year—growing by as much as 16% since the 2009 ZE implementation. Figure 17 presents the surface area differences across various strike zone boundary definitions, comparing changes by year. A steeper slope identifies a larger area of uncertainty over whether a pitch is called a ball or a strike. A perfectly horizontal slope on the y-axis at zero would identify a strike zone boundary with no uncertainty—a perfect drop-off from 100% strike certainty to 0% strike certainty. Note that in Figure 17 the lines for both right and left handed batters are both flattening and dropping down the y-axis. Based on this data, both sensitivity and specificity seem to be improving, allowing for more accurate strike and ball calls even at the edges of the strike zone.

[14] No statistical tests were used for this portion of the analysis, as including the measurement error for the surface area of the strike zone from the non-parametric estimation is not well addressed. Therefore, this exhibition should be taken together with the accuracy rate analysis.

Further supporting the notion that strike rates have increased even after controlling for location, Table 11 presents a semi-parametric regressions with year, pitch type, and ball-strike count dummy fixed effects.[15] These models allow for the evaluation of the statistical significance of the changes taking place in the semi-parametric model across years, and include umpire fixed effects for strike call rates. As shown in the regression table, strike rates show statistically significant increases shortly after the implementation of the ZE system in MLB.

Lastly, the left panel of Figure 18 identifies the percentage of area of the umpires' strike zone that lies within the rulebook strike zone. Alternatively, the right panel of Figure 18 identifies the percentage of the rulebook strike zone that is filled with the umpires' strike zone. In other words, this panel identifies how closely the edges of the umpires' strike zone lie to the edges of the rulebook strike zone. Of most interest here is the stark increase in conformity of the edges of the umpires' strike zone after the implementation of the ZE system. Not only is more of the empirically derived strike zone contained within the rulebook strikes zone, but the empirically derived strike zone also fills more surface area of the rulebook zone. Taken together, these results identify clear performance improvements among umpires during this time on both ball and strike calls on pitches closer to the edge of the strike zone.

6 Umpire Contributions to Offensive Decline in MLB

6.1 Data and Modeling Strategy

MLB implemented new performance enhancing drug (PED) testing and enforcement in 2006. It is well-documented that run scoring has decreased substantially in MLB since these new policies were put in place (Rymer, 2013; Henderson, 2011). Presumably, MLB intended to rid the league of PEDs in response to fan interest in this endeavor, despite the apparent surge in revenues at the end of the 20th Century when the use of alternative performance enhancement is now known to be common among many players. The expectation was likely that, without some semblance of fairness with respect to these enhancements, fans would no longer maintain interest in the games. The subsequent decreases in run scoring and in home runs resulted in the new PED policies being trumpeted as a resounding success. Ultimately, the 2010 season was labeled the "Year of the Pitcher" (Dubner, 2010). However, issues with superstar players and PED use have continued to surface since the implementation of these policies. Additionally, the "Year of

[15] Results for the non-parametric estimation and degrees of freedom are available upon request.

the Pitcher" was concurrent with implementation of the ZE system for umpires—shown to have important effects on the strike zone—and reduced maximum barrel size for baseball bats (MLB Rule 1.10(a)).[16] There also exist reports that variation in baseball manufacturing may have induced significant increases in home runs and run scoring during the late 20th Century (Jaffe, 2012). Finally, PED policies were also put in place for the minor leagues; however, there have been at best only small decreases in minor league run scoring since 2006. This raises the question as to whether these policies truly have had any effect on the game of baseball, and if not, what role that changes in umpire ball-strike calling behavior have had in the offensive decrease across the league. Therefore, in this section, I estimate the number of additional called strikes in the league that are likely due to umpire behavioral performance improvements, and map this to the associated changes in league offense.

To begin, Albert (2010) estimates how run scoring is expected to change when ball-strike calls by an umpire are reversed. His simulations identify ball-strike count based run expectancies, which were translated to a weighted average using Pitch f/x in Mills (2014) as approximately 0.146 runs per call change. I take this estimate and apply it to the number of additional called strikes in each year relative to the 1999 resignation strategy. Run totals for MLB come from Baseball Almanac (2014), and are presented in Table 12 alongside the change in called strikes, and expected change in run scoring from this increase in called strikes. As you can see from the table, as much as 40 percent of the decline in run scoring may be directly attributed to the increases in called strikes across the league. However, this increase may be due to a combination of pitchers throwing more to the strike zone in addition to umpires making more strike calls on the same pitches.

If pitchers have in fact changed their approach in response to more help from umpires, this strategy may have contributed even further to the decline in run scoring. Therefore, Table 13 presents the rate at which pitchers have thrown pitches within the rulebook strike zone from 2007 through 2013, using all pitches in the Pitch f/x database with locational data included, leaving a sample size of approximately 4.77 million pitches. As you can see, from 2007 to 2010, pitchers gradually increased their propensity to throw inside the zone. However, after the 2010 season, the rate at which pitchers threw within the strike zone *decreased* to a point lower than in

[16] Neyer (2011) reports that this rule did not impact any hitters, as no players had bat barrel diameters above the new maximum width.

2008, indicating that a large portion of the increase in strike calls can be attributed to increases in the umpires' propensity to call more strikes than pitchers throwing to the zone more often.

Additionally, the last two columns in Table 13 exhibit that pitchers have realized the lower half of the strike zone is being called more often by umpires, and are therefore are throwing there more often. The proportion of low pitches—defined as those below 1.75 feet from the ground—has increased by approximately 18.5 percent, from a rate of 22.1 percent to 26.2 percent across these years. [17] The average pitch height has decreased from just over 2.4 feet, to approximately 2.3 feet. It is a common belief within baseball that lower pitches are more difficult to hit, indicating that this umpire-induced change in pitch location could be contributing to the decline in offense in the league as well.

This hypothesis is tested using all pitches which umpires did not subjectively judge during the Pitch f/x era: pitches at which the batter swings. Table 14 presents estimates of the impact of vertical locational changes from 2007 through 2013 on selected batter outcomes using a logistic regression with dummy fixed effects for ball-strike count, pitch type, and individual umpire. Using a dummy variable to indicate a low pitch as defined for Table 13, there were large, statistically significant decreases in contact, in-play, and hit rate relative to pitches above 1.75 feet, making clear the additional impact of the changes in the size and shape of the zone could also have on balls at which the batter swings. Specifically, when a batter swings at a pitch that is below this low-pitch boundary, the odds of making contact decrease by approximately 73.3 percent, the odds of putting the ball in play decrease by 48.9 percent, and the odds of a hit decrease by 26.9 percent. Taken together with the increased propensity with which pitchers throw to this area, the disadvantage to hitters due to the institutional changes are clear.

Lastly, to further identify the impact of umpire behavior on offensive changes in MLB, I exploit the variation in individual umpire accuracy rates to assess their respective impact on a number of statistical measures of pitching and offense in MLB. Table 15 presents regression models of the relationship between these individual umpire accuracy rates and game offense across MLB for the years that Pitch f/x data are available. These models use yearly earned run average (*ERA*), on-base plus slugging (*OPS*), strikeouts per 9 innings pitched (*K9*), and walks per 9 innings pitched (*BB9*) as dependent variables specific to games each umpire worked behind the plate. These regression models are weighted by each umpire's number of games worked in

[17] Similar results hold for other definitions of a low pitch.

each season, and include a control variable for the season to adjust for known decreases in offense independent of umpire behavior across seasons.

Each of the four models estimates effects of correct strike and correct ball call rates of umpires in each year in the expected directions. As the rate of correct strikes increases by one percentage point, *ERA* is reduced (fewer runs allowed by pitchers per game), *OPS* decreases (more hits, extra base hits, and walks), *K9* increases (more batters strike out per game), and *BB9* decreases (fewer bases on balls per game). Noting that the rate correct strike calls have increased by substantially more than correct ball calls, the net effect on offensive statistics appears to be negative as they relate to umpire accuracy improvements over the sample period for this empirical analysis (Table 16). With an increase in the rate of correct strikes of 8.96 percentage points and increase in correct balls of 2.63 percentage points from 2007 to 2013, *ERA* is estimated to decrease by between 0.14 and 0.24 runs, or 24.1 to 41.4 percent of the total change in *ERA* in MLB during this time for the year trend, and year fixed effects models, respectively. Additionally, approximately 22.7 to 43.2 percent of the reduction in *OPS*, 3.3 to 8.9 percent of the increase in *K9*, and 51.6 to 71.0 percent of the reduction in *BB9* may be attributed to umpire accuracy improvements. The low proportion of K9 attributable to umpire accuracy changes is interesting; however, it is likely that the large change in this measure is a result of changes in the way relief pitchers are used. High velocity relievers have been used more often and used in specific matchups more often than in the past. These pitchers are much more likely to strike batters out due to their high velocity, therefore reducing the umpires' own impacts on this metric. Taken together, there is clear evidence that umpire behavior changes have strongly influenced offensive output in MLB. The simple changing of a ball to a strike has direct impacts on the game, but also induces other behaviors among pitchers and hitters. Pitchers continue to increase the rate at which they throw pitches low in the zone—where they know they are now more likely to receive strike calls from the umpire—and batters are induced to swing at these pitches, shown to have negative effects on the likelihood of making contact or getting a hit.

It is likely that MLB has experienced increases in fan support based on the idea that new PED policies have removed drug use from the game. However, based on recent news of continued PED use among popular players, as well as the evidence presented here, any claim of success in PED testing should be considered with some skepticism. Whether or not MLB has truly rid the game of PED use is a question beyond the scope of this work. The league clearly

wants fans to believe the game is clean, and this effort to appear clean has seemingly been supported by league-demanded changes in umpire behavior whether or not that was the intent of the commissioner's office. Therefore, any discussion of the impact of crackdowns on performance enhancing drug use in MLB should ensure consideration of these institutional changes within the game—along with other contributing factors—before making grand conclusions regarding the impacts of drug prevention programs. Evaluation of the strategy and successes of drug testing programs in the workplace must account for other contributing factors to changes in productivity that may not be directly related to the ability level of the agents (athletes) involved.

7 Summary and Conclusions

This paper adds to the literature by empirically evaluating performance incentive effects among an elite group of professionals—MLB umpires—as they related to incentive pay and technological improvements in monitoring and training. I find that employees (umpires) have substantially changed their ball-strike calling behavior concurrent with a number of labor disputes, merit implementation, disciplinary actions, and monitoring and training improvements. These results are apparent both in terms of increased strike calling and increased accuracy on ball and strike calls. The evidence presented here identifies three main effects: 1) MLB was able to incentivize umpires to put forth effort in conforming to the rulebook strike zone, 2) they have been able to train its umpires to improve their ability over time, and 3) the league has seen improvements in the ability of its labor pool of umpires, as those that have arrived in MLB have increased their ability levels more quickly and above those with more experience.

However, contrary to past predictions and accompanying empirical work estimating the effects of incentive pay, I find *reductions* in performance variability across time among umpires within MLB. This may be due to the sample used in this study, which includes only a few employees that have already been sorted into the top of their profession. Nevertheless, this highlights the importance of identifying limits in productivity, the role of training in changing these limits, and how predictions from theory may or may not apply to certain groups of employees already known to be elite performers. If the number of positions is capped, but performance pay and job security from collective bargaining increases the labor pool due to the prospect of higher wages, then the variability in performance of the employees at the top of the

profession may become more homogenous. While monitoring, training, and merit policies put in place by MLB have resulted in very few if any terminations of umpires since the 1999 failed resignation strategy[18], the continued increase in umpire performance indicates that there are likely believable threats to retaining one's job if minimum standards are not met. This was perhaps most clearly exhibited with the release of three umpires after 2009, and the subsequent improvements among the rest of the umpires in MLB.

Most recently, in the 2014 season, the league has 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 success rates and optimal challenge strategies within the sport. Additionally, tracking rates at which calls are overturned for specific umpires over time may also shed light on the heterogeneity in these workers in other areas of their job, and whether this heterogeneity is changing across time if and when new performance incentives are put in place through collective bargaining. This could extend the literature beyond the myopic scope here that focuses only on ball-strike calls. Finally, the implementation of replay in 2014 has 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.

[18] Though, there have been increases in discipline through suspensions.

References

Albert, Jim. 2010. Using the count to measure pitching performance. *Journal of Quantitative Analysis in Sports* 6, no. 4.

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.

Armour, Mark. 2009. A tale of two umpires: When Al Salerno and Bill Valentine got thrown out of the game. *The Baseball Research Journal* 38, no. 2:126-130.

Bai, Jushan & Perron, Pierre. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66:47-78.

Bai, Jushan & Perron, Pierre. 2003. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18:1-22.

Bai, Jushan & Perron, Pierre. 2006. Multiple structural change models: A simulation analysis. In *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, Eds. D. Corbae, S.N. Durlaff, and B.E. Hansen. New York: Cambridge University Press, 2006, 212-237.

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 Almanac. 2014. League by league totals for runs scores. *Baseball Almanac*. http://www.baseball-almanac.com/hitting/hiruns4.shtml (accessed January 20, 2014).

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/2002-pitches-batting.shtml (Accessed April 8, 2014).

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).

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.

Dubner, Stephen J. 2010. Who stole all the runs in Major League Baseball? *Freakonomics Blog*. http://freakonomics.com/2010/10/07/freakonomics-radio-who-stole-all-the-runs-in-major-league-baseball/ (accessed June 19, 2014).

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.

Fort, Rodney & Lee, Young Hoon. 2006. Stationarity and MLB attendance analysis. *Journal of Sports Economics* 7:408-415.

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.

Gu, Chong & Wahba, Grace. 1993. Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society, Series B (Methodological)* 55:353-368.

Henderson, Joe. 2011. Scoring down across baseball as game emerges from steroid era. *The Tampa Tribune*. http://tbo.com/list/columns-jhenderson/scoring-down-across-baseball-as-game-emerges-from-steroid-era-242357 (accessed June 19, 2014).

Jaffe, Jay. 2012. What really happened in the juiced era? In *Extra Innings: More baseball between the numbers from the team at Baseball Prospectus*, ed. Steven Goldman, New York: Basic Books.

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.

Lee, Junsoo & Strazicich, Mark C. 2001. Break point estimation and spurious rejections with endogenous unit root tests. *Oxford Bulletin of Economics and Statistics* 63:535-558.

Lee, Junsoo & Strazicich, Mark C. 2003. Minimum LM unit root test with two structural breaks. *Review of Economics and Statistics* 85:1082-1089.

Lee, Junsoo & Strazicich, Mark C. 2004. Minimum LM unit root test with one structural break. Working Paper, Department of Economics, Appalachian State University.

Lindenauer, Peter K., Remus, Denise, Roman, Sheila, Rothberg, Michael B., Benjamin, Evan M., Ma, Allen, & Bratzler, Dale W. 2007. Public reporting andpay 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.

Mathewson, Jesse Douglas. 2010. Benefit of the doubt: Odd patterns in umpire compensation. *Beyond the Boxscore*. http://www.beyondtheboxscore.com/ 2010/12/24/1892898/benefit-of-the-doubt-odd-patterns-inumpire-compensation (accessed July 15, 2011).

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.

MLB.com. 2010. Index of game components. http://gd2.mlb.com/components/game/mlb/ (accessed November 11, 2013).

Moore, Matt. 2009. Resigned to their fate. *Referee Magazine (October 2009).* https://www.arbitersports.com/MyRefereeApp/Print.aspx?mod= PrintArticle&pid=110139 (accessed January 21, 2013).

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.

Los Angeles Times. 1995. Picketing umpires protest replacements: Lockout: Union, management await hearing before Ontario Labor Relations Board. *Los Angeles Times*. http://articles.latimes.com/1995-04-26/sports/sp-58960_1_scab-umpires (accessed March 25, 2014).

SDABU. 2014. San Diego Adult Baseball Umpires: History of umpiring. http://www.sdabu.com/history_main.htm (accessed April 1, 2014).

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.

Neyer, Rob. 2011. 2010's year of the pitcher can't be explained by thinner bats. *SB Nation*. http://www.sbnation.com/mlb/2011/2/8/1981731/year-of-the-pitcher-not-explained-by-bats (accessed August 10, 2014).

Nightengale, Bob. 2010. Yer out! Three umpire bosses fired over blown 2009 playoffs calls. *USA Today*. http://usatoday30.usatoday.com/sports/ baseball/2010-03-07-umpire-supervisors-fired_N.htm (accessed June 24, 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.

Perron, P. (1988). Trends and random walks in macroeconomic time series: Further evidence from a new approach. *Journal of Economic Dynamics and Control, 12*, 297-332.

Perron, Pierre. 1989. The Great Crash, the oil price shock and the unit root hypothesis. *Econometrica* 57:1361-1401.

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.

Roegele, Jon. 2014. The strike zone during the Pitch f/x era. *The Hardball Times*. http://www.hardballtimes.com/the-strike-zone-during-the-pitchfx-era/ (accessed July 5, 2014).

Rymer, Zachary D. 2013. Analyzing how PED testing has impacted offensive stats. *Bleacher Report*. http://bleacherreport.com/articles/1486347-analyzing-how-ped-testing-has-impacted-offensive-stats (accessed June 19, 2014).

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.

Stark, Jayson. 1995. Umpires settle off-field argument, five-year contract puts end to lockout. *Philly.com*. http://articles.philly.com/1995-05-02/sports/25675625_1_umpires-union-head-picket-line-richie-phillips (accessed March 20, 2014).

Tainsky, Scott, Mills, Brian M., & Winfree, Jason A. 2013. An examination of potential discrimination among MLB umpires. *Journal of Sports Economics*. DOI: 10.1177/1527002513487740.

Walsh, John. 2010. The compassionate umpire. *The Hardball Times*. http://www.hardballtimes.com/the-compassionate-umpire/ (accessed November 15, 2011).

Weber, Bruce. 2009. *As They See 'Em: A Fan's Travels in the Land of Umpires*. New York: Scribner.

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

Wood, Simon N. 2000. Modeling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society (B)* 62:413-428.

Wood, Simon N. 2003. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* 65:95-114.

Wood, Simon N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 99:673-686.

Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R.* FL: Chapman Hall, Taylor & Francis Group, LLP.

Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73:3-36.

Table 1: Umpire Historical Events of Interest

Year	Changes
1991	Strike and subsequent merit pay initiated through postseason assignment at beginning of season
1995	Restructured compensation, base raises for umpires; lockout ends shortly after season start
1999	Failed Resignation Strategy; Umpires merged at MLB level after 1999 season
2001	QuesTec installed in 4 parks prior to season
2004	New agreement on use of QuesTec as performance standard prior to season
2009	Implementation of ZE System as performance standard prior to season start
2010	Firing of 3 senior umpires after 2009 season

Table 2: Yearly Changes in Strike Rate, Called Strike Rate, and Variability in Strike Rate[a]

Year	Umps	Pitches	*TotalStrikeRate*	*wCV*	Called Pitches	*CalledStrikeRate*
1988	65	578,644	0.61770	0.01561	298,728	0.28665
1989	65	575,550	0.61473	0.01559	299,081	0.28381
1990	68	547,652	0.61216	0.01546	285,689	0.28243
1991	78	606,917	0.60974	0.01657	320,091	0.28632
1992	77	601,635	0.61215	0.01794	316,799	0.28904
1993	76	632,233	0.61030	0.01574	333,426	0.28561
1994	77	451,904	0.60839	0.01759	239,316	0.28567
1995	130	554,918	0.60817	0.02211	295,031	0.28704
1996	81	604,241	0.61004	0.01773	321,025	0.28934
1997	88	621,672	0.61134	0.01559	331,559	0.29541
1998	87	668,253	0.61351	0.01436	380,912	0.29746
1999	103	712,976	0.60778	0.01554	390,325	0.29236
2000	84	729,814	0.60764	0.01499	394,113	0.29402
2001	88	711,322	0.62234	0.01475	377,353	0.30976
2002	84	711,002	0.62033	0.01456	378,061	0.30612
2003	84	714,789	0.62232	0.01438	382,123	0.31250
2004	84	723,273	0.62082	0.01344	386,946	0.31046
2005	85	708,797	0.62652	0.01279	377,011	0.31629
2006	83	720,657	0.62330	0.01280	383,780	0.31195
2007	85	725,418	0.62273	0.01290	387,388	0.31307
2008	83	727,729	0.62206	0.01299	389,376	0.31313
2009	85	731,396	0.62077	0.01232	394,793	0.31743
2010	85	725,140	0.62427	0.01183	390,651	0.32307
2011	83	718,796	0.62619	0.01195	384,339	0.32250
2012	82	715,687	0.62820	0.01171	382,647	0.32632
2013	82	718,733	0.62948	0.01014	383,217	0.32500

a. Calculations here include temporary umpires in 1991 and 1995 labor disputes.

TABLE 3: Unit Root Results for Long-Term Measures

	CalledStrikeRate	***TotalStrikeRate***	***TotalStrikeRate***	***wRateCV***	***wRateCV***
Pre-Post 1999	**N**	**N**	**Y**	**N**	**Y**
ADF (*p*)[a]	-0.116 (1)	-0.508 (1)	-0.365 (1)	-0.652 (7)	-0.990 (1)
ADF (*p*)	-3.036 (1)	-2.892 (1)	-2.726 (1)	-3.006 (1)	-3.324 (1)*
P-P (*l*)[b]	-0.029 (2)	-1.729 (2)	-0.409 (2)	-0.994 (2)	-1.070 (2)
P-P (*l*)	-3.820 (2)**	-3.115 (2)	-3.379 (2)*	-2.606 (2)	-3.286 (2)*
LM-2[c] (*k*)	-24.85 (8)***	-11.13 (7)***	-8.03 (7)***	-9.06 (5)***	-7.36 (6)***
T_1	2001***	1999***	1999***	1998***	2002***
T_2	2008	2006	2006	2005*	2009**
λ	λ = (0.54, 0.81)	λ = (0.46, 0.73)	λ = (0.46, 0.73)	λ = (0.42, 0.69)	λ = (0.58, 0.85)
LM-1 (*k*)	-30.55 (8)***	-13.23 (7)***	-12.83 (7)***	-9.28 (5)***	-4.80 (5)**
T_1	2001***	1999***	1999***	2001***	2003**
λ	λ = 0.54	λ = 0.46	λ = 0.46	λ = 0.54	λ = 0.62

***, **, * denote statistical significance at the 99%, 95%, and 90% level, respectively. a. Lag for ADF test chosen by Schwarz Information Criteron. b. Lag for P-P test chosen by the Newey-West procedure. c. Lags chosen for LM-2 and LM-1 tests chosen using procedure from Lee & Strazicich (2003;2004).

TABLE 4: Sequential Structural Change Testing for Long-Term Measures

Measure	Pre-Post 1999	$SupF_t(1)$	$SupF_t(2)$	SupF(2/1)	Breaks
CalledStrikeRate	**N**	28.84***	22.62***	11.86*	**1**
TotalStrikeRate	**N**	67.72***	45.43***	11.18	**1**
TotaleStrikeRate	**Y**	59.24***	41.66***	13.43**	**2**
wRateCV	**N**	51.08***	57.35***	4.12	**1**
wRateCV	**Y**	20.00***	15.06***	6.58	**1**

***, **, * denote statistical significance at the 99%, 95%, and 90% level, respectively.

TABLE 5: Estimated Break Dates and Confidence Intervals for Long-Term Measures

Measure	Pre-Post 1999	T_1	T_2
CalledStrikeRate	**N**	**2000** [99, 01]	----- -----
TotalStrikeRate	**N**	**2000** [99, 01]	----- -----
TotalStrikeRate	**Y**	**2000** [99, 01]	**2008** [07, 09]
wRateCV	**N**	**1996** [95, 99]	----- -----
wRateCV	**Y**	**1996** [95, 00]	----- -----

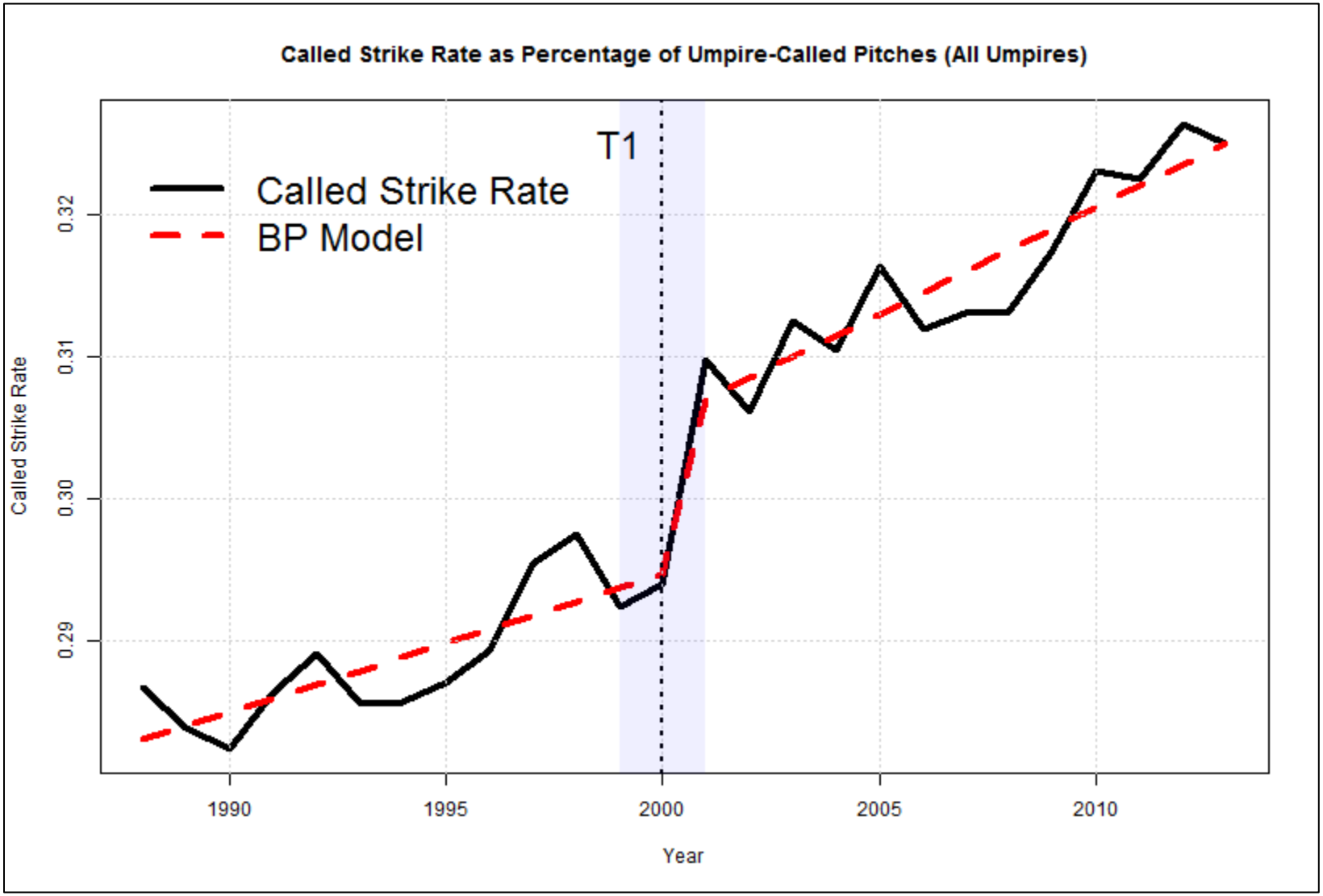
*Brackets denote 90% confidence interval for break date.

TABLE 6: Breakpoint Regression Models for Long-Term Measures

Measure	Pre-Post 1999	α_1[a]	β_1	α_2	β_2	α_3	β_3
CalledStrikeRate	***N***	0.00097***	0.28207***	0.00150***	0.28589***	-----	-----
t-value		*(4.70)*	*(172.38)*	*(7.30)*	*(68.16)*	-----	-----
TotalStrikeRate	***N***	-0.00050***	0.61457***	0.00052***	0.61332***	-----	-----
t-value		*(-2.95)*	*(460.76)*	*(3.12)*	*(179.38)*	-----	-----
TotalStrikeRate	***Y***	-0.00029*	0.61264***	0.00029	0.61770***	0.00220***	0.57326***
t-value		*(-2.00)*	*(534.42)*	*(0.96)*	*(116.41)*	*(3.58)*	*(38.70)*
wRateCV	***N***	0.00044***	0.01467***	-0.00029***	0.01854***	-----	-----
t-value		*(4.48)*	*(26.61)*	*(-7.67)*	*(26.46)*	-----	-----
wRateCV	***Y***	0.00043**	0.01486***	-0.00029***	0.01844***	-----	-----
t-value		*(2.74)*	*(16.89)*	*(-4.76)*	*(16.49)*	-----	-----

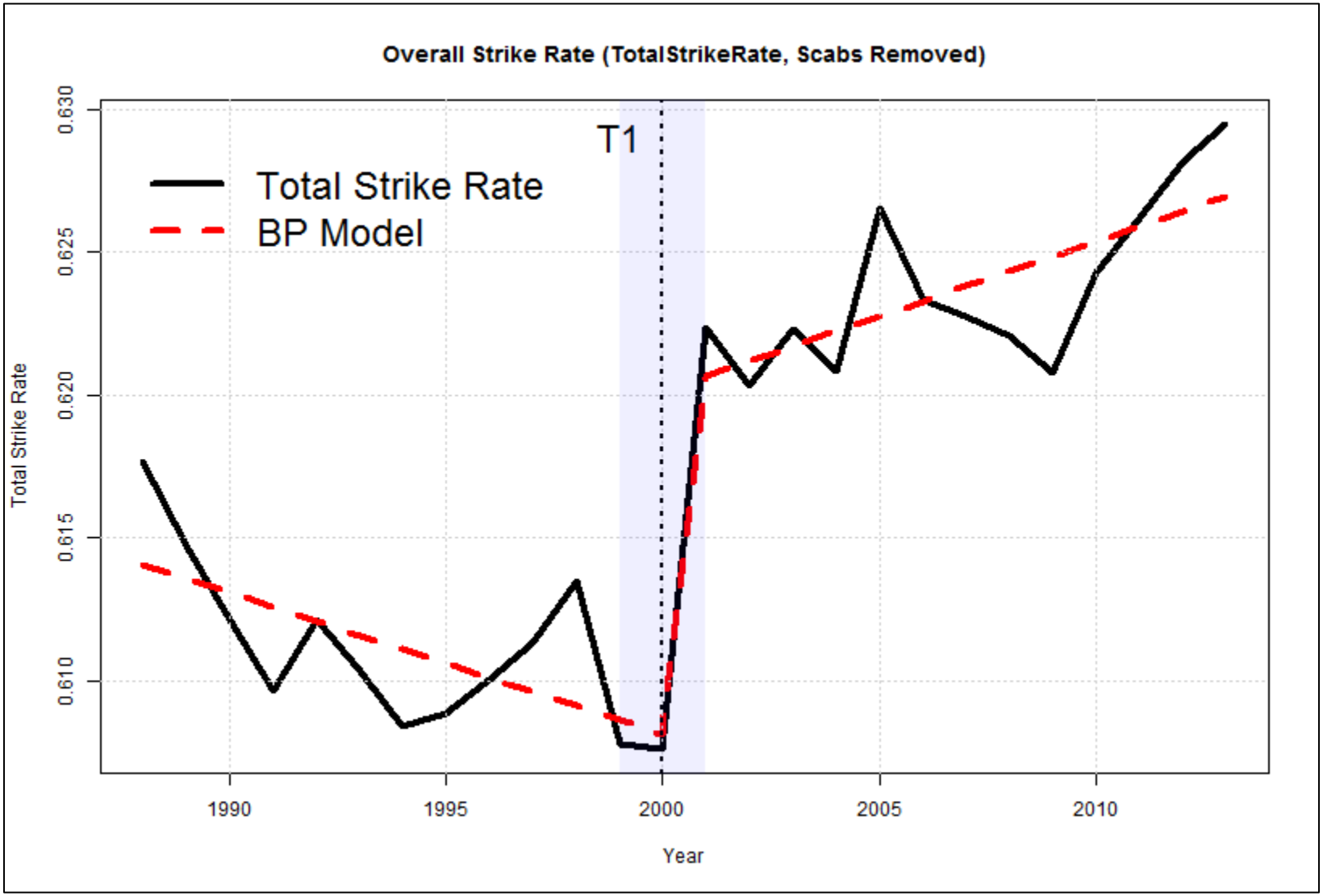
***, **, * denote statistical significance at the 99%, 95%, and 90% level, respectively. a. α_j and β_j refer to coefficient estimates the trend and intercept for regime j, respectively.

FIGURE 1: Structural Change Model of *CalledStrikeRate* (All Umpires—Includes Scabs)



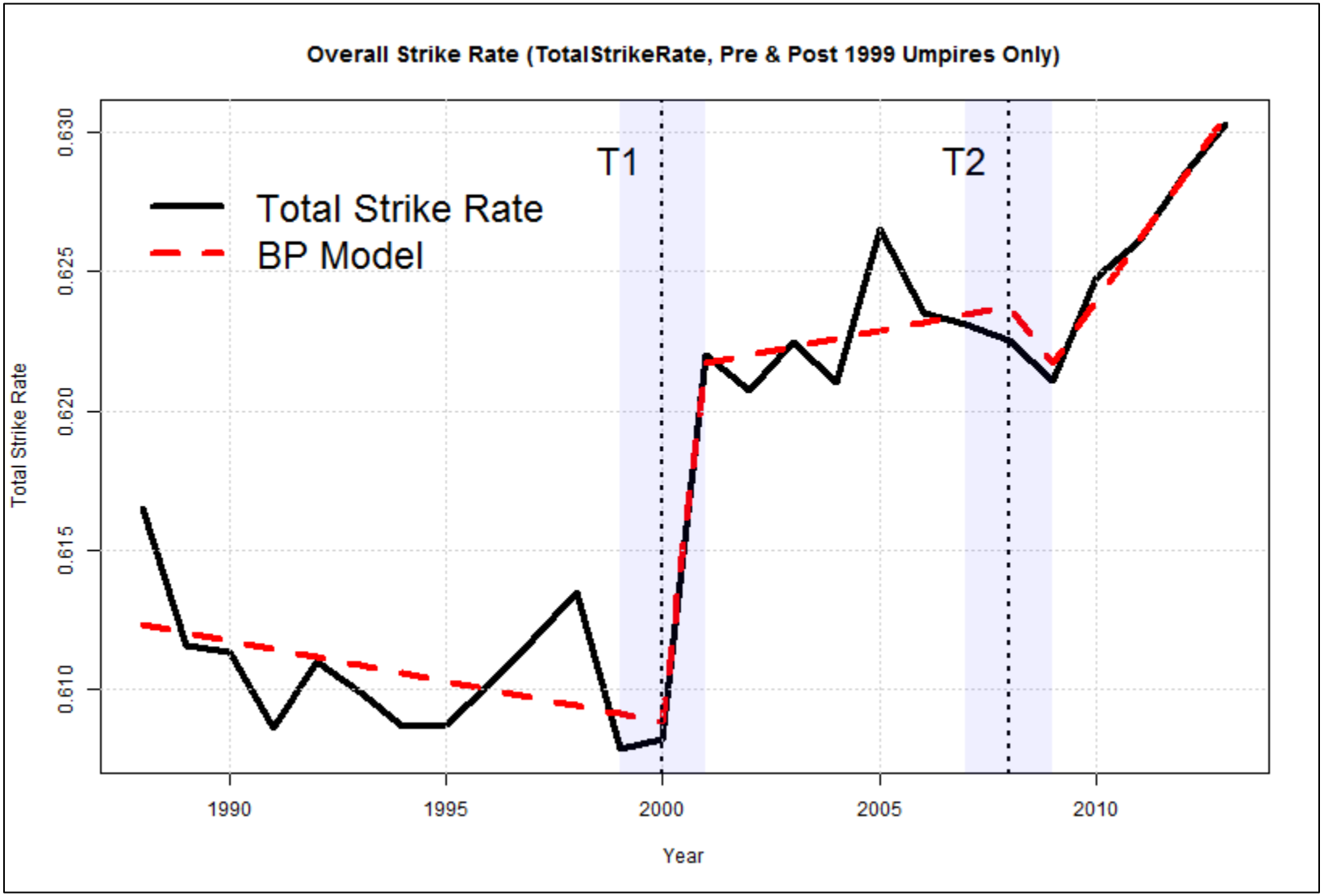
a. Shaded area presents the 90% confidence interval for the respective structural change date.

FIGURE 2: Structural Change Model of *TotalStrikeRate* (All Non-Scab Umpires in Sample)



a. Shaded area presents the 90% confidence interval for the respective structural change date.

FIGURE 3: Structural Change Model of *TotalStrikeRate* (Umpires Working Before and After 1999 Resignation)



a. Shaded area presents the 90% confidence interval for the respective structural change date.

FIGURE 4: Pre-2000 Strike Rate by Post-1999 Employment

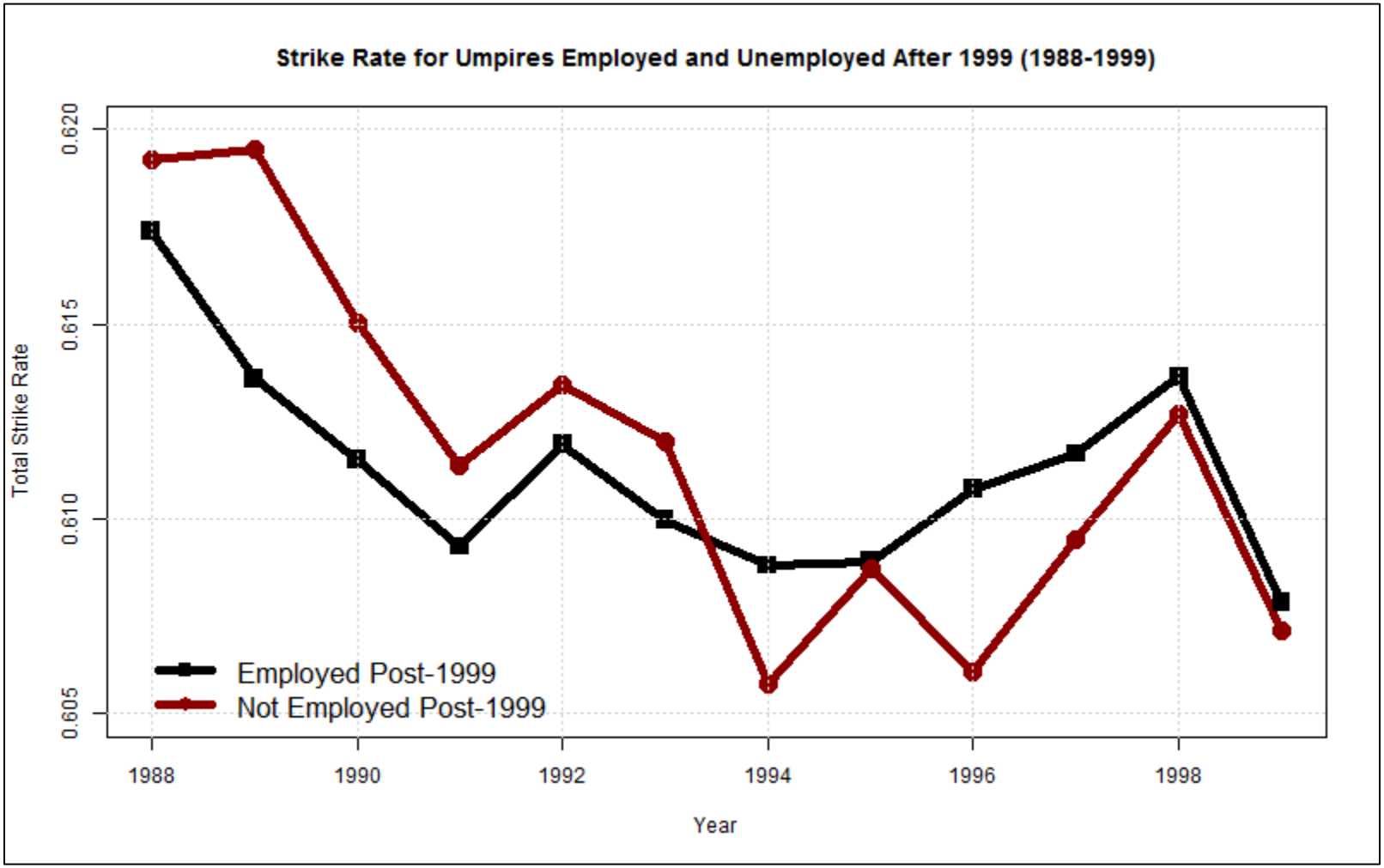


FIGURE 5: Pre-2000 Strike Rate Difference by Post-1999 Employment

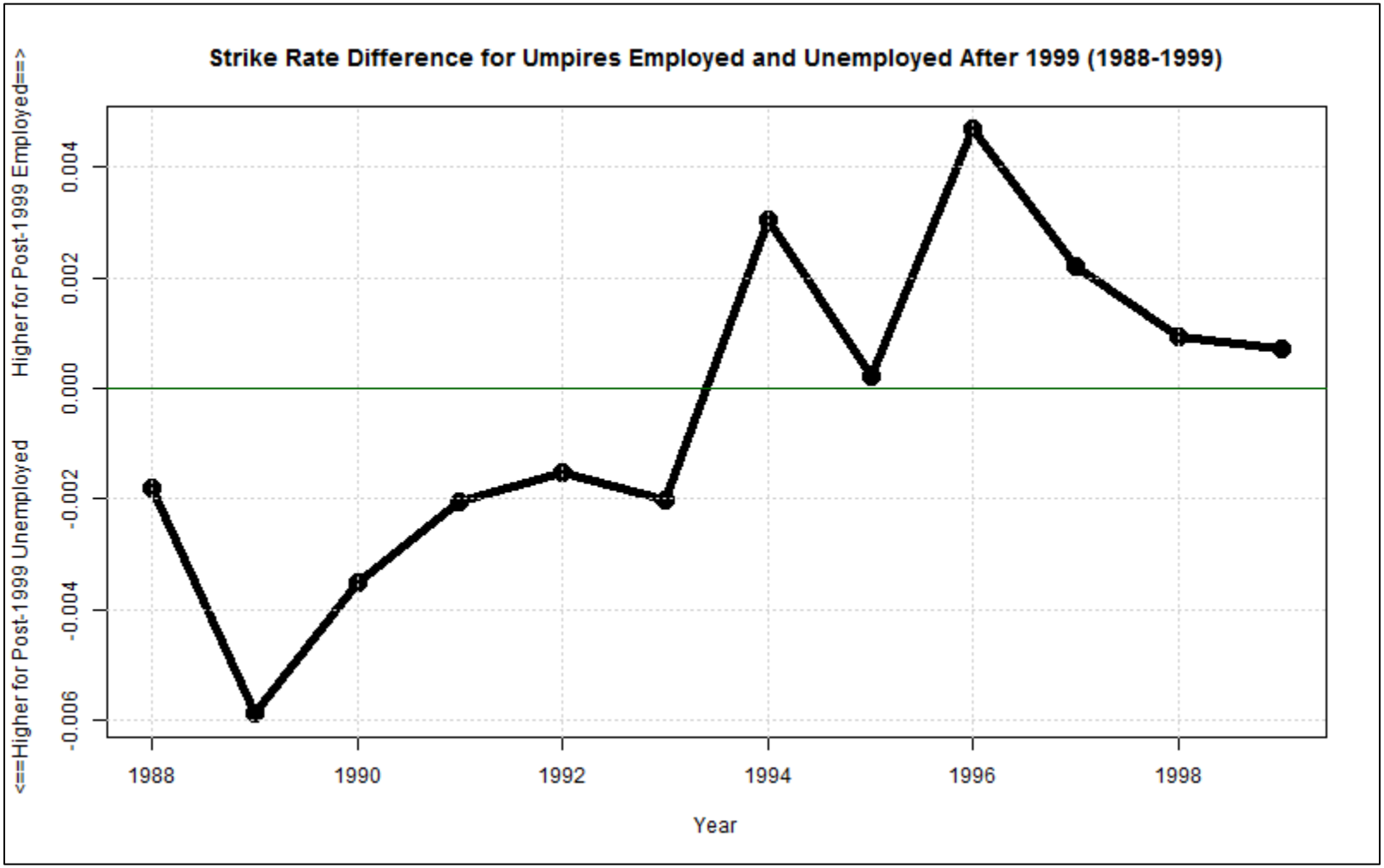
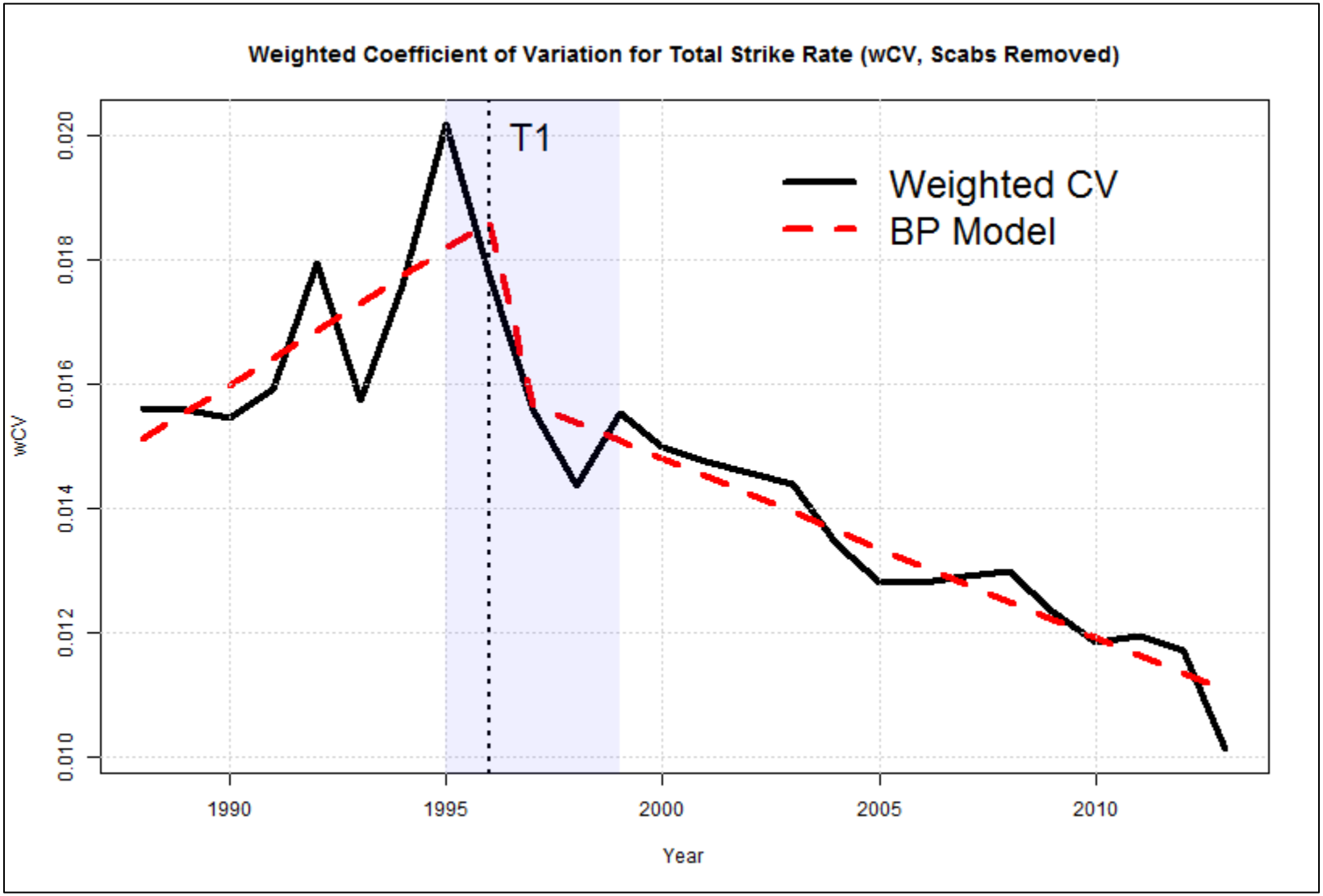


TABLE 7: Weighted OLS Regression Results for Umpire Pre-1999 Strike Rates by Post-1999 Employment

Var.	Coef. Estimate
Constant	0.61375***
S.E.	*(0.00083)*
Year[c]	-0.00034***
S.E.	*(0.00011)*
Not Employed After 1999	0.00444**
S.E.	*(0.00203)*
Year*Not Employed After 1999	-0.00065**
S.E.	*(0.00028)*

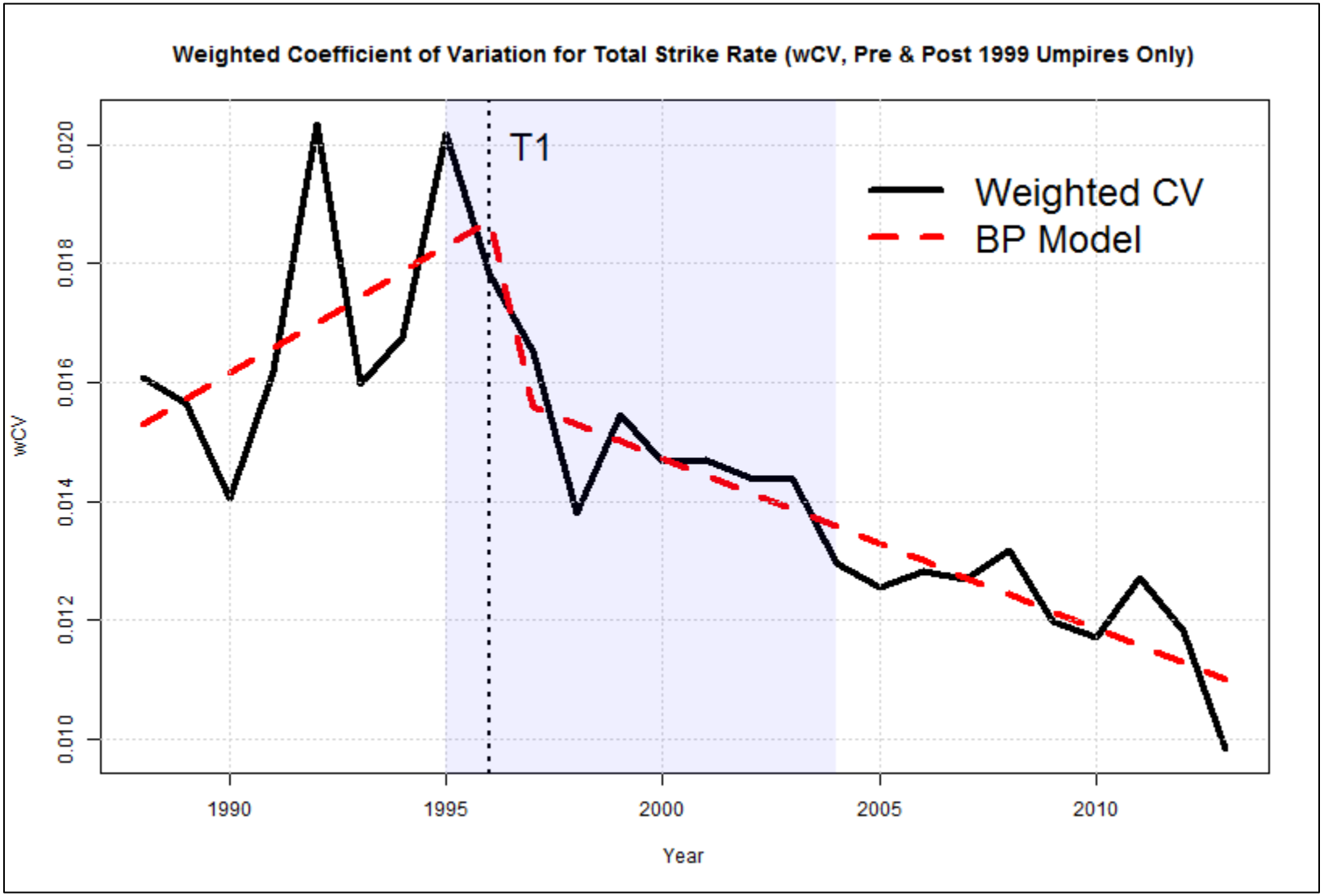
a. Least squares regression of yearly individual umpire strike rates. b. Regression weighted by $\sqrt{n_i}$, where n_i is the number of called pitches for each umpire-year observation.

FIGURE 6: Structural Change Model of *wCV* (All Non-Scab Umpires in Sample)



a. Shaded area presents the 90% confidence interval for the respective structural change date.

FIGURE 7: Structural Change Model of *wCV* (Umpires Working Before and After 1999 Resignation)



a. Shaded area presents the 90% confidence interval for the respective structural change date.

TABLE 8: Called Strike Rate by Monitored and Non-Monitored Conditions (1998-2008)

Var.	Coef. Estimate
Constant	0.28920***
S.E.	*(0.01287)*
Year	0.00260***
S.E.	*(0.00010)*
QuesTec	0.01920***
S.E.	*(0.00271)*
Year*QuesTec	-0.00198***
S.E.	*(0.00032)*

a. Least squares regression of strike rates clustered by umpire-pitcher-game match-up. b. Regression weighted by $\sqrt{n_i}$, where n_i is the number of called pitches for each pitcher-umpire-game observation. c. Umpire fixed effects included. d. Data come directly from Tainsky et al. (2013). c. Year is operationalized as a time variable from $(Year - 1997)$.

FIGURE 8: Strike Rate Difference by QuesTec Presence

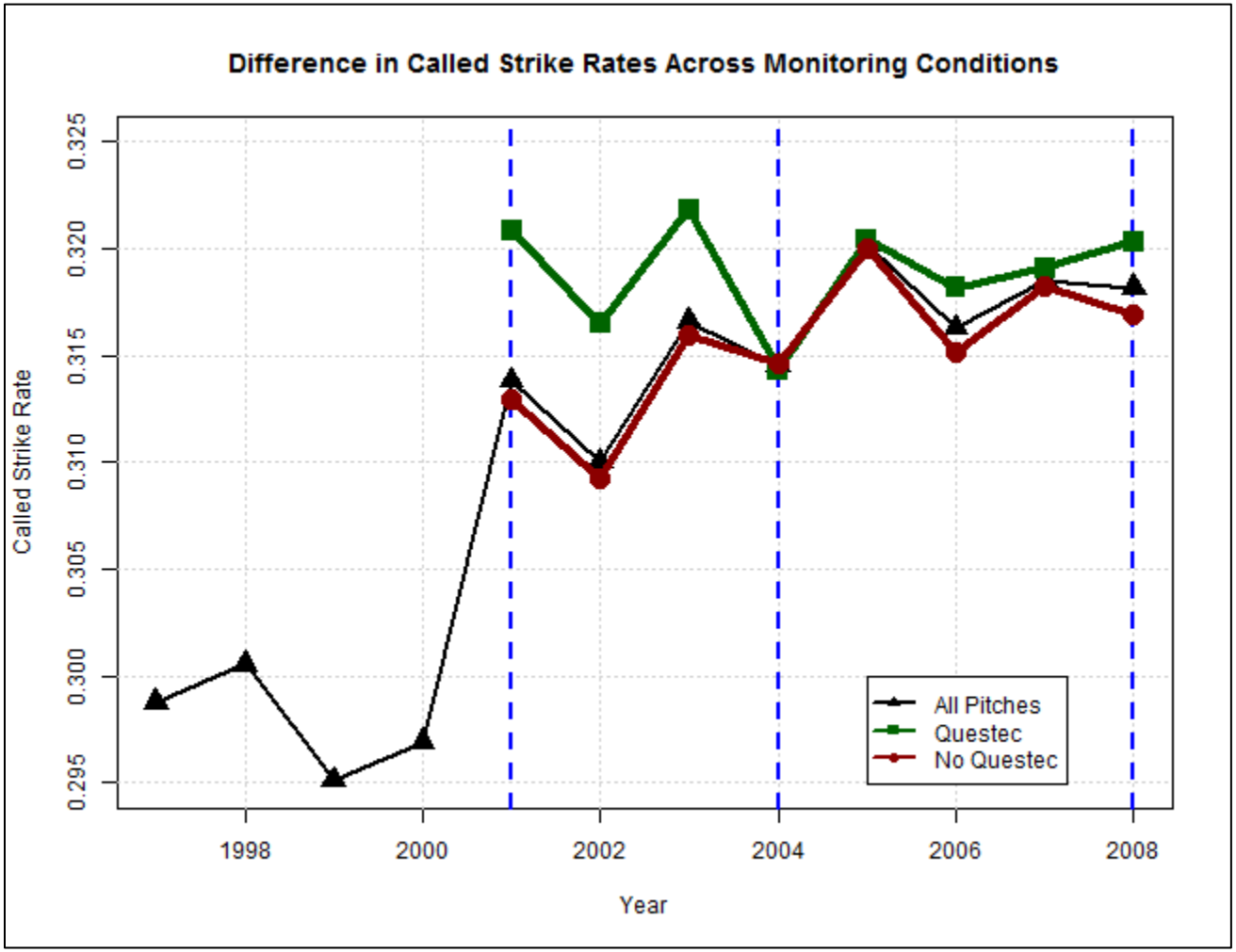


FIGURE 9: Exhibition of Correct Call and Incorrect Call Classifications

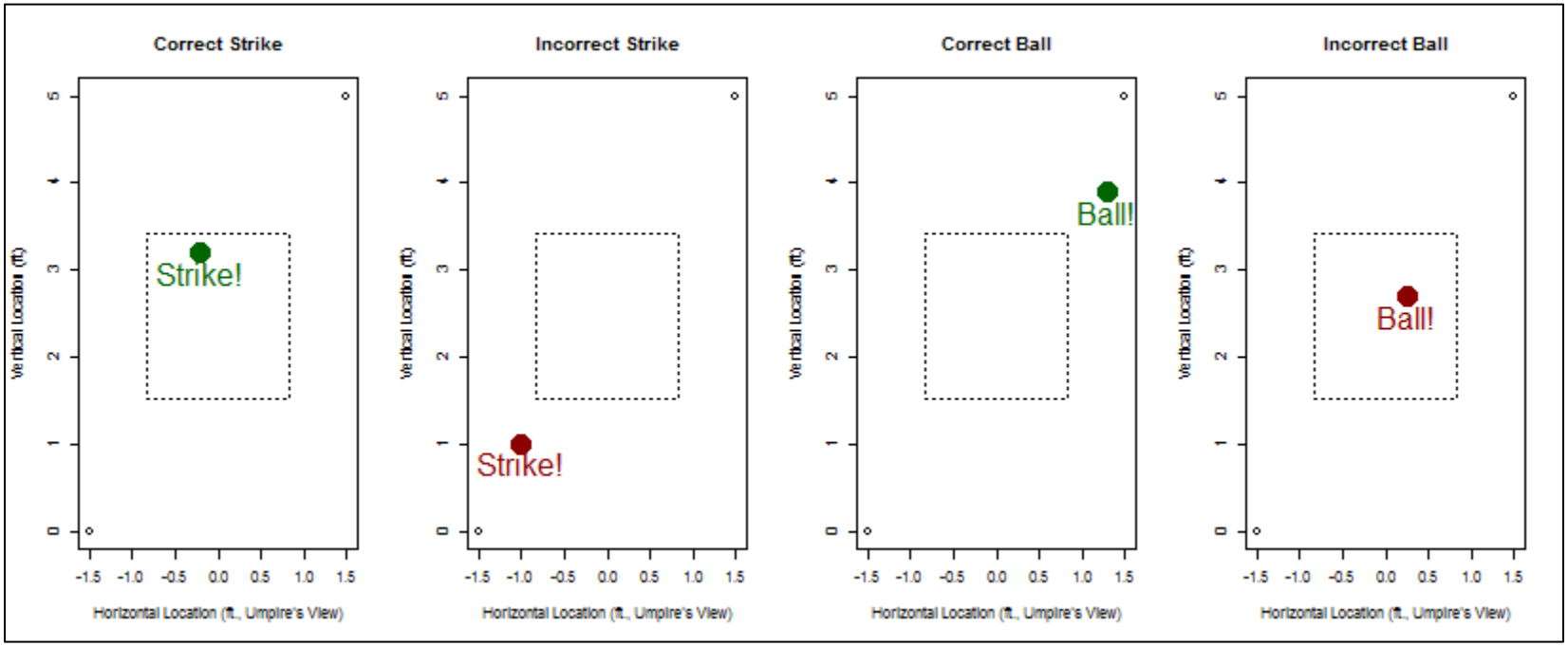


FIGURE 10: Sensitivity and Specificity of Umpire Strike Calls

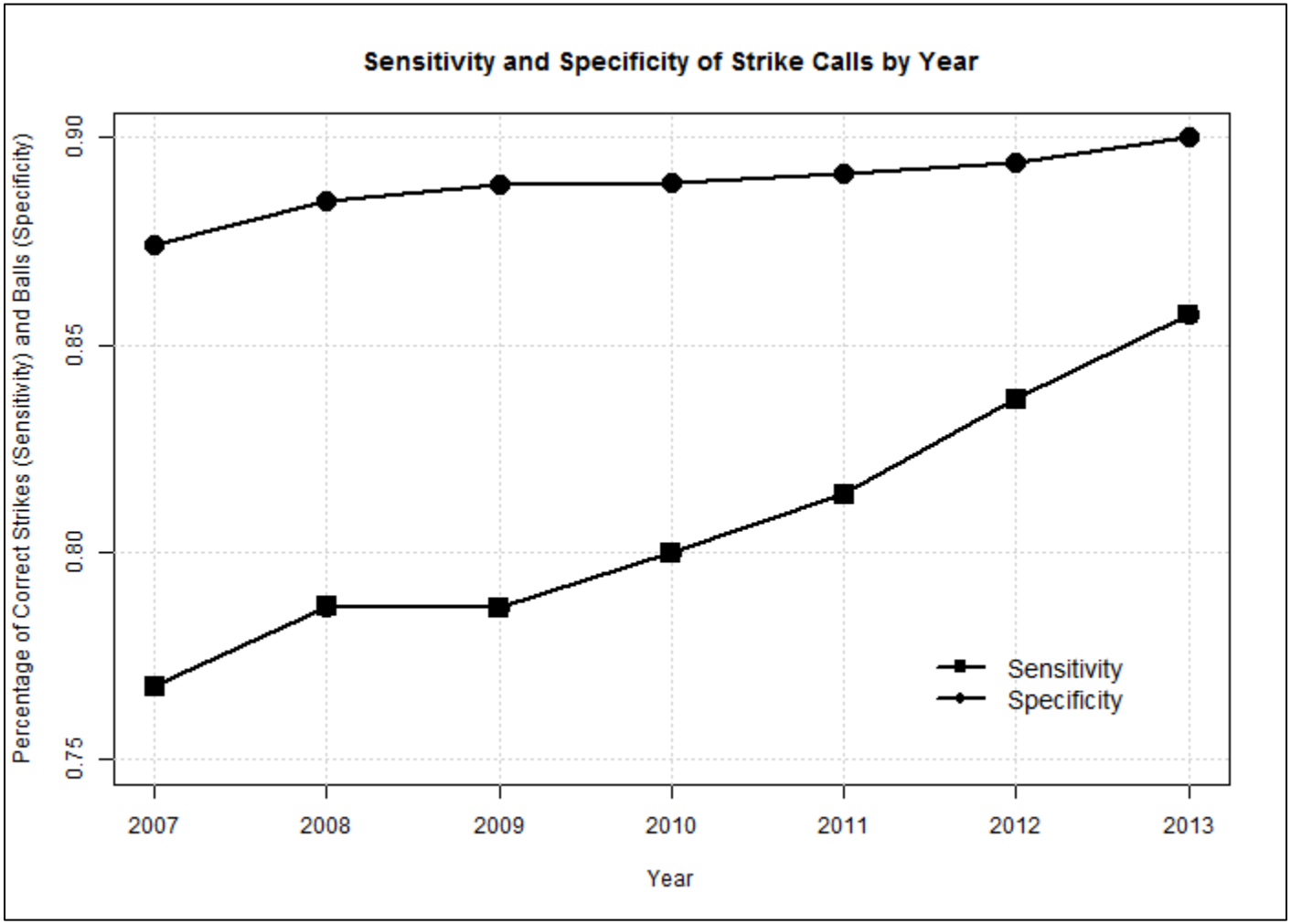
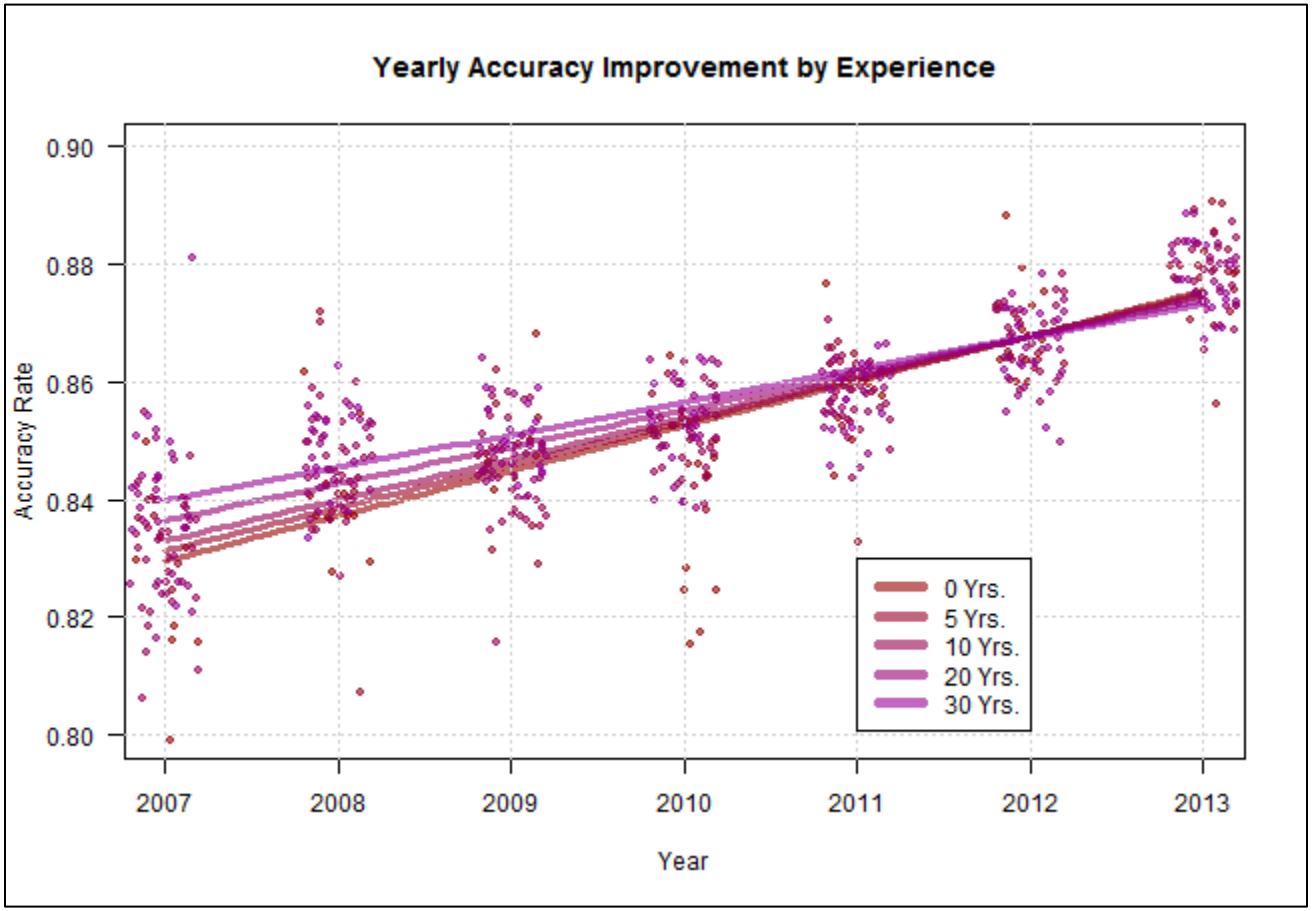


TABLE 9: Weighted OLS Regressions for Umpire Accuracy Rate Changes by Experience

	(1)	(2)	(3)	(4)
Umpire F.E.s	N	Y	N	Y
Constant	-13.100***	-12.280***	-15.340***	-14.520***
S.E.	*(0.50840)*	*(0.57090)*	*(1.0730)*	*(1.0780)*
Year	0.00695***	0.00654***	0.00806***	0.00765***
S.E.	*(0.00025)*	*(0.00029)*	*(0.00053)*	*(0.00054)*
Experience	-0.00017***	0.00016	0.14630**	0.14000**
S.E.	*(0.00006)*	*(0.00022)*	*(0.06169)*	*(0.05725)*
Year*Experience	-----	-----	-0.00007**	-0.00007**
S.E.	-----	-----	*(0.00003)*	*(0.00003)*

***, **, * refer to statistical significance at the 99%, 95%, and 90% levels, respectively. Panel regressions weighted by $\sqrt{n_{i,t}}$.

FIGURE 11: Umpire Accuracy Rates by Experience[a]



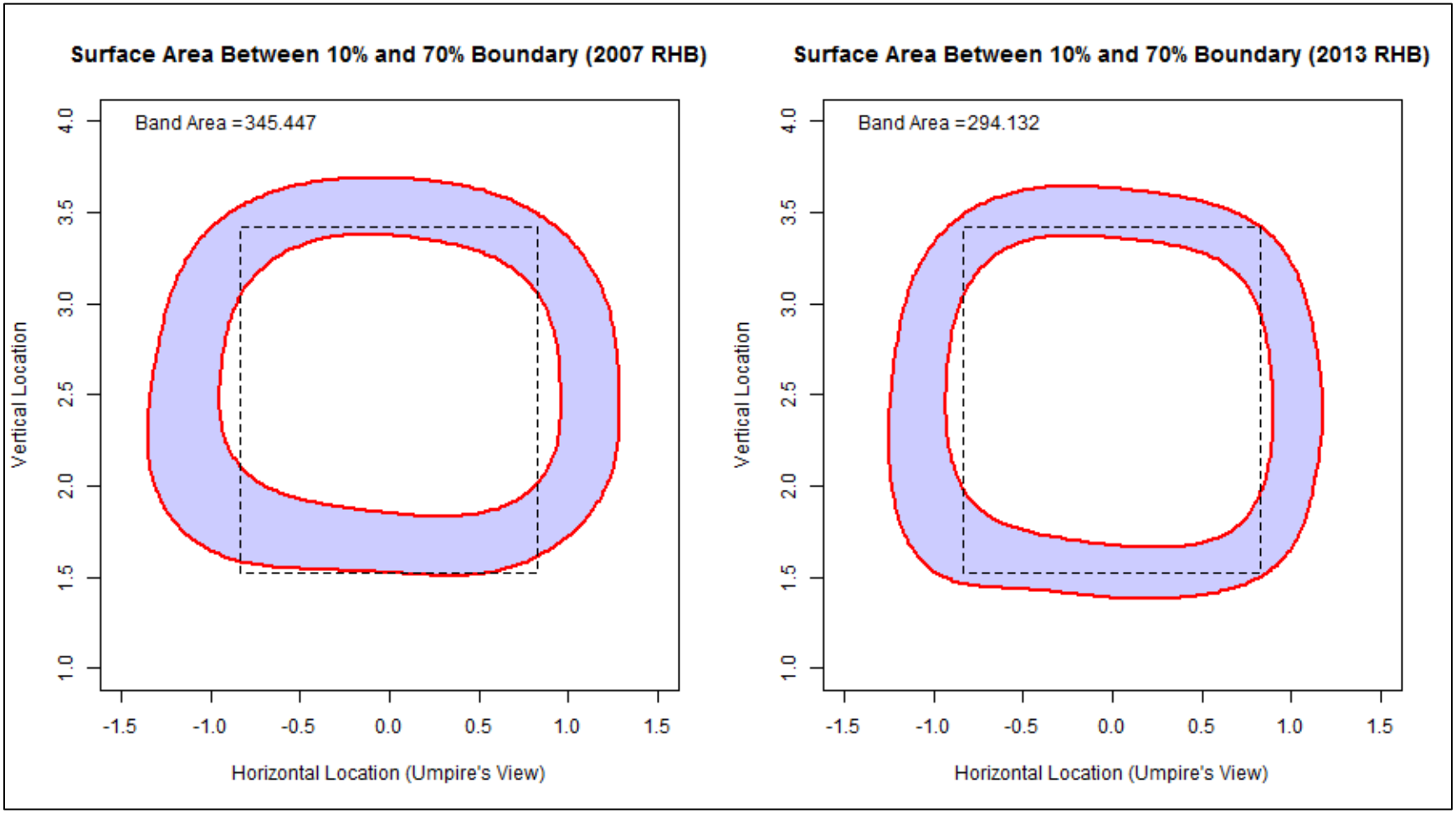
a. Figure refers to estimates in Column 4 of Table 9.

TABLE 10: Weighted OLS Regressions for Umpire Accuracy Rate Changes by Debut

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Umpire F.E.s	N	N	Y	N	N	Y	N	N	Y
Constant	-12.760***	-14.190***	-14.150***	-12.830***	-15.340***	-15.490***	-1.2900***	-26.114***	-25.170***
S.E.	*(0.50710)*	*(1.1580)*	*(1.0440)*	*(0.51360)*	*(1.4290)*	*(1.3150)*	*(0.52190)*	*(4.4130)*	*(3.5670)*
Year	0.00678***	0.00749***	0.00747***	0.00681***	0.00806***	0.00813***	0.00685***	0.01342***	-0.01294***
S.E.	*(0.00025)*	*(0.00058)*	*(0.00052)*	*(0.00026*	*(0.00071)*	*(0.00065)*	*(0.00026)*	*(0.00219)*	*(0.00177)*
Pre-2000 Debut	-0.00511***	1.7650	1.8540	-----	-----	-----	-----	-----	-----
S.E.	*(0.00119)*	*(1.2880)*	*(1.1400)*	-----	-----	-----	-----	-----	-----
Pre-2004 Debut	-----	-----	-----	-0.00358***	2.8830*	3.2180**	-----	-----	-----
S.E.	-----	-----	-----	*(0.00136)*	*(1.5310)*	*(1.3870)*	-----	-----	-----
Pre-2009 Debut	-----	-----	-----	-----	-----	-----	-0.00269	13.397***	12.740**
S.E.	-----	-----	-----	-----	-----	-----	*(0.00231)*	*(4.4437)*	*(3.5910)*
Year*Pre-2000	-----	-0.00088	-0.00092	-----	-----	-----	-----	-----	-----
S.E.	-----	*(0.00064)*	*(0.00057)*	-----	-----	-----	-----	-----	-----
Year*Pre-2004	-----	-----	-----	-----	-0.00144*	-0.00160**	-----	-----	-----
S.E.	-----	-----	-----	-----	*(0.00076)*	*(0.00069)*	-----	-----	-----
Year*Pre-2009	-----	-----	-----	-----	-----	-----	-----	-0.00666***	-0.00633***
S.E.	-----	-----	-----	-----	-----	-----	-----	*(0.00221)*	*(0.00179)*

***, **, * refer to statistical significance at the 99%, 95%, and 90% levels, respectively. Panel regressions weighted by $\sqrt{n_{i,t}}$.

FIGURE 12: Exhibition of Difference Bands Between Strike Zone Boundary Definitions



*Band area reported in square inches.

FIGURE 13: Exhibition of Strike Zone Overlap Measurement

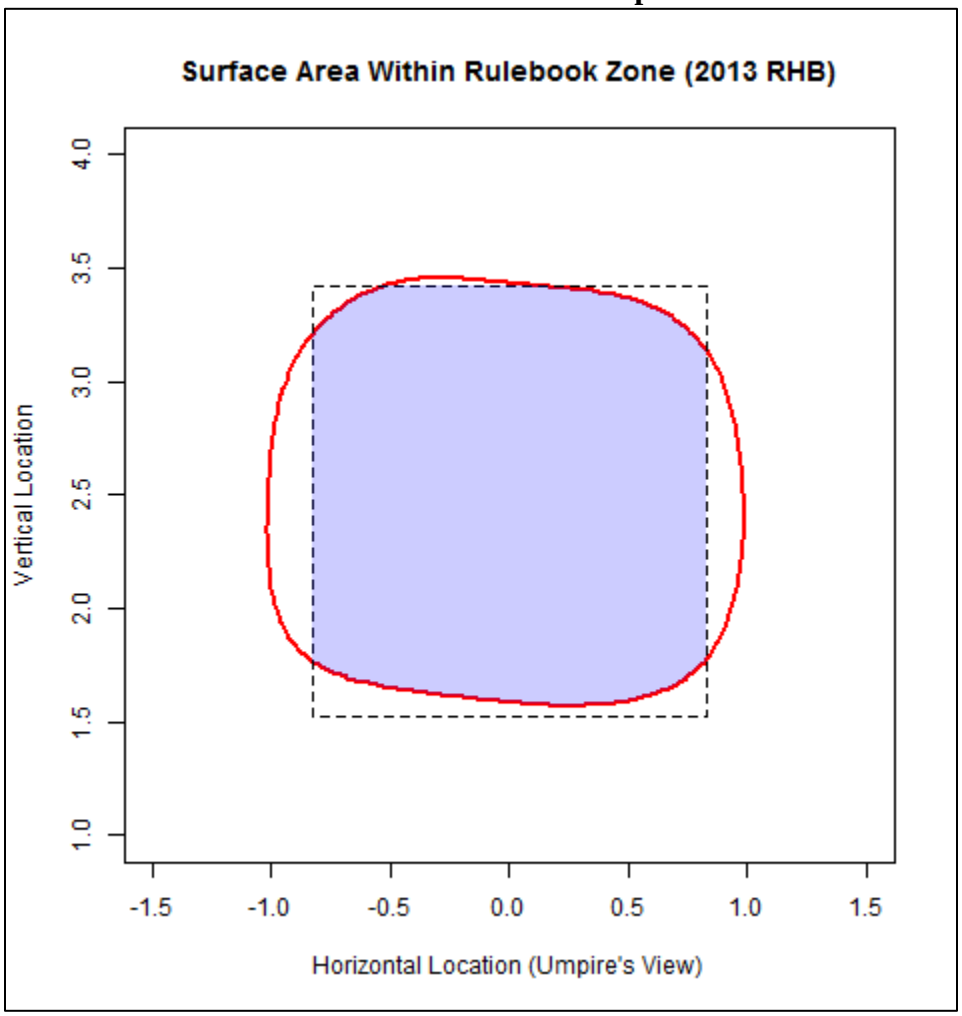


FIGURE 14: Visualization of 50% Strike Zone

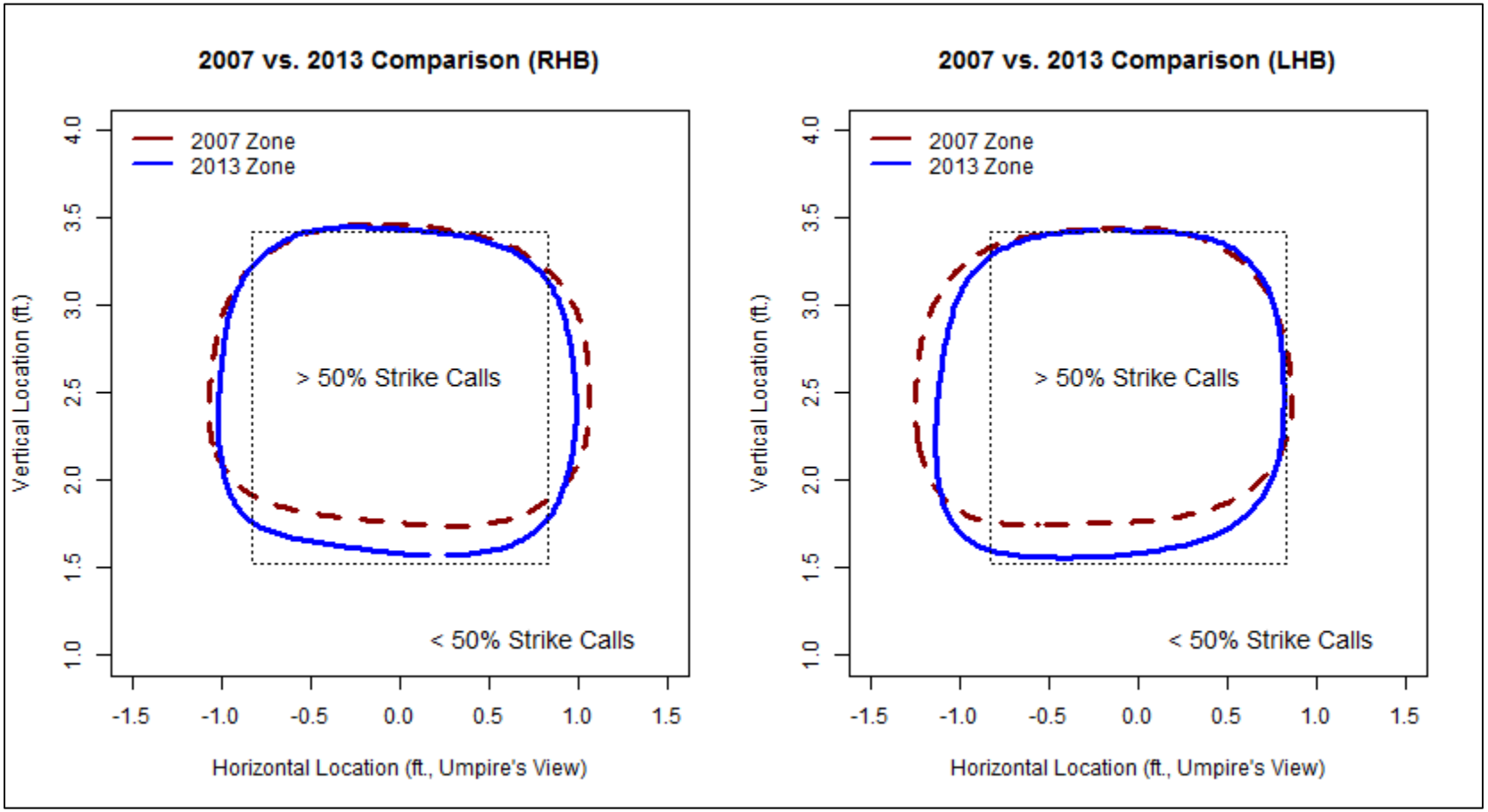


FIGURE 15: Surface Area Growth of 50% Strike Zone Boundary by Year

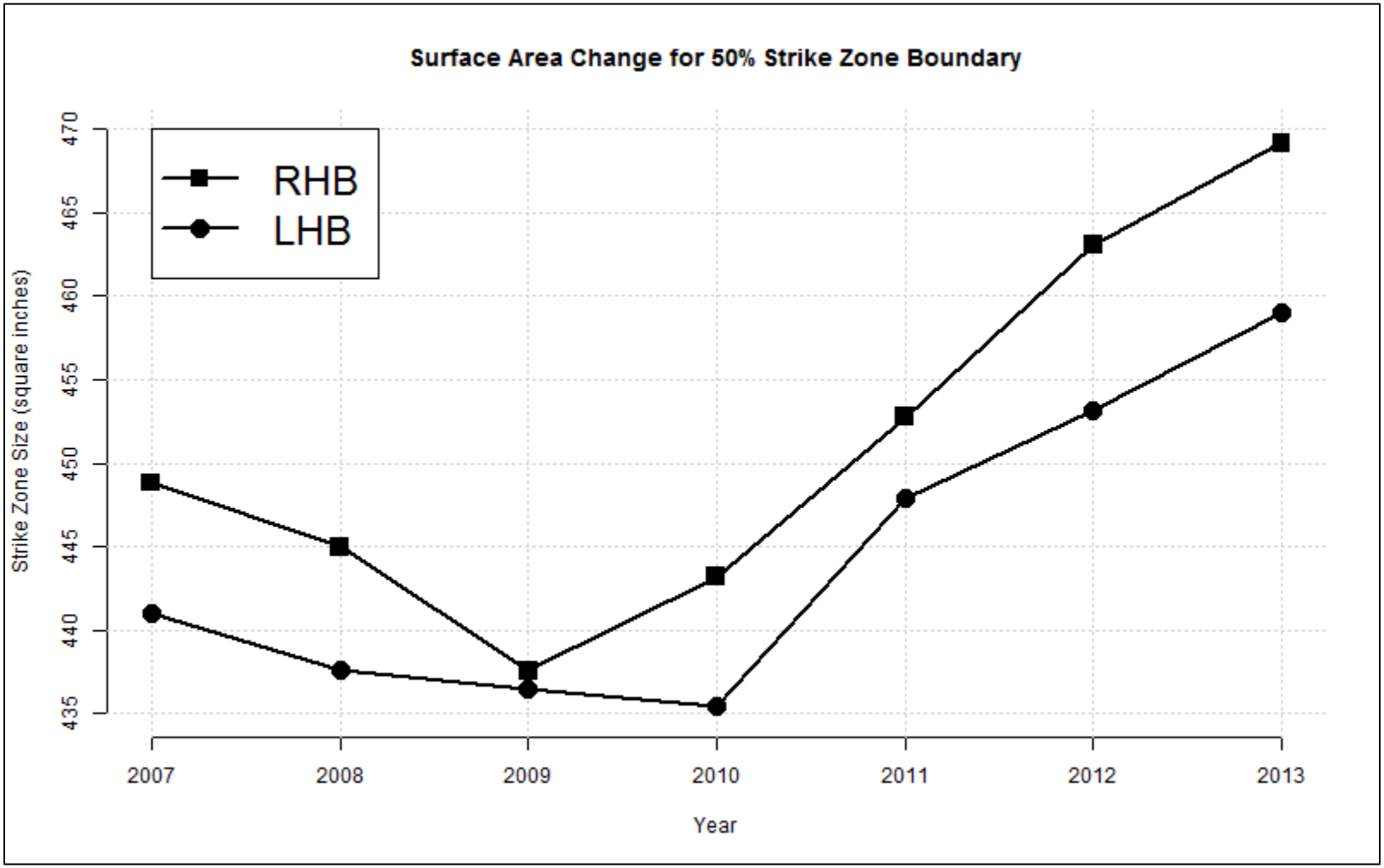


FIGURE 16: Surface Area Percent Growth by Probability Boundary (2008-2013)

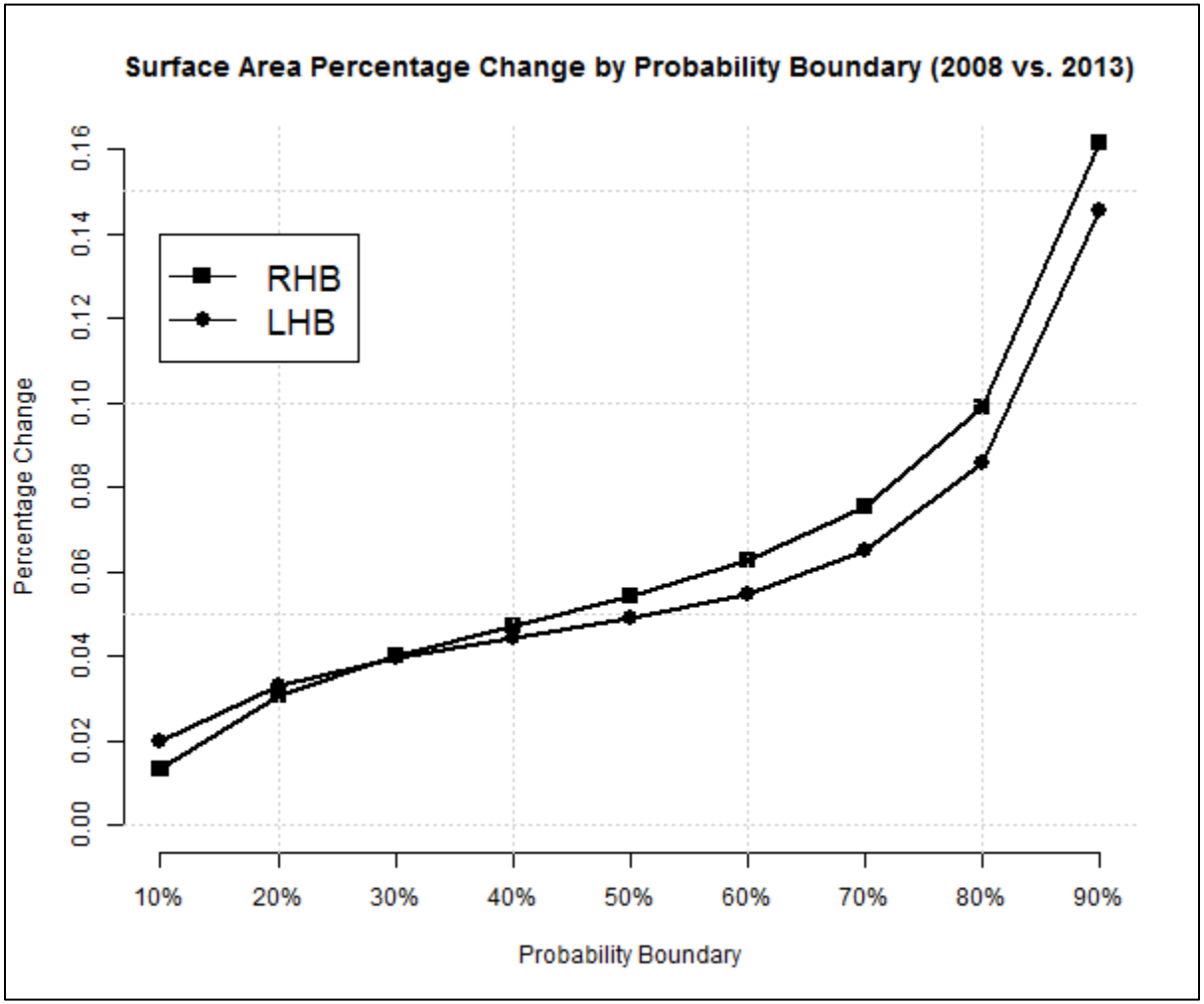
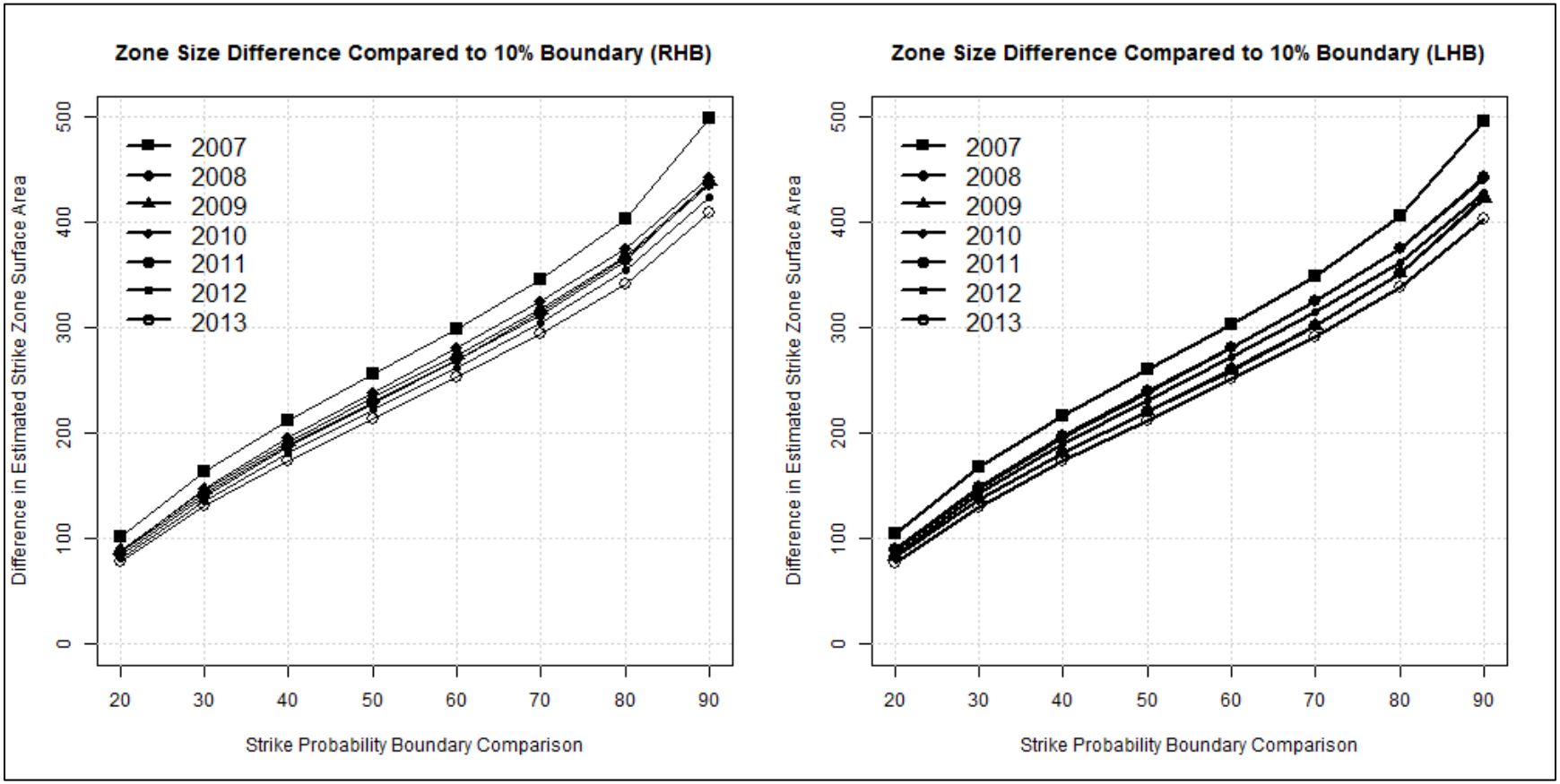


FIGURE 17: Surface Area Difference Between 10% Boundary and Various Strike Zone Boundary Definitions



*Y-axis area reported in square inches.

TABLE 11: Semi-Parametric Logistic GAM for Probability of a Strike Call

Var.	Year F.E.	Year Trend
Constant	-2.5437***	-106.30***
S.E.	*(0.04406)*	*(3.4750)*
Year	-----	0.05167***
S.E.	-----	*(0.00173)*
2008	-0.01385	-----
S.E.	*(0.01153)*	-----
2009	-0.08757***	-----
S.E.	*(0.01325)*	-----
2010	-0.04123***	-----
S.E.	*(0.01355)*	-----
2011	0.02723**	-----
S.E.	*(0.01369)*	-----
2012	0.13823***	-----
S.E.	*(0.01378)*	-----
2013	0.18306***	-----
S.E.	*(0.01379)*	-----

a. Includes umpire, pitch type, and ball-strike count dummy fixed effects, with separate surfaces fit for right and left handed batters.

FIGURE 18: Surface Overlap Between 50% Boundary Zone and MLB Rulebook Zone

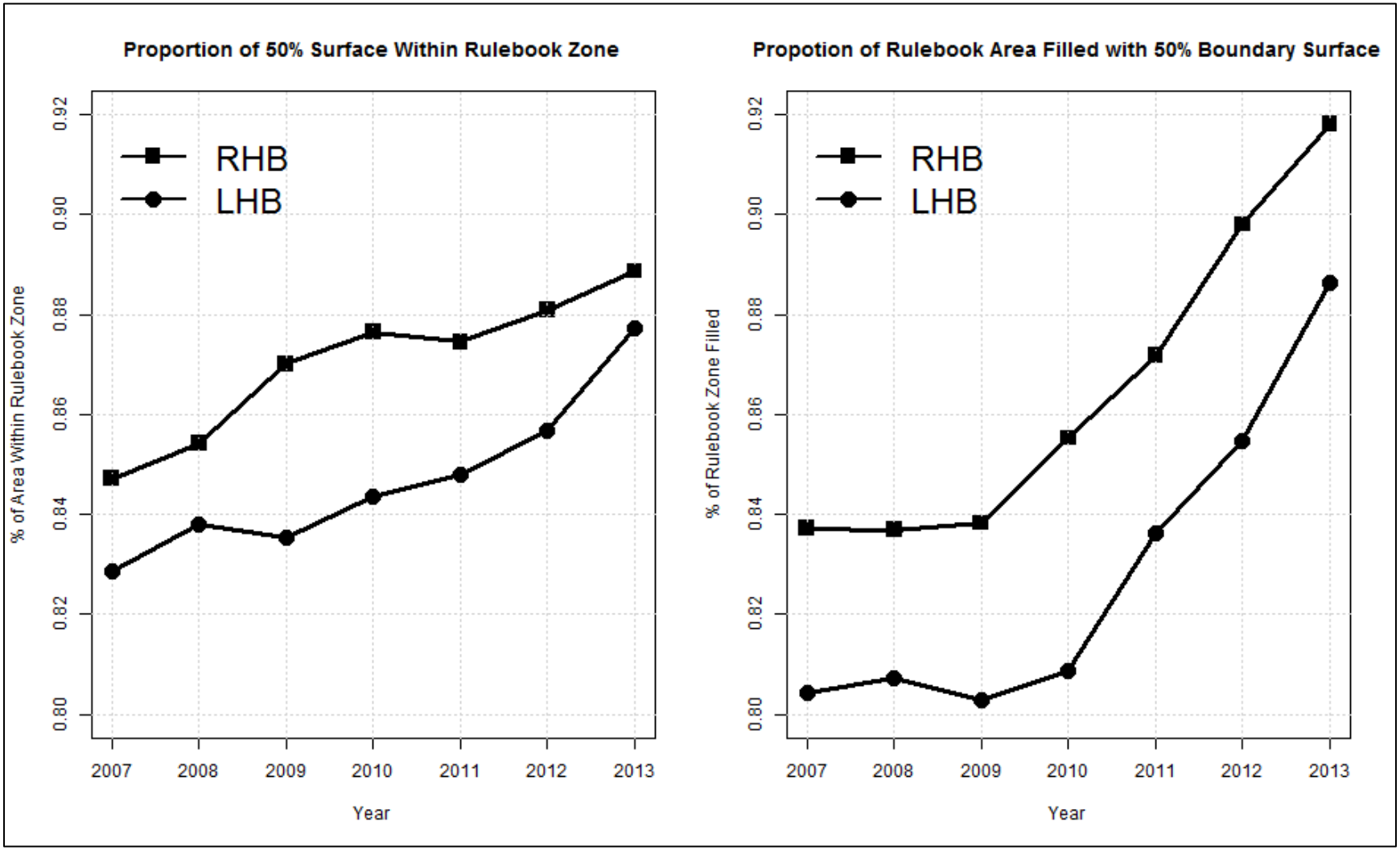


TABLE 12: Expected Runs from Umpire Strike Changes Relative to 1999[a]

Year	Total Runs Scored	Called Pitches	Called Strike Rate	Δ Called Strikes	Δ Expected Runs from Strike Increase	Δ True Runs
1999	24,691	390,325	0.29236	-----	-----	-----
2000	24,971	394,113	0.29402	653	-95	+280
2001	23,199	377,353	0.30976	6,565	-960	-1,492
2002	22,408	378,061	0.30612	5,201	-760	-2,283
2003	22,978	382,123	0.31250	7,697	-1,125	-1,713
2004	23,376	386,946	0.31046	7,005	-1,024	-1,315
2005	22,325	377,011	0.31629	9,021	-1,319	-2,366
2006	23,599	383,780	0.31195	7,518	-1,099	-1,092
2007	23,322	387,388	0.31307	8,021	-1,172	-1,369
2008	22,585	389,376	0.31313	8,085	-1,182	-2,106
2009	22,419	394,793	0.31743	9,896	-1,447	-2,272
2010	21,308	390,651	0.32307	11,994	-1,753	-3,383
2011	20,808	384,339	0.32250	11,581	-1,693	-3,883
2012	21,017	382,647	0.32632	12,994	-1,899	-3,674
2013	20,255	383,217	0.32500	12,509	-1,828	-4,436

[a]Data gleaned from Baseball Almanac (2014).

TABLE 13: Pitcher Behavioral Changes from Expanded Strike Zone

Year	Pitches Thrown	Within Zone	Percent to Rule Zone	Low Pitches	Low Pitch Rate	Avg. Pitch Height (in.)
2007[a]	339,397	156,000	45.96	74,999	22.10	28.92
2008	732,234	342,024	46.71	162,028	22.13	28.87
2009	773,799	362,893	46.90	172,384	22.28	28.85
2010	742,057	341,428	47.15	168,128	23.22	28.50
2011	733,913	345,122	47.02	174,588	23.79	28.39
2012	718,686	336,354	46.80	183,502	25.53	27.79
2013	748,150	348,648	46.60	196,942	26.32	27.61

a. Only part of the 2007 season was recorded by Pitch f/x. Some pitches did not record locational data and were removed from the analysis.

TABLE 14: Pitch Height Impacts on Contact, In-Play, and Hit Rate on Batter Swings[a,b]

	Contact Likelihood	**In-Play Likelihood**	**Hit Likelihood**
2007 Rate % (Low Pitch)	*80.80 (57.40)*	*42.01 (29.56)*	*23.10 (18.32)*
2013 Rate % (Low Pitch)	*79.35 (56.67)*	*40.30 (29.01)*	*21.98 (17.53)*
	(1)	**(2)**	**(3)**
Constant	1.27153***	-0.19425***	-1.05518***
S.E.	*(0.00824)*	*(0.00686)*	*(0.00796)*
Low Pitch[c]	-1.32279***	-0.67109***	-0.31271***
S.E.	*(0.00531)*	*(0.00519)*	*(0.00617)*

a. Logistic regression with β estimates reported. b. Model includes ball-strike count and pitch type dummy fixed effects and includes data from 2010 through 2013. c. Dummy variable indicating pitch is less than 1.75 feet off the ground, or approximately the bottom of the empirically derived strike zone in 2007.

TABLE 15: Umpire Ball and Strike Accuracy Impacts on Selected Reported Statistics[a]

	Earned Run Average (ERA)		On Base Plus Slugging (OPS)		Strikeouts Per 9 Innings (K/9)		Walks Per 9 Innings (BB/9)	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Constant	153.08***	3.6119	12.460***	0.62962***	-263.20***	11.556***	75.528***	1.0025
S.E.	*(38.551)*	*(2.2298)*	*(2.7956)*	*(0.16078)*	*(32.420)*	*(1.8558)*	*(25.434)*	*(1.4624)*
Year	-0.07482***	-----	-0.00593***	-----	0.13640***	-----	-0.03783***	-----
S.E.	*(0.19953)*	-----	*(0.00145)*	-----	*(0.01678)*	-----	*(0.01316)*	-----
2008	-----	-0.11424	-----	-0.00693	-----	0.22136***	-----	0.06426
S.E.	-----	*(0.07566)*	-----	*(0.00546)*	-----	*(0.06297)*	-----	*(0.04962)*
2009	-----	-0.13741*	-----	-0.00816	-----	0.40074***	-----	0.10619**
S.E.	-----	*(0.07866)*	-----	*(0.00567)*	-----	*(0.06547)*	-----	*(0.05159)*
2010	-----	-0.32836***	-----	-0.02586***	-----	0.50306***	-----	-0.01790
S.E.	-----	*(0.08629)*	-----	*(0.00622)*	-----	*(0.07182)*	-----	*(0.05660)*
2011	-----	-0.41798***	-----	-0.03149***	-----	0.48415***	-----	-0.12567**
S.E.	-----	*(0.09645)*	-----	*(0.00696)*	-----	*(0.08028)*	-----	*(0.06326)*
2012	-----	-0.28483**	-----	-0.02130**	-----	0.86910***	-----	-0.10152
S.E.	-----	*(0.11677)*	-----	*(0.00842)*	-----	*(0.09718)*	-----	*(0.07658)*
2013	-----	-0.36228**	-----	-0.02613**	-----	0.88163***	-----	-0.09501
S.E.	-----	*(0.14201)*	-----	*(0.01024)*	-----	*(0.11820)*	-----	*(0.09314)*
Correct Strike Rate[c]	-0.02908***	-0.04010***	-0.00246***	-0.00341***	0.03037***	0.02681***	-0.04144***	-0.04321***
S.E.	*(0.01115)*	*(0.01199)*	*(0.00081)*	*(0.00086)*	*(0.00938)*	*(0.00998)*	*(0.00736)*	*(0.00787)*
Correct Ball Rate	0.04487**	0.04694**	0.00473***	0.00467***	-0.07058***	-0.07806***	0.07810***	0.06339***
S.E.	*(0.02096)*	*(0.02161)*	*(0.00152)*	*(0.00156)*	*(0.01763)*	*(0.01799)*	*(0.01383)*	*(0.01417)*

a. Regression limited to umpire-year observations with more than 5 games worked behind the plate, and are then weighted by $\sqrt{n_{i,t}}$, where n_i is the number of games worked behind the plate in season t for umpire i. b. Umpire dummy fixed effects included in all models. c. Correct Strike Rate and Correct Ball Rate are scaled from 0 to 100—rather than 0 to 1—for interpretability.

TABLE 16: Changes in ERA, OPS, K9, and BB9 Attributed to Umpire Accuracy

Year	*ERA*	*OPS*	*K/9*	*BB/9*
2007	4.46	0.758	6.67	3.33
2008	4.32	0.749	6.83	3.39
2009	4.31	0.751	6.99	3.46
2010	4.07	0.728	7.13	3.28
2011	3.94	0.720	7.13	3.11
2012	4.01	0.724	7.56	3.05
2013	3.88	0.714	7.57	3.02
Δ *from 2007 to 2013*	**- 0.58**	**- 0.044**	**+ 0.90**	**- 0.31**
Year Trend Models				
Δ *from Umpire Strikes*[a]	- 0.26	- 0.022	+ 0.27	- 0.37
Δ *from Umpire Balls*[b]	+ 0.12	+ 0.012	- 0.19	+ 0.21
Year F.E. Models				
Δ *from Umpire Strikes*	- 0.36	- 0.031	+ 0.24	- 0.39
Δ *from Umpire Balls*	+ 0.12	+ 0.012	- 0.21	+ 0.17
Net* Δ *from Umpire Accuracy				
Year Trend	- 0.14	- 0.010	+ 0.08	- 0.16
Year F.E.	- 0.24	- 0.019	+ 0.03	- 0.22
% Attributable to* Δ *in Accuracy				
Year Trend	24.1	22.7	8.9	51.6
Year F.E.	41.4	43.2	3.3	71.0

a. Uses 8.96 percentage point change in correct strike rate from 2007 to 2013 and coefficients from Table 15.
b. Uses 2.63 percentage point change in correct ball rate from 2007 to 2013 and coefficients from Table 15.