Can Decision Biases Increase with the Stakes?  
Evidence from the Field*

Etan Green    David P. Daniels  
Stanford University  

October 5, 2014

Abstract

High stakes are thought to reduce decision biases. This paper identifies a decision bias that increases with the stakes. We examine over one million decisions made by 75 home plate umpires in Major League Baseball, who are directed and incentivized to make binary calls according to a single, objective criterion: the location of the pitch. Using state-of-the-art camera measurements provided by MLB itself, we find that every umpire distorts this directive by avoiding the option that would more strongly change the expected outcome of the game. This impact aversion produces distortions that grow larger as the stakes increase.

*Please direct all correspondence to eagreen@stanford.edu. The authors wish to thank Doug Bernheim, Nir Halevy, David Hausman, Dorothy Kronick, Jonathan Levav, Max Mishkin, Muriel Niederle, Roger Noll, Justin Rao, Peter Reiss, Al Roth, and Charlie Sprenger for helpful comments and suggestions on previous drafts. Green and Daniels also thank the Stanford University Graduate School of Business and a National Science Foundation Graduate Research Fellowship, respectively, for generous financial support. Previous versions of this paper were presented at the 2014 MIT Sloan Sports Analytics Conference in Boston and the 2014 Behavioral Decision Research in Management Conference in London. Earlier versions of this paper, under various titles, date to February, 2014.
1 Introduction

High stakes are thought to reduce decision biases, especially in real-world settings with experienced agents (e.g. Levitt and List, 2008). This paper uses data from the real-world decisions of experienced agents to identify a decision bias that increases with the stakes.

We examine over one million decisions made by 75 professional umpires in Major League Baseball, who are directed and incentivized to make binary calls, either a ball or a strike, according to a single, objective criterion: the location of the pitch. Using state-of-the-art camera measurements provided by MLB itself, we find that every umpire distorts this directive by avoiding the option that would more strongly change the expected outcome of the game. We call this bias impact aversion. As we explain below, impact aversion produces larger distortions as the stakes increase and the decision becomes more consequential.

Over the course of a game, the expected outcome changes in response to the actions of the players and the calls of the umpires. For the home plate umpire, each call offers a choice between an option that would shift the expected outcome of the game in favor of one team and an alternative that would shift the expected outcome in favor of the other team. An option is pivotal, or has a large impact, when it would significantly shift the expected outcome. An option is non-pivotal, or has a small impact, when it would preserve the current expected outcome.

Given a choice between a pivotal option and a non-pivotal alternative, an impact-averse umpire will choose the pivotal option less frequently than when his options are equally (non-)pivotal. To illustrate, consider a situation in which both of the umpire’s options are equally (non-)pivotal and the umpire is indifferent between them, selecting balls and strikes each 50% of the time. When the situation changes such that one option becomes pivotal, we find that the average umpire will now distort his decisions by choosing the pivotal option as much as 25 percentage points less frequently, selecting the pivotal option only 25% of
the time and the non-pivotal option 75% of the time. In general, a larger difference in the impacts of the umpire's options induces more bias towards the non-pivotal option.

High-stakes decisions are those in which the umpire can significantly change the expected outcome of the game—i.e., in which at least one option is pivotal. Some high-stakes decisions feature two pivotal options—one that would likely award the game to one team, and an alternative that would likely award the game to the other team—but these decisions are rare. More commonly, high-stakes decisions feature one option that is pivotal and one option that is non-pivotal. Thus, the difference between the impacts of a ball and a strike tends to increase as the stakes rise, implying that impact aversion produces larger distortions at higher stakes.

It is often argued that high stakes will reduce behavioral anomalies, especially in settings characterized by experienced agents and intense competition (List, 2003; Hart, 2005; Levitt and List, 2008). Impact aversion provides a counterexample to this claim, since it distorts high-stakes decisions by professionals in the field. Thus, this paper contributes to a growing body of field studies that identifies systematic ways in which individuals violate standard economic assumptions (for a review, see DellaVigna, 2009), even in settings with experienced agents, intense competition, and high stakes (e.g. Northcraft and Neale, 1987; Berger and Pope, 2011; Pope and Simonsohn, 2011; Pope and Schweitzer, 2011). Impact aversion not only distorts high-stakes decisions, but actually induces larger distortions at higher stakes. At least one decision bias in the real world grows in importance as the stakes increase.¹

Major League Baseball directs umpires to call balls and strikes solely based on the location of the pitch, and it offers umpires strong incentives to follow its directive. MLB uses cameras to monitor the consistency of umpires' calls. Furthermore, as we show in Section 4.3,

¹Experimental evidence has shown that in some circumstances, greater monetary incentives produce more bias (for a review, see Camerer and Hogarth, 1999), such as when they cause the participant to “choke” (Ariely, Gneezy, Loewenstein and Mazar, 2009).
MLB withholds lucrative assignments from umpires who are more impact averse. Thus, impact aversion is inconsistent with simple agency models in which the principal effectively incentivizes the agent to follow its directive (Laffont and Martimort, 2002). Evidence of such (ir)regularities is rare in principal-agent contexts, because it is typically difficult for the analyst to observe what the agent is contracted to do or what he actually does (Prendergast, 1999). Indeed, it is often hard for the principal herself to specify what the agent should do or to evaluate what he actually does (e.g. Bertrand and Mullainathan, 2001). An advantage of our setting is that the principal offers a clear directive and precisely evaluates the agent’s decisions with state-of-the-art monitoring technology—and even so, the agent violates the principal’s directive.

If impact aversion does not result from incentives provided by the principal, then it must derive from other sources (e.g. Myerson, 1982; Holmstrom and Milgrom, 1991; Kamenica, 2012). One possibility is that public pressure biases umpires’ decisions in favor of the less pivotal option. Umpires face public criticism after making mistakes that influence important game outcomes. As we describe in Section 3.6, umpires who make a close call that prevents a perfect game2 endure criticism of their choices—in one umpire’s case, for the remaining 35 years of his career. In order to avoid public criticism, umpires may avoid making pivotal calls that the public could view as mistaken. If so, umpires will display greater impact aversion when they face greater scrutiny. Indeed, as we show in Section 3.6, impact aversion increases dramatically when a game is televised nationally rather than locally—i.e., when scrutiny is more intense. The more visible the game, the more invisible umpires try to become.

To account for these findings, we propose a simple model in which agents balance two goals: first, to make the correct choice, and second, to avoid making pivotal mistakes. In the context of umpire decision making, the first motivation is aligned with MLB’s directive and incentives, and the second motivation may derive from preferences or public pressure.

---

2A perfect game is a game in which the pitcher does not allow a single batter to reach base.
The model predicts that decision makers will choose the correct option when it is obvious but will avoid the more pivotal option when the correct option is not obvious. We structurally estimate a coefficient of impact aversion for each umpire (allowing for possible impact neutrality or impact seeking), and we find that every umpire in our sample is impact averse.

The remainder of the paper is organized as follows. Section 2 describes the directive and incentives faced by umpires in Major League Baseball. Section 3 presents evidence of impact aversion and shows that the bias is larger when umpires face greater scrutiny. Section 4 presents and estimates a simple model of impact aversion; these estimates demonstrate that every umpire in our sample is impact averse, that impact aversion is not correlated with experience, and that more impact-averse umpires are less likely to receive lucrative assignments. Section 5 shows that impact aversion is larger for more difficult decisions; this result is predicted by an extension of our model in which umpires are unsure of the true location of the pitch and are risk averse in this second-order probability. Section 6 estimates that impact aversion distorts $3,000 of team value each call. Section 7 concludes, comparing impact aversion to other decision biases and discussing possible impact aversion in judicial decision making.

2 Background

2.1 The Strike Zone

Most plays in baseball begin with the pitcher throwing a pitch to the batter. If the batter chooses not to swing, the home plate umpire makes a call—either a ball or a strike. The home plate umpire has a simple job: to decide whether the pitch intersects the strike zone. Pitches that intersect the strike zone should be called strikes; pitches that do not intersect the strike zone should be called balls.

There are two strike zone definitions of interest. The first is the official strike zone, which
Major League Baseball defines as “that area over home plate the upper limit of which is a horizontal line at the midpoint between the top of the shoulders and the top of the uniform pants, and the lower level is a line at the hollow beneath the kneecap.” The second is the enforced strike zone, which varies from umpire to umpire. Conventional wisdom holds that MLB tolerates small deviations between the official strike and an umpire’s enforced strike zone so long as the umpire enforces his strike zone consistently (Sullivan, 2001). We find evidence in support of this claim. As we show in Section 4.3, umpires that are more consistent in their calls are more likely to receive lucrative assignments, but umpires that are more correct in their calls vis-à-vis the official strike zone are not more likely to receive those assignments. Accordingly, we evaluate umpires on their consistency, not on their correctness.

2.2 Incentives from Major League Baseball

Deviations between enforced strike zones and the official strike zone have not always been small. As recently as the 1990s, pitches far beyond the side of home plate—that hitters would have to lunge for—were often called strikes, while high strikes—over the plate and above the hitter’s belt—were almost always called balls. Major League Baseball could not remedy the problem by rewarding the least egregious violators, because the umpires union mandated that all umpires split both postseason assignments and the extra pay—as much as half an umpire’s base salary over the entire postseason—equally among all umpires (Callahan, 1998).

In 1999, MLB initiated three small measures aimed at reducing discrepancies between enforced strike zones and the official strike zone: first, reminding all umpires of the definition of the official strike zone; second, instructing team officials to monitor each umpire’s enforced strike zone; and third, suspending an umpire who physically confronted a player—the first suspension ever given to an umpire. A clumsy response by the umpires union paved the way for baseball to strengthen the formal incentives faced by umpires. First, the union authorized

\[^3\text{http://mlb.mlb.com/mlb/official_info/umpires/rules_interest.jsp.}\]
a strike. Then, when it realized that its contract with MLB forbade a strike, the union tried to dissolve itself—convincing 57 of the 66 union umpires to resign—so as to negotiate a new contract. When a federal court ruled the attempted dissolution null and void, Major League Baseball accepted the resignations of 22 umpires and hired 30 new umpires (Callan, 2012).

Home plate umpires in Major League Baseball now operate under a high degree of monitoring, incentives for good performance, possible punishment for poor performance, considerable training, and stringent screening. MLB employs over a dozen officials to monitor and evaluate umpire performance. Most games are overseen in person by a representative from the league, who files a report detailing blown calls. The league uses pitch-tracking technology to evaluate the calls of home plate umpires. In the early 2000s, MLB installed the QuesTec system in half of its stadiums, which tracked the location of each pitch as it crossed the region above home plate. Prior to the 2009 season, MLB installed the more accurate PITCH f/x system in every park, which captures the location of each pitch 20 times along its trajectory. After each game, the home-plate umpire receives a breakdown of his performance, including a score that measures the accuracy of his calls (Drellich, 2012).

Rewards and discipline are closely tied to performance. Umpires are evaluated twice each season; evaluations are based on reports from umpire observers and analysis of the camera data. MLB claims that the best umpires are assigned to postseason games, and our analysis in Section 4.3 supports this claim. “There have been situations where umpires have been disciplined” as a result of poor evaluations, according to Joe Torre, the Executive Vice President of Baseball Operations (Callan, 2012). After the 2009 season, baseball fired three of its umpire observers after a number of important missed calls during the postseason (Nightengale, 2010). Since 2000, a handful of umpires have been suspended for inappropriate confrontations with players and managers. In 2013, baseball suspended a home plate umpire for forgetting a rule (Hoffman, 2013).

Selection of Major League umpires is stringent and performance-based. To become a
Major League umpire, a candidate must attend umpire schools, graduate in the top fifth of his class, and then rise through four levels of the minor leagues before qualifying to fill in for a Major League umpire on vacation (Caple, 2011). MLB employs 70 full-time umpires at any one time and 8 to 12 fill-ins from the minor leagues. Typically, only one fill-in is hired as a full-time MLB umpire after each season (O’Connell, 2007).

2.3 Data and descriptives

Umpires are supposed to call balls and strikes based solely on the location of the pitch. We measure umpires’ adherence to this normative benchmark with precise pitch location data from the PITCH f/x cameras—the same system used by MLB monitor the calls of home plate umpires. We define the location of the pitch by its coordinates when it intersects the plane rising from the front of home plate, on which the official strike zone is defined. The PITCH f/x system also provides estimates of the top and bottom borders of the official strike zone based on the batter’s stance prior to each pitch. We use these measurements to normalize the vertical location of the pitch. We merge pitch location data from MLB.com with pitch and game data from Retrosheet.org, including the number of balls and strikes in the count, the number of outs, whether there is a runner on each base, the identity of the home plate umpire, and the game’s start time and attendance.

Our data comprise every pitch recorded by the cameras during the 2009-11 regular seasons—over 2 million pitches. Umpires make calls on 53% of pitches in the sample. After eliminating the 47% of pitches that are swung at, the 13,000 balls that were thrown intentionally, and the 50,000 calls made by the 21 umpires who each make fewer than 7,500 calls during the window, our sample contains 1,036,355 calls made by 75 umpires. About

---

4 About 1% of pitches are not captured by the cameras.
5 While the width of the official strike zone is fixed, the height of the official strike zone varies with the height and stance of the batter. According to Major League Baseball, “The strike zone shall be determined from the batter’s stance as the batter is prepared to swing at a pitched ball.” [http://mlb.mlb.com/mlb/official_info/umpires/rules_interest.jsp](http://mlb.mlb.com/mlb/official_info/umpires/rules_interest.jsp)
two-thirds of calls are balls and the remaining third are called strikes. 6% of calls occur in three-ball counts, 19% of calls occur in two-strike counts, and 2% of calls occur in full counts (three balls and two strikes).^6^

**Figure 1:** (a) $\hat{m}(X)$: the probability of a strike call when the batter does not swing, and (b) $\hat{f}(X)$: the distribution of calls. The dotted lines denote the boundaries of the official strike zone on the plane that rises from the front of home plate (seen from the umpire’s view). (a) Pitches that cross the plane in the middle of the official strike zone are almost always called strikes; those that cross well outside the official strike zone are almost always called balls. Pitches that cross near the boundaries of the official strike zone are sometimes called strikes and sometimes called balls. (b) Pitches along the bottom of the official strike zone comprise a disproportionate share of calls.

From our sample of over a million calls, we non-parametrically estimate the probability of a called strike conditional on the location of the pitch. Figure 1a shows this estimate of the enforced strike zone. The dotted lines denote the boundaries of the official strike

---

^6^The count keeps track of the prior balls and strikes in the at-bat, or the sequence of consecutive pitches to the batter. Every at-bat begins with a count of zero balls and zero strikes. A ball is added when the umpire makes a ball call. A strike call is added when the umpire makes a strike call or when the batter swings—unless the count has two strikes and he makes contact with the pitch but does not put it in the field of play, in which case the count remains at two strikes. At-bats end most commonly when the batter swings and hits the pitch in the field of play, when the count reaches four balls, or when the count reaches three strikes.
zone—the width of home plate on the horizontal axis and the normalized distance from knees to chest on the vertical axis—on the plane that rises from the front of home plate. The umpire stands behind home plate and looks through the plane, over the catcher’s head, and towards the pitcher. A right-handed batter would stand to the umpire’s left. The contour lines denote \( \hat{m}(X) \): our estimate of the probability of a called strike conditional on \( X = (x_1, x_2) \), the location of the pitch. This estimate is the prediction from a kernel regression of an indicator for whether the call is a strike.\(^7\) Pitches that intersect the middle of the official strike zone are obvious strikes, and umpires call them strikes more than 90% of the time; pitches that cross far outside the official strike zone are obvious balls, and umpires call them balls more than 90% of the time. In between, pitches that intersect the plane at the same location are sometimes called strikes and sometimes called balls. This band of inconsistency is wide: more than half a foot separates pitches that are called strikes 90% of the time and those that are called strikes 10% of the time.\(^8\) Figure 1b shows \( \hat{f}(X) \): our estimate of the distribution of calls by location.\(^9\) Since calls disproportionately cluster near the lower boundary of the official strike zone, the band of inconsistency plays an outsized role in determining the outcomes of pitches, at-bats, and even games.\(^10\)

3 Evidence of Impact Aversion

3.1 Pivotal situations: non-parametric estimates

An umpire is *inconsistent* if he makes different calls on pitches that cross the plane at the same location; an umpire is *biased* if these differences correlate with normatively extraneous

\(^7\)We use a bivariate Gaussian kernel and Silverman’s rule of thumb bandwidth for each axis.
\(^8\)The smoothing nature of the estimator may obscure a sharper boundary, though the bandwidth is small enough to minimize this concern.
\(^9\)For the density estimate, we also use a bivariate Gaussian kernel and Silverman’s rule of thumb bandwidth for each axis.
\(^10\)The modal pitch location for all batters is the bottom outside corner. Hence, the bimodality in Figure 1b is a consequence of pooling calls for right- and left-handed batters.
(non-location) factors. In baseball, the count tracks previous pitches in the at-bat, or the sequence of pitches between pitcher and batter. We first look for bias in two asymmetrically pivotal situations: when the count has three balls or two strikes. A fourth ball would end the at-bat by walking the batter; a third strike would end the at-bat by striking him out. Unless there are three balls and two strikes (a full count), the umpire can extend the at-bat by calling a strike to avoid a walk or by calling a ball to avoid a strike-out. The count should not influence an umpire’s calls. According to Peter Woodfork, who oversees umpires as MLB Senior Vice President for Baseball Operations, Major League Baseball “strives[s] to make sure [umpires] are consistent throughout all at-bats, no matter the count” (Baumbach, 2014).

To visualize bias for a particular situation, we plot the difference between two non-parametric estimates of the enforced strike zone, \( \hat{m}(X|S) - \hat{m}(X|<3 \text{ balls} & <2 \text{ strikes}) \): the first estimated on a subset of pitches for which the situation \( S \) is true (e.g. 3 balls & <2 strikes), and the second estimated on pitches in baseline counts with fewer than three balls and fewer than two strikes. Since the situations we consider are extraneous to the location of the pitch, the two enforced strike zones should be identical, and their difference should be zero across the plane.

Instead, Figures 2a and 2b show dramatic changes in the enforced strike zone when the count has three balls (2a), and when the count has two strikes (2b). In both graphs, pitches in full counts are excluded from the underlying strike zone estimates. If the enforced strike zones are the same, their difference will be a flat plane at zero.\(^{11}\) In each graph, the difference is near zero in both the center of the official strike zone and far outside of it. Even in three-ball or two-strike counts, obvious strikes are still called strikes, and obvious balls are still called balls. But where calls are not obvious, umpires enforce different strike zones. In

\(^{11}\)To attain a smoothed measure of the difference, we estimate each non-parametric strike zone using the maximum bandwidth along each axis across the two plots.
Figure 2: $\hat{m}(X|S) - \hat{m}(X|< 3 \text{ balls } \& < 2 \text{ strikes})$, for situation $S$ listed in figure titles. The change in the probability of a called strike when the count has (a) three balls, and (b) two strikes. The baseline case comprises calls in counts with fewer than three balls and fewer than two strikes. The enforced strike zone expands in three-ball counts and contracts in two-strike counts, particularly at the top and bottom of the official strike zone (which is displayed as a red box). In each subfigure, both the contour plot and the 3-D heat map show exactly the same information.

(a) 3 balls & < 2 strikes

(b) 2 strikes & < 3 balls
three-ball counts, Figure 2a shows that the probability of a strike increases along the band of inconsistency—the strike zone expands. In two-strike counts, Figure 2b shows that the probability of a strike decreases along the band of inconsistency—the strike zone contracts.\footnote{Baseball commentators have previously shown that the enforced strike zone expands in three-ball counts and contracts in two-strike counts. For instance, see Moskowitz and Wertheim (2011) or Carruth (2012). Our findings go beyond these in at least seven ways. First, we measure the extent of the biases non-parametrically, semi-parametrically, and structurally. Second, we show that the bias represents an aversion to changing the expected outcome of the game, \textit{not} to ending the at-bat as has been thought. Third, we show that impact aversion is larger when umpires are subject to greater public scrutiny, suggesting that impact aversion is an attempt to avoid public criticism. Fourth, we show that \textit{every} umpire exhibits impact aversion. Fifth, we show that Major League Baseball rewards the least impact-averse umpires with lucrative assignments, implying that high levels of impact aversion contradict the league’s goals. Sixth, we show that calls for which pitch location is less certain induce even more impact aversion than comparable calls for which pitch location is more certain; we show that this behavior is consistent with second-order risk aversion. Seventh, we generalize impact aversion to other decision makers.}

Figure 3: $\hat{m}(X|3 \text{ balls} \& 2 \text{ strikes}) - \hat{m}(X|<3 \text{ balls} \& <2 \text{ strikes})$. The change in the probability of a called strike when the count has three balls and two strikes (full counts). The baseline case comprises calls in counts with fewer than three balls and fewer than two strikes. In full counts, the enforced strike zone contracts more moderately than with just two strikes (2b). Both the contour plot and the 3-D heat map show exactly the same information.

With three balls and fewer than two strikes, a ball would be more pivotal than a strike. Similarly, with two strikes and fewer than three balls, a strike ends the at-bat while a ball prolongs it. The expansion of the enforced strike zone in three-ball counts and the contraction of the strike zone in two-strike counts suggest that umpires are averse to making the more pivotal call. By this logic, full counts (three balls and two strikes) should induce an
intermediate effect—either a smaller expansion of the enforced strike zone than with three balls, or a smaller contraction than with two strikes. Because the umpire cannot avoid a pivotal call in a full count, he will distort the strike zone less than when he chooses between a pivotal call and a non-pivotal one. Figure 3 shows the difference in the enforced strike zone between calls in full counts and calls in counts with fewer than three balls and fewer than two strikes. Full counts induce a more moderate contraction of the strike zone than with just two strikes. The fact that the enforced strike zone contracts in full counts relative to situations where neither choice is pivotal is consistent with our observation in Section 3.3 that third strikes tend to be more pivotal than fourth balls.

### 3.2 Semi-parametric estimates

The non-parametric estimates in Figures 2 and 3 assume that all calls in a given situation (e.g. counts with three balls) are independent draws from an identical distribution. However, the enforced strike zone varies across umpires and shifts left when a left-handed hitter is at bat. To account for these sources of variation, we estimate the semi-parametric model

\[ y_i = p_i + \omega(p_i) \cdot S_i \beta + \epsilon_i, \]  

where \( y_i \) is an indicator for a strike on call \( i \), \( p_i \) is the baseline probability of a strike call based on pitch location alone, \( S_i \beta \) is a linear term of situation-specific distortions, \( \omega(p_i) \) is a scalar weight that accounts for the shape of the bias, and \( \epsilon_i \) is a mean-zero error. We are interested in \( \beta \), the amount of distortion associated with situation \( S_i \) (e.g. three-ball count) being true.  

\[ \text{Umpires position themselves differently behind the catcher based on the handedness of the hitter, according to a former umpire. The empirical strike zone is horizontally symmetric about the midline of home plate for right-handed hitters, but for left-handed hitters, umpires call strikes on outside pitches more frequently than on inside pitches. This asymmetry accounts for the left-shift of the enforced strike zone relative to official strike zone in Figure 1a.} \]
The baseline probability \( p_i \) is a measure of the probability of a strike call in the absence of distortion. We measure the baseline probability \( p_i \) as \( \hat{m}_{u(i),h(i),\neg S}(X_i) \) using a kernel regression of \( y_i \) on pitch location. For call \( i \), we estimate \( m_{u(i),h(i),\neg S} \) only on pitches called by umpire \( u(i) \), pitched to batters of handedness \( h(i) \), and for which none of the states \( S_i \) are true—i.e. for counts with fewer than three balls and fewer than two strikes. If only three-ball and two-strike counts induce bias, then the baseline probability identifies the likelihood of a strike call based solely on the location of the pitch.

We weight the distortion term \( S_i \beta \) by a function \( \omega \) of the baseline probability. As Figure 2 shows, distortion is greatest when pitches are borderline, and distortion is nonexistent when pitches are obvious balls or obvious strikes. Accordingly, we define \( \omega(p_i) \equiv 1 - 2|p_i - 0.5| \).\(^{14}\) We then estimate \( \beta \) by regressing \( y_i - p_i \), the component of the observed call not explained by pitch location, on \( \omega(p_i) \cdot S_i \beta \). We interpret \( \hat{\beta} \) as the percentage point change in the probability of a strike from a baseline probability of 0.5—i.e. the bias on a borderline pitch.\(^{15}\)

Table 1 reports \( \hat{\beta} \), with standard errors clustered by \((u,h)\) tuple. The semi-parametric estimates echo the effects depicted non-parametrically in Figure 2. In three-ball counts, borderline pitches are called strikes more than 58% of the time; in two-strike counts, borderline pitches are called strikes only 31% of the time. In full counts (Model 2), the probability of a strike decreases by about 12 percentage points (0.09 – 0.19 – 0.02): 50/50 calls become 38/62 calls. The strike zone expands in three-ball counts, contracts in two-strike counts, and contracts to a lesser extent in full counts.

These estimates show that umpires violate their directive to call balls and strikes based solely on pitch location. However, the claim that asymmetrically pivotal counts cause changes in the enforced strike zone rests on an assumption of exogeneity with respect to

\(^{14}\)For certain balls or strikes \((p_i \in \{0, 1\})\), \( \omega(p_i) = 0 \). For \textit{borderline pitches}, or locations in which balls and strikes are equally probable \((p_i = 0.5)\), \( \omega(p_i) = 1 \). For \( 0 < |p_i - 0.5| < 0.5 \), \( 0 < \omega(p_i) < 1 \). Note that because the biases are greater at the top and bottom of the official strike zone than along the sides, \( \hat{\beta} \) will overstate the bias on the sides and understate the bias along the top and bottom.

\(^{15}\)More generally, one can interpret \( \omega(p_i) \cdot \hat{\beta} \) as the distortion on a call for which \( S_i \) is true.
<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-ball count</td>
<td>0.082***</td>
<td>0.088***</td>
<td>0.075***</td>
<td>0.081***</td>
</tr>
<tr>
<td></td>
<td>(17.21)</td>
<td>(15.67)</td>
<td>(12.25)</td>
<td>(17.01)</td>
</tr>
<tr>
<td>2-strike count</td>
<td>-0.19***</td>
<td>-0.19***</td>
<td>-0.20***</td>
<td>-0.18***</td>
</tr>
<tr>
<td></td>
<td>(-37.20)</td>
<td>(-35.40)</td>
<td>(-35.97)</td>
<td>(-34.54)</td>
</tr>
<tr>
<td>Full count</td>
<td>-0.022</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-2.07)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pitching team losing * 3-ball count</td>
<td>0.019</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.04)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batting team losing * 2-strike count</td>
<td>0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.60)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Called strike on last pitch in at-bat * 2-strike count</td>
<td></td>
<td>-0.057***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-7.33)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1036335</td>
<td>1036335</td>
<td>1036335</td>
<td>1036335</td>
</tr>
</tbody>
</table>

$t$ statistics in parentheses

* $p < 0.01$, ** $p < 0.001$, *** $p < 0.0001$

**Table 1:** Semi-parametric regression on strike call from Equation 1. Coefficients of weighted linear component ($\hat{\beta}$) reported. Coefficient is percentage point change on the probability of a called strike for a borderline pitch under the given situation. Standard errors clustered by umpire–batter handedness ($75 \times 2 = 150$ clusters).

omitted situational variables. While the structural context of umpire decision making is relatively simple, we address some potential confounds by including additional situational variables interacted with the three-ball and two-strike indicator variables.\(^{16}\)

First, we address the alternative explanation that our estimates can be explained by favoritism of underdogs or a desire to keep the game close. Price et al. (2012) show that referees in the National Basketball Association disproportionately call discretionary fouls on the leading team. In three-ball counts, umpires may view the pitcher as the underdog and favor him by expanding the strike zone. In two-strike counts, umpires may view the batter as the underdog and favor him by contracting the strike zone. If so, we should observe a

\(^{16}\text{We cannot include these situational variables directly because the distortion term is assumed to be zero when there are fewer than three balls and fewer than two strikes.}\)
greater distortion when the underdog is trailing, which would also help keep the game close. Model 3 includes two indicator variables: one for three-ball counts in which the pitching team is trailing, and one for two-strike counts in which the batting team is trailing. The first interaction explains a small component of the three-ball effect with marginal significance \((p = 0.043)\); the second interaction suggests that if anything, umpires contract the strike zone less when the batting team is trailing. Favoritism of underdogs or a desire to keep the game close are unlikely to explain umpires’ aversion to pivotal calls.

Second, we address the possibility that negative autocorrelation, or the gambler’s fallacy (Tversky and Kahneman, 1974; Rabin, 2002), can explain our results. After calling a strike, umpires are less likely to call a strike on the subsequent pitch, controlling for the count and the location of the pitch (Green and Daniels, 2014). By contrast, ball calls are no less likely after a ball.\(^{17}\) If negative autocorrelation does explain the contraction of the strike zone with two strikes, we should observe the contraction only in two-strike counts preceded by a called strike. Model 4 includes an indicator variable for two-strike counts preceded by a called strike, which explains a small component of the two-strike effect. When a two-strike count is preceded by a ball or a swing, borderline pitches become 32/68 calls; when a two-strike count is preceded by a called strike, borderline pitches become 26/74 calls. Negative autocorrelation cannot fully account for umpires’ aversion to calling third strikes.\(^{18}\)

Additional alternative explanations are addressed in the Appendix, which considers the possibility that impact aversion might be a response to the umpire’s rational expectations of the forthcoming pitch.

\(^{17}\)For both of these effects, the base case comprises the first pitch in the at-bat and calls that follow swings.\(^{18}\)Interestingly, the strike zone expands only in three-ball counts preceded by a ball, and not in three-ball counts preceded by a swing or a called strike. However, it is impossible to say whether this is due to negative autocorrelation, as a three-ball count preceded by a ball is also the first three-ball count faced by the batter in the at-bat. There is no autocorrelation (negative or positive) following balls when the count has fewer than three balls.
3.3 A continuous measure of call impact

By expanding the strike zone in three-ball counts and shrinking it in two-strike counts, umpires reveal an aversion to calls that end at-bats. But do they avoid these calls because they are pivotal to the outcome of the at-bat, or because they are pivotal to the outcome of the game? If umpires are averse to impacting the game, then the three-ball strike zone should expand more when the bases are loaded (and a walk would score a run), and the two-strike strike zone should contract more when there are two outs (and a strike-out would end the inning).

To determine whether umpires avoid calls that affect the outcomes of games over and above the outcomes of at-bats, we consider a continuous measure of how each call (ball or strike) impacts the outcome of the half-inning.\(^{19}\) A baseball game comprises a series of half-innings in which one team pitches and the other team bats. When three outs are recorded, the half-inning ends and the teams switch roles in the next half-inning. Before a pitch, the state of the half-inning can be summarized by the expected number of runs the batting team will score over the remainder of the half-inning. We define a half-inning state as the tuple of the count, outs, and runners on base, of which there are \((4 \times 3) \times 3 \times 2^3 = 288\) combinations.

We estimate \(\hat{E}[r_s]\), the expected number of runs to be scored over the remainder of each half-inning state \(s\), as \(\hat{R}_s = \frac{1}{||s||} \sum_{i \in s} r_i\), the empirical average in corresponding states using 26 years and 16 million pitches of data.\(^{20}\) Table 2 lists properties for select half-inning states. Generally, \(\hat{R}_s\) increases with the number of balls, decreases with the number of strikes, increases with men on base, and decreases with the number of outs.\(^{21}\)

\(^{19}\)Research shows that the number of runs a team scores closely tracks its probability of winning (Goldstein, 2014). In addition, the effect of a call on the outcome of the game cannot be measured reliably because the state sparse is too sparse for many states.

\(^{20}\)These data comprise almost every pitch thrown during the 1988-2013 regular seasons. We observe the least common half-inning state 688 times.

\(^{21}\)In some three-ball and zero-strike counts with a runner on third and fewer than two outs, calling a strike increases the expected number of runs to be scored. We suspect that this is because hitters are instructed not to swing with three balls and zero strikes, but are allowed to swing with three balls and one strike. Since pitches in both counts are likely to be in the strike zone, swings with runners on third are likely to be
Table 2: The expected run measure $\hat{R}_s$, the call impact measures $\delta_{\text{ball}}$ & $\delta_{\text{strike}}$, and the differential impact measure $\Delta$ for selected half-inning states.

We measure the impact of calling a ball or a strike as the change in the expected number of runs to be scored over the remainder of the half-inning as a result of the call:

$$\delta_{\text{ball}} = \hat{R}_{s'}_{\text{ball}} - \hat{R}_s$$

$$\delta_{\text{strike}} = \hat{R}_{s'}_{\text{strike}} - \hat{R}_s$$

where $\delta_{\text{ball}}$ is the impact of calling a ball, $\delta_{\text{strike}}$ is the impact of calling a strike, $s$ is the current half-inning state, and $s'$ is the half-inning state brought about by the call.\(^\text{22}\)

In Table 2, $\delta_{\text{ball}}$ is positive and large in three-ball counts and even more positive with runners on base. Similarly, $\delta_{\text{strike}}$ is negative and large in two-strike counts with zero outs and the bases empty ($c$ & $d$) and even more negative with two outs and the bases loaded ($g$ & $h$). In high-stakes states—two outs, bases loaded—a second strike ($e$ & $f$) decreases the expected number of runs nearly as much as a third strike with the bases empty and zero outs ($c$ & $d$).

Figure 4a shows the distribution of $\delta_{\text{ball}}$ and $\delta_{\text{strike}}$ in our sample of over a million calls. The graph contains one circle for each half-inning state, sized according to the relative beneficial for the batting team.

\(^{22}\hat{R}_{s'}_{\text{strike}} \equiv 0 \text{ when a strike ends the half-inning.}\)
**Figure 4:** Distribution of half-inning states by strike and ball impact, $\delta_{\text{strike}} \& \delta_{\text{ball}}$, for calls made by umpires in our sample. The impact of a ball or a strike is the difference in the expected number of runs to be scored over the remained of the half-inning from making that call. Sizes of circles (a) represent the relative incidence of states with associated impact. The *differential impact* of a call, $\Delta$, is $\delta_{\text{ball}} + \delta_{\text{strike}}$. For most calls, a strike and a ball are equally non-pivotal, creating a peak in the distribution of $\Delta$ at zero (b). But for some states, ball and strike impacts are asymmetric: one call is more pivotal than the other. When one call is pivotal, the alternative tends to be less pivotal (c).

Incidence of that state. Most decisions are relatively non-pivotal regardless of whether a ball or a strike is called; these calls have strike and ball impacts near zero. However, a number of states are more pivotal for strike calls than for ball calls, or more pivotal for ball calls than for strike calls. States in which at least one option is pivotal tend to be asymmetrically pivotal, or lie off of the diagonal.

An asymmetry in the impacts of his options may induce the umpire to avoid the more pivotal option when the correct call is not obvious. We measure how asymmetrically pivotal a call is according to its *differential impact*, the sum of its ball and strike impacts: $\Delta = \delta_{\text{ball}} + \delta_{\text{strike}}$. For states that lie on the diagonal in Figure 4a (for which $\delta_{\text{ball}} = -\delta_{\text{strike}}$), $\Delta = 0$. For asymmetrically ball-pivotal calls, $\Delta > 0$; for asymmetrically strike-pivotal calls, $\Delta < 0$. Figure 4b shows the distribution of differential impact in our sample. The distribution peaks at zero: many calls are non-pivotal. There is more mass in the negative
domain than in the positive domain: strikes tend to be more pivotal than balls (every state with a full count is asymmetrically strike-pivotal). And the distribution has long tails: some calls are asymmetrically strike-pivotal by more than half a run ($\Delta < -0.5$), and some calls are asymmetrically ball-pivotal by more than half a run ($\Delta > 0.5$).

High-stakes decisions are those in which at least one option is pivotal—in which the umpire can greatly change the course of the game. When one option becomes pivotal, the alternative tends to remain non-pivotal. Hence, high-stakes decisions are typically characterized by asymmetrically pivotal options. Figure 4c shows the relationship between the impact of the more pivotal option, $\max(\delta_{\text{ball}}, -\delta_{\text{strike}})$, and the absolute asymmetry of the call, $|\Delta| = |\delta_{\text{ball}} + \delta_{\text{strike}}|$. The slope of the regression line is positive and significant ($\hat{\beta} = 0.64$, $p < 0.001$): states with at least one pivotal option tend to be asymmetrically pivotal. This implies that the impact-averse umpire, who is averse to choosing the more pivotal option, will exhibit a stronger aversion when the stakes are higher because the impacts of his options will be more asymmetric.

3.4 Umpires are averse to making the more pivotal call

We investigate whether umpires are impact averse by observing how the probability of a called strike changes with our differential impact measure $\Delta$. If umpires are averse to making the more pivotal call, we should observe that conditional on the location of the pitch, the probability of a called strike increases monotonically with $\Delta$. When $\Delta < 0$, a strike call is asymmetrically pivotal, and the probability of a strike call should decline; when $\Delta > 0$, a ball call is asymmetrically pivotal, and the probability of a strike call should increase. We estimate a variation of the semi-parametric model in Equation 1, in which the distortion is
a non-linear function of $\Delta$:

$$y_i = p_i + g(\omega(p_i) \cdot \Delta_i) + \epsilon_i$$ (2)

We are interested in the shape of $g$, which we estimate from a kernel regression of $y_i - p_i$ on $\omega(p_i) \cdot \Delta_i$.\(^{24}\) We interpret $\hat{g}(z)$ as the change in the probability of a strike call from a baseline probability of 0.5 when $z = \Delta$—i.e. the bias on a borderline pitch with differential impact $\Delta$.\(^{25}\) If umpires avoid making the more pivotal call, we will observe $\hat{g} > 0$ when $\Delta > 0$ and $\hat{g} < 0$ when $\Delta < 0$.\(^{26}\)

Figure 5 shows $\hat{g}$, the distortion on a borderline pitch. The distortion is consistent with impact aversion: negative for asymmetrically strike-pivotal calls ($\Delta < 0$) and positive for asymmetrically ball-pivotal calls ($\Delta > 0$). $\hat{g}$ is generally increasing in $\Delta$, but the steepest increases occur in a narrow band around zero. For highly asymmetric calls, $\hat{g}$ is flat. When balls are asymmetrically pivotal, $\hat{g}$ peaks at just $\Delta = 0.05$. This corresponds to half-inning state $a$ in Table 2, in which the bases are empty, there are zero outs, and the count has two balls and one strike. Here, a ball is pivotal because it creates a count favorable to the hitter, not because it walks the batter. Hence, large distortions may occur when the count

\(^{23}\)Unlike the baseline probability in Equation 1, which is calculated on the subset of calls with fewer than three balls and fewer than two strikes in the count, $p_i$ here is calculated when the umpire’s calls are symmetrically pivotal, or when $\Delta = 0$. This construction ensures that $\hat{g} = 0$ when $\Delta = 0$, or that the baseline probability alone explains the call when the impacts of the umpire’s options are symmetric. Specifically, we measure this baseline probability $p_i$ as $\hat{m}_{\mu(\mathbb{R}),h(\mathbb{R})}(\mathbb{X}_i, \Delta = 0)$, the prediction from a three-dimensional kernel regression of $y_i$ for calls made by umpire $u(\mathbb{R})$ on batters of handedness $h(\mathbb{R})$. $p_i$ is the two-dimensional slice of $\hat{m}$ where $\Delta = 0$—the strike zone that the umpire would enforce if the impacts of calling a ball and a strike were symmetric. Since the distribution of $\Delta$ is concentrated at zero, our estimates are not meaningfully distorted by the curse of dimensionality. The correlation between the baseline probabilities as calculated in Equation 1 and here is 0.98.

\(^{24}\)Since the distribution of $\Delta$ is highly uneven (see Figure 4b), we use an adaptive bandwidth with a local bandwidth factor of the form $(f(x) / \exp(\frac{1}{N} \sum_{i=1}^{N} \log \hat{f}(X_i)))^{-\alpha}$, where $\hat{f}(x)$ is a density estimate using Silverman’s rule of thumb bandwidth. We use $\alpha = 0.5$ to balance smoothness and detail in the visual appearance of the function.

\(^{25}\)More generally, one can interpret $\hat{g}(z)$ as the bias on call with $\omega(p_i) \cdot \Delta_i = z$.

\(^{26}\)When $\Delta = 0$, $g = 0$ by assumption: when the impacts of the umpire’s options are symmetric, the probability of a strike call depends on pitch location alone.
Figure 5: \( \hat{g} \): the change in the probability of a called strike from the baseline probability. States with slightly asymmetric call impacts produce sizable distortions, beyond which the effect of differential impact (\( \Delta \)) is largely stable. Annotations refer to the states described in Table 2. Dotted lines denote 95% confidence intervals.

has fewer than three balls and fewer than two strikes. When \( \Delta = 0.05 \), borderline pitches are called strikes about 55% of the time. More positive asymmetries induce similar amounts of distortion.

When strikes are asymmetrically pivotal, \( \hat{g} \) falls quickly from \( \Delta = 0 \) until \( \Delta = -0.1 \). A differential impact of \(-0.1\) corresponds to state \( c \) in Table 2, in which the bases are empty, there are zero outs, and the count has two balls and two strikes. Here, a strike is pivotal because it ends the at-bat, even though it decreases the expected runs measure by just a fifth of a run. When \( \Delta = -0.1 \), borderline pitches are called strikes only 35% of the time. Further decreases in \( \Delta \) induce similar amounts of distortion.

These patterns confirm that umpires are impact averse, and they show that even a small asymmetry in the impacts of the umpire’s options significantly distorts his decisions. This implies that impact aversion distorts many decisions, not just the most asymmetrically pivotal ones.
3.5 Narrow versus broad framing

The relative steepness of $\hat{g}$ around zero suggests that umpires are as sensitive to moderate asymmetries as they are to large asymmetries. But this pattern may also arise if umpires greatly avoid making an impact on the at-bat but are less concerned about making an impact on the game—i.e., if umpires narrowly frame their decisions.\textsuperscript{27}

To determine whether impact aversion is narrowly framed around at-bats, we estimate $\hat{g}$ separately each of the twelve possible counts. As Table 2 shows, the same count can have varying differential impacts depending on the number of outs and whether there are runners on base. If umpires define impact wholly by the count—if impact aversion is narrowly framed—then $\hat{g}$ will be independent of $\Delta$ in each count. By contrast, if umpires are averse to making an impact on the half-inning over and above their impact on the at-bat—if impact aversion is broadly framed—then $\hat{g}$ will increase with $\Delta$ in every count. If umpires are averse to making an impact on the at-bat only by virtue of its impact on the half-inning, then $\hat{g}$ will resemble Figure 5 for all counts.

Figure 6 shows that umpires reveal an aversion to making the call that more greatly changes the outcome of the half-inning, rather than the call that more greatly changes the outcome of the at-bat; their decisions are broadly framed around the outcome of the half-inning. In eleven of twelve counts, $\hat{g}$ sharply increases with $\Delta$ around $\Delta = 0$ for borderline pitches. Moreover, the amount of distortion is similar across counts for moderate asymmetries; when $\Delta = -0.1$, for instance, the distortion is between $-10$ and $-20$ percentage points for six of the seven counts in which $\Delta \leq -0.1$ is observed.\textsuperscript{28}

\textsuperscript{27}Narrow framing refers to the tendency to evaluate new gambles in isolation from other risks (Kahneman, 2003; Barberis, Huang and Thaler, 2006).

\textsuperscript{28}These figures reveal other interesting patterns. As in Figure 5, umpires appear not to differentiate between calls that are moderately asymmetric in their impacts and those that are extremely asymmetric. In full counts (Figure 6i), $\Delta = -0.1$ and $\Delta = -0.5$ both imply about a 15 percentage point decrease in the probability of a strike call on a borderline pitch, even though these states portend considerably different outcomes for the half-inning. With three balls, the strike zone only expands when the count has zero strikes (Figure 6h), and then only at moderate levels of differential impact. When the count has three balls and one strike (Figure 6j), we estimate the distortions as a precise zero across the observed range of $\Delta$. In addition,
Figure 6: $\hat{g}$: the change in the probability of a called strike from the baseline probability.
3.6 Impact aversion under varying scrutiny

Umpires face public pressure to avoid making mistakes that greatly influence important outcomes. In 2010, umpire Jim Joyce’s erroneous safe call at first base thwarted what would have been only the 21st perfect game in baseball history. “He simply called the play as he saw it,” said The New York Times. “The problem, of course, is that Joyce’s decision is easily the most egregious blown call in baseball over the last 25 years.” After watching the replay, Joyce told reporters, “I just cost that kid a perfect game...It was the biggest call of my career” (Kepner, 2010). Influential decisions often attract negative publicity even when it is not clear ex post that the decision was mistaken. In 1972, the home plate umpire Bruce Froemming broke up Milt Pappas’ bid for a perfect game by calling ball four on a full-count pitch with two outs in the ninth inning. During Froemming’s final season 35 years later, a still-upset Pappas told ESPN that the last two pitches “were strikes or ‘that close’ to being strikes that he should’ve raised his right hand (to signal a strike)” (Weinbaum, 2007).

In order to avoid such criticism, umpires may avoid making pivotal calls that could, ex post, be mistaken. If so, umpires will display greater impact aversion when they face greater public scrutiny. We observe how impact aversion varies between games that are broadcast nationally, for which public scrutiny is high, and games that are broadcast locally, for which public scrutiny is lower. We find that impact aversion is considerably greater for games televised nationally than for games televised locally, suggesting that impact aversion is motivated at least in part by public pressure.29

the most asymmetrically strike-pivotal calls, which occur in two-strike counts, induce dramatically different distortions depending on the number of balls. With zero balls (Figure 6c), the strike zone contracts by as much as 25 percentage points—50/50 calls become 25/75 calls. But with just one ball (Figure 6f), the bias is not statistically different from zero for the most asymmetrically strike-pivotal calls. For moderately strike asymmetric states in two-strike counts, the strike zone contacts by 10 to 20 percentage points regardless of the number of balls.

29Two obvious comparisons of scrutiny—low versus high attendance and regular season versus playoffs—are problematic. Home fans are biased in favor of the home team, and they might pressure the umpire to display impact aversion only when such behavior is beneficial to their side. However, umpires appear immune to such pressure: there is no meaningful difference in the enforced strike zones for the home and away teams, umpires are equally impact averse when the home and away teams are at bat, and larger crowds

26
The vast majority of regular season games are televised locally. Two notable exceptions are Fox Saturday Baseball and ESPN Sunday Night Baseball. Most Saturdays during the regular season, Fox televises three simultaneous afternoon games. One game is a national telecast, while the other two are regional. Every Sunday night, ESPN televises a single game nationwide. These national telecasts oversample on heated rivalries, teams from large markets, and teams with winning records. Compared to local telecasts, Fox Saturday Baseball and ESPN Sunday Night Baseball attract larger television audiences, offer more compelling match-ups, and have greater postseason implications than other regular season games. Presumably, umpires face greater public scrutiny on Fox and ESPN.

Figure 7a shows \( \hat{g} \) for ESPN Sunday night games (red), Fox Saturday games (green), and other weekend games (blue). At moderate and extreme asymmetries, the lines separate, showing greater impact aversion on national telecasts. When a ball would be more pivotal than a strike, umpires on ESPN call balls on borderline pitches as little as 25% of the time, compared to a rate as low as 45% during local telecasts. When a strike would be more pivotal than a ball, umpires on Fox call strikes on borderline pitches as little as 20% of the time, are associated with only marginally higher levels of impact aversion. This last result may be a consequence of our finding in this section: games with high attendance are more likely to be broadcast nationally. Playoff games pose as another high scrutiny setting, but the effect of scrutiny on impact aversion is confounded by the selection process for playoff officiating which, as we show in Section 4.3, rewards the least impact-averse umpires. By contrast, regular season assignments are based only on considerations of logistics and fairness: minimizing travel and ensuring that umpires officiate each team a similar number of times (Trick, Yildiz and Yunes, 2012).

30 A third is TBS Sunday Baseball. Identifying games televised by Fox and ESPN is straightforward. Fox publishes the games they will televise prior to the season, while ESPN televises the only game played on Sunday at 8pm Eastern. By contrast, TBS updates its schedule as the season progresses and chooses games in non-exclusive time slots, making it difficult to identify which games were aired by TBS. Fox does not disclose its schedule for the final two Saturdays of the regular season in advance, and we exclude these dates from the analysis.

31 The national telecast is shown to viewers who do not reside where the regional telecasts are shown. Regional telecasts reach more viewers than local telecasts because a region covers multiple local markets, and because Fox reaches more homes than the cable channels that typically televise games locally.

32 ESPN cannot air a single team more than five times in a season (Newman, 2012); among its scheduled games, Fox did not air a single team more than 9 times in one season.

33 There is no meaningful difference in impact aversion between local telecasts on the weekend and local telecasts during the week.
Figure 7: \( \hat{g} \): the change in the probability of a called strike from the baseline probability, separately for national, regional, and local weekend telecasts (a), and relevant differences (b,c). Distortions induced by impact aversion are greater (i.e. farther from zero) for national and regional telecasts than for local telecasts.

(a) Distortion for ESPN Sunday night games, Fox Saturday games, and local weekend telecasts

(b) Difference in distortion between Fox telecasts and other Saturday games

c) Difference in distortion between ESPN telecasts and other Sunday games
compared to a rate as low as 35% during local telecasts. Figures 7b and 7c show differences between the red and green lines and other Saturday and Sunday games, respectively, with 95% confidence intervals. Impact aversion is greater for both the Fox and ESPN telecasts: for asymmetrically strike-pivotal calls, the distortion is more negative; for asymmetrically ball-pivotal calls, the distortion is more positive.

4 A Simple Model of Impact Aversion

We propose and estimate a single parameter, state-based utility model of umpire decision making. We use this model to characterize the heterogeneity in impact aversion among umpires. In our model, umpires derive utility from making calls that are consistent with their interpretations of the strike zone. Umpires gain utility when they make self-consistent calls, and they lose utility when they make self-inconsistent calls.

Our model presumes that umpires prefer to make the self-consistent call. But it also allows for umpires to have preferences about making pivotal calls in error. If an umpire calls a self-consistent ball or strike, he receives a fixed amount of utility regardless of the impact of that call. But if he calls a self-inconsistent ball or strike, the amount of disutility he receives depends on how pivotal that call is. Consider the hypothetical utilities in Figure 8. If the umpire’s call is correct according to his idiosyncratic strike zone, he receives a utility of 1. The impact of his call does not affect the utility he gains from making the self-consistent call. But if he calls a self-inconsistent ball or strike, his disutility rises in proportion to the impact he makes. Our model measures the slope of this disutility, which can be interpreted

---

34 The variance of the difference between two random variables is the sum of the variances of each random variable minus twice the covariance. Since the estimates follow each other closely, the true covariance is almost certainly positive. We assume the covariance to be zero, which means that our confidence intervals are overstated.

35 The range of the estimate is narrower than in Figure 5 because during the observation window, we do not observe instances of the most asymmetrically ball-pivotal calls on Fox Saturday games, nor do we observe instances of the most asymmetrically ball- or strike-pivotal calls on ESPN Sunday night games.

---

29
Figure 8: Hypothetical utilities. Calling a self-consistent ball or strike generates a fixed amount of utility regardless of the impact of that call. However, the disutility generated by calling a self-inconsistent ball or strike depends on how pivotal that call is (δ_{call}). In this example, a self-inconsistent ball or strike generates more disutility when the impact of that call is high. We estimate the slope of $U_{\text{self-inconsistent}}$ for each umpire.

as the disappointment the umpire anticipates when he makes a self-inconsistent call that changes the expected outcome of the game. If an umpire is impact neutral, this slope will be zero. But if he is impact averse, this slope will be negative. (If he is impact seeking, the slope will be positive.)

Prior to each pitch, the umpire forms beliefs about the impact of calling a ball and impact of calling a strike on the expected outcome of the game, which we measure as $\delta_{\text{ball}}$ and $\delta_{\text{strike}}$. The umpire then observes the location of the pitch, which signals the probability that the pitch is a strike according to his enforced strike zone.

We model umpires as maximizing the signal-weighted utilities of making the self-consistent and self-inconsistent calls. With probability $p$, a strike call is self-consistent; with probability
1 - p, a strike call is self-inconsistent:

\[ U_{\text{strike}} = p \cdot U_{\text{self-consistent}} + (1 - p) \cdot U_{\text{self-inconsistent}} \]

The reverse is true for calling a ball:

\[ U_{\text{ball}} = (1 - p) \cdot U_{\text{self-consistent}} + p \cdot U_{\text{self-inconsistent}} \]

Given these utilities, the umpire calls a strike if \( U_{\text{strike}} > U_{\text{ball}} \), and he calls a ball if \( U_{\text{strike}} < U_{\text{ball}} \).

We normalize \( U_{\text{self-consistent}} = 1 \). Symmetrically, we fix \( U_{\text{self-inconsistent}} = -1 \) when the impact of the associated call is zero. When a call is pivotal, we allow \( U_{\text{self-inconsistent}} \) to vary linearly with its impact. We also allow the slope of this relationship to vary by umpire.

\[
\begin{align*}
U_{\text{strike}}(p) &= p - (1 - p)(1 - \lambda_u \delta_{\text{strike}}) \\
U_{\text{ball}}(p) &= (1 - p) - p(1 + \lambda_u \delta_{\text{ball}})
\end{align*}
\] (3) (4)

If the umpire observes an obvious ball or strike (\( p \in \{0, 1\} \)), he receives a utility of 1 for making the obviously self-consistent call and a utility of 0 for making the obviously self-inconsistent call. He makes the self-consistent call. But if the signal is indeterminate (\( p \in (0, 1) \)), his call depends on the amount of disappointment he expects to feel when making the self-inconsistent call. If \( \lambda_u = 0 \), the umpire is not influenced by the impact of the call, and he receives a utility of \( 2p - 1 \) for calling a strike and \( 1 - 2p \) for calling a ball. Again, he makes the self-consistent call. But if \( \lambda_u > 0 \), he may choose the self-inconsistent call if it is the less pivotal choice. Consider a call for which \( p = 0.6 \), \( \delta_{\text{strike}} = -0.1 \) and \( \delta_{\text{ball}} = 0 \). Here, a strike is the self-consistent call, but the umpire calls a ball if \( \lambda_u > 10 \).

This model of impact aversion is distinct from standard economic models of decision
making under risk. In our model, decision makers receive positive utility for making a correct (e.g. self-consistent) choice and negative utility for making an incorrect (e.g. self-inconsistent) choice; impact averse arbitrators receive greater disutility when a mistake would prove consequential. Hence, impact aversion differs from standard assumptions of risk aversion because an impact averse utility curve is kinked at the reference point that divides correct and incorrect decisions, and because the angle of the kink is endogenous.36

Impact aversion is similar to loss aversion (Kahneman and Tversky, 1979) in that both describe kinked utility around a reference point, but they differ in two important ways. First, in loss aversion a variable reference point determines what is coded as a gain and what is coded as a loss, whereas in our model of impact aversion, a correct decision is always coded as a gain, and an incorrect decision is always coded as a loss. Second, the degree of loss aversion, defined as the ratio of the slopes of the utilities for losses and gains, is presumed to be exogenous to the model and is often estimated to be about 2.25 (Tversky and Kahneman, 1991). By contrast, the relative impacts of the decision maker’s options endogenously determine the degree of impact aversion she displays. In impact aversion, the reference point is fixed, and the disutility of a loss is endogenous.

36In a related paper, Romer (2006) shows that coaches in the National Football League avoid options that increase the likelihood of winning in expectation, but may result in large decreases in that probability. “The natural possibility,” Romer writes, “is that the actors care not just about winning and losing, but about the probability of winning during the game, and that they are risk-averse over this probability. That is, they may value decreases in the chances of winning from failed gambles and increases from successful gambles asymmetrically.” Risk aversion applies naturally to actors whose utility is function of a continuous and positive outcome, like the probability of winning during the game. However, exogenous risk aversion sits more uneasily with actors whose utility is a function of a binary and opposing outcome, like making the correct or incorrect choice, and is inconsistent with the disutility of an incorrect choice depending endogenously on attributes of the situation.
4.1 Structural estimates

We measure the signal $p$ as the baseline probability of a strike call, or the probability of a strike based solely on the location of the pitch.\textsuperscript{37} Next, we estimate $\lambda_u$ separately for each umpire: first adding an IID type I extreme value error term to each of the utilities, and then finding the $\hat{\lambda}_u$ that maximizes the resulting logistic likelihood function. Calls with asymmetric impact identify $\lambda_u$.

**Figure 9:** Distribution of the coefficient of impact aversion $\hat{\lambda}_u$ for the 75 umpires in our sample (a), and the relationship between these estimates and MLB umpiring experience (b). For an unbiased umpire, $\lambda_u = 0$. The smallest $\hat{\lambda}_u$ is 10. Impact aversion does not appear to be correlated with experience.

\textsuperscript{37}Unlike the baseline probability in Equation 2, which is calculated when the umpire’s calls are symmetrically pivotal, here the baseline probability is calculated when the impacts are not only symmetric but also both equal to zero, or when $\delta_{\text{ball}} = \delta_{\text{strike}} = 0$. Specifically, we measure $p_i$ as $\hat{m}_{u(i),h(i)}(X_i, \delta_{\text{ball}} = 0, \delta_{\text{strike}} = 0)$: the probability that umpire $u^{(i)}$ calls a strike on a batter with handedness $h^{(i)}$ when both options are non-pivotal. $\hat{m}$ is a kernel regression in four dimensions: two for the location of the pitch, one for $\delta_{\text{ball}}$, and one for $\delta_{\text{strike}}$. $p_i$ is the likelihood that the umpire would call a strike were he not influenced by the impact of either call. The correlation between the baseline probabilities as calculated in Equation 2 and here is 0.95.
has a $\hat{\lambda}_u = 10$ with a standard error of 0.55, and the largest standard error for any umpire’s $\hat{\lambda}$ is 1.1. Every umpire in our sample shades away from the more pivotal call when the self-consistent call is not obvious.

Heterogeneity in impact aversion among umpires can be explained by persistent, individual-level characteristics. We estimate $\lambda_{u,t}$, a coefficient of impact aversion for each umpire $u$ in each season $t$ from 2009-11, and we regress $\hat{\lambda}_{u,t}$ on $\alpha_u$, a set of umpire fixed effects. This regression has an $R^2$ of 0.63 (adjusted-$R^2 = 0.44$); stable differences among umpires account for much of the year-to-year variation in impact aversion. We also rank order $\hat{\lambda}_{u,t}$ by season and observe a correlation of 0.56 between the orderings in 2009 and 2011; relative levels of impact aversion are persistent across the observation window. Impact aversion appears persistent over longer time horizons, as well. Figure 9b shows the relationship between $\hat{\lambda}_u$ and tenure, which we define as the year in which the umpire first officiates a Major League game. Though a causal interpretation is likely confounded by unobserved selection, there does not appear to be a relationship between tenure and impact aversion.

4.2 Strike thresholds

To see the distortion of the strike zone implied by a particular $\hat{\lambda}_u$, consider a counterfactual prediction: the signal an umpire would need to receive in order to be indifferent between calling a ball and a strike. An unbiased umpire is indifferent when he receives a signal of $p = 0.5$, but a biased umpire ($\lambda_u > 0$) may require a different signal when choosing between options with asymmetric impacts. Let $\tilde{p}_u$ be a strike threshold: the signal $p$ at which umpire $u$ with parameter $\lambda_u$ is indifferent between calling a ball and calling a strike:

$$\tilde{p}_u = \{p : U_{\text{strike}} = U_{\text{ball}}; \lambda_u\}$$

We weight each observation of $\hat{\lambda}_u$ by the inverse of its variance.
Substituting from Equations 3 & 4 and solving for $\tilde{p}_u$:

$$\tilde{p}_u = \frac{2 - \lambda_u \delta_{\text{strike}}}{4 + \lambda_u (\delta_{\text{ball}} - \delta_{\text{strike}})}$$  \hspace{1cm} (5)

For an umpire averse to making pivotal calls ($\hat{\lambda}_u > 0$), $\tilde{p}_u > 0.5$ when $\delta_{\text{ball}} < -\delta_{\text{strike}}$, and $\tilde{p}_u < 0.5$ when $\delta_{\text{ball}} > -\delta_{\text{strike}}$. When a strike is more pivotal than a ball, the biased umpire needs a signal greater than 50% in order to call a strike; he is ball-biased. But when a ball is more pivotal than a strike, the biased umpire calls strikes when he is less than 50% sure that the pitch is actually a strike; he is strike-biased. By construction, $\tilde{p}_u = 0.5$ when $\delta_{\text{ball}} = -\delta_{\text{strike}}$: the umpire is unbiased when the impacts of his options are symmetric.

Figure 10: Strike thresholds $\tilde{p}_u$ for the minimum (a) and maximum (b) $\lambda_u$ as computed using Equation 5. By construction the $\tilde{p} = 0.5$ for calls with symmetric impact. When a strike is asymmetrically pivotal, $\tilde{p} > 0.5$: both the least biased and the most biased umpires need a signal of greater than 50% to call a strike 50% of the time. Annotated letters correspond to half-inning states from Table 2.
served $\hat{\lambda}_u$ (10b). For both the least and most impact-averse umpires, the strike threshold deviates greatly from 0.5 with moderate amounts of asymmetry. Half-inning state $a$ has nearly symmetric call impacts with $\Delta < 0.05$ (see Table 2). Even so, the strike threshold ranges from 37% to 42% in the population—no umpire needs to be more than 42% confident that a pitch is a strike in order to call a strike 50% of the time. Heterogeneity in impact aversion is small relative to the magnitude of impact aversion for the least biased umpire. For each of the five half-inning states plotted on the figures, the difference in the strike thresholds between the most and least biased umpires is smaller than the difference between the strike threshold of the least biased umpire and the unbiased threshold of 0.5.

4.3 Lucrative assignments

Assignments to officiate the All-Star Game or the playoffs constitute the largest incentives umpires face. These lucrative assignments can raise the annual income of an experienced umpire from $300,000 (Hines, 2007) to $400,000 (Thompson, 2014). “Assignments for the All-Star Game as well as post-season games are based on merit,” says former MLB VP of Umpiring Ralph Nelson, “Each year we attempt to choose ‘all-star’ umpires who have shown exemplary work during the season.”

Major League Baseball appears to view the most impact-averse umpires as the least exemplary: more impact-averse umpires are less likely to receive lucrative assignments.

We measure $\hat{\lambda}_u$ during the 2009-11 regular seasons. We then observe whether this measure of impact aversion predicts All-Star Game or playoff assignments in subsequent seasons. The regression results reported in Table 3 predict an umpire’s chances of officiating an All-Star Game in 2012-14 (Models 1-3), or at least one series during the 2011-13 postseasons (Models 4-6).

Model 1 shows that 21% of umpires in our sample officiate an All-Star Game in 2012,

---

40 A crew of six umpires is assigned to each postseason series. The umpires rotate positions (home plate; first, second, and third base; right and left field) each game. In results not shown, assignment to initial
\begin{table}
\centering
\begin{tabular}{lcccccc}
\hline
 & \multicolumn{3}{c}{All-Star Game} & \multicolumn{3}{c}{Playoff series} \\
 & (1) & (2) & (3) & (4) & (5) & (6) \\
\hline
$\hat{\lambda}$ (standardized) & -0.073* & -0.10** & -0.10* & -0.15*** & -0.20*** & -0.15*** \\
 & (-1.69) & (-2.06) & (-1.83) & (-3.43) & (-3.86) & (-3.00) \\
Consistency (standardized) & 0.072 & 0.078 & 0.12* & 0.085 & 0.085 \\
 & (1.34) & (1.35) & (1.87) & (1.65) & \\
Correctness (standardized) & -0.027 & -0.054 & -0.019 & 0.011 & 0.011 \\
 & (-0.43) & (-0.87) & (-0.37) & (0.22) & \\
Hired in 1999 & -0.094 & 0.044 \\
 & (-0.80) & (0.46) & \\
Hired after 1999 & 0.14 & -0.20 \\
 & (0.90) & (-1.57) & \\
Full-time & 0.038 & 0.43*** \\
 & (0.34) & (3.49) & \\
Constant & 0.21*** & 0.21*** & 0.19 & 0.73*** & 0.73*** & 0.43*** \\
 & (4.52) & (4.50) & (1.58) & (15.05) & (15.24) & (3.57) & \\
Observations & 75 & 75 & 75 & 75 & 75 & 75 \\
$R^2$ & 0.031 & 0.049 & 0.088 & 0.113 & 0.159 & 0.375 \\
\hline
\end{tabular}
\begin{tabular}{l}
\textit{t} statistics in parentheses \\
* $p<0.1$, ** $p<0.05$, *** $p<0.01$
\end{tabular}
\end{table}

Table 3: Linear probability model of officiating an All-Star Game between 2012 and 2014 or at least one playoff series between 2011 and 2013, with Huber-White standard errors.

An umpire whose $\hat{\lambda}_u$ is one standard deviation below the mean—i.e. less impact averse than average—receives an All-Star Game assignment with 28% probability. But an umpire who is one standard deviation more impact averse than average receives a postseason assignment with only 14% probability. A kernel regression (not reported) shows this relationship to be approximately linear.

Major League Baseball may penalize more impact-averse umpires because they are inaccurate in making their calls. We predict lucrative assignments using two measures of positions in the rotation is uncorrelated to the independent variables in Table 3.

41Because $\hat{\gamma}_u$ is an estimate with known variance, precise standard errors can be calculated by sampling from the distribution of each estimate. These standard errors are similar to the analytic standard errors presented because the variances of the estimates are similarly small.

42A kernel regression (not reported) shows this relationship to be approximately linear.

37
accuracy. The first, consistency, measures the percent of an umpire’s calls that are correct according to his own strike zone.\textsuperscript{43} The second, correctness, is the share of a umpire’s calls that are correct according to the official strike zone.\textsuperscript{44} Model 2 shows no statistically significant relationship (at the 10% level) between consistency or correctness and assignment to an All-Star Game. Moreover, neither measure of accuracy can explain the negative relationship between impact aversion and playoff assignment, which strengthens when the consistency and correctness measures are included in the regression. Punishment for impact aversion cannot be explained as punishment for inaccuracy.

Model 3 shows that the effects of impact aversion on All-Star Game assignment persists when we control for the period in which the umpire was hired and his experience during the observation window. The variable ‘Full-time’ is an indicator for whether the umpire officiate at least 89 games behind home-plate during the observation window, which divides the 16 umpires who officiate 48 to 86 games over this window from the 59 umpires who officiate 89 to 105; umpires who officiate fewer games are less likely to have been employed full-time during the observation window. None of these variables is statistically significant at conventional levels, and their inclusion in the model does not affect the magnitude of the coefficient on $\hat{\lambda}_u$.

The determinants of playoff selection appear similar to those of All-Star Game selection. Model 4 shows that a one standard deviation increase in impact aversion is associated with a 15 percentage point decrease in the probability of receiving a postseason series assignment between 2011 and 2013.\textsuperscript{45} Model 5 shows that more consistent umpires are more likely to receive at least one playoff assignment, but more correct umpires are not more likely. This

\textsuperscript{43}To calculate consistency, we identify 50% contour lines for each umpire–batter handedness tuple. For reference, Figure 1a shows the 50% contour line for all calls in the data. Strike calls inside the 50% contour line and ball calls outside the 50% contour line are considered consistent; other calls are considered inconsistent.

\textsuperscript{44}Consistency and correctness are positively correlated ($\sigma = 0.54$), and umpires are more consistent than correct: the average umpire is consistent on 89.8% of calls (s.d. 0.5%) and correct on 84.6% of calls (s.d. 0.9%).

\textsuperscript{45}As before, a kernel regression (not reported) shows this relationship to be approximately linear.
finding is consistent with anecdotal evidence that the league tolerates deviations from the official strike zone as long as umpires enforce those deviations consistently. Again, neither measure of accuracy can explain the negative relationship between impact aversion and playoff assignment. Model 6 shows that, in contrast to All-Star Game selection, Major League Baseball rewards more experienced umpires with playoff assignments—in addition to punishing the most impact-averse umpires.\footnote{Conditional on receiving at least one postseason assignment during 2011-13, the number of assignments and their prestige (i.e. whether the umpire officiates a World Series) depends (positively) on the umpire’s tenure, but not his level of impact aversion, consistency, or correctness.}

These estimates suggest that a one standard deviation increase in impact aversion—a difference in $\lambda$ of just 2.2—reduces an umpire’s expected annual earnings by as much as $20,000.\footnote{This estimate is 20\% ($\hat{\lambda}$ from Model 5) of the difference between the experienced umpire’s base salary ($300,000) and the experienced umpire’s total compensation with assignment bonuses ($400,000).}

5 Noisy Signals

We extend the model from Section 4 to address situations in which umpires observe a noisy signal rather than a point probability. In doing so, we assume that umpires are second-order risk averse (e.g. Nau, 2006; Abdellaoui, Klibanoff and Placido, Forthcoming), or \textit{ceteris paribus} prefer the option with the less noisy signal.\footnote{Abdellaoui, Klibanoff and Placido (Forthcoming) provides experimental evidence that individuals are second-order risk averse. Nau (2006) presents a theoretical analysis of second-order risk aversion.} Section 5.1 shows that under second-order risk aversion, impact aversion increases with signal noise. Section 5.2 examines three empirical situations characterized by noisy signals. In all three cases, consistent with the extended model’s predictions, impact aversion is greater for decisions with noisy signals than for comparable decisions with non-noisy signals.
5.1 Extended model with second-order risk aversion

A non-noisy signal \( p \) is the point probability that a strike is consistent with the umpire’s idiosyncratic strike zone. Let a noisy signal \( F_p \) be a symmetric distribution around \( p \)—i.e. a mean-preserving probability spread. In the model from Section 4, a noisy signal \( F_p \) and a non-noisy signal \( p \) produce the same behavior. Because the utilities in Equations 3 and 4 are linear in \( p \), \( \int_{p-\epsilon}^{p+\epsilon} U(q)dF(q) = U(p) \) for both \( U_{\text{strike}} \) and \( U_{\text{ball}} \).

This is no longer the case when the umpire is second-order risk averse. We incorporate second-order risk aversion by introducing the concave and strictly increasing function \( v(\cdot) \):

\[
U_{\text{strike}}(p) = v\left(p - (1 - p) \cdot \gamma_{\text{strike}}\right) \tag{6}
\]

\[
U_{\text{ball}}(p) = v\left(1 - p - p \cdot \gamma_{\text{ball}}\right) \tag{7}
\]

where \( \gamma_{\text{strike}} \) and \( \gamma_{\text{ball}} \) are the option-specific coefficients of impact aversion \( 1 - \lambda\delta_{\text{strike}} \) and \( 1 + \lambda\delta_{\text{ball}} \), respectively.

The concavity of the utilities affects choice when the signal becomes noisy. Consider a simple noisy signal \( F_p \) that realizes \( p - \epsilon \) with probability \( \frac{1}{2} \) and realizes \( p + \epsilon \) with probability \( \frac{1}{2} \), for \( \epsilon > 0 \). Under this Bernoulli noisy signal,

\[
\int_{p-\epsilon}^{p+\epsilon} U(q)dF(q) = \frac{1}{2} U(p - \epsilon) + \frac{1}{2} U(p + \epsilon)
\]

\[
= \frac{1}{2} v(a - b\epsilon) + \frac{1}{2} v(a + b\epsilon)
\]

\[
< v(a) = \mathbb{E}[U(p)],
\]

where \( a \) is an option-specific function of \( p \) and \( \gamma \), and \( b = 1 + \gamma \). Because \( v \) is concave, the utility of an option decreases as signal noise \( \epsilon \) increases. Moreover, this decrease is sharper for more pivotal options, or those with higher \( \gamma \) (since \( b = 1 + \gamma \)). A noisy signal introduces
the symmetric second-order risks that a pivotal option may be certainly right and that it may be certainly wrong. With concave second-order utility, an impact-averse umpire will overweight the second-order risk that a pivotal option is certainly wrong relative to the second-order risk that it is certainly right—and he will overweight the downside risk more for more pivotal options. Hence, second-order risk aversion makes an impact-averse umpire err even more towards the less pivotal option when the signal is noisy than when the signal is not noisy.

5.2 Impact aversion is increasing in the noisiness of the signal

We assume that the signal is more noisy when the location of the pitch is more difficult to observe. We examine three situations in the data characterized by noisy signals: pitches near the top and bottom borders of the official strike zone, which move up and down based on the hitter’s height and stance; off-speed pitches, which follow a curved trajectory rather than a straight line; and pitches in which the umpire must make his call instantaneously, rather than be allowed to take his time. In all three cases, impact aversion is greater under noisy signals than under comparable non-noisy signals. These findings are consistent with the predictions of the extended model in Section 5.1.

5.2.1 The top and bottom of the official strike zone

The location of the pitch with respect to the official strike zone is more uncertain at the top and bottom of the official strike zone than along its sides for two reasons. First, the width of the official strike zone is fixed, but the height varies both with the height of the batter and with the stance the batter takes for each pitch. Second, the vertical location of the pitch is more difficult to observe than its horizontal location. Standing behind home plate, the umpire can more easily tell whether a pitch passes over the white of the plate than whether it crosses between the bottom of the batter’s knees and the midline of his chest.
Difficulty in determining the location of the pitch with respect to the top and bottom of the official strike zone creates uncertainty about the probability that the pitch is a strike. If a pitch passes over the edge of home plate at the hitter’s belt, it is likely borderline, or a strike about 50% of the time. But if it passes over the center of home plate at the level of the batter’s knees, it might be borderline, but depending on the batter’s stance and the umpire’s perception, it might be a certain strike or a certain ball instead. Pitches near the top and bottom of the official strike zone carry noisier signals than pitches along the sides.

If umpires become more impact averse as the signal becomes noisier, then we should observe greater bias at the top and bottom of the official strike zone than along the sides. This is exactly what we see in Figures 2a and 2b: the expansion of the strike zone in three-ball counts and the contraction of the strike zone in two-strike counts are both greater at the top and bottom of the official strike zone than along its sides. In both figures, the distortions along the top and bottom are twice as large as along the sides. Where the location of the pitch is more uncertain, umpires display greater impact aversion.

5.2.2 Off-speed pitches

The ease of identifying the location of the pitch also varies by the type of pitch. The locations of off-speed pitches, which tend to move vertically or laterally from the umpire’s perspective, are more difficult to observe than the locations of fastballs, which trace a more linear path from the pitcher’s hand to the catcher’s mitt. Using the PITCH f/x data, MLB classifies each pitch into one of more than a dozen types. We reduce this taxonomy to two types: fastballs, which comprise 64% of calls, and off-speed pitches, which comprise the remaining 36%. About two-thirds of off-speed pitches are either curveballs or sliders, two pitch types that pitchers spin upon release in order to induce vertical or lateral movement. On average, fastballs drop 5.0 vertical inches from release until crossing home plate, while
off-speed pitches fall 9.5 inches.49

**Figure 11:** $\hat{g}$: the change in the probability of a called strike from the baseline probability, for off-speed pitches and fastballs (a), and their difference (b). Distortions induced by impact aversion are generally greater (i.e. farther from zero) for off-speed pitches.

---

Figure 11 shows $\hat{g}_{\text{Offspeed}}$ and $\hat{g}_{\text{Fastball}}$ separately (11a) as well as the difference $\hat{g}_{\text{Offspeed}} - \hat{g}_{\text{Fastball}}$ (11b). Impact aversion is larger for off-speed pitches than for fastballs: the bias is more negative when the call is asymmetrically strike-pivotal and generally more positive when the call is asymmetrically ball-pivotal. Noisier signals induce greater impact aversion.

### 5.2.3 Time pressure

For most calls, play stops, and the umpire renders his verdict about a second after the catcher catches the pitch. But for 1.5% of calls, the umpire must announce his choice immediately, because the call tells the catcher whether to make a play on a potential baserunner. These calls occur in three-ball counts with a runner on first, except for calls with two strikes and two

49The $t$-statistic for this difference is of the order $10^3$.
outs. Time pressure increases uncertainty about the location of the pitch. If noisy signals induce greater impact aversion, we should observe more bias for calls with time pressure than for calls in three-ball counts without time pressure.

**Figure 12:** \( \hat{g} \): the change in the probability of a called strike from the baseline probability, for calls with time pressure and in three-ball counts without time pressure (a), and their difference (b). Distortions induced by impact aversion are generally greater (i.e. farther from zero) under time pressure.

---

Figure 12 shows \( \hat{g}_{TP} \) and \( \hat{g}_{\neg TP} \) (3 balls) separately (12a) as well as the difference \( \hat{g}_{TP} - \hat{g}_{\neg TP} \) (12b). Calls under time pressure generally exhibit greater impact aversion than calls not under time pressure. In all three cases, noisy signals induce greater impact aversion than non-noisy signals.

---

50When the count has three balls and a walk would advance the runner(s) but a called strike would not end the inning, the call tells the catcher how he should address a potential steal. If the call is a strike, the catcher should make a play on the runner. But if the call is a ball, the runner advances and the catcher can only err by trying to make a play. Since the home plate umpire’s focus is on the pitch rather than the runners, he must make his call immediately in case a play needs to be made, even if no runners are trying to advance.

51We compare calls under time pressure to calls in three-ball counts without time pressure because time pressure implies three balls.
6 Economic Significance of Umpires’ Impact Aversion

On the free agent labor market, MLB teams spend an average of $6.5M for each win that the acquired player is expected to contribute (Silver, 2014). With nearly 2500 games in each season, impact aversion need only affect the outcomes of a small number of games in order to significantly affect the economic fortunes of teams. Section 6.1 measures the number of calls that reverse in expectation as a result of the bias. Section 6.2 measures the mean distortion, and its corresponding dollar value, induced by each call.

6.1 Call reversals

A call reverses in expectation if \[ p < 0.5 \] and \[ p + \hat{g}(\omega(p) \cdot \Delta) > 0.5 \] or \[ p > 0.5 \] and \[ p + \hat{g}(\omega(p) \cdot \Delta) < 0.5 \]. In the first case, the pitch is a ball in expectation according to its location, but the umpire calls it a strike more than half the time; in the second case, the pitch is a strike in expectation according to its location, but the umpire calls it a ball more than half the time.

In an average game, impact aversion reverses four calls in expectation, or one call in every forty. Calls in counts with zero balls and one strike flip most frequently, at 5.4%, followed by calls in counts with three balls and zero strikes, which reverse in 4.4% of calls. In two-strike counts, calls flip between 3.4% and 4.1% of the time. An average game comprises eighty at-bats. About half of these reach a two-strike count, and about half of those include a call with two strikes. Among at-bats in which a call is made in a two-strike count, 5.8% include at least one call that flips in expectation from a strike to a ball. Absent impact aversion, these at-bats likely would have ended in strikeouts; but 42% of these at-bats end in something other than a strikeout. Once a game, on average, an expected third strike is called a ball because of impact aversion. And once every other game, an at-bat ends in something other

\[ p \] refers to the baseline probability from Equation 2. Here, \( \hat{g} \) refers to the count-specific distortion estimates from Figure 6.
than a strikeout after a third strike should have been called.

6.2 Mean distortion

Our estimate of the distortion induced by impact aversion is \( \hat{g}(\omega(p_i) \cdot \Delta_i) \).\(^{53}\) We define the mean distortion as \( \frac{1}{N} \sum_{i=1}^{N} |\hat{g}(\omega(p_i) \cdot \Delta_i)| \), or the average absolute deviation in the observed calls from their baseline probabilities. The mean absolute distortion is 2.9 percentage points for all calls, which implies that the rate at which the average pitch is called a strike is 2.9 percentage points from an unbiased rate based on pitch location alone. This figure is higher in more asymmetrically pivotal counts. When the count has zero balls and one strike or three balls and zero strikes, the mean distortion is 5.7 percentage points. In two-strike counts, the mean distortion varies from 4.9 to 5.4 percentage points; on average, the looming impact of a strikeout makes umpires about five percentage points less likely to call strike three.

We use the mean distortion measure to quantify the financial consequences of impact aversion. If teams are willing to pay $6.5M to turn a loss into a win, then a risk-neutral team is willing to pay up to \$(dp \cdot 6.5)M\) for a call that increases its probability of winning by \( dp \) over the opposite call. Assume that the probability of winning is a linear function of the number of runs a team scores. We regress an indicator for whether a team wins on how many runs it scores using 26 years of game data.\(^{54}\) According to this model, an extra run increases the probability of winning by 8.6 percentage points. Hence, \( dp = 0.086 \times |\hat{R}_{s'} - \hat{R}_{s}\) ball - \( \hat{R}_{s}\) strike\),

where \( \hat{R}_s \) is the expected runs measure in half-inning state \( s \), and \( s' \) is the state that follows the associated call. For the average call, the absolute difference in the win probability resulting from the umpire’s choices, or \( \frac{1}{N} \sum_{i=1}^{N} dp_i \), is 1.2 percentage points. This implies that on average, \$75,000\) hangs in the balance for each call.

We are interested in the fraction of this amount that is attributable to impact aversion.

\(^{53}\)As in Section 6.1, we use the count-specific distortion estimates from Figure 6.

\(^{54}\)These data are the same data used to calculate the expected runs measure for each half-inning state: almost every pitch thrown during the 1988-2013 seasons.
We calculate this quantity as:

$$\frac{6.5M}{N} \sum_{i=1}^{N} \left| dp_i \cdot \hat{g}(\omega(p'_i) \cdot \Delta_i) \right| \approx 3,000$$

Here, we weight the change in the win probability by the amount of distortion induced by \( \omega(p'_i) \cdot \Delta_i \). If \( dp \) and \( \hat{g} \) were independent, this figure would be the product of $75,000 and the mean distortion estimate of 2.9%, or about $2,000. The true figure is higher because the calls that greatly affect which team is likely to win are subject to higher levels of distortion. On average, impact aversion distorts about $3,000, or 4%, of team value every call.

7 Conclusion

Major League Baseball umpires are impact averse. Despite a directive from MLB to call balls and strikes based solely on pitch location, every umpire reveals an aversion to the option that more greatly changes the expected outcome of the game.

One possible explanation for our findings is that contrary to its official directive, Major League Baseball informally incentivizes impact aversion. This explanation is unlikely. MLB uses stereoscopic cameras to precisely observe the consistency of each umpire’s calls, and these observations feed directly into umpires’ evaluations: umpires who display greater impact aversion are less likely to receive lucrative assignments to officiate the All-Star Game or postseason series. The bias also appears to be at odds with the league’s interests. Impact aversion likely prolongs games (by reducing strike-outs at a higher rate than reducing walks), and Major League Baseball began taking steps to shorten games just before our observation window (Bloom, 2008).

The alternative explanation is that impact aversion comes from sources other than Major League Baseball. One possibility is that impact aversion represents a response to public
pressure to avoid pivotal mistakes. If umpires try to avoid public criticism for pivotal mistakes, they will shy away from the more pivotal option when they are unsure which option is correct, and they will do so more strongly when they are under greater scrutiny. Consistent with this interpretation, we find considerably higher levels of impact aversion during national broadcasts compared to local broadcasts. The most impact-averse umpires appear to be trading compensation for less public scrutiny.

In our context, impact aversion is a decision bias. Impact aversion implies an avoidance of potentially consequential options, which makes it similar to a class of decision biases in which individuals avoid potentially consequential decisions. These include status quo bias (Samuelson and Zeckhauser, 1988; Choi, Laibson, Madrian and Metrick, 2003; Johnson and Goldstein, 2003), omission bias (Ritov and Baron, 1992; Schweitzer, 1994), and choice deferral (Tversky and Shafir, 1992); see Anderson (2003) for a review. Impact aversion is a bias towards options that better preserve the status quo. While active choice, or requiring a decision to be made, has been proposed as a policy to prevent status quo bias (Carroll, Choi, Laibson, Madrian and Metrick, 2009; Keller, Harlam, Loewenstein and Volpp, 2011; Schrift and Parker, 2014), we find that umpires display impact aversion even under active choice.

Impact aversion is distinct from arbitrator biases previously identified in the empirical literature. Studies in sports settings have documented evidence of player favoritism by arbitrators (Sutter and Kocher, 2004; Zitzewitz, 2006; Price and Wolfers, 2010; Parsons, Sulaeman, Yates and Hamermesh, 2011; Mills, 2013; Kim and King, 2014; Zitzewitz, 2014). Though some experiments document evidence of action bias, or a bias towards making consequential decisions, other experiments show that omission bias is more prevalent than action bias (Baron and Ritov, 2004).

However, our findings suggest that active choice does reduce impact aversion. Choosing a strike is more “active” than choosing a ball, in the sense that an arm motion signals a strike (and a full-body motion signals a third strike), whereas no motion signals a ball. Our main finding is that umpires shade towards balls when a strike would be more pivotal, and they shade towards strikes when a ball would be more pivotal. But they shade more towards balls when a strike would be more pivotal than they shade towards strikes when a ball would be more pivotal.

48
In contrast, an impact-averse umpire will favor particular choices, not particular players. A notable exception is the finding by Price, Remer and Stone (2012) that referees in the National Basketball Association favor choices that are more profitable for the league. However, impact aversion among umpires does not appear to be a manifestation of profit-seeking by Major League Baseball. External incentives to appear evenhanded motivate labor arbitrators to violate their directive (Bloom and Cavanagh, 1986; Klement and Neeman, 2013); by contrast, a desire to appear invisible appears to motivate the violations in our setting. Much of the empirical literature on judicial decision making focuses on how the political ideology of the judge biases her rulings (e.g. Epstein, Landes and Posner, 2011). However, recent evidence shows that judges display decision biases as well: experienced parole judges become discontinuously more likely to grant merciful rulings after food breaks (Danziger, Levav and Avnaim-Pesso, 2011).

Judges are often compared to umpires. “Judges are like umpires,” said Supreme Court Chief Justice John Roberts at his confirmation hearing, “it’s my job to call balls and strikes.” However, umpires do not simply call balls and strikes, and neither do judges, according to legal scholars. “Supreme Court justices,” says constitutional law scholar Erwin Chemerinsky, “determine the rules and the strike zone” (quoted in Lithwick, 2014). In this sense, judges truly are like umpires.

Since umpires are impact averse, judges may be impact averse as well. An emerging literature on the psychology of judges argues that salient, but legally irrelevant, information distorts judicial rulings (Bordalo, Gennaioli and Shleifer, Forthcoming). One such salient factor might be the repercussions from making a mistake that proves consequential to the outcome of the case. Relative to non-pivotal mistakes, pivotal mistakes may make a case more likely to be overturned on review; they may reduce the judge’s chances of winning an election, an appointment, or a confirmation; and they may make the judge feel regret.

“Justices have tremendous discretion,” says Chemerinsky, which may produce even greater impact aversion among judges than among umpires.

Impact aversion may be especially important in procedural decisions, such as decisions about the admissibility of evidence. For example, a request to admit the defendant’s confession presents the judge with asymmetrically pivotal options. If the judge admits the confession, the jury will likely convict the defendant; if the judge excludes the confession, the jury’s perception of the defendant’s guilt remains unchanged. As with balls and strikes, the admissibility of a confession is supposed to be decided on a single, objective criterion—whether the confession was made voluntarily (Leo, 2013)—and without regard to the impact of the confession on the jury. If judges are impact averse, they will distort these procedural rulings by too often excluding confessions when they would sway the jury. In this way, impact aversion would reverse case outcomes.
A Alternative Explanations: Rational Expectations

We consider the possibility that evidence of impact aversion can be explained by umpires’ rational expectations of the forthcoming pitch. Umpires might form expectations from the long-run distribution of pitches thrown in particular counts. If pitchers tend to throw strikes in three-ball counts, umpires might expect a strike in those counts; if pitchers tend to throw balls in two-strike counts, umpires might expect balls in those counts.

**Figure 13:** \( \hat{f}(X|S) - \hat{f}(X|<3 \text{ balls } \& <2 \text{ strikes}) \), for situation \( S \) listed in figure titles. The change in pitch density when the count has (a) three balls and fewer than two strikes, and (b) two strikes and fewer than three balls. The base case comprises pitches in counts with fewer than three balls and fewer than two strikes.

\( \begin{align*}
(a) & \quad 3 \text{ balls, } <2 \text{ strikes} \\
(b) & \quad 2 \text{ strikes, } <3 \text{ balls}
\end{align*} \)

Indeed, pitchers do throw more strikes in three-ball counts, and fewer strikes in two-strike counts. But these deviations are limited to the center of the official strike zone, where the call is obvious. As Figure 13 shows, pitches on the edge of the official strike zone—where the biases are largest in Figure 2—are thrown just as frequently in pivotal counts as in non-pivotal counts. Umpires may expect more strikes in three-ball counts and fewer strikes in two-strike counts, but they can rationally expect those deviations only where strikes are obvious. Where the correct call is uncertain—i.e. where umpires display the greatest bias—
pitcher tendencies do not inform umpires’ rational expectations about the forthcoming pitch.

Rational expectations may also be informed by whether the batter swings. Specifically, a batter’s decision not to swing may signal to the umpire that the pitch is a ball. Our results cannot be explained by swing signaling directly because umpires only make calls when the batter does not swing; the enforced strike zone varies, but the signal does not. Still, the rate at which batters swing in certain states may inform the umpire of the likelihood of a strike in those states. If in asymmetrically strike-pivotal states, batters swing more often, then the decision not to swing may signal that the pitch is a ball. However, the argument is uni-directional: choosing not to swing can only signal that the pitch is a ball, but in asymmetrically ball-pivotal states, we find that umpires are more likely to call strikes. Swing rates cannot explain the expansion of the strike zone when a ball would be pivotal. As with pitch location, swing rates cannot fully account for impact aversion.
References


Lithwick, Dahlia (2014) “This Court Erred,” *Slate*, URL: http://slate.me/1rZ22iW.


Moskowitz, Tobias and L Jon Wertheim (2011) *Scorecasting: The hidden influences behind how sports are played and games are won*: Random House LLC.


