Title: Accusations of Unfairness Bias Subsequent Decisions: A Study of Major League Umpires

Authors: Travis J. Carter[1]*, Erik G. Helzer[2].

Affiliations:

[1]Department of Psychology, Colby College.

[2]The Johns Hopkins Carey Business School.

*Correspondence to: travis.carter@colby.edu

Abstract: What happens when decision-makers are accused of bias by an aggrieved party? We examined the ball-and-strike calls of Major League Baseball umpires before and after arguments from players or managers resulting in ejection. Prior to ejection, the accusing team was, in fact, disadvantaged by the home plate umpire's calls. After the ejection, umpires did not revert to neutrality—they exhibited the *opposite* bias, advantaging the accusing team. This pattern was only evident when the ejection was related to pitch location, not other kinds of ejections. Using a laboratory analogue of the umpires' situation, we replicated this post-accusation tendency with experimental participants. This study further revealed that decision-makers were unaware of the shifts in their behavior in response to the accusations, and another survey indicated that this tendency violates beliefs about fairness. These results suggest that performance following accusations may unwittingly succumb to this insidious tendency to favor the accusing party.

One Sentence Summary: After being (rightly) accused of biased behavior toward one team, MLB Umpires responded by committing the opposite bias, now giving more favorable calls to the accuser's team.

MANUSCRIPT UNDER REVIEW—PLEASE DO NOT CITE WITHOUT PERMISSION

Main Text:

Among the many responsibilities leaders bear is a commitment to fairness. Perceptions of fairness are important for many organizational and interpersonal outcomes (*1–3*), and leaders, as decision-makers, find themselves in the unique position to uphold fairness standards. Even the best-intentioned leaders, however, will occasionally have their decisions questioned on the grounds of fairness. In this paper, we ask how accusations of bias (one type of fairness violation) affect the subsequent judgments of decision-makers.

Herein, we examine the perceived fairness of repeated judgments. In such cases, a decision-maker is expected to base her decisions on an evaluation of the evidence, applying some pre-ordained standard consistently to each case. Conducted correctly, this procedure should result in fair outcomes on average. Common examples include judges' rulings for courtroom objections, managers' application of company standards to different job candidates, and (most relevant to the present studies) baseball umpires' application of a common strike zone to batters on both teams.

Systematic bias in serial decisions may be particularly insidious: over time, small procedural biases can compound into large absolute differences in the distribution of outcomes (*4*). Although much research has examined the factors that influence judgments of fairness, as well as recipients' reactions to decisions that are judged as fair or unfair (*5*), little is known about the effects of accusations of unfairness on decision-makers' subsequent decisions.

In the present analysis, we view such accusations as a form of performance feedback: the decision-maker is made aware of a perceived pattern of uneven decisions, presumably indicative of a flawed (or biased) process. Specifically, we examined fairness-related feedback delivered to an evaluator by a self-interested party—someone directly (and negatively) affected by those decisions. For instance, an employee might complain that his performance reviews are unfairly lower than others', implying that the manager is exhibiting a bias against him.

In such cases, decision-makers are keenly aware of others' biases (*6*), and may be particularly apt to dismiss an accusation of bias as being self-interested, consequently making no attempts to investigate or alter subsequent decisions. This reaction may not be warranted, however, given that even self-interested assessments are constrained by reality (*7*). Even if the accusation is considered seriously, the decision-maker may search for evidence of bias and find none—not because it does not exist, but because the process by which the judgment is formed is impervious to introspection (*8*). All of this suggests that even a pattern of unfairness exists, decision-makers will remain unaware of the presence of bias.

We considered three major possibilities for how decision-makers would respond to an accusation from an interested party. First, it could have no systematic effect on subsequent decisions, either because of a successful attempt to ignore the feedback or because the decision-making process is immune to conscious intervention (*9*). Second, it could exacerbate existing biases, strengthening the existing response tendency, either due to motivational processes, such as reactance (*10*), or inherent cognitive biases, such as escalation of commitment (*11*). Third, it could lead to overcorrection, producing a new bias in the opposite direction, either resulting from overzealous conscious attempts to correct for past errors or from a non-conscious correction mechanism. In testing these possibilities, we sought to understand the role that conscious processes, as well as explicit beliefs about feedback accuracy, play in shaping decision-makers' responses.

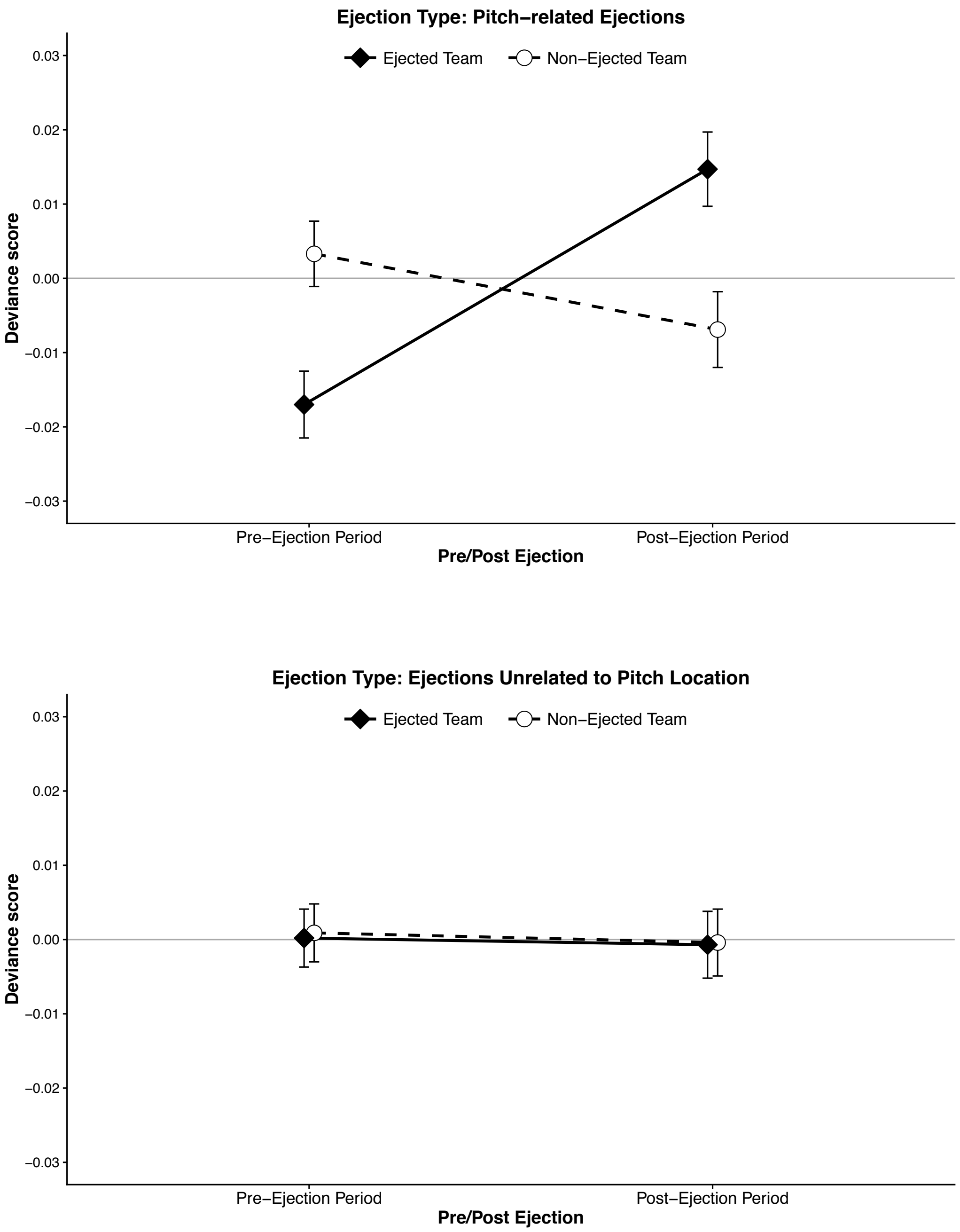
We began by examining a near-ideal decision context in Study 1: the ball-and-strike judgments of Major League Baseball (MLB) umpires. Hundreds of times in each game, the home plate umpire declares a pitch a ball or a strike based on their perception of whether it was inside or outside the strike zone, a decision that must be made immediately, and that carries opposite consequences for the two teams involved. Umpires are regularly accused of bias in their decisions, but we focused on the clearest and most discernible cases: when a player or manager is ejected from a game as a result of arguing a call with the umpire. Ultimately we sought to compare the relative favorability of the umpire's calls toward the ejected and the non-ejected team both before and after the ejection. We further expected that accusations of unfairness would exert the most direct influence on subsequent judgments in the same domain. In this case, only arguments related to pitch location (i.e. ball-and-strike calls) should lead to shifts in umpires' subsequent ball-and-strike calls; ejections prompted by other circumstances (e.g. a close play at third base) should lead to no such shifts, thus providing an important point of comparison, and a crucial benchmark and for our predictions.

In order to measure the relative favorability of umpires' ball-and-strike judgments, we employed a data-driven approach that allows for both absolute and relative comparisons, making use of the PITCHf/x pitch location data from 2008-2013. Specifically, we divided the *x*-*z* coordinate plane into bins, then calculated a *deviance* score for each called pitch by comparing the actual ball or strike call made by the umpire to the long-run probability of pitches in that same location being called a ball or a strike. This approach not only allows for the aggregation and comparison of pitches regardless of their location, it also minimizes the impact of shifts in players' behavior as a result of an ejection; as long as the umpire is the arbiter of judgment for a given pitch, the prior probability should serve as a neutral baseline against which to compare any individual judgment. The details of the calculation can be found in the Supplemental Materials, but put simply, positive deviance scores reflect calls favorable to the batting team, and negative scores reflect calls unfavorable to the batting team.

Examining deviance scores for pitches thrown during the pre- and post-ejection periods of games featuring a single ejection using linear mixed-effects models, it was clear that for ejections unrelated to pitch location (396 games, n = 42,414 pitches), the ejection had no impact on the relative favorability of the umpire's calls, b = 0.000, 95% CI [–0.012, 0.013], t(402.79) = 0.06, p = .954 (Fig. 1, bottom panel).

Pitch-related ejections (311 games, n = 34,563 pitches), however, did lead to different patterns of favorability before vs. after the ejection, b = 0.042, 95% CI [0.026, 0.057], t(314.61) = 5.37, p < .001 (Fig. 1, top panel). The accusation of bias was apparently made with good reason: prior to the ejection, the ejected team received *less* favorable calls than the non-ejected team, t(442.10) = 3.99, p_{hb} < .001.[1] In the post-ejection period, however, umpires reversed this bias; the ejected team received *more* favorable calls than the non-ejected team, t(744.20) = –3.52, p_{hb} < .001. Further examinations indicated that this reversal remained for the rest of the game, rather than fading shortly after the ejection (see Supplemental Materials). Thus, the data are consistent with the third possibility outlined above: in response to an accusation of unfairness, umpires *overcorrected*, introducing the opposite bias—but only when the argument was relevant to the domain of judgment.

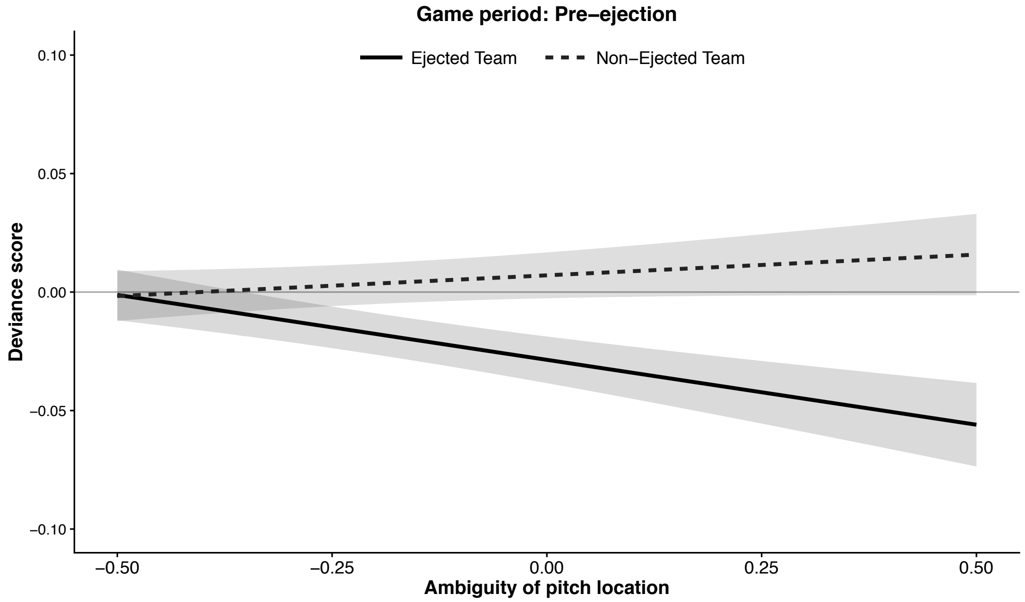
[1] The subscript *hb* (i.e. P_{hb}) indicates a p value that was adjusted using the Holm-Bonferroni method (*12*) in order to account for multiple comparisons.


Ejection Type: Pitch–related Ejections

Ejection Type: Ejections Unrelated to Pitch Location

Fig. 1. Deviance scores by batting team, period of game, and type of ejection (Study 1). Top panel: Ejections resulting from arguments related to pitch-location (396 games, *n* = 42,414 pitches). Bottom panel: Ejections unrelated to pitch location (311 games, *n* = 34,563 pitches). Error bars represent +/– 1 standard error.

Given that umpires' primary goal is accurate judgments, it seems likely that the bias would be most pronounced when there is some ambiguity about whether a given pitch should be called a ball or strike—pitches near the edge of the strike zone. To test this hypothesis, we created a proxy variable for ambiguity based on each pitch's prior probability of being called a ball or strike. Indeed, the bias exhibited by umpires was moderated by the ambiguity of the pitch location, *b* = 0.134, 95% CI [0.086, 0.181], *t*(34429.69) = 5.49, *p* < .001 (see Table S5). The bias against the ejected team pre-ejection (Fig. 2, top panel), and in favor of the ejected team post-ejection (Fig. 2, bottom panel), was strongest for the most ambiguous pitches.



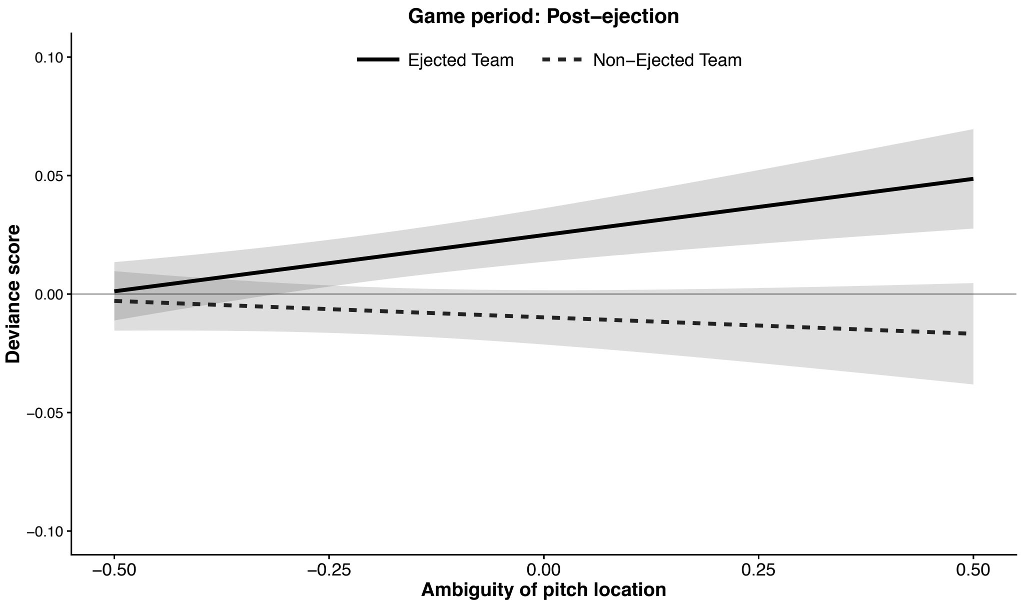


Fig. 2. Deviance scores by batting team, ambiguity of pitch location, and period of game (Study 1). Top panel: Pre-ejection period. Bottom panel: Post-ejection period. Depicted scores and 95% confidence intervals (shaded regions) derived from model predictions.

Given that umpires have the most control over the ball and strike calls, we consider the analysis of deviance scores to be the primary test of our hypothesis. It is, of course, interesting to consider whether the observed changes in umpires' post-ejection judgments had downstream consequences, such as changing the likelihood of batters getting on base (as measured by on-base percentage; OBP) and scoring runs. However, because these outcomes are less directly under the influence of the home plate umpire, we would expect to observe smaller effects compared to deviance scores.

Using the at-bat as the unit of analysis, we tested for the effects of batting team and period of game on the likelihood of getting on base (OBP; Fig. S4) and runs scored per at-bat (R/AB; Fig. S5) using linear mixed-effects models. In both cases, the ejected team experienced poorer outcomes than the non-ejected team in the pre-ejection period (OBP: $z = 5.44$, $p_{hb} < .001$;

R/AB: $z = 6.31$, $p_{hb} < .001$), just like deviance scores. However, unlike deviance scores, which showed a reversal in fortunes in the post-ejection period, the ejected team's disadvantage relative to the non-ejected team was merely eliminated for OBP ($z = 0.21$, $p_{hb} = .831$), and merely attenuated for R/AB ($z = 2.31$, $p_{hb} = .021$). Baseball games are certainly dynamic systems in that players and managers act and react as circumstances change, but the lack of a reversal in the post-ejection period for these two measures cannot be written off as simple regression to the mean. Indeed, a mediation analysis indicated that the improvements in the ejected teams' OBP (indirect effect: 0.039, 95% CI [0.022, 0.058]) and R/AB indirect effect: 0.016, 95% CI [0.008, 0.027]) following pitch-related ejections were at least partially due to changes in the umpires' balls-and-strikes calls after the ejection, even if those improvements were insufficient to overtake the non-ejected team.

In order to examine the effects of fairness accusations in a setting that allows for causal inferences, in Study 2 we created a laboratory task mimicking the situation that umpires face, with 100 participants randomly assigned to receive accusatory feedback (or not). The task involved viewing a series of images and judging whether the number of dots on each image was higher or lower than a target number (see Fig. S6). Participants were all told they had been assigned to the role of Judge, and would actually perform the dot-estimation task. Ostensibly, they had been partnered with another participant (the Observer) whose job was to observe the Judge's performance and provide feedback after each block of trials. Both the Judge and the Observer were paid a bonus based on the Judge's performance, but with misaligned incentives. That is, the Judge (participant) was paid based on the number of accurate responses she gave, whereas the Observer (partner) was paid based on the number of *directional* responses (i.e. the number of "higher" vs. "lower" responses, counterbalanced) given by the Judge, regardless of accuracy. Thus, any feedback by the Observer suggesting more directional responses could be seen as purely self-serving—to the detriment of the participant's own interests.

After a period of relatively neutral "feedback" from the Observer, participants in the Critical Feedback condition began receiving feedback accusing them of giving too few directional responses—easily interpreted as an attempt to garner a more favorable outcome for themselves. Participants in the Control condition received neutral feedback throughout the experiment.

Mimicking the umpires' response to an accusation of bias, the feedback manipulation impacted the proportion of participants' directional responses, $F(1,91) = 4.09$, $p = .046$, $\eta_p^2 = .043$ (see Fig. 3, top panel). Participants in the critical feedback condition shifted their judgments to be more favorable to the observer after the critical feedback began, $t(91) = –2.26$, $p_{hb} = .052$, whereas participants in the control condition, showed no systematic change, $t(91) = 0.66$, $p_{hb} = .513$. What's more, examining participants' explicit estimates after each block of trials showed no evidence that they were aware of the shifts in their directional responses (all *F*s < 1.05; see Fig. 3, bottom panel) or in accuracy (see Supplemental Materials). Those findings, coupled with participants' strong belief that the Observer's judgment was both biased ($p < .001$) and inferior to their own ($p < .001$), strongly suggest that the shifts were not intentional (see Supplemental Materials).

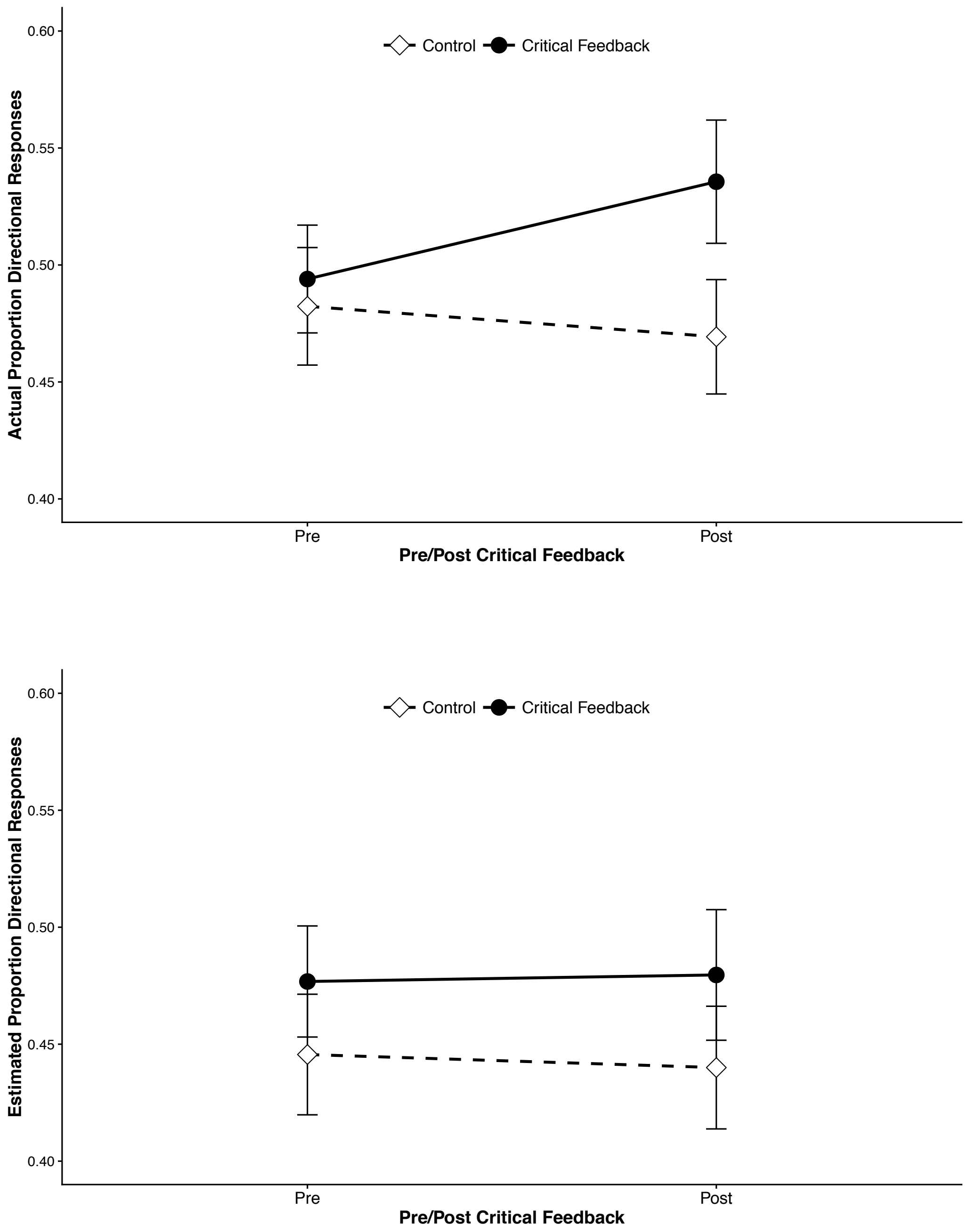


Fig. 3. Actual (top panel) and Estimated (bottom panel) proportion of directional responses by period of the study and feedback condition (*N* = 93; see Materials and Methods). Error bars represent +/– 1 standard error.

The tendency to respond to accusations of unfairness by offsetting one bias with another poses interesting questions about the nature of fairness. To see how the pattern of behavior we observed comports with folk intuitions about fairness we presented another group of participants with a scenario where there was a clear pattern of bias in a decision-maker's past decisions, and asked them what would constitute a fair outcome going forward. The vast majority (73%) indicated that the fairest response would be to *eliminate* bias toward both parties, and only 19% of indicated that the pattern we observed in both studies (instituting a new bias in favor of the aggrieved party) would be fair. Thus, there may be important implications for this work with regard to perceptions of procedural fairness following an accusation of bias.

To be sure, the precise mechanisms underlying changes in cognitive and behavioral responses following accusations of bias are not easily identified by the present studies. As such, it is unclear how decision-makers might avoid the negative consequences of feedback. Our experimental participants appeared to have minimal access to the effects feedback had on their judgments, suggesting that these biases may be particularly insidious and resistant to conscious control (*8*). Objective performance feedback following accusations of bias (such as a computer-generated report on umpires' accuracy calling balls and strikes), followed by recalibration and practice may help reduce these biases in the long-run (*13*); however, it is unclear how feasible such interventions would be in the context of real-world serial decisions.

All in all, the present studies provide some insight into an overlooked aspect of the psychology of fairness. Despite their best intentions, decision-makers charged with upholding fairness may from time to time slip in their duties. These studies suggest that the most obvious resolution to this problem—making decision-makers aware of such slips through informal channels—may promote new patterns of unfairness instead of eliminating the underlying problem. This may be very welcome news to the aggrieved party, though certainly not to the decision-maker who must hear a new round of complaints.

References and Notes:

1. Y. Cohen-Charash, P. E. Spector, The role of justice in organizations: A meta-analysis. *Organ. Behav. Hum. Decis. Process.* **86**, 278–321 (2001).

2. B. A. Mellers, J. Baron, Eds., *Psychological perspectives on justice: Theory and applications* (Cambridge University Press, Cambridge, 1993; http://ebooks.cambridge.org/ref/id/CBO9780511552069).

3. D. T. Miller, Disrespect and the experience of injustice. *Annu. Rev. Psychol.* **52**, 527–553 (2001).

4. L. Babcock, G. Loewenstein, Explaining bargaining impasse: The role of self-serving biases. *J. Econ. Perspect.* **11**, 109–126 (1997).

5. T. R. Tyler, What is procedural justice?: Criteria used by citizens to assess the fairness of legal procedures. *Law Soc. Rev.* **22**, 103 (1988).

6. E. Pronin, D. Y. Lin, L. Ross, The bias blind spot: Perceptions of bias in self versus others. *Pers. Soc. Psychol. Bull.* **28**, 369–381 (2002).

7. Z. Kunda, The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498 (1990).

8. T. D. Wilson, N. Brekke, Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychol. Bull.* **116**, 117–142 (1994).

9. R. E. Nisbett, T. D. Wilson, Telling more than we can know: Verbal reports on mental processes. *Psychol. Rev.* **84**, 231–259 (1977).

10. J. Brehm, *A theory of psychological reactance.* (Academic Press, Oxford, England, 1966).

11. B. M. Staw, Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organ. Behav. Hum. Perform.* **16**, 27–44 (1976).

12. S. Holm, A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).

13. Y.-W. Chien, D. T. Wegener, R. E. Petty, C.-C. Hsiao, The flexible correction model: Bias correction guided by naïve theories of bias: theory-based bias correction. *Soc. Personal. Psychol. Compass*. **8**, 275–286 (2014).

14. Major League Baseball (Organization), *The official rules of Major League Baseball.* (Triumph Books, Chicago, 2014).

15. B. M. Mills, Technological innovations in monitoring and evaluation: Evidence of performance impacts among Major League Baseball umpires (2015), (available at http://www.brianmmills.com/uploads/2/3/9/3/23936510/full_revised_manuscript.pdf).

16. R Development Core Team, *R: A language and environment for statistical computing* (2016; https://www.r-project.org/index.html).

17. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).

18. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, *lmerTest: Tests in linear mixed effects models* (2016; https://CRAN.R-project.org/package=lmerTest).

19. D. J. Barr, R. Levy, C. Scheepers, H. J. Tily, Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**, 255–278 (2013).

20. T. J. Moskowitz, L. J. Wertheim, *Scorecasting: The hidden influences behind how sports are played and games are won* (Three Rivers Press, New York, First paperback edítion., 2011).

21. R. P. Larrick, T. A. Timmerman, A. M. Carton, J. Abrevaya, Temper, temperature, and temptation: Heat-related retaliation in baseball. *Psychol. Sci.* **22**, 423–428 (2011).

Acknowledgments:

The authors declare no conflicts of interest. The data used in all studies will be made available via the Open Science Framework (OSF). The complete PITCHf/x data are available from the MLB website (http://mlb.mlb.com/gdcross/components/game/mlb/). The authors wish to thank Devin Pope for his advice on the calculation of the deviance measure.

Supplementary Materials for

Accusations of Unfairness Bias Subsequent Decisions: A Study of Major League Umpires

Travis J. Carter, Erik G. Helzer

correspondence to: travis.carter@colby.edu

Materials and Methods (Study 1)

Data Sources and Experimental Design

Data pertaining to pitch location and game conditions for every regular season Major League Baseball game from 2008-2013 were drawn from the several sources. The Major League Baseball website makes available data on the location (*x* and *z* coordinates) of each pitch (PITCHf/x), game event information (runners on base, steals, and ejections at each point in each game), and game personnel (batters, pitchers, home plate umpires). We obtained data on Win Expectancy (WE; the calculated probability of a team winning the game based upon the score, inning, number of outs, runners on base, and game environment) and Leverage Index (LI; an index of the amount of pressure in a game based upon its current conditions, such as score and inning) from Fangraphs.com. Attendance and temperature information for each game was obtained from Retrosheet.com. Many of the analyses that control for these secondary game-level variables can be found in the supplemental materials.

Critically, a list of ejections was populated from these data, including all relevant information about who was ejected by whom and at what point in the game the ejection occurred. This list was then compared against the data and descriptions found on the Umpire Ejection Fantasy League (UEFL) website (http://portal.closecallsports.com), which catalogues ejections, to create a list of 1,126 ejections occurring across regular season games between 2008 and 2013, ranging from 164 (2009) to 207 (2008) per year.

Ultimately, this created a 2 (Ejection Type: Pitch-related vs. Other) × 2 (Batting Team: Ejected vs. Non-ejected) × 2 (Period of Game: pre- vs. post-ejection) with the latter two factors varying within each game, and the first factor varying between games. The primary dependent variable is the pitch deviance measure (described in detail below), which occurred at the level of the individual pitch. We also examined two major offensive outcomes: on-base percentage (OBP) and runs scored per at-bat (R/AB), which occurred at the level of the at-bat.

Calculation of Pitch Deviance Measure:

The rulebook strike zone is defined as a box with horizontal limits (on the *x*-axis) that range from one edge of home plate to the other, with vertical (on the *z*-axis) limits defined based on the batter: the upper limit is defined as “a horizontal line at the midpoint between the top of the shoulders and the top of the uniform pants,” and the lower limit is defined as “a line at the hollow beneath the kneecap” (*14*). In practice, however, umpires call pitches according to a strike zone that differs from the rulebook strike zone in systematic ways, ignoring the corners and making adjustments for right- and left-handed batters, for instance (*15*). Thus, using the rulebook strike zone is clearly inadequate as a neutral and valid reference point to assess umpires’ judgment. Although it might seem straightforward to simply use the “practical strike zone” instead, there is simply too much ambiguity in the data to identify hard boundaries between strike and ball. Additionally, it is important to account for the possibility that, if umpires are altering their practical strike zone based on arguments leading to ejections, then batters and pitchers might similarly be altering their behavior to adapt to that new strike zone. For instance, if a batter argues after being called out on a third strike that was well above the strike zone, a pitcher may attempt to throw more pitches in the same location expecting the same result, and batters may feel it necessary to swing at those pitches to avoid a similar called-strikeout. Although it may be impossible to completely account for these changes in players' behavior, we employed a data-driven approach that should minimize its influence while simultaneously defining the practical strike zone in a much more fluid way: calculating the probability that a pitch in a given location would be called a ball or a strike based on what call umpires typically make for a pitch in that specific location. Thus, regardless of any shifts in behavior on the part of batters or pitchers as a result of an ejection or perceptions of an umpire's shifting strike zone, as long as the umpire is the arbiter of judgment for a given pitch, the prior probability should serve as a neutral baseline against which to compare any individual judgment.

In order to group pitches based on their location, the *x*-*z* coordinate plane was divided into a grid of bins. To calculate the bin size, we used the size of a baseball as a guide. The size of an official MLB baseball is defined as measuring "not less than nine nor more than 9 1/4 inches in circumference" (*14*). With the circumference defined by the rulebook as a range, we used the midpoint of that range (9.00-9.25 in., 22.86-23.50 cm) as the value for the circumference, 9.125 in. (23.178 cm). Using basic geometry to find the diameter from the circumference, $9.125/\pi = 2.905$ in. (7.379 cm), we calculated bins based on one-quarter (0.726 in., 1.844 cm) the diameter of a baseball. This size was intended to reflect a balance between the precision of the location and the precision of the probability. Smaller bins provide more meaningful discrimination based on location, but larger bins provide more confident estimates of the prior probability by ensuring a sufficient sample size of pitches located within the bin. It is worth noting, however, that the analyses were robust to different bin sizes. The results were the same using both larger (one-half baseball diameter; 1.452 in., 3.689 cm) or smaller (one-eighth baseball diameter; 0.363 in., 0.922 cm) bins.

Translating the PITCHf/x data in its raw form onto a grid of bins of a fixed size required further calculations. To understand how we accomplished this, it's helpful to know a bit more about how the raw data reflect the rulebook strike zone (defined above). Having a fixed definition for the horizontal dimension (the width of home plate) allows for all bins to have an equal width. The PITCHf/x data scale the *x*-axis relative to the strike zone's horizontal boundaries (-1 and +1 correspond to the left and right edges of the strike zone). In order to determine the horizontal boundaries of individual bins based on an absolute size (fractional diameter of a baseball, as described above), we translated this into an absolute width. Because a pitch is still considered a strike if only part of the baseball crosses the plate, we calculated this as the width of home plate (17.00 in., 43.18 cm) plus the full width of one baseball (2.905 in., 7.379 cm)—allowing for pitches where the middle of the ball touches any part of the plate to be within the horizontal boundaries of the strike zone—for a total of 19.905 in. (50.559 cm). Thus, knowing that 2.0 units on the relative scale (the range of –1 to +1) corresponds to 19.905 in. (50.559 cm) on the absolute scale, allows for an easy translation of the absolute bin widths to the relative scale.

This same basic approach was applied to the *z*-axis, but because the vertical dimension of the strike zone is defined based on each player's height and stance, which can even vary from pitch to pitch, it was problematic both practically and theoretically to keep a constant bin height.

The most reasonable solution, in our minds, was to ensure that *x*-*z* coordinate plane would be divided into the same number of bins for each batter, even if that meant the absolute bin height would vary from pitch to pitch. Fortunately, the PITCHf/x system identifies and reports the absolute height of the strike zone on each pitch, which can be used to normalize the *z*-axis to be on the same relative scale as the *x*-axis, with the vertical boundaries of the strike zone defined as –1 and +1. We used the average absolute height of the strike zone across the entire data set (21.761 in., 55.272 cm) to translate the intended absolute bin heights to the relative scale.

Having defined the bin boundaries, we identified which bin each pitch in the entire data set of 2,212,150 called pitches (2008-2013) fell into. Next, we calculated the percentage of pitches located in each bin that were called *balls*, effectively identifying the probability that a pitch in that location would be called a ball. Each bin's probability ranged from 0 (100% strikes) to 1 (100% balls; see Fig. S1, which depicts a heat-map of these prior probabilities). Note that the probabilities were calculated separately for left and right-handed batters, given known differences in the practical strike zones based on handedness (*15*). Thus, for any given pitch, the actual call made by the umpire can be compared to the long-run probability of pitches in that same location being called a ball or a strike. This allows us to calculate the degree to which a call deviated from that probability by subtracting the probability of a called ball or strike from the actual call (called ball = 1; called strike = 0). For instance, a called ball located in a bin where 80% of the called pitches were called balls (bin-probability = .80) would have a deviance value of 0.20 (1.0 – 0.80 = 0.20). A called strike located in that same bin would have a deviance value of –0.80. Thus, the range of possible deviance scores for a pitch goes from –0.999 (an extremely unlikely called strike) to +0.999 (an extremely unlikely called ball). Put more simply, positive deviance values reflect calls that were favorable to the batting team, and negative deviance values reflect calls that were unfavorable to the batting team, relative to what would be expected.

Across the entire sample of pitches, the deviance scores form a distribution that is, by definition, centered on zero, meaning that zero is the expected deviance value for any given pitch or collection of pitches, and provides an absolute baseline to test for evidence of bias. That is, if the average deviance score for a given collection of pitches is significantly different from zero, that would be evidence of systematic bias in favor of one team, with the magnitude and valence of the average indicating the amount and direction of bias, respectively. For instance, if the average deviance for pitches thrown to the visiting team over the course of a single game was +.03, that would roughly translate to ball-and-strike calls being an average of 3% more favorable than expected.

Having an absolute comparison point is helpful, but relative comparisons are the most relevant for the present analyses—particularly if a given umpire has a slightly more expansive (or constrictive) strike zone than the league average. For instance, imagine a situation where the visiting and home teams had average deviance scores of +.03 and +.06, respectively, for a single game. The fact that both scores are positive would indicate that the umpire generally employed a slightly smaller strike zone (more called balls) than the league average, but the home team's relatively larger score indicates that it received more favorable calls than the visitors, on average. Thus, regardless of how a given umpire's strike zone compares to the league as a whole, comparing the deviation scores of opposing teams can reveal which team, if any, was the recipient of undue generosity. (We also deal with umpire- and game-level variation statistically.)

Thus, examining these deviance values in the aggregate allows us to detect systematic shifts in the umpires' calls before and after an ejection, and whether they favor one team over another. It is worth noting that, because umpires are generally quite accurate in their calls—as depicted by the relatively small band of ambiguity in Fig. S1—the vast majority of pitches show a fairly small amount of deviation from their prior probability. Thus, examining aggregated outcomes represents a very conservative test of a hypothesis of shifting favorability.

Calculation of Bin Ambiguity

For each pitch, the prior probability (P) of being called a ball ranges from 0 (all strikes) to 1 (all balls), with 0.5 representing equal likelihood of being called a ball or strike. Thus, pitches in bins with a prior probability closer to 0.5 would be considered more ambiguous. The ambiguity index was thus calculated using the following formula:

$$(|P - 0.5| * 2) + 0.5$$

Although this index would theoretically range from –0.5 (no ambiguity) to +0.5 (complete ambiguity—equal likelihood of called ball or strike), because pitches located in bins with zero variability were excluded from the analyses, the lower end of the range of possible ambiguity scores was actually –0.4967. It is worth noting that the prior probability for each pitch factors into the calculation of both the ambiguity index and the deviation measure, and thus factors into both sides of the regression equation. By its very definition, there are limits to the amount of variation that can be observed for very high and low probability values (i.e. the highly unambiguous pitches). Nonetheless, observing differences in the predictive power of that pitch's ambiguity based on whether the ejected or non-ejected team was batting and whether it occurred in the pre- or post-ejection period should be informative.

Calculation of Offensive Statistics

To calculate on-base percentage (OBP), each at-bat was categorized as resulting in the batter getting on base (coded as 1) or not (coded as 0). This allows us to examine OBP at the level of the at-bat—essentially the likelihood that batter reached base. According to standard scorekeeping procedure, some at-bats are not figured into the calculation of OBP, such as when the batter successfully executes a sacrifice bunt, or is given a base due to catcher's interference. These instances were coded as missing, and were thus not included in the analysis.

We also examined the number of runs scored per at-bat (R/AB), which is simply the number of runs scored by the batting team during a given at-bat. Note that we counted runs scored for any reason, not just runs resulting from the batter getting a hit, or even a sacrifice (bunt or fly). A run scored from a player stealing home on a passed ball, or from the umpire issuing a walk with the bases loaded, were also counted.

Coding Ejection Type

Although baseball players and managers frequently express their displeasure with unfavorable calls, it is not obvious exactly how one should quantify such an expression, nor how to identify those instances where the umpire is specifically accused of being biased. We believe that the least ambiguous examples of such expressions are cases when a player or manager argues a call with the umpire, and is subsequently ejected from the game. Although it is not possible to know the exact content of the arguments that lead to ejections, the authors' collective lifetime of experience watching baseball[2] certainly suggests that the arguments typically involve the player or manager accusing the umpire of a pattern of unfavorable (biased) calls, with the latest pitch being only the most recent example.

In order to identify the underlying reason for each ejection, which is not included in the data provided by MLB, an independent coder categorized each ejection based on the UEFL descriptions of the game context in which the ejection was made. Each ejection was coded as

[2] The Atlanta Braves and Seattle Mariners, respectively.

resulting from an argument that was either related or unrelated to pitch location. The coder was not blind to the hypotheses, but because the descriptions of the ejection event were completely separate from the PITCHf/x data—meaning that the coder had no knowledge of the location of the antecedent or subsequent pitches when making the determination—there was virtually no risk of this knowledge biasing the coder's judgments. The vast majority of the ejections were unambiguous, such as a batter arguing that a called third strike should have been called a ball (pitch-related ejection), or the manager arguing that a runner called safe at first base should have been called out (other ejection). Some cases, however, introduced some difficulty, such as an argument about a batter who begins to swing and then attempts to hold back the swing (a "check swing"). In this case, if the umpire rules the ambiguous motion as a swing, then it is a strike regardless of the location. If instead it is ruled as a non-swing, then the umpire must judge it a ball or a strike based on the pitch location—meaning that the pitching team could argue about the lack of a check swing call, and either team could argue about a called ball or strike. In these cases, the coder attempted to discern, based on the context of the ejection and the description of the ejection, whether the true crux of the argument was about the location of the pitch or about something else. Any ambiguity was resolved through discussion with the authors prior to examining the actual pitch data.

Selection of Games and Pitches:

Of the 1,126 ejections, 489 (43.4%) were coded as being related to pitch-location, and 637 (56.6%) were coded as "other." We excluded from the analysis any game with more than one ejection event, or when members of both teams were ejected, so that it would be clear which team suffered the ejection, and exactly when it occurred. That is, if several members of the same team were ejected as a result of the same at-bat (e.g. the umpire ejects both the batter and manager of the batting team for arguing the same called third strike), this would be considered a single ejection event. If two members of the same team were ejected after different at-bats (e.g. a batter was ejected immediately after arguing a called third strike, but the manager was ejected after arguing about a call several at-bats later), then it would be considered two ejection events. Of the 815 (72.4%) games that met the definition of a single ejection event, we further excluded games where either team did not have at least one at-bat with a called pitch before and after the ejection, leaving a total of 707 games in the final data set (n = 311 involving pitch-related ejections, n = 396 involving other kinds of ejections). It is worth noting that the results do not change if the full set of 815 games is included in the analyses.

From that set of games, there were 110,806 called pitches with valid x and z coordinates.[3] To ensure the robustness of the analysis, we excluded pitches falling in bins that contained too few pitches (fewer than 100) to be confident that the prior probability was reasonably accurate (15,398 pitches). These pitches were virtually all called balls (99.94%), indicating that they were well outside the strike zone. We also excluded pitches from the analysis where there was zero variability in the bin (100% called balls or 100% called strikes; 17,195 pitches), the vast majority of which (95.00%) were called balls. After these exclusions, the final data set we examined consisted of 79,220 pitches, which we can be certain required at least some interpretation on the part of the umpire, defined quite conservatively.

Statistical Analysis:

[3] Some data were missing due to a combination of the occasional technical error on the part of the PITCHf/x system, and because the system was not yet fully deployed in every stadium until mid-way through the 2008 season.

Examining the effects of these independent variables involves comparing the outcomes of opposing teams within the same game, with many of the outcomes determined by the umpire behind home plate, who also officiated other games in the data set. By using outcomes between opposing teams within the same game, any fixed bias in outcomes for that particular game or that particular umpire should not be problematic (such as a given umpire's tendency to have a larger or smaller strike zone than the league average). However, because observations from the same game came from the same umpire, they would violate the assumption of independence underlying standard linear models or ANOVA. Thus, we employed linear mixed-effects models for all analyses for Study 1, treating the three main independent variables as fixed effects.

To aid the interpretation of model parameters, we used contrast coding rather than dummy coding for all categorical variables (i.e. Pitch-related ejections: +0.5, Other types of ejections: –0.5; Ejected team: +0.5, Non-ejected team: –0.5; Post-ejection period: +0.5; Pre-ejection period: –0.5).

All analyses were conducted in R (*16*), using the lme4 package (*17*) for the linear mixed-effects models and confidence intervals, and the lmerTest package (*18*) to calculate *p* values (using Satterthwaite's approximations for the degrees of freedom), estimate cell means, and to perform any post hoc or pairwise comparisons. For any post hoc or pairwise comparisons, *p* values were corrected for multiple comparisons using the Holm-Bonferroni procedure (*12*); corrected *p* values are indicated with a subscript (i.e. p_{hb}).

For the random-effects structure, we began with a maximal model (*19*), which we identified as having random intercepts for home plate umpire and for game (nested within umpire), allowing for the effects of Batting Team, Period of Game, and their interaction to vary within individual games (i.e. random slopes). Although the maximal model also allowed for correlated slopes and intercepts, the model would not reliably converge unless those were dropped from the model. It is worth noting that in the few models with correlated slopes and intercepts that did converge, the model fit was no better than a model without correlated slopes and intercepts, as evidenced by a likelihood ratio test, χ^2 (6) = 6.86, *p* = .334. The final model we employed should thus account for any non-independence of the individual observations, while also allowing us to examine or control for the impact of pitch, at-bat, and game-level variables (e.g. whether there was an impact of game attendance on deviance scores).

The structure defined above was used for all analyses of the pitch-level deviance scores. Because the two at-bat-level measures (OBP and R/AB) involved non-normal data, the analyses for those measures involved generalized linear mixed-effects models, specifically a mixed-effects binary logistic regression for OBP, and mixed-effects Poisson regression for R/AB. In both cases, even the slightly reduced version of the maximal model failed to converge, so the complexity of the random-effects structure was selectively reduced until convergence could be achieved. This resulted in a model with random intercepts for game, and random slopes for Batting Team, Period of Game, and their interaction within games (without allowing for correlated slopes and intercepts). In other words, only the random intercepts for umpires were dropped from the model.

For the pitch-level deviance measure, we first tested a linear mixed-effects model featuring a 2 (Batting Team: Non-ejected team vs. Ejected team) × 2 (Game Period: Pre-ejection vs. Post-ejection) × 2 (Ejection Type: Pitch-related vs. Other kinds of ejections) design for the fixed effects. The results of this model are presented in Table S1. Based on the significant three-way interaction, we conducted a linear mixed-effects model testing the effects of Batting Team and Game Period separately for pitch-related ejections (see Table S2) and other kinds of ejections (see Table S3), which is what is reported in the main text. Based on the results of the deviance measure, for the at-bat-level measures, we only conducted the analyses on the subset of data with pitch-related ejections.

Mediation analysis. As reported in the main text, we conducted a mediation analysis to confirm that the observed relative improvements in at-bat-level offensive outcomes (OBP and R/AB) by the ejected team after the ejection do in fact result from shifts in the umpires' behavior (as measured by the deviance scores), rather than being solely due to regression to the mean. That is, the observed pattern of effects for OBP and R/AB—a mere attenuation of the pre-ejection bias—could be explained by regression to the mean. The pitch-level deviance scores, however, show a *reversal* (not attenuation) of the pre-ejection bias in the post-ejection period, so a regression-to-the-mean explanation does not apply. Thus, if we can demonstrate that the effect of the ejection on the at-bat-level variables was statistically mediated by deviance scores, then we can be reasonably sure that those effects were not merely a regression to the mean. In order to ensure that the mediator (deviance scores) and the dependent variables (OBP and R/AB) were all operating at the level of the at-bat, we calculated at-bat-level deviance scores as the mean deviance score for each at-bat. For this analysis, we treated the batting team × period of game interaction as the independent variable. Consistent with the pitch-level analysis, there was a significant effect of the independent variable (batting team × period of game) on the mediator (at-bat-level deviance scores), $b = 0.040$, 95% CI [0.023, 0.058], $t(328.20) = 4.52$, $p < .001$. When the mediator was included in the same analyses described above predicting OBP (mixed-effects binary logistic regression) and R/AB (mixed-effects Poisson regression), it was a strong predictor in both cases (ps < .001). To test the significance of the indirect effect (the product of the effect of the independent variable on the mediator, and the effect of the mediator on the dependent variable, controlling for the independent variable), we calculated Monte Carlo confidence intervals with 50,000 repetitions (MacKinnon, Lockwood, & Williams, 2004; Preacher & Selig, 2012), which did not include zero in either case, as reported in the main text (OBP: 0.039, 95% CI [0.022, 0.058]; R/AB: 0.016, 95% CI [0.008, 0.027]).

Supplementary Text (Study 1)

Examining the At-Bat Prompting the Ejection.

For all analyses, we examined only outcomes occurring during the pre- and post-ejection periods, excluding the at-bat that prompted the ejection (hereafter referred to as the ejection at-bat). To confirm that the ejection at-bat featured particularly egregious calls from the perspective of the team that suffered the ejection, at least for pitch-related ejections, we examined the deviance scores of pitches thrown during those at-bats ($n = 1,236$ pitches). This would most clearly be evident in highly unlikely strike (vs. ball) calls when the ejected team was batting (vs. pitching) during the ejection at-bat. Indeed, there was a significant interaction between Ejection Type (Pitch-related vs. Other) and Ejected Team Role (Batting vs. Pitching) on the deviance measure, $b = 0.384$, 95% CI [0.298, 0.470], $t(901.90) = 8.74$, $p < .001$ (see Fig. S2). For pitch-related ejections, the average deviance score was considerably lower when the ejected team was batting ($M = –0.210$, 95% CI [–0.241, –0.179]) compared to when the ejected team was pitching ($M = 0.194$, 95% CI [0.145, 0.243]), $t(944.0) = –13.80$, $p_{hb} < .001$. For other types of ejections, there was no difference in deviation scores depending on whether the ejected team was batting ($M = –0.007$, 95% CI [–0.050, 0.036]) or pitching ($M = 0.013$, 95% CI [–0.035, 0.061]) at the time of the ejection, $t(1007.7) = –0.62$, $p_{hb} = .535$. As expected, the at-bat prior to a pitch-related ejection featured calls that were highly unfavorable for the team that was ultimately ejected (unlikely strike calls for the batting team, unlikely ball calls for the pitching team), suggesting that it was these egregious calls that ultimately prompted the ejection-inducing argument.

Control variables.

The interaction between batting team (ejected vs. non-ejected) and period of game (pre- vs. post-ejection) holds when controlling for characteristics of the pitch location (prior probability within the bin, number of pitches in the bin; $p < .0001$), characteristics of the current at-bat and game situation (current count of balls and strikes, number of outs, number of runners in scoring position; $p < .0001$), as well as other metrics of situational pressure and importance, such as Win Expectancy (WE; probability the batting team would win), Run Expectancy (RE; probability the batting team would score a run during this at-bat), and Leverage Index (LI; an index of the importance of the situation), $p < .0001$). Although the home plate umpire issued pitch-related ejections in all but two cases (99.4% of games), the interaction also held when limiting the analysis to ejections issued by the home plate umpire ($p < .0001$).

Considering a continuous measure of game period.

The models described above and reported in the main text treat Period of Game as a categorical variable (i.e. pre- vs. post-ejection), largely as a matter of simplicity. Such an approach implicitly assumes that any bias exhibited by the umpire (i.e. favoring one team over another) is constant within the pre- and post-ejection periods, but it's worth considering whether the reality is more complicated than can be accommodated by a dichotomous variable. For instance, it could be that the apparent reversal in bias in the post-ejection period is limited to a few at-bats immediately following the ejection—the umpire could issue a few make up calls to assuage the ejected team's anger before reverting to a more neutral (unbiased) baseline, or perhaps even revert back to the previous pattern of bias.

To test this possibility, we first created a continuous measure of game period by calculating the distance from the current outcome to the ejection event (hereafter referred to as AB distance) by subtracting the current at-bat from the ejection at-bat. For instance, if the ejection occurred during the 25th at-bat of the game, then pitches thrown during the 29th at-bat would all have a distance score of +4 (29 – 25), and all pitches thrown during the 12th at-bat would have a distance score of –17 (29 – 12). Thus, all events in the pre-ejection period have negative values, and all events in the post-ejection period have positive values.

First, in a model with batting team, AB distance, and their interaction predicting deviance scores (pitch-related ejections only), the batting team × AB distance interaction was significant, $b = 0.0005$, 95% CI [0.0002, 0.0007], $t(237.13) = 3.62$, $p < .001$, which is consistent with the categorical variable (see Fig. S3, top panel). However, because it's likely that any irregularities would be non-linear, we also considered a model testing linear, quadratic, and cubic versions of the AB distance measure (and each version's interaction with batting team). Intriguingly, all three of the interaction terms were significant (Linear: $p < .001$; Quadratic: $p = .033$; Cubic: $p = .012$; see Table S4). As can be seen in Fig. S3 (bottom panel), which depicts the predicted values from this model 50 at-bats before and after the ejection, there is some evidence of non-linearity in the relative favorability of the umpire's calls, but those seem to occur primarily at more extreme values of AB distance, which are also the least represented in the data (hence the increasingly large confidence intervals).

To take a slightly different approach, we examined the post-ejection period separately to see if the bias in favor of the ejected team (relative to the non-ejected team) remained constant as the distance from the ejection at-bat increased. Consistent with a constant effect, in a model with batting team, AB distance, and their interaction, there was the expected main effect of batting team, *b* = 0.021, 95% CI [0.033, 0.009], *t*(255.95) = 3.51, *p* < .001, but no interaction between batting team and AB distance, *b* = –0.0002, 95% CI [–0.0009, 0.0005], *t*(2482.39) = –0.50, *p* = .615. A similar model including linear, cubic, and quadratic terms for AB distance showed the same result: a main effect for batting team (*p* < .001), but not for any terms involving AB distance (all *p*s > .14). Based on these two results, it appears that the shift in bias after the ejection remains—and remains relatively constant—throughout the game, and that treating game period as a categorical variable is a valid approach.

Testing potential moderators.

We began by considering whether properties of the ejection itself may have moderated the observed reversal in bias exhibited by the umpires in games featuring pitch-related ejections. Although there is evidence that much of the proverbial home team advantage is due to more favorable calls from the umpires (*20*), there was no evidence that the shift in the umpires' favor after an ejection was solely granted to the home team. Indeed, the shift in favor of the ejected team was evident whether it was the home team or the away team that was ejected (both *p*s < .001). None of the other properties of the ejection we examined moderated the effect, such as whether the ejected team was batting or pitching at the time of the ejection, or whether the person ejected was a manager or a player. In each case, the interaction between batting team (Ejected vs. Non-Ejected) and period of game (Pre vs. Post Ejection) remained significant (all *p*s < .001), but there was no significant three-way interaction with the moderator (all *p*s > .17).

We also tested whether variables related to the game itself might have mattered, including the ambient temperature (see *21*), game duration (in minutes; log transformed due to right skew), and game attendance. Neither temperature nor game duration moderated the effect (both *p*s > .11), nor diminished it (interaction: all *p*s < .0001). Game attendance (centered on the grand mean, 31,311.74), however, did show some promise as a moderator. Although the interaction between batting team and period of game remained significant (*p* < .0001), there was also a three-way interaction with game attendance (*p* = .018). The interaction was such that the basic effect—bias against the ejected team prior to the ejection, and in favor of the ejected team after the ejection—was larger when attendance was high, and nearly absent when it was low. However, we hesitate to draw strong conclusions from this finding, as attendance is no doubt correlated with other relevant variables, such as the home team's current record, or the importance of the game.

We also tested whether variables related to the importance and impact of the situation surrounding the ejection at-bat. That is, it's possible that umpires may exhibit a stronger bias in favor of the ejected team when the ejection came at a particularly bad time for that team. Specifically, we tested the score differential prior to the ejection, and situational importance metrics associated with the ejection at-bat, including RE, RE24 (the change in RE as a result of the ejection at-bat), LI, WE, and WPA (Win Probability Added; the change in WE as a result of the ejection at-bat). In every case, the interaction between batting team (ejected vs. non-ejected) and period of game (pre- vs. post-ejection) remained significant (all *p*s < .0001), but there was no significant three-way interaction with the moderator (all *p*s > .34).

Finally, we tested whether the umpires' calls might be sensitive to the importance of the *current* situation (specifically the current at-bat's WE, LI, and RE), especially depending on which team is batting. For instance, the shift in bias from the pre- to post-ejection period might only occur for relatively unimportant situations, perhaps to avoid having an undue influence on the game in very important situations. However, there was no evidence of any sensitivity to context. For all three of the variables we tested, the interaction between batting team (ejected vs.

non-ejected) and period of game (pre- vs. post-ejection) remained significant (all *p*s < .0001), but the three-way interaction was not significant (all *p*s > .16).

Materials and Methods (Study 2)

Experimental Design

The study employed a 2 (Feedback: Control vs. Critical) × 2 (Period of Study: Pre- vs. Post-Critical Feedback) factorial design, with the first factor manipulated between-participants, and the last factor manipulated within-subjects.

Participants.

We recruited 100 participants (64 male, 36 female) from Amazon.com's Mechanical Turk to play a "visual estimation game" with incentives for accuracy. Data collection stopped upon reaching the target sample size ($N = 100$), and the data were not examined prior to that point.

We excluded participants who did not appear to make a reasonable effort at the task by setting minimum standards for accuracy (at least 55%) and variable responding (at most 90% of responses could be in the same direction). Based on these criteria, which were the only criteria we considered, seven participants were excluded from the analyses, though including them does not change the outcome of any analysis.

Dot-Estimation Task.

After consenting to participate, a game designed to test "perceptual acuity" was introduced to participants. The game required participants to make perceptual judgments similar to those made by umpires in Study 1. Participants completed 10 blocks of 10 trials. To begin, for each block, participants were assigned a target number that varied randomly between 12 and 20. Then, for each of the ten trials within the block, an array of dots was flashed on the screen and participants had to judge whether the number of dots in the array was higher or lower than the target number (indicating their response with a key press). The dot arrays were randomly generated for each trial such that the actual number of dots was within a certain range (between 5% and 25%) of the target number (but never equal to the target number), ensuring variability in difficulty. Because the number of dots was generated randomly, the number of trials in each block where the correct response was "higher" also varied (ranging from 0-10, but typically 3-7). This was made explicit to participants to ensure they were not deliberately giving an equal number of higher and lower responses.

On each trial, the target number was displayed on the screen for 600ms, followed by the dot image for 400ms. Participants were required to respond within 2 seconds, or they had to repeat the trial with a newly generated dot array. This was intended both to make the task somewhat difficult and to encourage snap judgments. Prior to starting the game, participants completed a practice block of easy trials, on which they were given feedback about their performance, to ensure they understood the game.

Over the 100 trials, participants averaged 75.91% accuracy, 95% CI [74.46%, 77.37%], which was significantly greater than chance, $t(92) = 35.40$, $p < .001$, but nowhere near a ceiling effect. Thus, the difficulty of the task was neither impossible nor trivial.

Partner Description and Incentive Structure.

Participants were told that they had been assigned the role of Judge, and were paired with a partner who had been assigned to the role of Observer, whose role was to watch the Judge's performance and provide feedback after each block of trials. In truth, the partner did not exist, and all feedback provided by the Observer was bogus.

The incentives for participants assigned to the role of Judge (i.e. all actual participants) were based on accuracy. In addition to their base pay, participants received an additional \$0.02 for each correct response. To penalize guessing, this bonus was only awarded for the number of correct responses above chance (50%). Thus, over the course of 100 trials, participants with perfect accuracy would earn a bonus of \$1.00 (\$0.02 for each of 50 correct responses above chance), and participants with a mere 60% accuracy would earn \$0.20 (\$0.02 for each of 10 correct responses above chance). Accuracy of 50% or less would earn no bonus. Participants were not given feedback about their performance, and therefore their bonus, until the very end of the experiment.

The Observer's monetary incentives were also explained to participants. Whereas the Judge (participant) was paid for accuracy, the Observer (partner) was paid based on the *direction* of the response (i.e. higher or lower) given by the Judge, regardless of its accuracy. For instance, for each "higher" response, the Observer's bonus would increase by \$0.01, and for each "lower" response, it would decrease by \$0.01. The particular response associated with a positive or negative outcome was counterbalanced. As in the main text, the response that yielded a higher monetary bonus for the Observer is referred to as the "directional" response. Thus, if the participant gave 64 directional responses out of 100 trials, the Observer would earn a bonus of \$0.14. With 50 directional responses or fewer, the Observer's bonus would be zero. This misaligned incentive structure was designed to make it clear that the Observer's feedback, particularly any instruction to give more directional responses, might be purely self-serving—to the detriment of the participant's own interests.

To ensure that participants clearly understood both their own and their partner's incentives, participants were required to pass a "quiz" about the incentive structure before the first block of trials.

Feedback Manipulation.

At the end of each block of trials, participants received some feedback ostensibly written by their partner. In the beginning, the feedback was relatively neutral (e.g. "man those dots move fast. Nice!") or served to remind the participant of the partner's incentives (e.g. "ur getting quick! Remember, higher is better! jk"). After block 5, the feedback diverged depending on condition. Participants in the critical feedback condition began receiving feedback critical of their performance, but always suggesting that the participant's errors were systematically biased in a direction that hurt the partner's bonus (e.g. "too many lows! I think you might have missed a couple of highs there! help us both out!"). This critical feedback continued and intensified (e.g. "what, do you have something against highs??" and "you're killing me here!") until the final block of trials. Participants in the control condition received feedback that did not point to a particular directional bias (e.g. "so many highs and lows! it's hard to keep up with all of them. sorry i can't be more help!" and "i'm getting tired! keep your head in the game, ur almost done"). The final feedback, after the last block of trials, was somewhat neutral and identical in both conditions. Thus, participants responded to 50 trials before and 50 trials after the critical feedback began, allowing us to compare responses not just between conditions, but also over time.

Explicit Beliefs and Manipulation Checks

After each block of trials, following the partner feedback, participants estimated the number of correct responses (from 0 to 10) they gave, as well as the number of higher/lower responses they gave in that block. These explicit estimates allowed us to ascertain whether participants were aware of any bias creeping into their responses.

After the final block of trials and the last round of feedback, participants evaluated both their own and their partner's overall ability at the dot estimation task (1 = Very Poor; 6 = Average; 11 = Very Good), and estimated the number of trials, out of 100, that they thought their partner would have answered correctly. These items were intended to rule out the possibility that participants began to doubt their own abilities in the face of critical feedback. To confirm that this is not a likely explanation for the results, we conducted a 2 (Feedback: Control vs. Critical) × 2 (Target: Self vs. Other) mixed-model ANOVA, with feedback as a between-participants variable, and target as a within-participants variable. As expected, there was a strong main effect of target, $F(1,91) = 29.18$, $p < .001$, $\eta_p^2 = .092$, indicating that participants clearly thought that they were more skilled than their partner at the game. More importantly, this main effect was qualified by a significant feedback × target interaction, $F(1,91) = 4.47$, $p = .037$, $\eta_p^2 = .047$ (see Fig. S7). Specifically, although participants in the control condition did indeed think more highly of their own abilities, $t(91) = 2.12$, $p_{hb} = .036$, this tendency was exaggerated in the critical feedback condition, $t(91) = 5.40$, $p_{hb} < .001$, suggesting that participants only became more sure of their own abilities (and more skeptical of their partner's abilities) as a result of the critical feedback. Furthermore, there was no difference in participants' estimates of how many trials their partner would have gotten correct (Critical feedback: $M = 61.92$, 95% CI [57.17, 66.67]; Control: $M = 60.00$, 95% CI [55.72, 64.28]), $t < 1$, *ns*. Thus, across multiple measures, we found no support for the idea that critical feedback led participants to doubt their abilities. In fact, just the opposite appeared to be true: as described above, participants gave higher estimates of their own abilities in the critical feedback condition compared to participants in the control condition, $t(91) = 2.86$, $p_{hb} = .010$.

During the last block of questions, participants also indicated the degree to which they thought their "partner was judging the dot images accurately, or had a biased perspective" (1 = Definitely biased to judge Lower; 6 = Partner's judgment was accurate; 11 = Definitely biased to judge Higher). This item, intended as a check on the incentive structure, was reverse-scored for participants whose partner's incentive was for "lower" responses, so that higher numbers always indicated greater bias. (The direction of the partner's incentive did not impact perceptions of bias, $p = .106$.) Overall, participants did perceive a great deal of bias in their partner ($M = 7.58$, 95% CI [7.07, 8.09]), as confirmed by a one-sample *t*-test against the scale midpoint, $t(92) = 6.21$, $p < .001$. However, this belief was stronger for participants in the critical feedback condition ($M = 8.44$, 95% CI [7.67, 9.21]) than participants in the control condition ($M = 6.58$, 95% CI [6.07, 7.09]), $t(82.47) = –4.05$, $p < .001$, $d = 0.814$. This confirms that participants did in fact perceive a bias consistent with the partner's monetary incentives, and that the perception of bias was especially large when the partner appeared to give feedback consistent with her own self-interest.

Participants also answered two questions related to the cover story, one about the degree to which their partner motivated them (1 = Not at all; 11 = A great deal), and one about how pleasant it was to have someone watching their performance (1 = Extremely Unpleasant; 11 = Extremely Pleasant). Although it was not explicitly intended as such, participants' responses to this last question can speak to the possibility that participants liked their partner, and made more directional responses in order to ensure that she got a reasonable bonus. Contradicting that account, participants in the critical feedback condition reported that it was less pleasant to have someone else watching their performance ($M = 5.06$, 95% CI [4.28, 5.84]), compared to participants in the control condition ($M = 6.33$, 95% CI [5.75, 6.90]), $t(91) = 2.56$, $p = .012$, $d = .680$.

Finally, after providing basic demographic information (gender, age, race, household income), participants were asked for "any comments or thoughts you might have about your experience in the task, the experiment, the incentives, or about your partner." Some participants expressed mild suspicion about whether their partner actually existed, but none with certainty, so no one was excluded from the analyses based on suspicion.

Statistical Analysis

All analyses were conducted in R (*16*). For all tests of the simple-effects, p values were corrected for multiple comparisons using the Holm-Bonferroni procedure (*12*); corrected p values are indicated with a subscript (i.e. p_{hb}).

As reported in the main text, we tested the effect of the feedback manipulation by conducting a 2 (Feedback: Critical Feedback vs. Control) × 2 (Timing: Pre- vs. Post-Feedback) mixed-model ANOVA, with feedback as a between-subjects variable and timing as a within-subjects variable.

Supplementary Text (Study 2)

Effects on Accuracy

Although our main prediction was about the shift in directional responses, we considered whether the shift in directional responses had an impact on participants' accuracy. The results of a similar mixed-model ANOVA predicting accuracy of responses also yielded an interaction between feedback condition and timing, $F(1,91) = 4.97$, $p = .028$, $\eta_p^2 = .052$. Participants in the critical feedback condition became less accurate after the introduction of the critical feedback ($M_{pre} = 0.772$, 95% CI [0.747, 0.798]; $M_{post} = 0.746$, 95% CI [0.720, 0.771]), though this difference was not statistically significant, $t(91) = –1.76$, $p_{hb} = .164$. Conversely, participants in the control condition showed a non-significant improvement in accuracy ($M_{pre} = 0.747$, 95% CI [0.720, 0.775]; $M_{post} = 0.771$, 95% CI [0.744, 0.798]), $t(91) = 1.44$, $p_{hb} = .164$.

To see if participants were able to detect the shifts in accuracy, we conducted a 2 (Feedback: Critical Feedback vs. Control) × 2 (Timing: Pre- vs. Post-Feedback) × 2 (Response Type: Actual vs. Estimated) mixed-model ANOVA on the proportion of correct responses, with feedback as a between-subjects variable, and timing and response type as within-subjects variables. This analysis revealed main effects of feedback condition ($p = .016$) and response type ($p < .001$, indicating a general tendency to underestimate accuracy), a two-way interaction between feedback condition and response type ($p = .022$), all of which were qualified by a three-way interaction, $F(1,91) = 9.22$, $p = .003$, $\eta_p^2 = .092$. The interaction was such that, despite actually becoming less accurate after the critical feedback started (as described above), participants in the critical feedback condition estimated that their accuracy increased, $t(91) = 2.18$, $p_{hb} = .064$. Participants in the control condition, however, did not change their estimates over time, $t(91) = 0.60$, $p_{hb} = .547$.

Materials and Methods (Explicit Beliefs Study)

We recruited 90 participants from MTurk for a brief study about fairness. Participants were randomly assigned to read one of three versions of the scenario: Disadvantage-to-self, Advantage-to-self, or Self-as-reviewer. Each version described a situation where a manager is found to be exhibiting a biased pattern of behavior (similar to that of the umpires in Study 1). As an example, the disadvantage-to-self scenario is presented below:

> *Suppose you and a coworker are being evaluated for a promotion at work. Only one of you will be promoted. To decide who gets the promotion, an independent reviewer from*

Human Resources (HR) will be monitoring both your and your coworker's performance over a two-week period and judging it for its quality.

The reviewer evaluates your and your coworker's work in this way: She keeps close tabs on the amount that each of you worked and what you accomplished each day, and assigns a score for each of you based on both criteria. Then, at the end of each day, you and your coworker receive a "scorecard," which displays the rating that each of you received for the day. For example, on a particular day, you might receive a score of 4 to 3, indicating that you outscored your coworker on that day. This scorecard system is intended to keep both of you performing at your best. It's also worth noting that the reviewer's only goal is to render a fair and accurate judgment.

By the halfway point in the judging period, you have noticed a bias in the daily scores. It seems to you that your coworker and you are performing at more or less at the same level, yet at the end of most days, your coworker receives a higher score than you. You are concerned that if this continues you will be passed up for the promotion based upon these skewed scores.

You don't want to get the reviewer in trouble, but you decide to bring up the issue of bias to the reviewer's boss in HR. The boss takes a look at the scores, alongside the concrete work that each of you has done, and decides that he, too, sees a bias that disfavors you. He decides to get in touch with the reviewer and let her know that her reviews appear to be biased.

Suppose that after talking to her boss, the reviewer agrees that her ratings have shown bias. Because of the way the system works, she cannot change her past ratings, so she is faced with a question about what is the right thing to do. There is one week remaining in the review process and then the promotion decision will be made. What is the fair thing for the reviewer to do?

a. *To be fair, the reviewer should simply attempt to rid herself of bias, judging you and your coworker on the merits of the work you both do for the next week*
b. *To be fair, the reviewer should keep using the same criteria to judge you and your coworker, even if there is a bias in those criteria*
c. *To be fair, the reviewer should "reverse" her bias, so that she now favors you over your coworker for an equal number of evaluations*

What varied between conditions was the role in which participants imagined themselves to be. In the disadvantage-to-self condition (above), participants imagined that they had been disadvantaged as a result of the manager's decisions; in the advantage-to-self condition, participants imagined themselves to be the co-worker who had benefitted from the manager's bias; in the self-as-reviewer condition, participants imagined themselves in the role of the manager. In every case, they answered a version of the question (above) appropriate to their role.

The different conditions were intended to allow for the possibility that self-interest would color participants' views of what was fair. Although the pattern of responses is consistent with that notion, because the three conditions did not significantly differ, we collapsed across the three versions of the scenario.

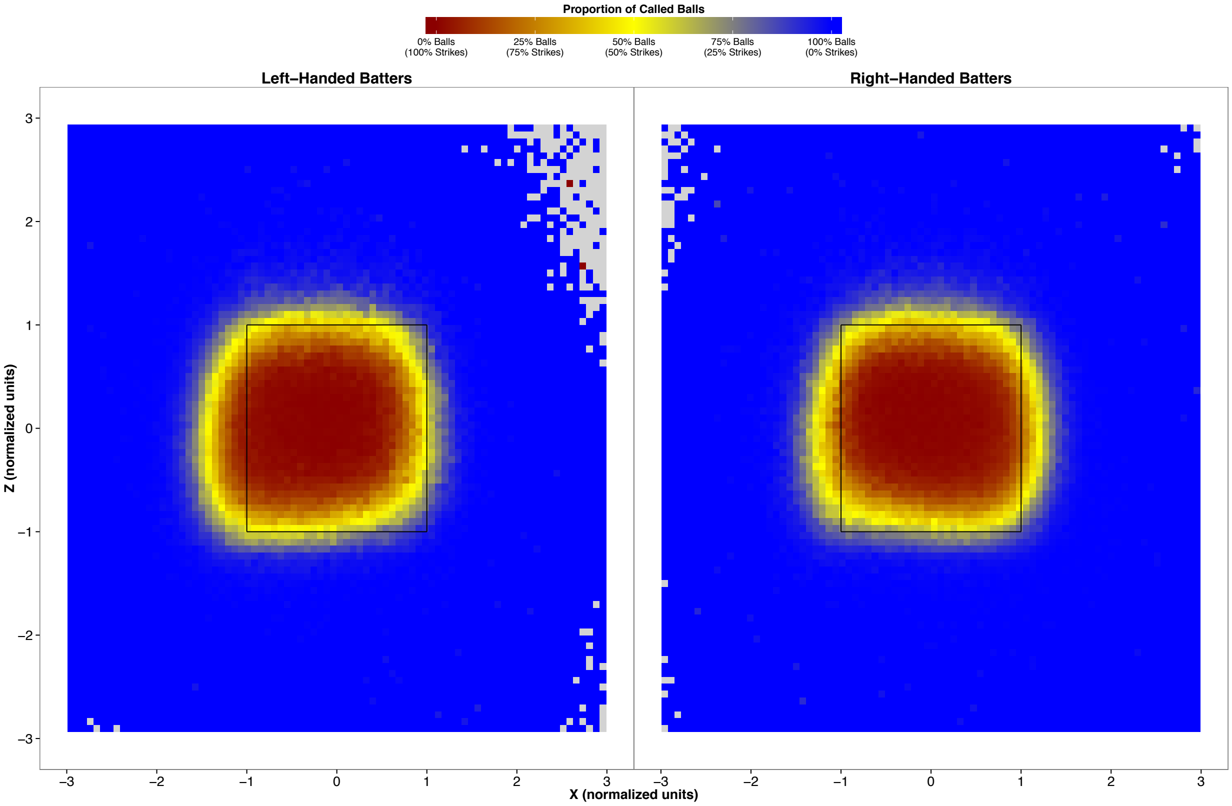


Fig. S1.

Heat map depicting the likelihood of ball or strike call based on pitch location as viewed from the umpire's perspective (Study 1). Probabilities are calculated from 2,212,150 called pitches in regular season games between 2008-2013.

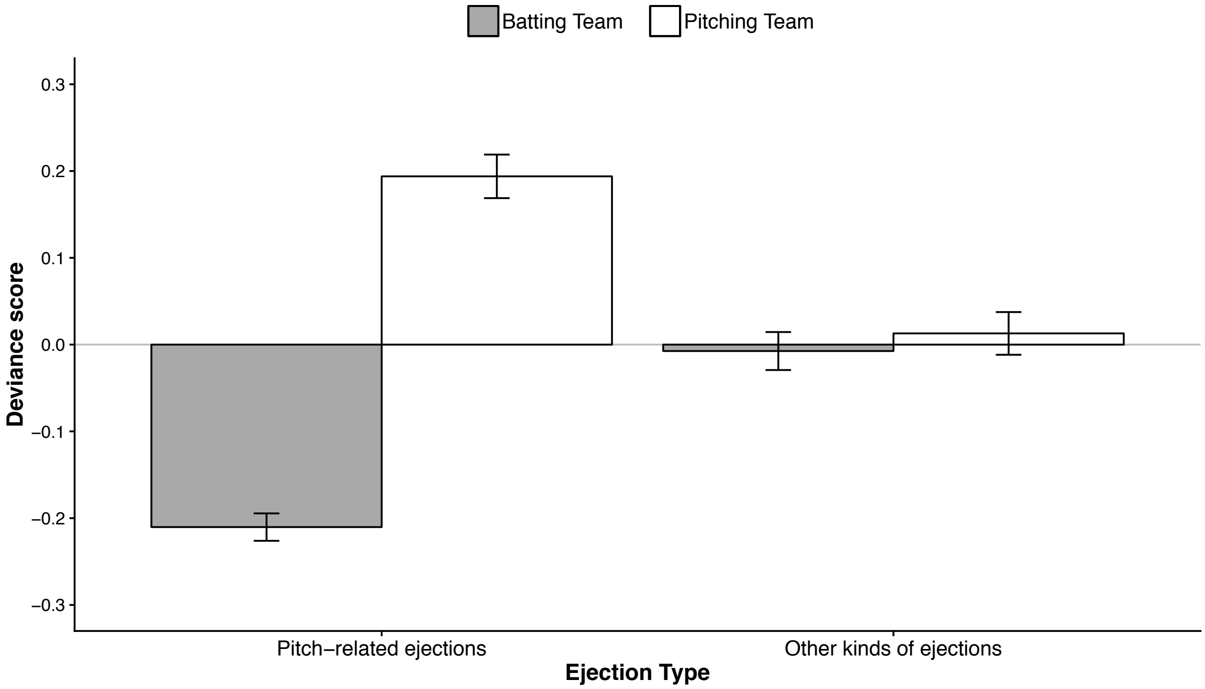


Fig. S2

Pitch-level deviance scores from pitches thrown during the ejection at-bat (Study 1). Error bars represent +/– 1 standard error.

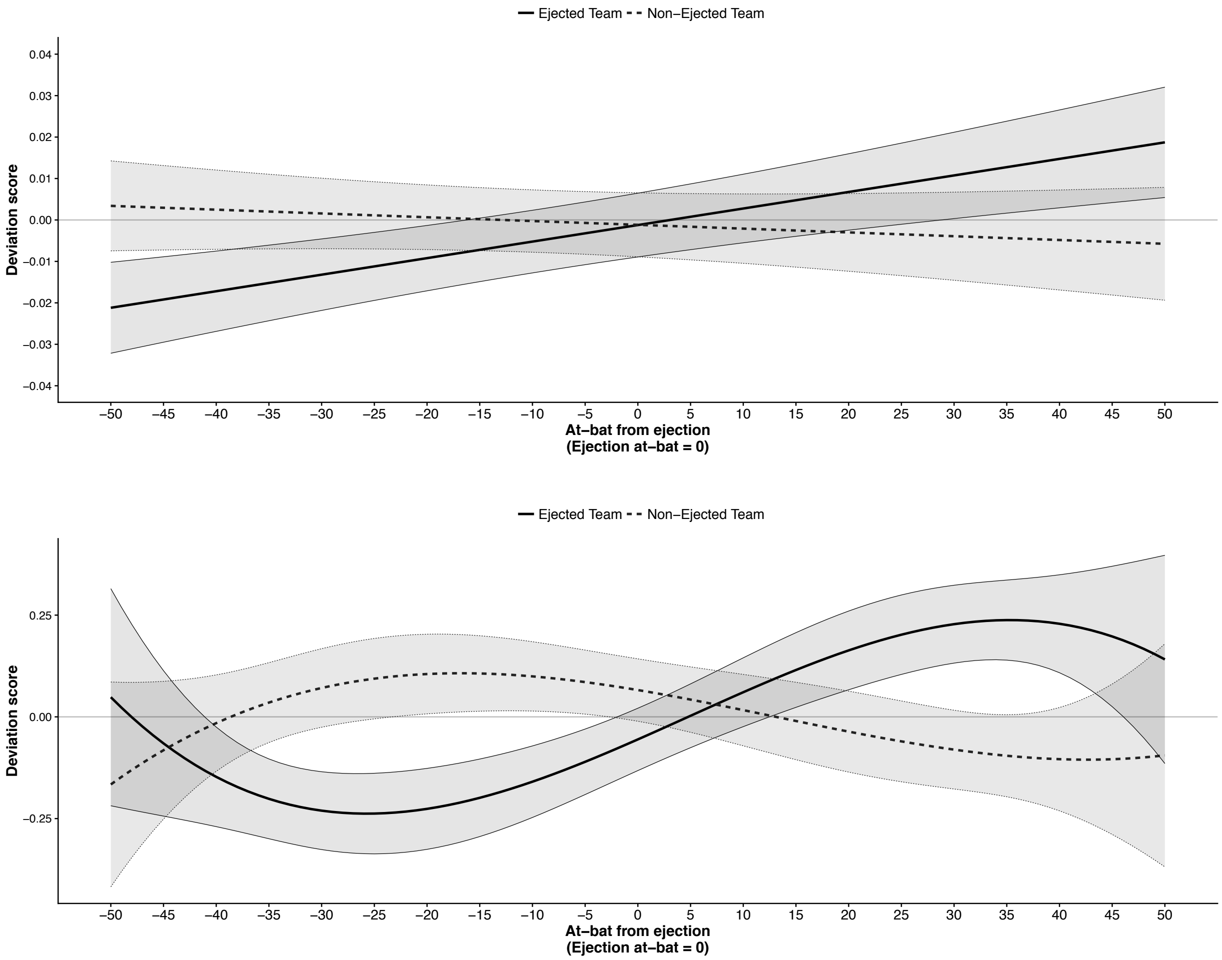


Fig. S3

Predicted values from a linear mixed-effects model predicting pitch-level deviance scores by Batting Team and AB Distance (top panel; Study 1). The bottom panel depicts a model that includes linear, quadratic, and cubic terms for AB distance. Shaded regions represent 95% confidence intervals.

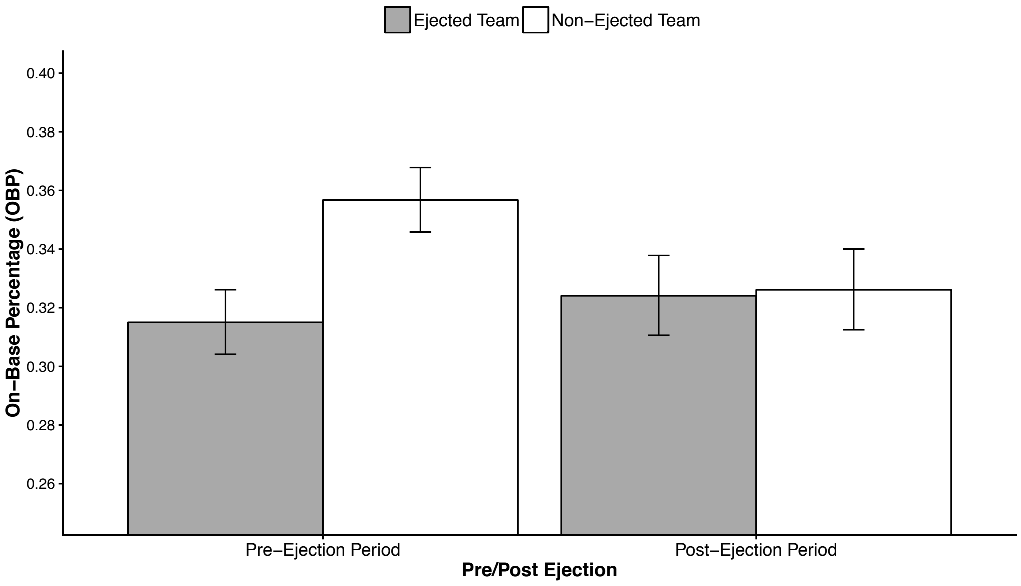


Fig. S4

On-base percentage (OBP; Study 1). Error bars represent +/- 1 standard error.

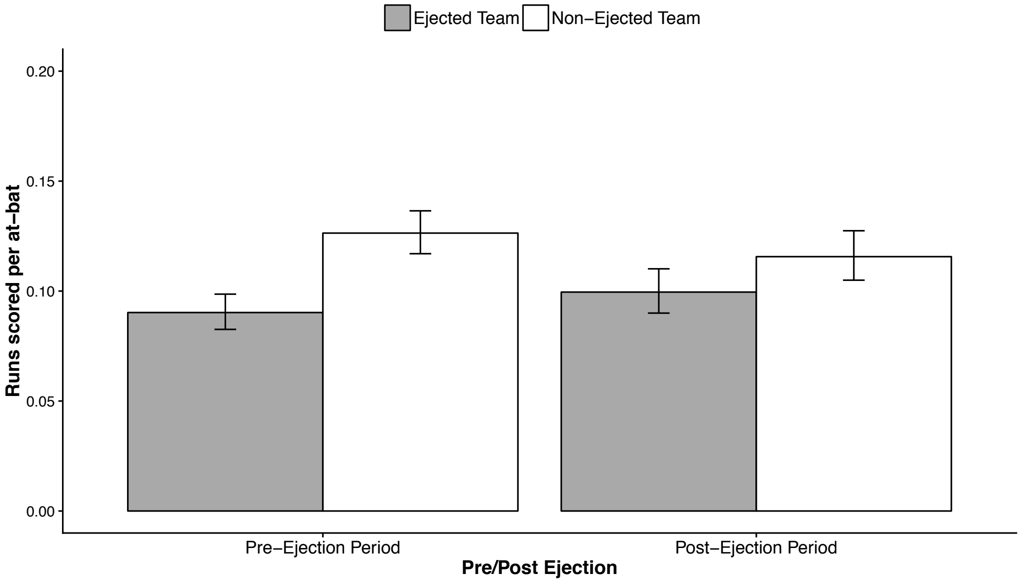


Fig. S5

Runs scored per at-bat (R/AB; Study 1). Error bars represent +/- 1 standard error.

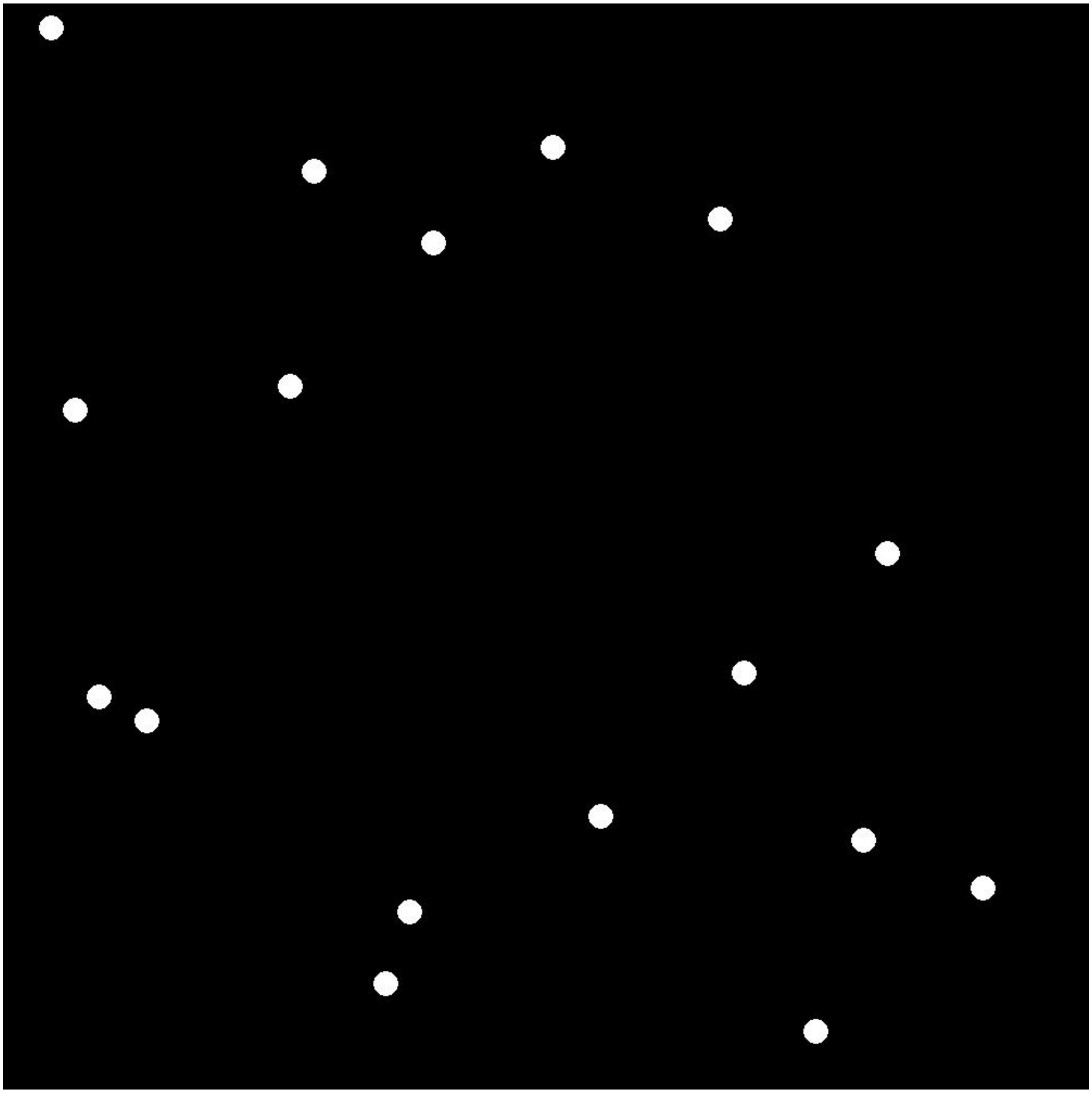


Fig. S6
Example image used in the dot-estimation task (Study 2).

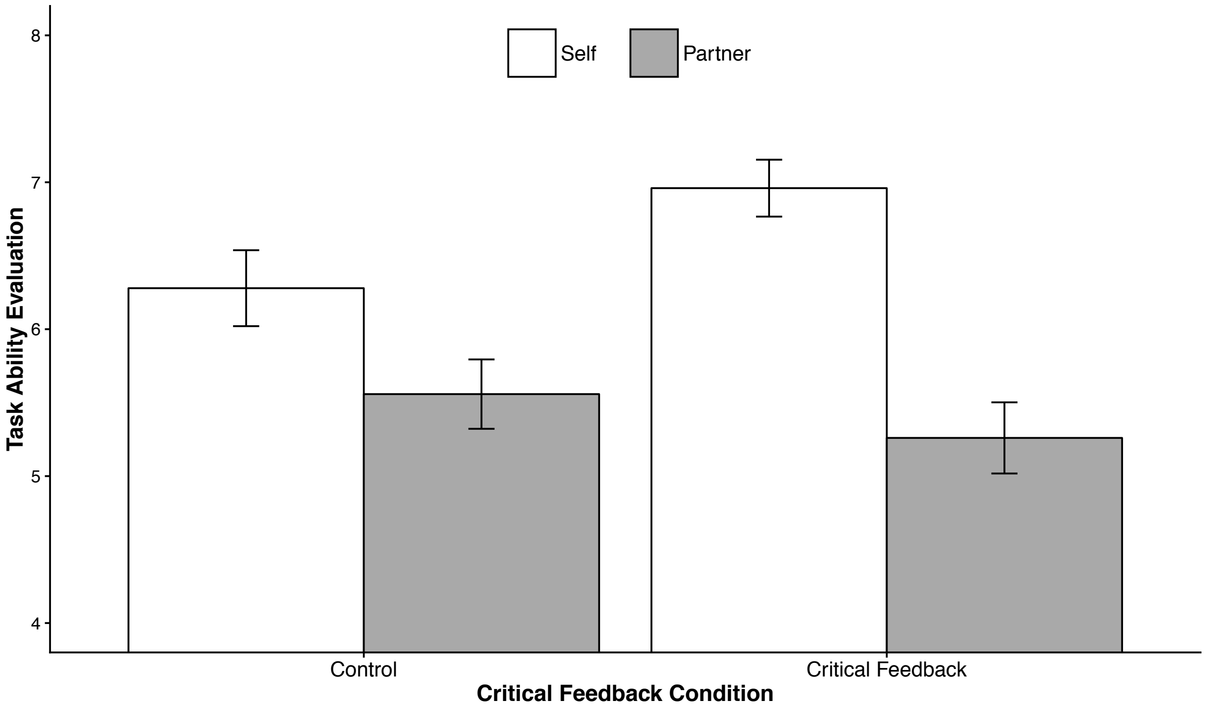


Fig. S7
Participants' perceptions of their own and their partner's ability in the game by feedback condition (Study 2). Error bars represent +/– 1 standard error.

Table S1.

Linear mixed-effects model predicting pitch-level deviance scores by Game Period, Batting Team, and Ejection Type (Study 1).

	SD		*b*		*t*	*df*	*p*
Random effects							
Home Plate Umpire							
Intercept	0.016	[0.011, 0.021]					
Game ID							
Intercept	0.036	[0.032, 0.039]					
Game Period (Pre- vs. Post-ejection)	0.014	[0.000, 0.027]					
Batting Team (Non-ejected vs. Ejected)	0.036	[0.027, 0.043]					
Game Period × Batting Team	0.034	[0.000, 0.055]					
Residual	0.318	[0.316, 0.320]					
Fixed effects							
Intercept			-0.001	[-0.006, 0.004]	-0.48	85.71	.634
Game Period (Pre- vs. Post-ejection)			0.005	[-0.000, 0.010]	1.84	676.05	.066
Batting Team (Non-ejected vs. Ejected)			0.000	[-0.005, 0.006]	0.09	716.83	.932
Ejection Type (Other vs. Pitch-related)			-0.002	[-0.009, 0.006]	-0.41	716.4	.680
Game Period × Batting Team			0.021	[0.011, 0.031]	4.23	712.72	< .001
Game Period × Ejection Type			0.012	[0.002, 0.022]	2.30	677.85	.022
Batting Team × Ejection Type			0.001	[-0.009, 0.012]	0.24	716.78	.808
Game Period × Batting Team × Ejection Type			0.042	[0.022, 0.061]	4.11	712.65	< .001

Table S2.

Linear mixed-effects model predicting pitch-level deviance scores by Game Period, Batting Team (pitch-related ejections only; Study 1).

	SD		*b*		*t*	*df*	*p*
Random effects							
Home Plate Umpire							
Intercept	0.014	[0.000, 0.022]					
Game ID							
Intercept[a]	0.037	[0.033, 0.042]					
Game Period (Pre- vs. Post-ejection)	0.016	[0.000, 0.033]					
Batting Team (Non-ejected vs. Ejected)	0.033	[0.018, 0.044]					
Game Period × Batting Team	0.049	[0.000, 0.075]					
Residual	0.321	[0.319, 0.324]					
Fixed effects							
Intercept			–0.001	[–0.008, 0.005]	–0.44	73.74	.660
Game Period (Pre- vs. Post-ejection)			0.011	[0.003, 0.018]	2.77	284.33	.006
Batting Team (Non-ejected vs. Ejected)			0.001	[–0.007, 0.009]	0.15	319.24	.878
Game Period × Batting Team			0.042	[0.026, 0.057]	5.37	314.61	< .001

Table S3.

Linear mixed-effects model predicting pitch-level deviance scores by Game Period, Batting Team (other kinds of ejections only; Study 1).

	SD		*b*		*t*	*df*	*p*
Random effects							
Home Plate Umpire							
Intercept	0.014	[0.004, 0.021]					
Game ID							
Intercept[a]	0.036	[0.031, 0.041]					
Game Period (Pre- vs. Post-ejection)	0.013	[0.000, 0.029]					
Batting Team (Non-ejected vs. Ejected)	0.038	[0.028, 0.048]					
Game Period × Batting Team	0.009	[0.000, 0.051]					
Residual	0.315	[0.313, 0.318]					
Fixed effects							
Intercept			-0.001	[-0.007, 0.004]	-0.49	78.36	0.624
Game Period (Pre- vs. Post-ejection)			0.001	[-0.006, 0.008]	0.32	398.67	0.751
Batting Team (Non-ejected vs. Ejected)			0.000	[-0.007, 0.008]	0.13	402.89	0.898
Game Period × Batting Team			0.000	[-0.012, 0.013]	0.06	402.79	0.954

Table S4.

The table below reports the results of a linear mixed-effects model predicting pitch-level deviance scores by Batting Team (non-ejected vs. ejected) and a continuous measure of game period (AB Distance). To allow for non-linear effects of AB Distance, the model included linear (L), quadratic (Q), and cubic (C) terms, as well as their interaction with Team.

	SD		*b*		*t*	*df*	*p*
Random effects							
Home Plate Umpire							
Intercept	0.014	[0.000, 0.022]					
Game ID							
Intercept[a]	0.037	[0.032, 0.043]					
Distance from Ejection AB (Linear)	1.743	[0.000, Inf]					
Batting Team (Non-ejected vs. Ejected)	0.033	[0.019, 0.044]					
Distance from Ejection AB × Batting Team	4.268	[0.000, 6.731]					
Residual							
Fixed effects							
Intercept			-0.003	[-0.009, 0.004]	-0.79	71.58	.430
Distance from Ejection AB (Linear)			0.776	[0.027, 1.526]	2.03	268.89	.043
Distance from Ejection AB (Quadratic)			-0.100	[-0.790, 0.593]	-0.28	568.81	.778
Distance from Ejection AB (Cubic)			-0.300	[-0.974, 0.373]	-0.87	910.19	.384
Batting Team (Non-ejected vs. Ejected)			0.004	[-0.004, 0.011]	0.88	301.60	.379
Distance from Ejection AB (L) × Batting Team			-2.588	[-4.021, -1.153]	-3.55	253.60	< .001
Distance from Ejection AB (Q) × Batting Team			-1.497	[-2.868, -0.127]	-2.14	543.20	.033
Distance from Ejection AB (C) × Batting Team			1.720	[0.374, 3.070]	2.50	919.98	.012

Table S5.

Linear mixed-effects model predicting pitch-level deviance scores by Game Period, Batting Team, and Bin Ambiguity (only pitch-related ejections; Study 1).

	SD		*b*		*t*	*df*	*p*
Random effects							
Home Plate Umpire							
Intercept	0.014	[0.000, 0.022]					
Game ID							
Intercept[a]	0.037	[0.033, 0.043]					
Game Period (Pre- vs. Post-ejection)	0.016	[0.000, 0.032]					
Batting Team (Non-ejected vs. Ejected)	0.033	[0.018, 0.044]					
Game Period × Batting Team	0.049	[0.000, 0.075]					
Residual	0.321	[0.319, 0.324]					
Fixed effects							
Intercept			-0.002	[-0.009, 0.005]	-0.46	98.36	.649
Bin Ambiguity			-0.001	[-0.013, 0.011]	-0.15	34430.77	.880
Game Period (Pre- vs. Post-ejection)			0.018	[0.009, 0.027]	3.92	591.79	< .001
Batting Team (Non-ejected vs. Ejected)			-0.000	[-0.010, 0.009]	-0.10	627.40	.923
Bin Ambiguity × Game Period			0.035	[0.012, 0.059]	2.91	34426.65	.004
Bin Ambiguity × Batting Team			-0.005	[-0.029, 0.018]	-0.45	34416.04	.653
Game Period × Batting Team			0.070	[0.052, 0.089]	7.51	648.52	< .001
Bin Ambiguity × Game Period × Batting Team			0.134	[0.086, 0.181]	5.49	34429.69	< .001