Let's Make Up!

A Study of the Dynamics of Makeup Calls by Umpires in Major League Baseball

Meghan Thornton-Lugo
Jonathan R. Clark
Matthew W. McCarter
Department of Management
College of Business
1 UTSA Circle
University of Texas – San Antonio
San Antonio, TX 78249
matthew.mccarter@utsa.edu

Extended Abstract

[Work in progress. Please do not cite or quote without permission.]

Most of us who watch professional sports hear about makeup calls – calls made by game referees that are intended to unofficially create equity because of a previously missed or bad call. In Major League Baseball, makeup calls are talked about all time among sports anchors, fans, and players. Many MLB umpires deny makeup calls. The MLB umpire Don Denkinger claimed,

> *If an umpire thinks he missed a call, he isn't going to make up for it by missing the next one, too. There are no make-up calls* (Skipper, 1997: 59).

The MLB umpire Terry Cooney agreed.

> *Now sometimes you hear people talk about a call and they ask: Was that a make-up call? That's ridiculous. No umpire in his right mind would make a make-up call because you try to get them all right every time* (Skipper, 1997: 66).

The current research examines whether makeup calls in the MLB exist and, if they do, what facilitates such calls. There is some evidence that a person will take steps to restore equity (also called justice) when an error in judgement has been made. In conflict management research, attempts to restore justice – through such things as an apology – occur from perpetrators who unintentionally harms another person (e.g. Goodstein & Aquino, 2010; Leunissen et al., 2013).

> ***Proposition 1 (P1):*** *Errors in judgement that have negative consequences for another person increase the likelihood of bias in future judgements in favor of the person initially harmed.*
>
> ***Hypothesis 1 (H1):*** *Missed calls made by the umpire increase the likelihood of that umpire giving a makeup call in favor of the player initially harmed.*

What could exacerbate makeup calls? There is some evidence to suggest that the stakes of the situation impact the likelihood of the perpetrator attempting to restore justice.

> ***Proposition 2 (P2):*** *The likelihood of bias in future judgements in favor of the person harmed by an initial error is greater when the stakes of the judgement are higher.*
>
> ***Hypothesis 2 (H2):*** *The umpire's likelihood of giving a makeup call in favor of the player harmed by an initial missed call is greater when the stakes of game situation are higher.*

Data and Empirical Strategy

Data Sources and Sample

We examine these issues using data on every pitch thrown during the Major League Baseball (MLB) playoffs from 2008 through 2014. Our primary data comes from the Baseball Savant database (https://baseballsavant.mlb.com/). Baseball Savant's database over the time period covered by our study was derived from MLB Advanced Media, LP, which has technology and personnel installed in each MLB stadium to chronicle every pitch of every game in a given season.

The data include information about each game, each pitch, the associated game situation, and the people involved in each pitch, including the plate umpire, pitcher, catcher, and batter. Using these data, we focus our attention on the decisions made by plate umpires concerning balls and strikes, and, accordingly, we limit our sample to those pitches which did not result in a swing of the bat or a batter hit by the pitch. The sample ensures we only evaluate pitches for which the plate umpire had an opportunity to make a decision.

We supplement this pitch-level data with at-bat-level, play-by-play data from Retrosheet. Retrosheet catalogues and archives data on every play of every game dating back to 1922 (www.retrosheet.org/). For our purposes, the Retrosheet data includes additional situational variables (e.g. the score) that we use in constructing our variables and empirical models. After accounting for missing data and lagged independent variables, our final sample includes 29,249 calls made by umpires during MLB playoff games from the 2008 season through the 2014 season.

To test our hypotheses on the sample, we aimed to identify game situations in which the umpire made an error in judgement (i.e., missed a call), hurting either the pitcher or the batter, and evaluate the extent to which those missed calls impact the umpire's decision-making on subsequent calls involving the pitcher or batter impacted by the missed call. Our empirical models include control variables for the characteristics of the pitch, the game situation, and the players involved. Most importantly, our models control for the pitch location and whether or not it was actually a strike based on MLB Advanced Media's

electronic zone evaluation system. Because the rules of baseball dictate that only pitch location – in or out of the strike zone as defined by Major League Baseball (and captured by the zone evaluation system) – should matter for the call made by the umpire, our empirical model can quantify umpire bias based on the other characteristics of the situation, including missed calls on previous pitches.

Dependent Variable

As each of the relevant calls made by the plate umpire is binary, our dependent variable is a binary indicator of the call made. Specifically, *CalledStrike* is a binary variable indicating whether the current pitch is called a strike (CalledStrike = 1) or a ball (CalledStrike = 0). This information is derived from the Baseball Savant data, which provides a categorical description of the outcome of each pitch.

Key Independent Variables

We employ variables representing missed calls of two types. First, *MissedPitcher* is a binary indicator of whether the last call made on a pitch thrown by the current pitcher was a missed call that hurt that pitcher, i.e., an actual strike that was called a ball. Second, *MissedBatter* is a binary indicator of whether the last call made on a pitch to the current batter was a missed call that hurt that batter, i.e., an actual ball that was called a strike.

To capture the stakes of the game situation in which each call is made, we employ a *Leverage* index based on the inning of the game, the current score differential, how many outs there are, and the number of the batting team's players on base (and which bases they have reached). The Leverage index has been used in prior research (Chen, Moskowitz, & Shue, 2016), and was developed by Tom Tango to distinguish situations that are potentially crucial to the game's outcome from those that are not[1]. The logic for the Leverage index is simple, a pitch in the first inning of a 0-0 game, with two outs and nobody on base, is much less likely to influence the outcome of the game than a pitch in the bottom of the ninth inning of a 2-1 game, with two outs and the bases loaded. The Leverage index is based on a widely-used measure of the

[1] For a simple explanation of the Leverage Index, see: http://www.hardballtimes.com/crucial-situations/ (accessed July 13, 2017)

probability that the home team will win the game for any given game situation (Win Probability). More specifically, the Leverage index is based on the potential change in win probability (sometimes called the "swing value") based on all the possible outcomes – and the probability distribution of those outcomes – of the current pitch; e.g. making an out versus getting a base hit versus hitting a home run. Every game situation has a "swing value" and the Leverage index is simply the situation's swing value divided by the average swing value across all situations. A *Leverage* of 1.00 means the situation is average, less than 1.00 means it is less crucial, and more than 1.00 means it is more crucial.

Control Variables

We include a number of controls to account for the key reasons that an umpire might call a pitch a strike. The variable *Zone* is a series of indicators representing locations in and around the strike zone and whether or not the current pitch passed through that location. The strike zone itself is divided into nine locations, with additional locations comprising the perimeter (and beyond) outside the strike zone. *Strike* is a binary indicator of whether the current pitch was within the strike zone based on its location. *Count* is a series of indicators capturing the ball-strike count as of the current pitch. There are eleven possible counts, from 0-0 to 3-2. *HomeBatter* is a binary indicator of whether the current batter is a member of the home team. *BattingDiff* is a variable capturing the run differential for the team currently batting; e.g., if the team is down by 8 runs, *BattingDiff* equals -8 and if the team is up by 3 runs, *BattingDiff* is 3. Finally, *PitchType* is a series of indicators capturing the type of pitch thrown by the pitcher on the current pitch. Major League Baseball defines thirteen generic pitch types, including: Changeup (CH), Curveball (CU), Cutter (FC), Forkball (FO), Four-Seam Fastball (FA), Knuckle-curve (KC), Screwball (SC), Sinker (SI), Slider (SL), Splitter (FS), and Two-Seam Fastball (FT). Summary statistics for our dependent, key independent and control variables are reported in Table 1.

Empirical Model

We estimate the probability the plate umpire calls a strike using a logit model. The logit is a generalized linear model in which the link function is the logit (natural log of the odds) of the event being

studied, in this case *CalledStrike*, and the distribution of the errors is binomial. We begin examining our hypotheses with a base specification testing Hypothesis 1:

$$ln\left(\frac{\Pr(CalledStrike_{hijkt})}{1-\Pr(CalledStrike_{hijkt})}\right) = \alpha + \rho_i + \delta_j + \lambda_k + \gamma_t + \beta_1 MissedPitcher_{h-1ikt} +$$

$$\beta_2 MissedBatter_{h-1jkt} + \beta_3 Leverage_{hkt} + \beta_4 X_{hkt} + \varepsilon_{hijkt} \qquad \textbf{(1)}$$

Where ρ_i, δ_j, λ_k and γ_t represent pitcher, batter, game and year fixed effects, respectively. The latter is included to control for unobserved factors that may drive the average likelihood of a called strike over time. Game effects are included to capture the invariant characteristics of the game (i.e., the teams involved, location, temperature, etc.), but most importantly to capture umpire effects, since the plate umpire remains the same throughout a given game. Pitcher and batter effects are included to capture characteristics of the pitcher (i.e., height, arm angle, leg kick) and the batter (i.e., height, posture, proximity to the plate, stance) that may influence the tendency of the umpire to call a strike. $MissedPitcher_{h-1ikt}$ represents whether the previous pitch thrown by pitcher i was a missed call that went against the pitcher. $MissedBatter_{h-1jkt}$ represents whether the previous pitch seen by batter j was a missed call that went against the batter. Finally, X_{hkt} represents a vector of pitch-level control variables as described previously.

To test Hypothesis 2, we build on (1), adding the interactions between the missed call variables and *Leverage*, as in (2):

$$ln\left(\frac{\Pr(CalledStrike_{hijkt})}{1-\Pr(CalledStrike_{hijkt})}\right) = \alpha + \rho_i + \delta_j + \lambda_k + \gamma_t + \beta_1 MissedPitcher_{h-1ikt} +$$

$$\beta_2 MissedBatter_{h-1jkt} + \beta_3 Leverage_{hkt} + \beta_4 (MissedPitcher_{h-1ikt} * Leverage_{hkt}) +$$

$$\beta_5 (MissedBatter_{h-1jkt} * Leverage_{hkt}) + \beta_6 X_{hkt} + \varepsilon_{hkt} \qquad \textbf{(2)}$$

Where ρ_i, δ_j, λ_k, γ_t, $MissedPitcher_{h-1ikt}$, $MissedBatter_{h-1jkt}$ and $Leverage_{hkt}$ are as presented in (1), and X_{hkt} represents a vector of control variables as described previously. We de-mean the *Leverage* variable to facilitate the interpretation of the main effects in our interaction model. In estimating these models, all standard errors are clustered by game. We note that the results of our logit estimation strategy are exponentiated and presented in terms of odds ratios, where values greater than one signify positive

effects and values less than one signify negative effects. By reporting odds ratios, we can directly interpret the magnitude and significance of interaction terms in our non-linear models (Ai & Norton, 2003; Buis 2010). Thus, our hypotheses can be restated as follows:

Hypothesis 1: $\beta_1 > 1$ *or* $\beta_2 < 1$*;*

Hypothesis 2: $\beta_4 > 1$ *or* $\beta_5 < 1$

Results

The results of our analysis are presented in Table 2. Column 1 of Table 2 presents the results of a model in which only the control variables are included. Column 2 contains the results of our estimation of equation (1), and column 3 contains the results of our estimation of equation (2).

Hypothesis Testing

With respect to hypothesis 1, the estimate on *MissedPitcher* in column 2 suggests that, on average, umpires are more likely to call a strike (regardless of location) after missing a call that went against the pitcher. However, this estimate is not statistically significant at conventional levels. Accordingly, we do not find support for hypothesis 1 with respect to calls that went against the pitcher. In contrast, the estimate on *MissedBatter* in column 2 suggests that, on average, umpires are less likely to call a strike (regardless of location) after missing a call that went against the batter. More specifically, upon examining the odds ratio, the estimate suggests that the odds of a plate umpire calling a strike are more than 20 percent less likely after the umpire has missed a call that went against the batter. These estimates are statistically significant at conventional levels ($p < 0.001$) representing strong support for Hypothesis 1 with respect to umpire calls that went against the batter.

The estimates in column 3 allow us to test Hypothesis 2. Because we demeaned *Leverage*, the estimates on the main effect of *MissedBatter* and *MissedPitcher* represent the impact of those variables at an average level of *Leverage*. As in column 2, the estimates on these main effects suggest that the odds of

a plate umpire calling a strike are less likely after the umpire has missed a call that went against the batter – supporting Hypothesis 1.

With respect to Hypothesis 2, the estimate on *MissedPitcher x Leverage* indicates that when leverage is higher than average, plate umpires become less likely to call a pitch a strike following a missed call that went against the pitcher. The estimate is in the opposite direction of our hypothesis. However, these estimates are not significant at conventional levels and, thus, we do not find support for Hypothesis 2 with respect to missed calls that went against the pitcher in high leverage situations. The estimate on *MissedBatter x Leverage* indicates that when leverage is higher than average, plate umpires become less likely to call a pitch a strike following a missed call that went against the batter. These estimates are significant at conventional levels ($p = 0.021$) and, thus, we find strong support for Hypothesis 2 when it comes to batters.

In practical terms, these results in support of Hypothesis 2 suggest that while umpires are about 20% less likely to call a pitch a strike following a missed call that went against the batter in situations of average Leverage, for each additional leverage point (e.g., Leverage = 2 versus Leverage = 1), umpires become about 9% more unlikely to call a strike following a missed call against the batter. For instance, in the crucial circumstance of the game (e.g., bottom of the 9^{th}, bases loaded, two outs, tie score) where Leverage is high (e.g., Leverage = 7), umpires are about 65% less likely to call a strike following a missed call that went against the batter.

The next steps in the current research is better understand why umpires give makeup calls and why these makeup calls are experienced by the batters and not the pitchers. Regarding the mechanism, research on conflict resolution suggest guilt is a good suspect for explaining the motivation for umpires giving makeup calls (e.g. Leunissen et al., 2013). Regarding why batters seem to enjoy makeup calls and not pitchers, there may be a two-party explanation. It could be that batters are more likely to give negative feedback to the umpire – e.g. rolling shoulders, exclamation, roll of eyes or other gestures – compared to pitchers. If this is true, then why would

batters be inclined to complain more than pitchers? One explanation could be that the batter's subjective performance is more sensitive to bad calls compared to pitchers, thereby motivating them to do whatever they can (without being thrown out of the game) to bias the umpire in their favor.

References

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and Probit models. *Economics Letters,* 80(1), 123-129.

Buis, M. L. (2010). Stata tip 87: Interpretation of interactions in non-linear models. *Stata Journal,* 10(2), 305-308.

Chen, D. L., Moskowitz, T. J., & Shue, K. (2016). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *Quarterly Journal of Economics,* 131(3), 1181-1242.

Goodstein, J., & Aquino, K. (2010). And restorative justice for all: Redemption, forgiveness, and reintegration in organizations. *Journal of Organizational Behavior,* 31(4), 624-628.

Leunissen, J. M., De Cremer, D., Folmer, C. P. R., & Van Dijke, M. (2013). The apology mismatch: Asymmetries between victim's need for apologies and perpetrator's willingness to apologize. *Journal of Experimental Social Psychology,* 49(3), 315-324.

Skipper, J.C. (1997). *Umpires: Classic baseball stories from the men who made the calls*. London, UK: McFarland.

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	Min	Max
CalledStrike	0.319	0.466	0	1
MissedBatter	0.119	0.324	0	1
MissedPitcher	0.030	0.170	0	1
Leverage	1.040	1.575	0	24
BattingDiff	-0.110	2.863	-11	11
HomeBatter	0.497	0.500	0	1
Strike	0.233	0.423	0	1
Count: 0-0	0.243	0.429	0	1
Count: 0-1	0.134	0.341	0	1
Count: 0-2	0.062	0.241	0	1
Count: 1-0	0.135	0.342	0	1
Count: 1-1	0.108	0.310	0	1
Count: 1-2	0.087	0.282	0	1
Count: 2-0	0.049	0.215	0	1
Count: 2-1	0.048	0.213	0	1
Count: 2-2	0.062	0.241	0	1
Count: 3-0	0.023	0.151	0	1
Count: 3-1	0.021	0.144	0	1
Count: 3-2	0.029	0.167	0	1
PitchType: Change	0.089	0.285	0	1
PitchType: Curveball	0.108	0.311	0	1
PitchType: Four-Seam	0.002	0.047	0	1
PitchType: Cutter	0.074	0.261	0	1
PitchType: Fastball	0.352	0.478	0	1
PitchType: Forkball	0.001	0.023	0	1
PitchType: Splitter	0.009	0.094	0	1
PitchType: Two-Seam	0.133	0.339	0	1
PitchType: Knuckle-curve	0.008	0.087	0	1
PitchType: Sinker	0.092	0.289	0	1
PitchType: Slider	0.132	0.338	0	1
Zone: Up-Left	0.020	0.140	0	1
Zone: Up-Center	0.019	0.137	0	1
Zone: Up-Right	0.018	0.133	0	1
Zone: Middle-Left	0.026	0.159	0	1
Zone: Middle-Center	0.026	0.159	0	1
Zone: Middle-Right	0.030	0.170	0	1
Zone: Low-Left	0.032	0.175	0	1
Zone: Low-Center	0.029	0.168	0	1
Zone: Low-Right	0.034	0.180	0	1
Zone: Out-UpLeft	0.183	0.387	0	1
Zone: Out-UpRight	0.128	0.335	0	1
Zone: Out-BotLeft	0.226	0.418	0	1
Zone: Out-BotRight[2]	0.229	0.420	0	1

[2] We note that due to collinearity, "Zone: Out-BotRight" is excluded from regression results

Table 2: Logit Regression Results

	(1) CalledStrike		(2) CalledStrike		(3) CalledStrike	
Key Independent Variables						
MissedBatter			**0.80*** **	**[0.70,0.90]**	**0.79*** **	**[0.70,0.90]**
MissedPitcher			**1.15**	**[0.92,1.44]**	**1.15**	**[0.92,1.43]**
Leverage			0.96***	[0.94,0.99]	0.97**	[0.95,1.00]
MissedBatter x Leverage					**0.89****	**[0.81,0.98]**
MissedPitcher x Leverage					**0.86**	**[0.67,1.11]**
Control Variables						
BattingDiff	1	[0.99,1.02]	1	[0.99,1.01]	1	[0.99,1.01]
HomeBatter	1	[0.93,1.08]	1.01	[0.94,1.09]	1.01	[0.94,1.09]
Strike	57.90***	[43.36,77.30]	58.18***	[43.56,77.71]	58.16***	[43.57,77.63]
Count: 0-1 (vs 0-0)	0.39***	[0.34,0.45]	0.40***	[0.35,0.46]	0.40***	[0.35,0.46]
Count: 0-2 (vs 0-0)	0.14***	[0.11,0.17]	0.14***	[0.11,0.18]	0.14***	[0.11,0.18]
Count: 1-0 (vs 0-0)	1.22***	[1.08,1.38]	1.18***	[1.04,1.33]	1.18***	[1.04,1.33]
Count: 1-1 (vs 0-0)	0.56***	[0.49,0.64]	0.55***	[0.48,0.63]	0.55***	[0.48,0.63]
Count: 1-2 (vs 0-0)	0.24***	[0.21,0.29]	0.24***	[0.21,0.28]	0.24***	[0.21,0.28]
Count: 2-0 (vs 0-0)	1.65***	[1.39,1.95]	1.60***	[1.35,1.90]	1.60***	[1.35,1.90]
Count: 2-1 (vs 0-0)	0.89	[0.74,1.07]	0.88	[0.73,1.05]	0.88	[0.73,1.05]
Count: 2-2 (vs 0-0)	0.35***	[0.30,0.42]	0.35***	[0.29,0.41]	0.35***	[0.29,0.41]
Count: 3-0 (vs 0-0)	3.52***	[2.77,4.49]	3.44***	[2.70,4.38]	3.43***	[2.69,4.37]
Count: 3-1 (vs 0-0)	1.06	[0.82,1.37]	1.06	[0.81,1.37]	1.05	[0.81,1.37]
Count: 3-2 (vs 0-0)	0.49***	[0.38,0.62]	0.48***	[0.38,0.61]	0.48***	[0.38,0.61]
PitchType: Curveball (vs Change)	1.45***	[1.21,1.73]	1.45***	[1.21,1.73]	1.45***	[1.21,1.73]
PitchType Four-Seam (vs Change)	0.95	[0.47,1.94]	0.97	[0.48,1.96]	0.96	[0.47,1.94]
PitchType Cutter (vs Change)	1.36***	[1.10,1.69]	1.37***	[1.10,1.70]	1.37***	[1.10,1.70]
PitchType Fastball (vs Change)	1.64***	[1.40,1.91]	1.64***	[1.40,1.91]	1.64***	[1.40,1.91]
PitchType Forkball (vs Change)	0.94	[0.32,2.80]	0.92	[0.31,2.77]	0.94	[0.31,2.81]
PitchType Splitter (vs Change)	1.09	[0.73,1.61]	1.1	[0.74,1.63]	1.1	[0.74,1.63]
PitchType Two-Seam (vs Change)	1.64***	[1.37,1.97]	1.64***	[1.36,1.96]	1.63***	[1.36,1.96]
PitchType Knuckle-curve (vs	1.76**	[1.08,2.88]	1.74**	[1.07,2.85]	1.74**	[1.06,2.86]
PitchType Sinker (vs Change)	1.37***	[1.10,1.70]	1.37***	[1.10,1.70]	1.37***	[1.11,1.70]
PitchType Slider (vs Change)	1.74***	[1.46,2.08]	1.74***	[1.46,2.08]	1.74***	[1.45,2.07]
Zone Up-Center (vs Up-Left)	1.54**	[1.04,2.30]	1.54**	[1.03,2.29]	1.54**	[1.04,2.30]
Zone Up-Right (vs Up-Left)	0.61***	[0.43,0.86]	0.60***	[0.43,0.85]	0.61***	[0.43,0.85]
Zone Middle-Left (vs Up-Left)	29.99***	[11.37,79.09]	29.93***	[11.38,78.69]	29.94***	[11.39,78.72]
Zone Middle-Center (vs Up-Left)	49.17***	[13.80,175.24	48.94***	[13.76,174.12	50.54***	[14.13,180.84
Zone Middle-Right (vs Up-Left)	3.22***	[2.07,5.00]	3.23***	[2.07,5.03]	3.23***	[2.07,5.02]
Zone Low-Left (vs Up-Left)	0.86	[0.60,1.22]	0.86	[0.60,1.22]	0.86	[0.60,1.22]
Zone Low-Center (vs Up-Left)	0.79	[0.56,1.11]	0.79	[0.56,1.11]	0.79	[0.56,1.11]
Zone Low-Right (vs Up-Left)	0.61***	[0.43,0.87]	0.61***	[0.43,0.87]	0.62***	[0.43,0.88]
Zone Out-UpLeft (vs Up-Left)	1.70***	[1.48,1.95]	1.70***	[1.48,1.96]	1.70***	[1.48,1.95]
Zone Out-UpRight (vs Up-Left)	1.33***	[1.13,1.56]	1.33***	[1.14,1.56]	1.33***	[1.13,1.56]
Zone Out BotLeft (vs Up-Left)	2.12***	[1.87,2.42]	2.13***	[1.87,2.42]	2.13***	[1.87,2.42]
N	*29,249*		*29,249*		*29,249*	

Note: Key relationships are **bolded.** Exponentiated coefficients (Odds Ratios) and confidence intervals (in brackets) reported. All models include pitcher, batter, game and year fixed effects not reported. Robust standard errors are clustered by game.
* p<0.10, ** p<0.05, *** p<0.01