Vigilante Justice: A Study of Makeup Calls in Organizations and the Behavioral Laboratory

Meghan A. Thornton-Lugo*
Matthew W. McCarter
Jonathan R. Clark
William Luse
Zahra Heydarifard
Lulu S. R. Huang

University of Texas at San Antonio
One UTSA Circle
San Antonio, TX 78249

*Corresponding author. Email: meghan.thornton@utsa.edu.

Author’s note: The fifth and sixth authors contributed equally. This research was funded by an Internal Research Award at the University of Texas at San Antonio.
Abstract

Makeup calls occur when an individual compensates for an error or mistake. By making a subsequent decision that favors the harmed party and is proportional to the original harm, individuals seek to balance the scales. However, minimal evidence has been presented regarding the existence of makeup calls. Moreover, there is little research explaining what affective experiences could account for the presence of makeup calls and what situational factors exacerbate the impact of bad calls on subsequent makeup calls. The current research reports two studies examining these issues. In the Study 1, we use archival data from Major League Baseball to show the existence of makeup calls and the moderating impact of stakes. In Study 2, we use a laboratory experiment with undergraduate students to replicate the findings of makeup calls, its moderation by stakes and other situational features, and we assess whether states account for the makeup call effect. Our studies provide evidence for makeup calls and provide insight about how poor decision-making can lead to potentially unethical behaviors.

Keywords: organizational justice; atonement; guilt; emotions; decision-making

JEL Codes: C91, D00, D80, D89
INTRODUCTION

Makeup calls—those decisions and behaviors that seek to make amends to an injured party for previous decisions that caused said injury (i.e. bad calls; see Hamilton, 2011)—are almost considered a given in the realm of sports (see Lowe, 2014). Authors such as Greg Wyshynski (2013) have stated: “[R]efs use ‘makeup calls’ after questionable penalties; also, water is wet,” while multiple empirical analyses hint at the existence of makeup calls (e.g. Davis & Lopez, 2015; Gift, 2012; Lopez & Snyder, 2013; Mills, in press). But makeup calls are not just limited to sports arenas, courts, and fields. Makeup calls can occur in the workplace, too, where employees may seek to make amends for the harm they or their organization have done to others. Consider another example from the domain of sports that occurred not on the field but in the back offices of teams and the league at large: Deflategate. The Patriots, a National Football League (NFL) team, and their star quarterback Tom Brady were punished for alleged intentional deflation of footballs during the American Football Conference Championship game. However, as Don Van Natta Jr. and Seth Wickersham of ESPN reported (2015), this punishment may have been in retaliation for failures to adequately deal with spying done on the part of the Patriots during the 2007 regular season. In other words, the punishment for one wrongdoing was influenced by the failure to admonish another.

Despite the ubiquity of errors and makeup calls across multiple organizations (see Hamilton, 2011; Lopez & Snyder, 2013), the research about them remains limited. There is even less research examining how makeup calls operate; specifically, what mechanisms account for the impact of making a bad call on subsequent behaviors? In this current research, we specifically explore how guilt mediates the relationship between bad calls and makeup calls. We also explore whether magnitude of consequences, as captured by stakes and interdependence of
outcomes, comes into play when deciding to make a makeup call. To address these questions, we examine data from the $9.5 billion industry of Major League Baseball (Study 1) and a laboratory experiment using undergraduate students (Study 2).

The findings from these studies present three major contributions. First, we find evidence suggesting the presence of makeup calls in both real-world and experimental contexts. Extant empirical evidence of the makeup call in sports is limited (e.g., Davis & Lopez, 2015; Lopez & Snyder, 2013), and experimental observations of makeup calls even more so. Our paper therefore provides evidence that supports popular narratives surrounding makeup calls and errors, and that such a tendency is not confined to the domain of sports and refereeing.

Second, we also identify potential mediating mechanisms for the effect of bad calls on makeup calls. Namely, we provide evidence that emotions—specifically, guilt—may account for this relationship. This provides insight that is not captured by current accounts for bias in calls that rely on fan influence (e.g. Pettersson-Lidbom & Priks, 2010) or social pressure from teams (e.g. Lopez, in press) to explain poor officiating.

Third, we also demonstrate how situational influences exacerbate or mitigate these effects. As every decision-making situation brings with it its own set of constraints and factors that would influence the emotions experienced or the choices made, we illuminate which features may particularly influence makeup calls. Specifically, we show that stakes and interdependence of outcomes may impact whether a makeup call is made in response to a bad call.

**THEORETICAL BACKGROUND AND RELATED RESEARCH**

**Bad Calls and Makeup Calls**

Umpires, referees, and other officials make calls that significantly impact the course of an athletic competition. Officials should be impartial, confident, and competent in their assessments
of game situations, which also function as “prescriptive utterances” about what happens during a
game (Griffioen, 2015). However, factors unique to those situations can bias the assessment of
officials and create the motivation to restore justice and fairness between competitors (see Lopez
& Snyder, 2013; Mills, in press). One such factor that can impact decision-making are bad calls.

Bad calls, though ubiquitous across multiple organizations and settings—particularly
sports—have been ill-defined when they have been explored or discussed in past literature.
Often, the understanding of what a bad call is has been implied (see Anshel & Delaney, 2001;
Hamilton, 2011; Rawstorne, Anshel, & Caputi, 2000; Russell, 2004). Of those researchers who
do define it, they choose to characterize bad calls as unavoidably common “mistakes” and
“errors” for officials (Philippe, Vallerand, Andrianarisoa, & Brunel, 2009). Though similar,
errors and mistakes are nevertheless not synonymous (Frese & Keith, 2015). Between the two,
we choose to characterize bad calls as errors that occur as a result of poor decision-making or
judgment, as they conform to Frese and Keith’s (2015) description of errors as “unintentional
deviations from goals, rules, and standards,” (p. 663).

In some cases, bad calls can be reversed or rectified. Much like a deserving employee
petitioning for reconsideration when overlooked for a promotion, players and teams may petition
officials to reconsider and reverse their calls (e.g. coaches’ challenges in the NFL). However,
not all bad calls allow for reversals or similar remedies. For instance, if a baseball umpire calls a
pitch a strike when it is really a ball resulting in the batter being called out, it cannot be undone
or reversed. If an HR manager fails to classify an incident as sexual harassment when
appropriate, there may be little recourse until another incident takes place. In such scenarios,
what options are available to the official who has made a bad call? Makeup calls are those
actions that allow individuals to recover from previous harm or bad calls (Hamilton, 2011).
Makeup calls are common in sports, such as baseball, that involve fast paced action and subjective judgements by official (Pizzi, 2008). In situations where reversals are not, individuals may be more likely to use makeup calls to rectify their wrongs. Makeup calls restore balance, harmony, or equilibrium to the extent possible (Hamilton, 2011).

**Evidence of makeup calls in sports.** Though the potential for their existence is often assumed (see Wyshynski, 2013; Lowe, 2014), there is only limited empirical evidence of makeup calls in sports. For example, Lopez and Snyder (2013) showed that NHL teams with more penalties in the first period received fewer in the second period more often compared to the other team with fewer penalties in the first period. The asymmetry suggested that referees aimed to even out penalties awarded to teams over the course of the game. Moreover, Lopez and Snyder (2013) found that for each additional penalty a team had against them relative to their opponent, there was a 6% reduction in the number of penalties called on that team in the second period. Similarly, Davis and Lopez (2015) found that the probability of a call being made against an NHL team depended on the number of penalties against that team. Namely, when more penalties had been called against a team, they were less likely to have penalties called on them in the future. Furthermore, this effect became stronger as the stakes rose. Gift (2012) found that the tendency to make certain calls in the NBA on one team increased the likelihood of subsequent judgment calls (e.g. 3-second violations) on the opposing team. Mills (in press) has also offered an empirical investigation of makeup calls, finding that pitch counts impacted subsequent calls and even changed the shape of the umpire’s perceived strike zone. Specifically, the findings suggested a tendency for equality in counts.
While there is an intuitive understanding as to why makeup calls exist, we have yet to articulate a theory for that existence. For this purpose, we turn to the justice and affect literatures to explain why bad calls lead to makeup calls.

**MAKEUP CALLS AND JUSTICE**

Officiating a game in sports is not just a matter of simple trial and error in which mistakes carry little weight, but rather a matter of moral decision-making (Hamilton, 2011; Russell, 2003). Consider fairness theory (Folger & Cropanzano, 2001), which argues that perceptions of fairness are contingent on the evaluation of three counterfactuals: would (harm), could (alternative action), and should (moral principle). Namely, if an individual has caused harm or injury, could have acted differently, and violated some moral principle in the execution of that harm, then s/he has behaved unfairly. The process by which an official makes a bad call involves similar elements: an error resulting in deprivation of some good (i.e. a chance to score, chance to get a player out), a judgment that could have been called differently, and a violation of equity (Deutsch, 1975; see also Adams, 1965) among other principles (see Hamilton, 2011; Russell, 2004). Though the original formulation of fairness theory described the actions of others (Folger & Cropanzano, 2001) and has more often than not been empirically explored as such (e.g. Brockner, Fishman, Reb, Goldman, Spiegel, & Garden, 2007; Ganegoda & Folger, 2015; Nicklin, Greenbaum, McNall, Folger, & Williams, 2011; see also, Rupp, 2011), the principles of fairness theory can also be applied to perceptions of one’s own wrongdoing. In the current research’s context of Major League Baseball, if the official concludes that s/he did harm to a player/team, had discretion, and violated moral principles, s/he may hold themselves accountable for that injustice.
In light of this, officials may want to engage in reparative or corrective action (see Hamilton, 2011) to right their wrongs. As argued by the deontic perspective (Folger, 2001), individuals have a strong response to observing injustice such that they seek to see that someone or something is held accountable for that wrongdoing. In doing so, they aim to “make sure that no one gets off Scot free,” (Folger, 2001; p. 6). Though this model, much like fairness theory, is more often than not applied to victims or observers (e.g. Beúgre, 2010; Mitchell, Vogel, & Folger, 2015; O’Reilly, Aquino, & Skarlicki, 2016; Reich & Hershcovis, 2015; Zoghbi-Manrique-de-Lara, 2010), the principles still extend to one’s own wrongdoing. Specifically, an individual who commits a wrongdoing may be compelled to ensure that the state of affairs that they have created does not continue or is addressed. Though the deontic perspective only makes predictions that people aim to restore justice that has been compromised, others provide insight as to why a makeup call might be the most likely course of action to restore justice, particularly with respect to officials in the context of the present research.

One reason for a makeup call is that many decisions cannot be undone once they are made or the process of undoing them would be burdensome to all parties involved. A supervisor who fails to fairly evaluate their employee’s performance for an annual merit raise may have little recourse until the next annual review. A second reason is an explanation (Bies, 1987; Bies & Shapiro, 1987) or an apology (Exline, Deshea, & Holeman, 2007; Granillo, 2013; Goodstein, Butterfield, & Neale, 2016) in which there is an admission of fault could potentially function to atone for harm done or improve perceptions of fairness, it comes with a risk. Admitting such an error can undermine the perceived competence on the part of players, coaches, staff, and fans (Hamilton, 2011). Additionally, the admission of an error may similarly undermine future opportunities for advancement in sports (e.g. officiating playoffs, crew chief roles), as those roles
are often filled based on call accuracy (Berg, 2016; NBA Official, 2018; cf. Davis & Lopez, 2015). As such, a makeup call may be the safest route for reestablishing justice without compromising one’s reputation or future opportunities.

A Caveat: Are Makeup Calls Morally Problematic?

When one considers the vehemence with which officials deny makeup calls (see Malinowski, 2011), it must be acknowledged that makeup calls may be morally problematic (Hamilton, 2011; Pizzi, 2008). Makeup calls, though they may function as an act of reconciliation (Hamilton, 2011) and restore fairness to a game (Hamilton, 2011; Russell, 2003), are still morally ambiguous. They do not adhere to the rules determined by governing bodies that dictate rules of play. As noted by Hamilton (2011), makeup calls are therefore a form of “vigilante justice” (p. 219) wherein officials compromise the rules governing the game. Indeed, consider the argument by Johnson and Taylor (2016): “[C]ompetitors are due opportunities enabled only by permissible actions and not those enabled by impermissible actions,” (p. 149). The advantages or disadvantages awarded to players should therefore only stem from legitimate and accurate evaluations of their performance. A makeup call—in bestowing advantage when not necessarily merited—would violate this principle and call into question the fairness of the game. Folger (2001) similarly argues that cheating, though it may not cause serious harm, is still an affront to the rules and is itself worthy of scrutiny and reproach. Extended to organizations, the violation of rules or failure to adhere to them may similarly compromise the mandates for behavior within a company (see Trevino, 1992). As such, the motivation for a makeup call may be deterred if it is perceived to compromise rules and standards. Indeed, recent theorizing has suggested that unfair behaviors may actually lead to disengagement from that behavior due to the
guilt they experience (Scott, Colquitt, & Paddock, 2009). In other words, a bad call should inhibit bad calls.

However, makeup calls may be justified. First, if the makeup call is made in the service of restoring to one party what was incorrectly given to another party—as opposed to implementing additional procedures that would violate the “spirit” of the game—then it may be considered just and fair (Hamilton, 2011). Second, as noted by Russell (2003; see also Hamilton, 2011): “Rules should not unduly prejudice the outcome from the beginning in favor of some of the participants,” (p. 101). Applied to makeup calls, if an official makes an error, the rules should not prohibit him/her from ensuring that the game is fair and that one team is not advantaged by those rules. Indeed, refereeing in the past has been more of an “art” that aimed not at accuracy alone, but also at “common sense, a desire not to interrupt the flow of the game…, and rough justice,” (Pedowitz, 2008; p. 42). This rough justice may not adhere necessarily to strict rules, but rather pursue “competitive balance” (p. 101). Finally, as Folger and Cropanzano (2001) note, “people are most likely to be held responsible for sins of omission when there is a clear normative expectation that a certain beneficial action be taken,” (p.14). In other words, the failure to intervene after a bad call may similarly compromise the official’s moral standing as s/he is the only person with the power to restore fairness in a game.

As such, though makeup calls may be morally problematic, and unethical in an absolute sense, we argue that bad calls due to their provocation of a deontic reaction on the part of officials, should be expected to produce future makeup calls.

_Hypothesis 1: Bad calls increase the likelihood of bias in future calls (i.e., makeup calls) in favor of the person initially harmed._
As similar logic applies to teams as well, i.e., bad calls will lead to future makeup calls that not only benefit the player but also benefit the team as a whole. Multi-foci justice suggests that the effects of injustice are target-specific (Cropanzano & Rupp, 2008). In other words, when responding to injustice, employees target their reactions to the party that harmed them (Rupp, Shao, Jones, & Liao, 2014). These ideas, which are grounded in social exchange theory (Blau, 1964), may similarly extend to actors, such that they target their makeup calls to the player impacted by their unfair behaviors, as we maintain in Hypothesis 1. However, we also maintain that bad calls may not affect specific team member alone. A bad call could have a spillover effect onto other players on the harmed player’s team (Lavelle, Rupp, & Brockner, 2007; see also Rupp & Cropanzano, 2009). For example, calling a ball a strike would not only prevent the player at bat from having as many opportunities to end up on base or score a run, but should that player be called out, the next player will face greater pressure to perform (or, in the extreme, will not have the opportunity to do so at all). The team now has fewer opportunities to earn points. As such, the party affected by a bad call may not be the player alone, but the team as a whole. If we consider other players on the team to be included in the focal party or an additional party to the bad call, research from multi-foci justice would similarly support our position. We would therefore expect that officials should engage in makeup calls with other players on the team.

**Hypothesis 2**: Bad calls increase the likelihood of bias in future calls (i.e., makeup calls) in favor of other people on the same team as the person initially harmed.

**GUILT AS A MEDIATOR ACCOUNTING FOR MAKEUP CALLS**

While we have hypothesized that makeup calls are likely and that their effects may target both the player harmed by the initial bad call and other players on the team, the mechanism accounting for this effect has yet to be determined. The deontic perspective to organizational
justice (Folger, 2001) argues that “deontic reactions” emerge in the face of injustice. Deontic reactions are based on moral principles and have the aim or goal of affecting or influencing those who have done wrong. Moral emotions constitute one type of deontic reaction, as they compel individuals to act in morally good ways and to avoid morally bad behaviors (Kroll & Egan 2004). Moral emotions include shame, guilt, and empathy (Eisenberg, 2000). Of these, guilt is the likeliest candidate to function as a deontic reaction, as it reflects a negative evaluation of a specific behavior (Lewis, 1971) rather than the person (Eisenberg, 2001; Tangney & Tracy, 2012). Guilt has also been defined as those emotions “that are linked to the interests or welfare either of society as a whole, or at least of persons other than the judge or agent” (Haidt, 2003, p. 853). When one considers that guilt is a moral emotion that stems not only from a negative evaluation of one’s own behavior, it seems logical to conclude that guilt may function as a self-sanctioning deontic reaction to one’s own unjust behaviors. Indeed, Folger (2001) acknowledges that individuals try to regulate their own (and others’) behaviors based on moral concerns. As such, we maintain that guilt is an affective deontic reaction that should be experienced by actors when they behave unfairly (see Scott et al., 2013).

Empirical work has supported this argument, with some scholars linking guilt to moral transgressions specifically and transgressions in the moral realm more generally (Ferguson, Stegge, & Damhuis, 1991, Sabini & Silver 1997; Smith, Webster, Parrott, & Eyre, 2002). Additionally, other scholars shift their attention to broader kinds of acts including violations of internalized standards (Tangney, 1990), prescriptive violations (Sheikh & Janoff-Bulman, 2010), disparity due to chance (i.e. survivor guilt; Brockner, Davy, & Carter, 1985; Brockner, Greenberg, Bortz, Davy, & Carter, 1986), and negative outcomes for others (Bohns & Flynn, 2013). Moreover, if guilt is a self-sanctioning moral emotion and a deontic reaction, it should
provoke behaviors that would support and affirm moral standards (Folger, 2001), particularly in behaviors toward others (Smith et al., 2002; Tangney, Stuewig, & Mashek, 2007). The other-oriented character of guilt should specifically lead to actions benefiting individuals and their relationships in a variety of ways through motivating reparatory behavior and atonement (Baumeister, Stillwell & Heatherton, 1994, 1995, 2001; Regan, 1971; Schmader & Lickel, 2006; Tangney, 1991, 1995). Indeed, the experience of guilt gives rise to motives and behaviors oriented toward reparation (e.g., Dunn, Farrar, & Hausserman, 2018; Morris & Keltner, 2000; see Tangney et al., 2007, for a review), cooperative behaviors (Ketelaar & Au, 2003), admitting one’s fault and wanting to make amends (Tangney, Miller, Flicker, & Barlow, 1996), and fewer self-interested decisions when given power (Oc, Bashshur & Moore, 2015).

Considering that bad calls compromise the rules of a game and subsequently invoke issues related to fairness, they should also provoke feelings of guilt as a moral emotion. In addition, given guilt’s impact on subsequent reparatory behaviors and atonement, it is also likely that guilt will account for the effect of bad calls on makeup calls. As such, we propose that:

Hypothesis 3: The influence of prior bad calls on the likelihood of bias in future calls (i.e., makeup calls) is mediated by guilt.

THE MODERATING EFFECT OF MAGNITUDE OF CONSEQUENCES

Though we expect that bad calls will lead to makeup calls, and that guilt will account for this effect, we also acknowledge—consistent with models of ethical decision-making (Trevino, 1986; Jones, 1991)—that other factors may influence this process. The time of a game within a season (Lopez & Snyder, 2013), timing within the game (Green & Daniels, 2014), physical location of games (Pettersson-Lidbom & Priks, 2010) and officials (Lopez, 2016), and the quality and type of player (Gift, in press; Mills, in press; Rainey, Larsen, & Stephenson, 1989)
may influence when makeup calls are given. Though the number of features that impact calls and subsequent reactions are numerous, considering the moral implications of bad calls and makeup calls (see Hamilton, 2011), we focus on features that are reflective of moral intensity. As noted by Jones (1991), features of the moral issue at hand may impact the ethical decision-making process. He specifically argued that moral intensity—“a construct that captures the extent of issue-related moral imperative in a situation” (p. 372)—would impact recognition, judgments, intentions, and behaviors in relation to moral issues. Though Jones (1991) discusses moral intensity as a single construct consisting of multiple dimensions, we have chosen to focus on magnitude of consequences, specifically (see Flannery & May, 2000; Mencl & May, 2016; Weber, 1996). Magnitude of consequences refers to the extent to which harm or benefit is done to another party (Jones, 1991). Others have similarly characterized magnitude of consequences as the sum of harm that impacts all possible victims of one’s wrongdoing (Kish-Gephart, Harrison, & Trevino, 2010; p. 5).

Generally, when features of moral intensity are high, issues are more likely to be perceived as moral, more advanced moral reasoning will be evoked, moral intentions will emerge, and ethical behavior will be more likely (Jones, 1991). Supporting this, research has found that moral intensity is negatively related to unethical intentions (Kish-Gephart et al., 2010), while magnitude of consequences has also been shown to have a positive impact on ethical issue recognition (May, Li, Mencl, & Huang, 2014), perceptions of behavior as unethical, and level of moral reasoning, and a negative impact on intentions to engage in unethical behaviors (Weber, 1996). When it comes to the specific role of affect in Jones’ (1991) model of ethical behavior and moral intensity, emotional responses play a role in two key ways. The first is that issues of high moral intensity are more “emotionally interesting,” (p. 381). In other
words, the features of a moral issue—especially magnitude of consequences and concentration of effect—will have a greater affective resonance with people observing the moral issue and may provoke stronger emotional responses. Second, moral intensity also influences moral intent via affect. Specifically, issues with greater moral intensity lead to strong emotional responses, which subsequently provoke moral behaviors through intentions.

To assess the impact of magnitude of consequences, we identified two features of sports contests that would impact the officials’ affective experience after a bad call and the subsequent influence of affect on behavior: stakes and outcome interdependence. These features, we believe, conform to descriptions of moral intensity and magnitude of consequences provided by Jones (1991). First, Jones notes that the death of a human is far greater than a minor injury. In the domain of sports, such degrees of import and impact can be captured by stakes. Consider an umpire who makes a bad call during the second inning of mid-season at a game that is poorly attended. Now, compare that to a bad call made during the World Series with the final inning approaching. Both of these situations should have differing effects on the experience of umpires. As captured by the inning and time of season in those examples, stakes should provoke stronger guilt among umpires. Indeed, and consistent with arguments grounded in moral intensity, when bad calls are made during critical times in a game or a season, it may create pressure and the desire to be fair may be amplified (see Lopez & Snyder, 2013; see also Bhal & Dadhich, 2011; Taylor & Curtis, 2010). Moreover, as noted previously, officials do not want to be a deciding factor in a game (see Russell, 2003). Rather, the players and their skill should determine who wins. When the championship title (as opposed to a mid-season game) is on the line, the guilt evoked should be stronger after a bad call.
Considering stakes and the arguments set forth by Jones regarding how moral intensity can impact behaviors, we propose:

*Hypothesis 4a: The stakes of the game will moderate the relationship between bad calls and the likelihood of bias in future calls (i.e., makeup calls), such that the relationship will become stronger as the stakes rise.*

We also consider how the moderating effects of stakes, as well as outcome interdependence, will operate in conjunction with our mediating effect proposed in Hypothesis 3. Pertaining to moral intensity, Jones (1991) notes that an act that impacts 1,000 people has more intensity than one that impacts 10. In light of this, we consider interdependence of outcomes (Wageman, 1995) —or, the number of parties affected by a bad call—when assessing moral intensity. Consider when a bad call is made in sports: Fans will boo or hurl insults (see Hennessy & Schwartz, 2007 for data on parents). The impacted player has a few words of objection. The coach comes out to confront the offending official (see Warneke & Ogden, 2012). An entire team peers from the dugout. All of these parties are impacted by the umpire’s bad call. As such, the magnitude of consequences as captured by interdependence of outcomes may exacerbate the impact of guilt on makeup calls. While Jones’s formulation seems to indicate that magnitude of consequences (i.e. moral intensity) should intensify guilt, when considering the number of people impacted, it may also intensify guilt’s impact on behaviors. Specifically, when an umpire makes a bad call, that call impacts potentially thousands of people in a ball park.

---

1 Wageman (1995) defined outcome interdependence as “the degree to which the significant outcomes an individual receives depend on the performance of others,” (p. 147) while van der Vegt, Emans, and van de Vliert (2000) described it as the extent to which group goals and group feedback are present. Our concept of outcome interdependence reflects to extent to which outcomes, feedback, or goals will impact multiple people.
including but not limited to the player, coach, the entire team and staff, and the fans. With the knowledge that his call can affect thousands, an umpire’s guilt may provoke more makeup calls than had his error impacted only a handful of people. Indeed, research on NHL referees showed that penalty differentials seemed to favor home teams when compared to away teams (Lopez & Snyder, 2013), suggesting that the salience of a home team crowd (i.e. more people) may have influenced the calls being made.

As such, we maintain that stakes will impact the guilt experienced by officials in response to making a bad call while interdependence of outcomes will impact the role of guilt in predicting makeup calls, both of which will operate to influence the mediating effect of guilt. We therefore hypothesize:

\textit{Hypothesis 4b: The stakes of the game will moderate the indirect effect of guilt on subsequent bias in future calls (i.e. makeup calls), such that the indirect effect of guilt is stronger when the stakes are higher.}

\textit{Hypothesis 4c: The interdependence of outcomes will moderate the indirect effect of guilt on subsequent bias in future calls (i.e. makeup calls) such that the indirect effect is stronger when outcome interdependence is higher.}

\textbf{OVERVIEW}

We tested our hypotheses through an archival study and an experiment. Study 1 involved an empirical investigation of a large sample of archival data available from Major League Baseball.\textsuperscript{2} This study was aimed at testing Hypotheses 1, 2, and 4a. Study 2 involved an

\textsuperscript{2} The utilization of sports as a domain in which to understand phenomena of interest to management scholars is not new (e.g. Bezrukova, Spell, Caldwell, & Burger, 2016; Campbell, Saxton, & Banerjee, 2014; Sieweke & Zhao, 2015; Smith & Hou, 2015; Staw & Hoang, 1995;
experimental design using data from undergraduate students enrolled in an introductory level business class. This study was aimed at testing Hypotheses 1, 3, 4b, and 4c. Specific details on the designs and results follow.

**STUDY 1- MAJOR LEAGUE BASEBALL**

**Methods**

*Data Sources and Sample.* We examined the issues related to Hypotheses 1, 2 and 4a using data on every pitch thrown during the 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 had technology and personnel installed in each MLB stadium to chronicle every pitch of every game in a given season. In this study, we focus specifically on the data derived from Major League Baseball’s PITCHf/x system. At the time our data set was cultivated from Baseball Savant, PITCHf/x was the system used by the MLB to track pitches. The data included information about each game, pitch, the associated game situation, and the people involved in each pitch, including the plate umpire, pitcher, catcher, and batter. It has since been replaced by a new system called Statcast (Arthur, 2017).

Using these PITCHf/x data, we focused our attention on the decisions made by plate umpires concerning balls and strikes and, accordingly, we limited our sample to those pitches which did not result in a swing of the bat or a batter hit by the pitch. The sample ensured we only evaluate pitches for which the plate umpire had an opportunity to make a decision.

---

Warneke & Ogden, 2012). While this context is specific, we believe that the principles and ideas we discuss still extend to the broader workplace.
We supplemented the pitch-level data with at-bat-level, play-by-play data from Retrosheet (http://www.retrosheet.org/). Retrosheet is a volunteer organization that catalogues and archives data on every play of every game dating back to 1922. For our purposes, the Retrosheet data included additional situational variables (e.g., the score) that we used in constructing our variables and empirical models. After accounting for missing data and lagged independent variables, our final sample included 29,249 calls made by umpires during MLB playoff games from the 2008 through the 2014 seasons.

To test our hypotheses, we identified game situations in which the umpire made an error in judgement (i.e., missed a call), hurting either the pitcher or the batter, and evaluated the extent to which those missed calls predicted bias in the umpire’s decision-making on subsequent calls involving the pitcher or batter impacted by the missed call. Our empirical models included control variables for the characteristics of the pitch, the game situation, and the players involved. Most importantly, our models controlled 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 the MLB (and captured by the zone evaluation system)—should matter for the call made by the umpire, by employing models that account for pitch location, our analysis should be able to quantify umpire bias based on the other characteristics of the situation, including missed calls on previous pitches.

Dependent Variables. 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**

**Calls.** We employ variables representing missed calls of three types. First, *MissedPitcher* is a count of how many of the last five calls 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 count of how many of the last five calls 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. Third, *MissedTeamatBat* is a count of the number of bad calls that went against the previous batter for the team currently at bat.

**Stakes.** To capture the stakes of the game situation in which each call is made, we employed a measure 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). Specifically, we employ *Leverage*. *Leverage* is an index that has been used in prior research (Chen, Moskowitz, & Shue, 2016), and was developed by Tom Tango (see Tango, Lichtman, & Dolphin, 2007) to distinguish situations that are potentially crucial to the game’s outcome from those that are not.³ The logic 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 index is based on a widely used measure of the probability that the home team will win the game for any given game situation (i.e. win probability). More specifically, the leverage index is based

---

³ For a simple explanation of the Leverage Index, see: http://www.hardballtimes.com/crucial-situations/.
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. An index 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. We note here that by limiting our sample to playoff games, we ensured that the stakes are actually meaningful to both teams, unlike some regular season games in which poorly performing teams may not have much to play for.

**Control Variables.** We included 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, with subsequent values after 0-0 (e.g. 1-0) approaching 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 = -8*; if the team is up by 3 runs, *BattingDiff = 3*). Finally, *PitchType* is a series of indicators capturing the type of pitch thrown by the pitcher on the current pitch. The MLB (n.d.) defines fourteen generic pitch types, including: Changeup (CH), Curveball (CU), Eephus (EP), Cutter (FC), Forkball (FO), Four-Seam Fastball (FA), Knuckle-curve (KC),
Knuckleball (KB), Screwball (SC), Sinker (SI), Slider (SL), Splitter (FS), Two-Seam Fastball (FT), and other. 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 $\text{CalledStrike}$, and the distribution of the errors is binomial.

We begin examining our hypotheses with a base specification testing Hypothesis 1:

$$\ln\left(\frac{\Pr(\text{CalledStrike}_{hijkt})}{1-\Pr(\text{CalledStrike}_{hijkt})}\right) = \alpha + \rho_i + \delta_j + \lambda_k + \gamma_t + \beta_1 \text{MissedPitcher}_{hijkt} +$$

$$\beta_2 \text{MissedBatter}_{hijkt} + \beta_4 \text{Leverage}_{hkt} + \beta_4 X_{hkt} + \epsilon_{hijkt}$$

(1)

Where $\rho_i$, $\delta_j$, $\lambda_k$ and $\gamma_t$ represent pitcher, batter, catcher and game fixed effects, respectively. Game effects are included to capture the invariant characteristics of the game (i.e., the teams involved, location, temperature, etc.), but perhaps most importantly to capture umpire effects, since the plate umpire remains constant throughout a given game. We note that game effects are also collinear with year effects, meaning that game effects also capture the general influence of time. Pitcher, batter, and catcher effects are included to capture characteristics of the pitcher (e.g., height, arm angle, leg kick), the batter (e.g., height, posture, proximity to the plate, stance) and catcher (e.g., posture and position, ability to frame the pitch) that may influence the tendency of the umpire to call a strike. $\text{MissedPitcher}_{hijkt}$ represents the count of how many of the previous five pitches thrown by pitcher $i$ were missed calls that went against the pitcher. $\text{MissedBatter}_{hjkt}$ represents a count of how many of the previous five pitches seen by batter $j$
were missed calls 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 term for the batter’s team, as in (2):\nn\frac{\ln\left(\frac{\Pr(Called\text{Strike}_{hjk}\)}{1-\Pr(Called\text{Strike}_{hjk}\)}\right)}{\Pr(Called\text{Strike}_{hjk}\)} = \alpha + \rho_t + \delta_j + \lambda_k + \gamma_t + \beta_1 MissedPitcher_{hikt} + \beta_2 MissedBatter_{hjk} + \beta_3 MissedTeamatBat_{hjk} + \beta_4 Leverage_{hkt} + \beta_5 X_{hkt} + \varepsilon_{hkt} \quad (2)\n
Where $MissedTeamatBat_{hjk}$ represents the number of missed calls that went against the previous batter for the team currently at bat, and all other variables are as described previously. To test Hypothesis 4a, we build on (2), adding the interactions between the missed call variables and Leverage, as seen in (3):\n\frac{\ln\left(\frac{\Pr(Called\text{Strike}_{hjk}\)}{1-\Pr(Called\text{Strike}_{hjk}\)}\right)}{\Pr(Called\text{Strike}_{hjk}\)} = \alpha + \rho_t + \delta_j + \lambda_k + \gamma_t + \beta_1 MissedPitcher_{hikt} + \beta_2 MissedBatter_{hjk} + \beta_3 MissedTeamatBat_{hjk} + \beta_4 Leverage_{hkt} + \beta_5 (MissedPitcher_{hikt} \times \text{Leverage}_{hkt}) + \beta_6 (MissedBatter_{hjk} \times \text{Leverage}_{hkt}) + \beta_7 (MissedTeamatBat_{hjk} \times \text{Leverage}_{hkt}) + \beta_8 X_{hkt} + \varepsilon_{hkt} \quad (3)\n
Where all variables are as described previously. In estimating these models, all standard errors are clustered by umpire. 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. While we report odds ratios, we have also estimated average marginal effects to confirm the direction and significance of interaction terms in our non-linear models (Ai & Norton, 2003; Buis 2010). We find that the direction and significance of marginal effects match our odds ratio estimates. Thus, our hypotheses can be restated as follows:
Hypothesis 1: $\beta_1 > 1$ or $\beta_2 < 1$;

Hypothesis 2: $\beta_3 > 1$

Hypothesis 4a: $\beta_5 > 1$ or $\beta_6 < 1$ or $\beta_7 > 1$

RESULTS

The results of our analysis are presented in Table 2. Column 1 contains the results of our estimation of equation (1), column 2 contains the results of equation (2), and column 3 contains the results of our estimation of equation (3).

| Insert Table 2 about here |

With respect to Hypothesis 1, the estimate on MissedPitcher in Column 1 suggests that, on average, umpires are more likely to call a strike (regardless of location) after missing a call that went against the pitcher. Similarly, the estimate on MissedBatter 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, the estimates suggest that the odds of a plate umpire calling a strike are nearly 15% more likely after the umpire missed a call against a pitcher and more than 20% less likely after the umpire has missed a call against the batter. These estimates are statistically significant at conventional levels ($p < 0.05$) representing strong support for Hypothesis 1 with respect to umpire calls that went against the batter and pitcher. When it comes to team effects, the estimate for MissedTeamatBat in Column 2 suggests that when a bad call is made against the batter’s team, a strike is less likely for subsequent batters at over 6%. This estimate is statistically significant at conventional levels ($p < 0.05$) representing strong support for Hypothesis 2, with regard to the impact of bad calls against the batter’s team on future strikes. In addition, though we did not predict any main effects of leverage, the estimates for Leverage suggest that as the stakes get higher an umpire is less likely to call a strike.
The estimates in Column 3 allow us to test Hypothesis 4a. Because of the interactions, the estimates on the main effect of MissedBatter, MissedPitcher and MissedTeamatBat represent the impact of those variables at the lowest level of Leverage. As in Column 1, the estimates on the main effects remain similar to those discussed above. With respect to Hypothesis 4a, the estimate on MissedPitcher x Leverage indicates that when leverage is high, plate umpires become less likely to call a pitch a strike following a missed call that went against the pitcher. In other words, as the stakes increase, umpires are less likely to call a strike—which would favor the pitcher—even after making a call that harmed them. This estimate is in the opposite direction of our hypothesis and, thus, we do not find support for Hypothesis 4a. Rather, we find strong evidence, significant at conventional levels (p < .05), for the opposite effect when it comes to pitchers.

In practical terms, these results suggest that while umpires are about 32% more likely to call a pitch a strike following a missed call that went against the pitcher in situations of low Leverage, for each additional leverage point (e.g., Leverage = 2 vs Leverage = 1), umpires become about 13% less likely to call a strike following a missed call against the pitcher. Overall, in the most crucial of circumstances (e.g., bottom of the 9th, bases loaded, two outs, tie score) where leverage is very high (e.g., Leverage = 7), our estimates suggest that following a missed call that went against the pitcher, umpires become about 51% less likely to call a strike relative to situations of low leverage (e.g., Leverage < 1).

Robustness

We ran several robustness checks on our findings. First, we acknowledge that because both our key independent and dependent variables are choice variables, our results may be subject to endogeneity concerns. Specifically, it may be the case that our findings are a function
of underlying patterns of bias that appear to be missed calls, but are in fact simply a function of patterns in an umpire’s willingness to call balls and strikes (e.g., an umpire may simply be less likely to call a strike after calling a strike on the previous pitch). To investigate this issue, we ran a supplemental analysis, including an autoregressive term (lagged dependent variable) in the model to account for the patterns, if any, we have described. We do not include such a term in our base analysis because of the potential for Nickell bias (Nickell, 1981), which cautions against including lagged dependent variables in fixed effects models like ours. However, Nickell (1981) suggests that the bias is a function of the number of within panel observations, and that the bias approaches zero as the number of observations approaches infinity. While not ideal, given that the number of calls made by umpires in a given game (our key panel variable) is high (mean = 150), we felt a robustness check was appropriate, given its comparative purpose and that the potential for Nickell bias should be somewhat reduced by our within panel sample size. Accordingly, we ran the model presented in equation (3), adding an autoregressive term representing \( \text{CalledStrike} \), lagged by one pitch. The results are shown in Table 3, column 1 and suggest that our findings are not unduly influenced by this particular endogeneity concern.

Second, in constructing our key independent variables related to missed calls, we decided to include any missed call that was made up to five pitches prior to the current pitch. We did so for two reasons: (1) for batters, five pitches prior was as far back as we observed a missed call, and (2) we wanted to allow for the possibility that, given a missed call, umpires might look for the right moment (the right pitch) to make up for it (e.g., they may prefer to do it on a pitch that is close). In assessing the robustness of our findings to this choice, we had two objectives in mind. First, we wanted to ensure that our findings were not a function of something anomalous
associated with looking back five pitches. Second, we wanted to know whether the length of
time (or, for how many pitches) missed calls against pitchers bias an umpire’s judgement. While
umpires only make an average of 7 to 8 calls on batters per game (and thus, the reason missed
calls only show up in a five-pitch window), the average for pitchers is much higher. To address
these issues, we ran four alternative models, with missed call windows ranging from 3 pitches to
7 pitches. We note that all four models indicate estimates that are similar in magnitude and
significance to our base results presented in Table 2 (supporting the robustness of our results),
with one notable exception: the magnitude and significance of the estimates related to pitchers
begin to diminish with a 7-pitch window (shown in Table 3, column 2). Accordingly, these
findings suggest that the window within which umpires may be willing to give makeup calls is
reasonably short, with the bias beginning to fade after six pitches.

Finally, in estimating our model, we ran a logit with fixed effects; four fixed effects to be
precise: game, pitcher, batter and catcher. In doing so, our findings may be subject to the
incidental parameters problem (Heckman, 1987). We were not able to run a conditional logit, as
preferred, because, to our knowledge, we are not aware of a method that allows for conditioning
on four different conditions. The incidental parameters problem is unique to non-linear models,
and accordingly, we check whether our findings may be subject to this issue by running a linear
probability model (LPM), proceeding in two parts. First, we ran an LPM on the complete sample
of data. Since one of the drawbacks of the LPM approach is the potential for non-sensical
(biased) predictions, we followed Horrace and Oaxaca (2006) and ran a second model excluding
observations with predicted values less than zero or greater than one based on the first model.
Results of the second model are reported in Table 3, Column 3. The results suggest that the
estimates are consistent with our base results in terms of direction and significance, suggesting that our logit estimates are unlikely biased by the incidental parameters problem.

**STUDY 1 DISCUSSION**

Overall, our findings provided support for Hypothesis 1. Specifically, when bad calls were made against pitchers (batters), umpires were more (less) likely to call strikes. Our estimates also suggested that umpires become less likely to call strikes on batters after making bad calls that went against the batter’s teammates, providing support for Hypothesis 2. However, when it came to Hypothesis 4a, we did not find support for the effects anticipated. Rather, we found that when stakes increased, umpires are less likely to make calls that favored pitchers. Notably, this effect was not replicated with batters, suggesting that there may be something conditional in the relationships between umpires, pitchers, and batters that make umpires less likely to give makeup calls (that would go against batters) to pitchers in moments of high stakes.

These findings suggest that organizational actors (e.g. supervisors) may engage in reparative behaviors when they have previously erred. These behaviors not only seem to target the person who was harmed, but others who might have been impacted by their harm. Moreover, when the potential impact of a makeup call is high, it is likely that the employee in question will err on the side of caution and *not* provide a makeup call. Applied to an example given earlier, if a supervisor failed to adequately pursue a claim of sexual harassment, s/he may choose to punish the perpetrator for a subsequent minor violation. This punishment may not extend necessarily to acts the perpetrator engaged in with the original victim, but potentially others, as well. However, if the consequences of pursuing such a claim might have strong negative impact on the party involved, the supervisor may refrain from pursuing it further.
To explore these findings, increase generalization, and uncover the mechanism for makeup calls, we replicated them in an experimental study.

**STUDY 2 – LABORATORY EXPERIMENT**

**Pilot Testing**

Study 2 involved the use of six jars containing approximately 300 items that participants would judge. Specifically, participants would judge whether each jar had over or under 300 items. To ensure that the jars were not consistently more or less likely to be perceived as containing over/under 300 items, we conducted pilot testing with 22 MBA students who provided their ratings of 6 sample jars. The jars contained 280, 288, 298, 305, 314, and 320 items, respectively. We assessed whether students could consistently identify whether these were over or under 300 items using a chi-square test. We found that the 320 and 314 item jars were consistently gauged as over 300 ($p < .05$). As such, we determined that our jars should have between 280 and 305 items to be equally likely to be judged as over or under 300 items.

**Methods**

**Participants and Design.** Participants were 207 students 43.9% female, 55.8% White in an undergraduate Introduction to Management course at a large state university in the Southwestern United States who participated in exchange for course credit and entry into a raffle for a $500 Amazon.com gift card. The average age was 21.6 years and 34.1% of participants had been enrolled for less than a year at the university. More than half (65.5%) of participants were currently employed and nearly three-fourths (75.7%) of participants spoke English as a first language.

**Materials and procedure.** One week before participation, students were given a brief information session regarding the experiment. They were provided general information
regarding the objectives and were told that a mobile device with Internet access would be required to participate. On the day of the experiment, students were briefed on study procedures in greater depth and told that they would be working with two other people in the room to make judgements about six jars filled with nuts, bolts, screws and other objects. Participants were also told that one of the three group members would be assigned as a judge, while the other two served as raters. Raters would be shown the jars, one at a time, and would rate whether they thought the number of items in the jar was greater than or less than 300. The judge’s responsibility was to determine which group member was correct. Unbeknownst to participants, everyone was assigned as a judge and the teammates were “virtual”. Participants were informed that they would be answering questions about how they felt after each round of the experiment, with six rounds in total.

In return for their participation, participants were told that they would receive extra course credit and entry into a raffle for one $500 Amazon.com gift card. Participants were told that they would receive 10 raffle entries every time they correctly determined that a team member made a correct judgement (i.e. determining that a team member correctly judged that a jar had over or under 300 items). Granting participants 10 raffle entries per evaluation offered the opportunity to earn up to 60 raffle entries for completing the study. Participants were also told that if they were a judge, the team member that they sided with (i.e. whoever’s judgment they agreed with) would also receive 10 raffle entries. In the interest of equality, however, each participant actually received 60 raffle entries for completing the study.

Participants accessed the study on their mobile devices with internet access, where they also completed consent forms and were given more specific information about the study, including information related to two manipulations (discussed below). At this point participants
received their role assignment as a judge and were shown the sample jar. Then, they proceeded
to evaluate whether each of the six presented jars contained more, less, or exactly 300 items. The
jars were presented in the same order for all participants, though this order was determined at
random. Participants were given 10 seconds to look at the picture of the jar. Participants were
given feedback from other team members after each jar was presented (e.g. “Subject 1 said it was
‘Over’”). Participants could then side with Subject 1, Subject 2, both or neither. They were then
told which subject was correct. As Subjects 1 and 2 were “virtual”, this feedback was
predetermined and not based on actual performance.

This procedure mirrors the experience of making a bad call, in that the decision-maker
was presented with information upon which he or she had to make a decision—in this case the
jar and the other team members’ evaluations. They were then given information as to who was
correct, which would provide an indication as to whether a bad call was made. Given that team
members who received the endorsement of the participant were also given raffle entries, there
was some explicit cost to making a bad call. They were, however, afforded subsequent
opportunities to make other evaluations, which would also present as opportunities for makeup
calls. This is not unlike a supervisor who picks from two sales pitch proposals from two
subordinates, only to have the one they selected fail to secure a client. When presented with
another opportunity between the two subordinates, the supervisor may err on the side of the
previously unselected subordinate to make up for their past mistake.

Following evaluation of each jar, participants completed survey items measuring their
feelings about their decision. After the final evaluation, participants completed survey items
measuring their liking of team members, a manipulation check item, and demographic
information. Lastly, participants were debriefed and dismissed.
**Experimental Manipulations**

**Stakes.** Participants were shown one of two screens, which were presented to subjects at random. One screen told participants what each raffle entry would be worth. For the low stakes manipulation, each raffle entry earned from a correct evaluation increased the chances of winning by 1%, meaning that if they performed well they could walk away with a 6% chance of winning. For the high stakes manipulation, each raffle earned from a correct evaluation increased the chances of winning by 2%, meaning that if they performed well they could walk away with a 12% chance of winning.

**Outcome interdependence.** Participants were presented with two teams: red or blue. They interacted with a subject (predetermined responses) from each team at the same time (i.e. Subject 1 was blue, Subject 2 was red). Participants were told that those on the blue team would not share raffle entries with team members. This meant that correct evaluations by one team member benefited only that team member, which created a low level of outcome interdependence. Participants were also told that those on the red team would share raffle entries with team members. This meant that correct evaluations by one team member benefited all members of that team. Shared raffle entries by participants on red teams created a high level of outcome interdependence by making incentives dependent on the success of team members.

**Measures**

**Guilt.** Guilt was assessed using the six guilt-related items from the Positive and Negative Affect Schedule (PANAS-X; Watson & Clark, 1994): “guilty,” “ashamed,” “blameworthy,” “angry at self,” “disgusted with self,” and “dissatisfied with self.” Participants were told to read each item and mark the appropriate answer in the space next to that word that would indicate the extent to which they felt this way during the judgement process they just went through. The
items were presented after each round of judgment and decision-making had ended. They were provided with a 5-point scale to respond to these items, with 1 = Very slightly or not at all, and 5 = Extremely. The items appeared with other PANAS-X items (e.g. delighted, alone) so participants could not guess what our specific outcome of interest was. Reliability measures for all guilt-related items were acceptable (α > .89).

**Liking.** Liking was also assessed as a control variable for both subjects 1 and 2 using items adapted from Wayne and Ferris (1990): “How much do you like this person?”, “I would get along well with this person,” “Working with this person is a pleasure,” “I think this person would make a good friend.” Reliability measures for liking were acceptable (Subject 1 α = .76, Subject 2 α = .82)

**Demographics.** Participants were asked to provide information regarding their age, time at the university, employment status, race/ethnicity, gender, and English-as-first-language status.

**RESULTS**

**Manipulation Checks.**

For the stakes manipulation, we asked participants how much raffle entries were worth: 1%, 2%, 3%, 4%. Participants assigned to each group responded consistently with the manipulation, with the low stakes group reporting a lower percentage (M = 1.03) than the high stakes group (M = 2.04; F(1, 203) = 720.2, p < .001). We also asked participants which team shared its raffle entries. 88.8% of participants correctly selected “Red” in response to this question (χ² = 123.322, p < .001). As such, we consider our manipulations to be successful.

**Hypothesis Testing**

We examined our hypothesis through three analysis sets. First, to calibrate our analysis, and test whether our experimental data reflects makeup call patterns similar to the data from
Study 1, we sought to replicate our analysis of Hypothesis 1 with the archival baseball data by constructing a model analogous to equation (1). Specifically, we assessed the likelihood that a judgment call would be made in favor of Subject 1 (see Table 4), with both prior bad calls against Subject 1 and prior bad calls against Subject 2 as predictors. As in the baseball data, the bad call variables represent the lagged (by one round) cumulative count of bad calls over the six rounds of the experiment (i.e., the number of prior bad calls up to and including the five previous rounds). All estimates were modeled within subjects with bootstrapped standard errors (1,000 replications). As shown in Table 4, the likelihood of a judgment in favor of Subject 1 was positively related to a bad call against Subject 1 ($p < .01$) and negatively related to a bad call against Subject 2 ($p < .01$). In other words, when a bad call was made against Subject 1, s/he was more likely to be favored in the subsequent call.

Having established the presence of the same effects observed in our baseball data, we turned our attention to testing our hypotheses related to the mechanisms behind makeup calls. Accordingly, we designed our second and third sets of analyses to model makeup calls directly. Specifically, in both sets of analyses, the outcome of interest was the presence of a makeup call (i.e., a bad call following another bad call) and the key independent variable was the lagged (by one round) cumulative count of bad calls over the six rounds of the experiment. With respect to Hypothesis 3, our second analysis tested guilt as a mediator in the relationship between bad calls and makeup calls using the bootstrapping methods recommended by Hayes (Hayes, 2009; Hayes, 2013) (see Table 5). To carry out the analysis we estimated two models, (i) one with guilt as the outcome and bad calls as the predictor, and (ii) one with makeup calls as the outcome and guilt
and bad calls as the predictors. To estimate indirect effects, we rely on the product of coefficients method, using bootstrapped and bias-corrected confidence intervals to test for statistical significance. All estimates were modeled within subjects. We found, consistent with our expectations, that the effect of bad calls on subsequent makeup calls was mediated by guilt. Namely, the indirect effect via guilt was statistically significant (Indirect effect = .008, $SE = .003$, 95% CI [.004, .015]). According to our results, the percent of the total effect mediated by the indirect effect was 21.8%. As such, we found support for Hypothesis 3 that the tendency towards makeup calls is a function of guilt.

With respect to Hypotheses 4b and 4c, we conducted a moderated mediation analysis, using the methods described by Preacher, Rucker and Hayes (2007) and others (Hayes, 2013). More specifically, Hypotheses 4b and 4c, taken together, contemplate the estimation of a series of models designed to test for what Preacher and colleagues described as “When the $a$ path is moderated by $W$ and the $b$ path is moderated by $Z$” (2007: 197). In our case, the “$a$ path” is the relationship between bad calls and guilt, and the $b$ path is the relationship between guilt and makeup calls, with $W$ represented by stakes, and $Z$ represented by interdependence. To test these relationships, we estimate a pair of models, (i) one with guilt as the outcome and bad calls and the bad calls/high stakes interaction as predictors, and (ii) one with makeup calls as the outcome, and bad calls, the bad calls/high stakes interaction, guilt, interdependence, and the interdependence/guilt interaction as predictors. These models allow us to test our hypotheses in isolation using (i) to test for the moderating effect of stakes and (ii) to test for the moderating effect of interdependence. But, taken together, these models allow us to take the analysis a step further, and model conditional indirect effects (i.e., the presence of moderated mediation) using
the modified product of coefficients methods described by Preacher and colleagues (2007). However, to test for the presence of moderated mediation, it is not enough to assess whether a conditional indirect effect is significant at a given level of the moderator. Rather, we need to be able to assess whether the conditional indirect effects at high levels of the moderator are statistically different from the conditional indirect effects at low levels of the moderator. To do so, we simply calculate the point estimate of the difference between the conditional indirect effect when stakes (interdependence) are high and when stakes (interdependence) are low. We then use bootstrapping methods to estimate the sampling distribution of this difference in conditional indirect effects, using that information to construct bootstrapped and bias-corrected confidence intervals.

As the coefficients in Table 6 suggest, the interaction between bad calls and stakes in terms of its predicted effect on guilt (Model 1) was negative and significant at conventional levels (95% CI [-.01715, -.0222]). Specifically, when stakes are higher, the impact of making a bad call on guilt is weaker. The negative direction of the estimated effect, however, is not consistent with our expectations. Although in the expected direction, the estimates examining the interaction of interdependence and guilt on makeup calls was not significant at conventional levels (95%CI [-.025, 0.186]). As noted previously, we also calculated estimates using a moderated mediation approach for both moderators. In terms of indirect effects, only the difference in conditional indirect effects for stakes was significant (95% CI [-.0194, -.0035]). This result suggests that as stakes get higher, the indirect effect of bad calls on makeup calls, via guilt, becomes weaker. This does not provide support for Hypothesis 4b, though it lends notable support to the important role that stakes and guilt play in this process. Notably, we did not have
similar findings with respect to a conditional indirect effect with respect to interdependence of outcomes and therefore failed to support Hypothesis 4c.

----------------------------------
Insert Table 6 about here
----------------------------------

STUDY 2 DISCUSSION

Our findings provided some support for our hypotheses. First, Hypothesis 1 was still supported: bad calls led to an increase in makeup calls that favored the party previously harmed. This evidence suggests that the findings from Study 1 are not just limited to the baseball diamond or even sports at large, but rather extend to other settings. Though our setup was slightly artificial as it was an experiment, it nevertheless provides evidence that makeup calls can take place in other organizational settings. Second, Hypothesis 3 was also supported, with guilt having a significant mediating effect when it came to accounting for bad calls’ effect on makeup calls. This suggests that actors who provide makeup calls may be doing so because of the guilt they experience, which is consistent with previous research on atonement (e.g. Baumeister et al., 1994, 1995, 2001; Regan, 1971; Schmader & Lickel, 2006; Tangney, 1991, 1995). Returning to another of our previous illustrations, this suggests that a manager who fails to adequately evaluate a subordinate thereby having an impact on their merit increase may encourage them to feel guilt. This guilt would in turn prompt them to find some way of making up for their error, even if it means compromising the rules or regulations, such as recommending the person for a promotion or title change even when not entirely merited.

Finally, we only provided some evidence for our moderating effects. Though stakes did moderate our outcomes, it was in the opposite direction. Namely, as stakes became higher, the indirect impact of guilt on makeup calls became weaker. This suggests, consistent with Study 1,
that while guilt does have an impact on behaviors, when stakes are high the impact is minimized. While a supervisor may feel guilty for neglecting to adequately evaluate a subordinate, that guilt’s effect may be quieted when the stakes of their decisions are much higher. This finding is not consistent with past findings that show that makeup calls are more likely as stakes get higher (Davis & Lopez, 2015).

**General Discussion**

The results from our two studies generally provide support for our hypotheses. First, with regard to our expectation that bad calls would predict makeup calls (i.e. bias in future calls favoring the harmed party), we found evidence across both studies to support this expectation. In our first study, archival data from the MLB suggested that umpires made calls that favored both pitchers and batters when they previously erred; our second study supported this with data from an experiment involving undergraduate students. Second, we also found some support for our second hypothesis, in which we argued that makeup calls could spill over to other individuals on teams impacted by the initial bad call. The data from our first study indeed suggested that umpires did make calls favorable to a batter’s teammate when the batter had received a bad call. Similarly, we also found support for our third hypothesis. We argued that this effect of bad calls on makeup calls would be accounted for by guilt. We found support for this in our second study, with our results showing that there was a significant indirect effect of guilt accounting for the impact of a bad call on makeup calls. However, this effect should be couched within the results that we found regarding our fourth and final hypothesis. Though we predicted that stakes would moderate the impact of bad calls on makeup calls such that the effect becomes stronger when stakes are higher, we actually found the opposite effect between our two studies. Not only did the effect become weaker as stakes became higher, but the indirect effect of guilt on makeup
calls also became weaker as stakes got higher. Finally, we failed to find an effect for interdependence of outcomes, which did not support one component of our fourth hypothesis.

**Theoretical Implications**

One of the most critical propositions that we put forth in our paper is that bad calls have a moral quality that provokes perceptions of injustice and subsequent deontic reactions among those who make the bad call. As noted earlier, it is unclear whether makeup calls would reflect error and feedback mechanisms (see Frese & Keith, 2015) rather than a deontic reaction (Folger, 2001). In other words, does the decision-making process pertaining to makeup calls have a moral or ethical quality to it? Additionally, though fairness theory (Folger & Cropanzano, 2001) and the deontic perspective generally (see Folger, 2001) are applied to others’ behaviors, we proposed that the principles could extend to one’s perceptions of one’s own behavior. Though our findings do not directly provide evidence regarding perceptions of fairness, we do have evidence supporting that deontic reactions are provoked on the part of actors. Guilt emerged when bad calls were made. Given guilt’s status as a moral emotion that is provoked in response to wrongdoing (see Eisenberg, 2000), this finding suggests that deontic reactions emerge even when one is the source of the perceived injustice. Our results also suggest that guilt provoked behavioral responses (i.e. makeup calls) that restored the state of fairness. This process—negative reactions in response to perceived injustice resulting in reparation—conforms to the processes outlined by Folger (2001) when it comes to deontic reactions. Our observation of subsequent makeup calls in response to bad calls as such suggests that moral sanctioning is not limited to those who harm us or those who harm others, but even our own harm of others. Furthermore, considering that deontic reactions are concerned with moral violations and moral
issues, these findings as such suggest, consistent with arguments made by Hamilton (2011) and Russell (2003), that makeup calls are a matter of moral decision-making.

Relatedly, our findings also have some bearing on how the deontic self-sanctioning process unfolds. While we have articulated that guilt would be provoked in response to perceived injustice at one’s hand and would prompt atonement, such a process does not conform to other theoretical perspectives regarding ethical decision making nor justice actors. The willingness to engage in objectively questionable behavior defies principles of ethical decision-making, such as those described by Trevino (1986). For example, Trevino noted that when there is a strong normative culture surrounding ethical behavior and codes of conduct there should be more agreement regarding what is ethical and presumably how people will behave. Considering that most—if not all—rules of play in sports do not explicitly allow makeup calls, the expectation would be that they should not emerge. In addition, Scott and colleagues (2013) argue that guilt may inhibit unfair behaviors rather than encourage them from those who are responsible for justice. However, we find that guilt as a result of a bad call can actually contribute to unfair behaviors (i.e. making a call that favors someone when not merited; see Adams, 1965). As such, the process of ethical behaviors, particularly as it pertains to guilt, may not always be one of inhibiting injustice or unethical behaviors, but rather may be one aimed at restoring balance as suggested by the deontic perspective (Folger, 2001) and others (see Pedowitz, 2008). As such, models of ethical decision-making and enactment of justice may need to be revised to account for the potential that unethical or normatively inappropriate behaviors could serve moral means (e.g. whistleblowing; Near & Miceli, 1985; see also, Watts & Buckley, 2017).
However, deontic reactions do seem to have its limits. Consistent with other hallmarks of ethical decision-making—namely, that situational influences can impact moral behaviors (Jones, 1991; Trevino, 1986)—stakes were found to influence both emotional experiences and decision making by setting constraints on how guilt impacts subsequent makeup calls. Specifically, situational factors that pertain to the potential impact of a decision (i.e. greater stakes) may result in more deliberate and cautious decision-making and game calling. This specifically suggests that when the actions that would re-establish justice present a disproportionate response, that actors may refrain from engaging in reparation. Such findings are consistent with other research showing that punishment and retribution favors proportionality and that stakes may limit the influence of the deontic reaction (Ball, Trevino, & Sims, 1994; Carlsmith, Darley, & Robinson, 2002; cf. Trevino & Ball, 1992). When taken into consideration with empirical evidence suggesting that greater social pressure leads to more favoritism (e.g. Lopez, 2016; Pettersson-Lidbom & Priks, 2010), our results, on the other hand, would suggest that greater situational pressure may lead to *less* favoritism considering the stifling effect we observed on makeup calls when stakes became higher. It must also be mentioned here that situational factors did not have a universal impact on our effects. Outcome interdependence was shown not to have an effect, which is inconsistent with both theory (Jones, 1991) nor empirical research (e.g. Lopez & Snyder, 2013). The lack of evidence however, may not suggest that the number of individuals affected does not figure into the moral calculus of a makeup call; rather, in conjunction with stakes, interdependence of outcomes simply may not have as strong an effect on decision-making. Taken together, the set of constraints applied by situational factors mitigates the power of ethic quality in calling makeup calls after bad calls and as such is helpful in predicting the act of atonement.
Practical Implications

There are discrepancies regarding whether makeup calls exist, with some journalists (e.g. Wyshynski, 2013) and researchers (e.g. Davis & Lopez, 2015; Gift, 2012; Lopez & Snyder, 2013; Mills, in press; Warneke & Ogden, 2012) suggesting they do while former referees (e.g. Javie, n.d.) and leagues (see Lowe, 2014) do not. Our research suggests that makeup calls do exist, and that the stakes of a situation may mitigate their emergence in response to bad calls. Such findings can be used by organizations broadly, and sports organizations in particular, as makeup calls emerged in our paper even outside the context of a sport competition. Specifically, our findings shed light on a neglected problem that if left ignored out of concerns for reputation or liability, could ultimately negatively impact an organization if proof of specific makeup calls are brought to light. For instance, should an upper-level manager or vice president become aware that an employee may be atoning for past mistakes through other non-sanctioned means, s/he may feel bolstered by the knowledge that makeup calls exist to step in and intervene rather than conveniently look away with the false sense of security that makeup calls are simply a fallacy. Such intervention could prevent more problematic behaviors that if left unchecked could harm the organization in the long run.

Related to previous points regarding ethical decision-making, this study has implications for the challenges faced by managers when engaging in ethical decision-making. Specifically, bad calls may produce other calls or decisions that we would label “bad” as a means to provide equity or equality. Rather than being based on absolute performance or specific criteria for judgment, however, decisions may be made based on a calculus that includes past errors on the part of a decision-maker. This then calls into question whether a makeup call—though an act of atonement—is actually ethical. Indeed, there is some legitimate criticism as to whether a
makeup call restores justice or impedes it (see Hamilton, 2011; Russell, 2003) and should the
decisions in question compromise behaviors that could be seen as illegal or unethical, may make
the decision-maker liable to punishment by other authorities no matter how pure their intentions
were. Consider, for example, a hiring manager whose attention is brought to racial bias and its
impact on hiring (see Jones, Sabat, King, Ahmad, McCausland, & Chen, in press). The manager,
due to the heightened awareness of this relationship and guilt over potentially having favored
White candidates in the past, may intentionally choose to hire a candidate who is a racial
minority. If such an intention is uncovered, though noble in its pursuit, it would be grounds for
discrimination lawsuit (see Pyburn, Ployhart, & Kravitz, 2008). As such, given our evidence of
makeup calls, it behooves organizations to identify other mechanisms for establishing fairness
that would not rely on subsequent bad calls or makeup calls that are liable to scrutiny.

Finally, our findings have some implications for performance evaluations and feedback.
Research has shown that positive affect has a positive impact on employee evaluations (Sinclair,
1988; see also Cardy & Dobbins, 1986). In other words, the more positive your feelings, the
more likely your subordinate will receive a positive review. On the other hand, the more
negative one’s feelings, the more likely one will provide a negative review. Our results show,
though, that this relationship may not apply to all affective experiences or all situations. Even
though there may be explicit rules governing evaluation of performance, particularly in the
context of sports competitions, it seems that negative emotions may produce positively biased
evaluations. Moreover, considering that performance evaluations may also be impacted by
positive attitudes toward evaluates influenced by preference, favoritism, and status, the addition
of negative emotion to the mix may create a situation ripe for bias and preferential treatment. As
such, organizations need to caution managers against heeding their guilt when evaluating performance, as it may also lead to bias.

**Limitations and Future Research**

Though our paper aimed to examine bad calls and their impact on makeup calls, we acknowledge that bad calls can be characterized as errors. Such a characterization, however, may suggest that the effects observed in our studies could be due to learning or damage control (see Frese & Keith, 2015) rather than atonement. Research has shown that when umpires make a bad call, coaches may remind them of their fault (see Warneke & Ogden, 2012). And, officials in the NBA may engage in higher scrutiny of opposing teams following a foul or a judgment call made against them (Gift, 2012). In fact, even announcers may influence the way in which calls are made. As Tim Donaghy (2009) noted, “It was a lot easier to hear from an announcer that you missed something…the than to get an email or a phone call from the league office saying that your blown call had cost a team a game,” (p.79). Indeed, research has shown that feedback, in addition to other reinforcements, can influence performance (Stajkovic & Luthans, 2003). Verbal feedback may also impact behavior, as research has shown that positive and negative verbal feedback can lead to fewer errors in performance (e.g. Kannappan, Yip, Lodhia, Morton, & Lau, 2012). Considering how often officials receive verbal feedback from players, coaches, and fans, such an effect may account for the impact of bad calls on makeup calls that we observed in our data. Moreover, officials’ performance may be impacted by the negative feedback they receive when making a bad call, especially considering their goals of avoiding making bad calls in the first place (see Idson & Higgins, 2000; Van-Dijk & Kluger, 2004). The inability to disentangle guilt’s effects from other possible explanations (e.g. feedback, learning) presents a limitation to our study that should be addressed in future research.
Related, our study did not test the impact of other emotions on makeup calls. Other moral emotions (e.g. pride; see Tangney, Stuewig, & Mashek, 2007) may serve to influence umpires and other decision makers, as well. Moral emotions emerge as a result of evaluations of conformity to both internal and external standards, generally speaking. Considering the influence of others in sports decisions (e.g. coaches; see Warneke & Ogden, 2012) as well as the broader organizational domain in which multiple parties may be privy to one’s decisions, other emotions could function to impact bad calls and makeup calls. For example, pride in good decision-making and judgment may lead umpires to course-correct in their pitch calls, rather than guilt alone. Future research should examine these other emotions that may come into play when accounting for the effect of a bad call on makeup call to determine the extent to which guilt is a unique contributing affective experience.

In addition, our study did not consider how social forces may impact decision-making. Researchers have shown evidence for officials’ preference toward home-field teams (Dohmen & Sauermann, 2016; Zitzewitz, 2014) or toward players (Parsons, Sulaeman, Yates, & Hamermesh, 2007). For example, after examining MLB games during 2004-2006, Parsons et al. (2007) found that when umpire and pitcher match race or ethnicity, strikes were more likely to be called. Officials may also be biased by star power. According to analyses of the NBA’s “Last Two Minute Reports (i.e. reports on the last two minutes of a game when the score was within 3 points in 2017-2018 season and applicable overtimes; NBA, 2018), the most bad calls went against the Brooklyn Nets—a team ranked 12th out of 15 teams in the Eastern conference (Herring & Paine, 2018). According to Herring and Paine, this may specifically be attributable to the lack of a star player on their team. Offering contrary evidence, Deutscher (2015) showed that there was no favoritism toward home-field team, superstars, and same nationality players after
analyzing “Last Two Minute Reports” of games from the 2014-2015 season. Although favoritism from an official is an ambiguous issue in sports, bias caused by favoritism has been shown to influence subjective performance evaluation in management scholarship more broadly (Bol & Smith, 2011; Du, Tang, & Young, 2012; Prendergast & Topel, 1993). Therefore, the influence of social factors on makeup calls (i.e. favoritism, preference, in-group/out-group bias) of umpires should be examined in future studies to delineate to what extent variance in makeup calls is attributable to interpersonal dynamics.

Finally, future research should explore how individual differences may impact this process. Specifically, given the moral nature of bad calls and makeup calls (see Hamilton, 2011), researchers should explore how traits pertaining to morality and moral behavior such as moral development (Kohlberg, 1958; see Trevino, 1986) and moral identity (Aquino & Reed, 2002) may impact this process. Specifically, those who operate at higher levels of moral reasoning may perform a different calculus when deciding whether to provide makeup calls when compared to those at lower levels. Though we did explore the situational factors that may influence the impact of bad calls on makeup calls, future work could illuminate how these features act in conjunction with individual differences and whether individual differences would mitigate against the tendency to call makeup calls and buffer against the situational features that amplify this effect.

CONCLUSION

The current research used archival Major League Baseball data and a laboratory study to explore the relationship between bad calls and makeup calls. First, our analysis of Major League Baseball data indicated that bad calls lead to an increase in makeup calls. Additionally, our laboratory study replicated the influence of bad calls on makeup calls and indicated that guilt
was a significant mediator of the relationship between bad calls and makeup calls whose influence decreased as stakes became higher. Together these results provide evidence that makeup calls occur in organizations and are related to guilt but diminish as the stakes of decision-making situations increase. Our findings show that decision making in organizations is a complex process that involves guilt and reparation for past retributions. It is our hope that this research will serve as a basis for future research on decision making, guilt, and retribution in organizations.
Acknowledgements: We would like to thank Feiou Zhang for his assistance with collecting data for this paper. The research for this paper was supported by an Internal Research Award from the University of Texas at San Antonio.
REFERENCES


Berg, T. 2016, November 1. Umpiring a World Series game is extremely hard and MLB umps are great at it. Available at: https://ftw.usatoday.com/2016/11/world-series-indians-cubs-umpires-umpiring-is-super-hard-mlb


Davis, N., & Lopez, M. 2015, May 15. Hockey refs are out to get you (if they already got the other guy). Available at: https://fivethirtyeight.com/features/hockey-refs-are-out-to-get-you-if-they-already-got-the-other-guy/


Du, F., Tang, G., & Young, S. M. 2012. Influence activities and favoritism in subjective
performance evaluation: Evidence from Chinese state-owned enterprises. The


Psychology, 51: 665-697.

outcomes, and ways to avoid regret. Journal of Social and Clinical Psychology, 26: 479-
504.


Flannery, B. L., & May, D. R. 2000. Environmental ethical decision making in the U.S. metal-

Folger, R. 2001. Fairness as deonance. In S. W. Gilliland, D. D. Steiner, & D. P. Skarlicki (Eds.),
Research in Social Issues in Management: 3-33. Greenwich, CT: Information Age.

& R. Cropanzano (Eds.), Advances in Organizational Justice: 1-55. Stanford, CA:
Stanford University Press.

Frese, M., & Keith, N. 2015. Action errors, error management, and learning in organizations.

Ganegoda, D. B., & Folger, R. 2015. Framing effects in justice perceptions: Prospect theory and

Gift, P. 2012. Do two wrongs make a right in NBA officiating? An analysis of referee bias in
make-up call situations. Paper presented at MIT Sloan Sports Analytics Conference,
Boston, MA.


Granillo, L. 2013, Aug 1. When umpires admit their mistakes. Available at:
https://www.sbnation.com/2013/8/1/4577394/when-umpires-admit-their-mistakes

Green, E., & Daniels, D. P. 2014. What does it take to call a strike? Three biases in umpire
decision-making. Paper presented at MIT Sloan Sports Analytics Conference, Boston,
MA.


Herring, C., & Paine, N. 2018, March 2. Which NBA team is wronged by the refs the most? Available at: https://fivethirtyeight.com/features/which-nba-team-is-wronged-by-the-refsthe-most/


Pitch Types. MLB. Available at: [http://m.mlb.com/glossary/pitch-types](http://m.mlb.com/glossary/pitch-types)


Wyshynski, G. 2013, April 9. *Kerry Fraser admits refs use ‘makeup calls’ after questionable penalties; also, water is wet*. Available at: https://sports.yahoo.com/blogs/nhl-puck-daddy/kerry-fraser-admits-refs-makeup-calls-questionable-penalties-205311386--nhl.html


**TABLE 1**

Study 1 Means, Standard Deviations, and Correlations for the Variables Measured

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>CalledStrike</td>
<td>0.34</td>
<td>0.47</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MissedPitcher</td>
<td>0.10</td>
<td>0.32</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>MissedBatter</td>
<td>0.12</td>
<td>0.34</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>MissedTeamatBat</td>
<td>0.29</td>
<td>0.51</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Leverage</td>
<td>0.96</td>
<td>0.89</td>
<td>0</td>
<td>9.27</td>
</tr>
<tr>
<td>BattingDiff</td>
<td>-0.10</td>
<td>2.81</td>
<td>-11</td>
<td>11</td>
</tr>
<tr>
<td>HomeBatter</td>
<td>0.50</td>
<td>0.50</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Strike</td>
<td>0.25</td>
<td>0.43</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 0-0</td>
<td>0.35</td>
<td>0.48</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 0-1</td>
<td>0.13</td>
<td>0.33</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 0-2</td>
<td>0.06</td>
<td>0.23</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 1-0</td>
<td>0.11</td>
<td>0.32</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 1-1</td>
<td>0.09</td>
<td>0.29</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 1-2</td>
<td>0.07</td>
<td>0.26</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 2-0</td>
<td>0.04</td>
<td>0.20</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 2-1</td>
<td>0.04</td>
<td>0.19</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 2-2</td>
<td>0.05</td>
<td>0.22</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 3-0</td>
<td>0.02</td>
<td>0.14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 3-1</td>
<td>0.02</td>
<td>0.13</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Count: 3-2</td>
<td>0.02</td>
<td>0.15</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Change</td>
<td>0.08</td>
<td>0.27</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Curveball</td>
<td>0.11</td>
<td>0.31</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Eephus</td>
<td>0.00</td>
<td>0.01</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Four-Seam</td>
<td>0.00</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Cutter</td>
<td>0.07</td>
<td>0.26</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Fastball</td>
<td>0.36</td>
<td>0.48</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Forkball</td>
<td>0.00</td>
<td>0.02</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Splitter</td>
<td>0.01</td>
<td>0.09</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Two-Seam</td>
<td>0.14</td>
<td>0.34</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Knuckle-curve</td>
<td>0.01</td>
<td>0.09</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Knuckleball</td>
<td>0.00</td>
<td>0.02</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Sinker</td>
<td>0.09</td>
<td>0.29</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Slider</td>
<td>0.13</td>
<td>0.34</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PitchType: Other</td>
<td>0.00</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Up-Left</td>
<td>0.02</td>
<td>0.15</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Up-Center</td>
<td>0.02</td>
<td>0.14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Up-Right</td>
<td>0.02</td>
<td>0.14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Middle-Left</td>
<td>0.03</td>
<td>0.17</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Middle-Center</td>
<td>0.03</td>
<td>0.17</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Middle-Right</td>
<td>0.03</td>
<td>0.18</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Variable</td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>-------------------</td>
<td>------</td>
<td>-----------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Zone: Low-Left</td>
<td>0.03</td>
<td>0.18</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Low-Center</td>
<td>0.03</td>
<td>0.17</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Low-Right</td>
<td>0.04</td>
<td>0.18</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Out-UpLeft</td>
<td>0.18</td>
<td>0.39</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Out-UpRight</td>
<td>0.13</td>
<td>0.33</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Out-BotLeft</td>
<td>0.22</td>
<td>0.41</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Zone: Out-BotRight</td>
<td>0.22</td>
<td>0.42</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

*Notes:* Due to collinearity, “Zone: Out-BotRight” is excluded from regression results.

### TABLE 2

**Study 1 Logit Results**

<table>
<thead>
<tr>
<th>Variable</th>
<th>(1) Called Strike</th>
<th>(2) Called Strike</th>
<th>(3) Called Strike</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modeling Likelihood of a Called Strike</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MissedPitcher</td>
<td>1.13***</td>
<td>1.15**</td>
<td>1.32***</td>
</tr>
<tr>
<td>MissedBatter</td>
<td>0.81***</td>
<td>0.77***</td>
<td>0.81**</td>
</tr>
<tr>
<td>MissedTeamatBat</td>
<td></td>
<td>0.90***</td>
<td>0.86***</td>
</tr>
<tr>
<td>Leverage</td>
<td>0.94***</td>
<td>0.94***</td>
<td>0.95**</td>
</tr>
<tr>
<td>MissedPitcher x Leverage</td>
<td></td>
<td></td>
<td>0.87**</td>
</tr>
<tr>
<td>MissedBatter x Leverage</td>
<td></td>
<td></td>
<td>0.95</td>
</tr>
<tr>
<td>MissedTeamatBat x Leverage</td>
<td></td>
<td></td>
<td>1.05</td>
</tr>
</tbody>
</table>

| Observations (Calls)     | 34,347            | 34,347            | 34,347            |
| Games                    | 239               | 239               | 239               |
| Umpires                  | 61                | 61                | 61                |
| Pseudo R2                | 0.43              | 0.43              | 0.43              |

*Notes:* Estimate are reported in the forms of odds ratios. All models include game, pitcher, and batter fixed effects. Controls: Inning, Score Differential of Batting Team, PITCHf/x Strike, Pitch Location, Pitch Type, Pitch Count, Home Batter. Standard Errors are clustered by umpire.

*p < .10

**p < .05

***p < .01
TABLE 3
Study 1 Robustness Checks

<table>
<thead>
<tr>
<th>Variable</th>
<th>(1) Called Strike</th>
<th>(2) Called Strike</th>
<th>(3) Called Strike</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modeling Likelihood of a Called Strike</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MissedPitcher</td>
<td>1.32***</td>
<td>1.26***</td>
<td>0.04***</td>
</tr>
<tr>
<td>MissedBatter</td>
<td>0.82**</td>
<td>0.81**</td>
<td>-0.04***</td>
</tr>
<tr>
<td>MissedTeamatBat</td>
<td>0.86***</td>
<td>0.86***</td>
<td>-0.02***</td>
</tr>
<tr>
<td>Leverage</td>
<td>0.94**</td>
<td>0.94**</td>
<td>-0.01**</td>
</tr>
<tr>
<td>MissedPitcher x Leverage</td>
<td>0.87*</td>
<td>0.92</td>
<td>-0.02**</td>
</tr>
<tr>
<td>MissedBatter x Leverage</td>
<td>0.94</td>
<td>0.94</td>
<td>-0.01</td>
</tr>
<tr>
<td>MissedTeamatBat x Leverage</td>
<td>1.04</td>
<td>1.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Model</td>
<td>Logit</td>
<td>Logit</td>
<td>LPM</td>
</tr>
</tbody>
</table>

Observations (Calls)                           | 34,341            | 34,341            | 30,097            |
Games                                          | 239               | 239               | 239               |
Umpires                                        | 61                | 61                | 61                |
Pseudo R2                                      | 0.43              | 0.43              | 0.43              |

Notes: Estimate are reported in the forms of odds ratios. All models include game, pitcher, and batter fixed effects. Controls: Inning, Score Differential of Batting Team, PITCHf/x Strike, Pitch Location, Pitch Type, Pitch Count, Home Batter. Standard Errors are clustered by umpire.

* $p < .10$
** $p < .05$
*** $p < .01$

TABLE 4
Study 2 Logit Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef.</th>
<th>SE</th>
<th>z</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modeling Likelihood of Judgment (in favor of player 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bad Calls (vs 1)</td>
<td>0.41</td>
<td>0.04</td>
<td>10.67</td>
<td>[0.33, 0.48]</td>
</tr>
<tr>
<td>Bad Calls (vs 2)</td>
<td>-0.31</td>
<td>0.02</td>
<td>-12.69</td>
<td>[-0.35, -0.26]</td>
</tr>
</tbody>
</table>
# TABLE 5

**Study 2 Test of Mediation by Guilt**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Outcome</th>
<th>Coef.</th>
<th>SE</th>
<th>z</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: Bad Calls</td>
<td>Guilt</td>
<td>0.08</td>
<td>0.02</td>
<td>4.33</td>
<td>[0.04, 0.12]</td>
</tr>
<tr>
<td><strong>Model 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B: Bad Calls</td>
<td>Makeup Calls</td>
<td>0.03</td>
<td>0.02</td>
<td>1.76</td>
<td>[-0.00, 0.06]</td>
</tr>
<tr>
<td>C: Guilt</td>
<td>Makeup Calls</td>
<td>0.10</td>
<td>0.03</td>
<td>3.47</td>
<td>[0.04, 0.15]</td>
</tr>
<tr>
<td>D: Indirect Effect (C x A)</td>
<td></td>
<td>0.01</td>
<td>0.00</td>
<td></td>
<td>[0.00, 0.02]</td>
</tr>
<tr>
<td><strong>Percent Mediated</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D / (D+B))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21.8%</td>
</tr>
</tbody>
</table>

*Notes:* Within Subjects Estimates (i.e., models include subject fixed effects)  
[BC] Bootstrapped, Bias-Corrected Confidence Interval  
*p < .10  
**p < .05  
***p < .01
# TABLE 6
Study 2 Moderated Mediation Analysis of Stakes and Interdependence on Makeup Calls

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Outcome</th>
<th>Coef.</th>
<th>SE</th>
<th>z</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bad Calls</td>
<td>Guilt</td>
<td>0.13</td>
<td>0.03</td>
<td>4.02</td>
<td>[ 0.06, 0.19]</td>
</tr>
<tr>
<td>Bad Calls x High Stakes</td>
<td>Guilt</td>
<td>-0.10</td>
<td>0.04</td>
<td>-2.54</td>
<td></td>
</tr>
<tr>
<td><strong>Model 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[-0.17, -0.02]</td>
</tr>
<tr>
<td>Bad Calls</td>
<td>Makeup Calls</td>
<td>0.06</td>
<td>0.02</td>
<td>2.86</td>
<td>[ 0.02, 0.11]</td>
</tr>
<tr>
<td>Bad Calls x High Stakes</td>
<td>Makeup Calls</td>
<td>-0.06</td>
<td>0.03</td>
<td>-1.95</td>
<td>[-0.12, 0.00]</td>
</tr>
<tr>
<td>Guilt</td>
<td>Makeup Calls</td>
<td>0.10</td>
<td>0.03</td>
<td>3.39</td>
<td>[ 0.04, 0.16]</td>
</tr>
<tr>
<td>Interdependence</td>
<td>Makeup Calls</td>
<td>0.11</td>
<td>0.02</td>
<td>5.20</td>
<td>[ 0.07, 0.16]</td>
</tr>
<tr>
<td>Guilt x Interdependence</td>
<td>Makeup Calls</td>
<td>0.08</td>
<td>0.05</td>
<td>1.48</td>
<td>[-0.03, 0.18]</td>
</tr>
<tr>
<td><strong>Moderated Mediation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional Indirect Effect</td>
<td></td>
<td>-0.01</td>
<td>0.00</td>
<td></td>
<td>[-0.02, 0.00]</td>
</tr>
<tr>
<td>(High vs. Low Stakes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional Indirect Effect</td>
<td></td>
<td>0.01</td>
<td>0.01</td>
<td></td>
<td>[ 0.00, 0.02]</td>
</tr>
<tr>
<td>(High vs. Low Interdependence)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Notes: Within Subjects Estimates (i.e., models include subject fixed effects)*

[BC] Bootstrapped, Bias-Corrected Confidence Interval

*p < .10

**p < .05

***p < .01