Organizational Behavior and Human Decision Processes 166 (2021) 149–165

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Using allegations to understand selection bias in organizations: Misconduct in the Chicago Police Department

Bryan K. Stroube

London Business School, Regent's Park, London NW1 4SA, United Kingdom

ARTICLE INFO

Keywords:
Selection biases
Misconduct
Allegations
Archival field data
Police

ABSTRACT

Selection biases present a fundamental challenge for research on ethics and misconduct. This issue is well understood at the individual level, where lab studies are often employed to sidestep it at the potential expense of external validity. However, much archival field data on ethics and misconduct are at risk of selection bias originating from *within* organizations, because organizations are typically responsible for evaluating and ultimately documenting who commits misconduct. In this paper I explore the nature and potential scope of this particular form of selection bias, its potential impact on the interpretation of extant findings from the literature, and how studying allegations may help detect it in specific contexts. Using detailed data on formal allegations of police misconduct in Chicago, I highlight how status characteristics such as race and gender may bias the creation of archival data. For example, black officers received allegations at rates similar to white officers but were more likely to have them sustained, and allegations made by black complainants were less likely to be sustained than those made by white complainants—even when including extensive sets of control variables. These findings indicate that accounting for allegations may be a fruitful methodological approach to better understand the optimal use of archival behavioral field data for research on ethics and misconduct.

1. Introduction

Archival field data are often readily available and represent measures of "natural" behavior that may be impossible to replicate in the laboratory. However, it can be difficult to effectively use such data for research on ethics and misconduct, because "dishonest acts are rarely randomly detected and recorded" (Pierce & Balasubramanian, 2015, p. 72). This problem is typically conceived at the individual level, where individual traits may be correlated with both behavior and detection. In this paper I highlight the organizational nature of this problem, given that organizations are ultimately responsible for evaluating behavior and creating the archival samples used by researchers. How likely is an organization to be unbiased in that process, and how might a researcher know if it was not?

I begin by outlining the nature of intra-organizational selection biases in archival behavioral field data. I then introduce allegations as a separate unit of analysis that may help detect the scope of organizational selection bias problems in specific archival data. This is because allegations are typically a necessary but not sufficient condition for individuals to be labeled "wrongdoers" in the type of official archival data created by organizations and studied by researchers. Studying allegations may help researchers detect whether an organization has non-randomly evaluated and recorded behavior, which would produce false negatives or false positives in the data.

The context of police misconduct is used to demonstrate the usefulness of this approach. Police misconduct has become one of the major social issues of the previous decade in the United States, closely related to one of the largest modern protest movements, Black Lives Matter (Chase, 2017; Milkman, 2017). I analyze a set of allegations made against officers in the Chicago Police Department (CPD), one of the largest police departments in the United States and one with a history of misconduct scandals (Hagedorn et al., 2013). The fatal shooting of Laquan McDonald in 2014 by a Chicago officer led to a national reaction and a re-examining of the department.[1] This included an extensive formal investigation by the Department of Justice.[2] It "confirmed that CPD's accountability systems are broadly ineffective at deterring or

E-mail address: bstroube@london.edu.

[1] The mayor of Chicago eventually fired the Police Superintendent for his handling of the incident (Davey & Smith, 2015), and the officer involved was later found guilty of second-degree murder (Smith, Williams, & Davey, 2018).

[2] U.S. Attorney Zachary Fardon noted of the Chicago Police Department investigation: "This is the largest pattern and practices investigation in the history of the Department of Justice" (Fielding, 2016).

https://doi.org/10.1016/j.obhdp.2020.03.003
Received 25 April 2018; Received in revised form 28 February 2020; Accepted 11 March 2020
Available online 1 September 2020

detecting misconduct, and at holding officers accountable when they violate the law or CPD policy" (United States Department of Justice, 2017b, p. 47). These conclusions summarize the challenge of using archival data in many settings: it is unlikely that the official archival samples of officers in Chicago that are punished for misconduct perfectly represent the samples of theoretical interest to researchers, i.e., officers that exhibited certain behaviors.

The 28,588 allegations of misconduct against the Chicago Police Department analyzed in this paper represent some of the most detailed accounts of police misconduct publicly available, including the nature of the allegations and characteristics of the complainants and officers. This allows for a focus on how selection biases may be impacted by three specific status characteristics: race, gender, and organizational affiliation (Berger, Rosenholtz, & Zelditch, 1980). The analysis begins with what is typically the most visible and unambiguous sample of misconduct in most archival field data: individuals that have been officially punished for misconduct. Using multinomial regressions with a range of controls, I find that contingent on an allegation being sustained, black officers were more likely to receive harsher punishment than white officers. Second, I calculate the likelihood that an allegation of misconduct was sustained in the first place, meaning that misconduct by an officer was officially documented. An allegation against a black or Hispanic officer was more likely to be sustained than an allegation against a white officer. However, an allegation coming from a black complainant was less likely to be sustained, and allegations coming from individuals within the Chicago Police Department were much more likely to be sustained than others. Finally, I predict the accrual of allegations at the officer level, where I find that black officers received allegations at rates similar to white officers and that Hispanic officers received allegations at rates lower than white officers, but that both groups were more likely than white officers to receive at least one sustained allegation. Female officers were less likely to receive allegations but were not more or less likely to receive a sustained allegation. Collectively, these multiple stages of analysis highlight how organizational selection biases might decouple the original evaluation of a behavior (i.e., an allegation) from the final consequence of the behavior (i.e., a punishment), so that the punishment documented in the archival sample is not necessarily representative of the behavior of interest. This indicates a potential problem for researchers.

This paper makes three main contributions to the literature on ethics and misconduct. First, it highlights the importance of intra-organizational selection biases as unique from the more frequently acknowledged selection biases related to individual behavior. This is critical for archival field research because data documenting official misconduct are often the outcome of an organizational process. Second, the paper provides guidance for how researchers using archival behavioral data may be able to make progress on understanding the scope of intra-organizational selection biases in specific settings by incorporating allegations as a discrete unit of analysis. With certain assumptions, the biases (or lack thereof) at one stage of the allegation process can be used to infer biases (or lack thereof) at other stages. Third, the paper highlights a number of potential directions for future research related to organization-level traits that may cause biased archival data. These include how variance in organizational culture and organizational policies may be related to the documentation of behavior within organizations. This is a particularly important concern for organization-level research that aims to build and test theory about the relationship between such traits and behavior itself.

2. The challenge of selection biases in organizations

Research on misconduct and ethical behavior faces multiple challenges related to selection bias. First, it faces the same challenge of individual-level selection bias found in any type of archival data. This bias is typically conceptualized as an omitted variable problem, where actors self-select into a behavior based on attributes unobserved by the researcher. This can create endogeneity problems that prevent causal interpretations, because the behavior and the outcome are linked via a mechanism that is unaccounted for by researchers (e.g., Shaver, 1998). The severity of the issue may be greater in studies of misconduct and ethical behavior, because individuals may also attempt to actively obfuscate behavior. Multiple empirical techniques have been developed to more accurately uncover behavior, including direct observation, comparison of multiple measures, examination of incentives that alter the value of obfuscation, and testing against models of "honest" behavior or otherwise efficient markets (Zitzewitz, 2012, p. 734-735). These techniques are often described as "forensic" because they involve determining the likelihood that something has occurred when there is no official record that it has actually occurred.

However, ethics and misconduct research faces a second type of selection bias specifically introduced during the documentation of the behavior. This is an organizational bias because it is not the actor who changes their behavior based on some set of factors but an intra-organizational process that alters the evaluation and documentation of the behavior. Palmer and Yenkey (2015, p. 894) succinctly summarized this challenge for research: "Most prior research on wrongdoing in organizations analyzes publicly available data on official prosecutions. But interpretation of such analyses is problematic because observed associations can reflect the operation of factors leading social control agents to selectively monitor and punish some rather than other wrongdoers as well as the factors leading organizational actors to engage in misconduct."

A bias in the documentation process will create problems for researchers whose goal is to develop general theory about the cause or consequence of a behavior. For example, a high-status actor may be more or less likely to abuse their power but also more or less likely to be punished by an organization when they do. The next section elaborates on the example of status to highlight how the combination of individual selection biases and organizational selection biases makes the use of archival data particularly challenging. Status also provides a theoretical basis for understanding why biases might exist in the empirical context of police misconduct, because race, gender, and occupation are salient individual status characteristics (Berger et al., 1980).

2.1. An illustrating example of status

The literature on status and misconduct presents compelling evidence that status may (1) alter the likelihood of ethical behavior, (2) alter others' evaluation of behavior, and (3) alter punishment for behavior. First, status itself may influence behavior. For example, using a vignette experiment, Bowles and Gelfand (2010) found that experimentally manipulating the experience of being high-status (to the extent possible) led to lower rule compliance. Using archival data, Mishina, Dykes, Block, and Pollock (2010) found evidence that the "prominence" of a firm (measured by inclusion in Fortune's Most Admired Companies list) positively moderated the relationship between performing above expectations and illegal corporate activity. More broadly, sociological theories of "middle status conformity" predict that the choice to deviate from conventional behavior is a function of an actor's status (Phillips & Zuckerman, 2001). Thus, there is evidence that status may influence actual behavior.

Second, status may influence others' evaluation of whether a behavior is right or wrong. This is partly because there is almost always some level of uncertainty involved in observing and interpreting behavior, and status is often used as a proxy to infer other traits such as quality (Simcoe & Waguespack, 2010). In the case of ethics and misconduct, status can lead to both more positive and more negative interpretations of the same behavior depending on context and moderating factors. For example, Fragale, Rosen, Xu, and Merideth (2009) found that evaluators attributed more intentionality to wrongdoing by high-status actors and therefore recommended harsher punishment for the same behavior. However, Kakkar, Sivanathan, and Gobel (2020)

found that the impact of being high-status was a function of the type of status (i.e., status based on "dominance" versus "prestige"). Bowles and Gelfand (2010) also found that the status of the evaluator factored into how much the status of others mattered when interpreting workplace deviance. At the industry level, Sharkey (2014) found that the status of a firm's industry lessened stock market reaction to earning restatements. Thus, there is also evidence that status may influence first-order evaluations of ethically questionable behavior.

Third, status may influence punishment. Although the line between evaluation and punishment can be blurry—and many evaluators (particularly in lab settings) do not have the power to dispense formal punishment—research has examined what leads to variance in punishment contingent on some definition of "guilt." For example, Polman, Pettit, and Wiesenfeld (2013) found that the level of ambiguity about misconduct shaped the level of punishment recommended to higher- versus lower-status individuals; people provided high-status individuals the benefit of the doubt when a transgression was ambiguous, but provided low-status individuals relative sympathy when transgressions were unambiguous. McDonnell and King (2018) found similar results using archival data, where higher-status firms, contingent on being found liable in employment discrimination lawsuits, were punished more severely (in punitive damage awards) by the courts than lower-status ones. Thus, there is also evidence that status may influence levels of punishment.

Collectively, these findings highlight the difficulty of testing and building theory directly from archival data. This is because at the first stage, status has the potential to influence actual behavior. But at the latter two stages—which often occur within organizations—status may alter how the behavior in the first stage is documented in official data. In their article on why sexual harassment claims often fall on "deaf ears," Peirce, Smolinski, and Rosen (1998) provide the example of the pharmaceutical company Astra, USA Inc. (later merged into AstraZeneca): "At Astra, everyone at the firm knew Bildman's [the CEO's] reputation and seventeen-year history of harassing women associates, but most overlooked it, knowing also of his status and record as a successful executive." As a consequence, there would be no official record of his behavior during that period.

This challenge is exacerbated because most studies examine these stages in at least partial isolation, meaning that assumptions are necessary about selection biases at other stages in order to draw conclusions about the results from any single stage. In archival data the most promising path is to create tests that directly address the prior stage, but this is naturally restricted by the limits of what can be observed. For example, McDonnell and King (2018) noted that one limitation of using discrimination lawsuits to measure how status affects juries' treatment of firms is that "discrimination charges are not assigned randomly, and we cannot rule out the possibility that an employer's status and reputation may affect the circumstances in which employees who are discriminated against will bring charges" (p. 81). This challenged is amplified for research that studies not lawsuits but rather wrongdoing recorded by organizations, because the biases in evaluation may be influenced by the organization itself.

2.2. *Organizational traits and the biased documentation of misconduct*

The example of status provides a mechanism by which an empirical sample of individuals might inaccurately represent underlying behavior; but why and how might this relate to the actual process by which organizations create archival data? First, "why" this might occur is because specific organizational traits may influence the likelihood that such biases are represented in the data. Organizational traits such as culture and policy are at potential risk of shaping how behavior is evaluated and recorded. This is problematic for organization-level research, because testing and building theory about an organizational trait that may both cause behavior (Greve, Palmer, & Pozner, 2010; Vaughan, 1999) and alter how behavior is documented will lead to biased estimates of the importance of those traits, an issue that will be revisited in the discussion section.

Second, "how" this documentation occurs within organizations is through a process of evaluating potential cases of questionable behavior and then documenting some but not others as misconduct. This indicates that it may be possible to understand the scope of these selection biases by looking for discrepancies between the "inputs" and "outputs" (final datasets) at the organizational level. In the next section I explore how incorporating allegations may be a promising methodological approach for doing this.

3. Allegations as a way to understand organizational biases

Allegations may help researchers detect selection biases within organizations because they are partly outside of an organization's control. This means they provide an alternative for understanding what an archival sample might look like in the absence of organizational involvement. An allegation is typically a necessary but not sufficient condition for an individual to be labeled a "wrongdoer" in the type of official archival data created by organizations and studied by researchers. This means allegations can help researchers understand the potential scope and source of selection bias within organizations that could lead to false negatives and false positives in archival data. By mapping out discrepancies between punishment for misconduct, the evaluation of alleged misconduct, and the creation of allegations, a researcher can document potential instances of organizational selection bias and the consequences for research that might attempt to use archival data from one stage of this process without acknowledging what occurs at the other stages.

I define an allegation as a specific claim made by a specific actor—to a specific organizational body—that a second actor has committed misconduct. In the sociology and organizational theory literatures, the organization that evaluates and acts on allegations is often called a "social control agent": "an actor that represents a collectivity and that can impose sanctions on that collectivity's behalf" (Greve et al., 2010, p. 56). Therefore, organizations are at the center of this process. An allegation lacking an organization is simply a dyadic dispute between individuals.

An allegation has a source, a specific target, and will lead to an official organizational outcome: an organization either (1) validates the allegation and the target is officially labeled a wrongdoer (and becomes a candidate for punishment), or (2) the allegation is not sustained and no official wrongdoing has occurred. Allegations are by nature a type of process data which occur after some behavior has allegedly taken place but before official judgement or punishment is passed on the target of the allegation. Further, official punishments are only possible via sustained allegations.

The specificity and organizational nature of allegations separate them from other phenomena such as rumor or gossip (Rosnow & Foster, 2005) and shift the focus away from the type of accusations considered by Faulkner (2011), which depended on media reports but not on formal evaluation by an organization. While allegations do not require the publicity of theoretical constructs such as scandals (Adut, 2005), allegations may both influence and be influenced by such public events. However, these related constructs generally lack the organizational aspect of allegations that is critical to the creation of archival data on misconduct.

What follows is a simple model that describes the role of allegations in how misconduct is evaluated by organizations. The model highlights three main stages in this labeling process: (1) a behavior is enacted by someone, (2) an allegation must then be leveled by another actor formally claiming that the previous actor has committed misconduct, and (3) an official evaluation of this allegation must be made by an organization to decide whether it will be sustained and the label of "wrongdoer" will be assigned, as well as the level of punishment to be administered. This intuition can be divided into the following

components: (1) A $behavior_i$ is enacted by actor A_j at time $t-1$. (2) In the subsequent period an $allegation_{i,j,k,t}$ is leveled against actor A_j by complainant B_k. It is a function of a specific unobserved behavior, $behavior_i$, enacted by actor A_j. (3) An organization's decision to label A_j a "wrongdoer" in period $t+1$ is then a direct function of this $allegation_{i,j,k,t}$ in combination with the characteristics of $behavior_i$, A_j, and B_k. These relationships are summarized in Fig. 1 and in the equations below, where $wrongdoer_{j,t+1}$ is the binary outcome of whether A_j is ultimately labeled a wrongdoer in archival data.

$$behavior_{i,t-1} = f(A_j) \tag{1}$$

$$allegation_{i,j,k,t} = f(A_j, B_k, behavior_{i,t-1}) \tag{2}$$

$$wrongdoer_{j,t+1} = f(allegation_{i,j,k,t}, A_j, B_k, behavior_{i,t-1}) \tag{3}$$

As illustrated with the example of status in the previous section, the individual traits of A and B are of interest because they may be simultaneously correlated with specific types of behavior by A_j, the propensity of B_k to consider a behavior wrongdoing, the propensity of B_k to allege wrongdoing against A_j, as well as the propensity of an organization to validate $allegation_{i,j,k,t}$, label A_j an official $wrongdoer_j$, and determine the level of punishment. It is the final stage that is most salient for organizational researchers, because it directly determines what is recorded in archival data. However, the earlier stages are also important insofar as expectations about later stages have the potential to influence earlier stages. For example, B_k may be less likely to make an allegation in the first place if they believe an organization is unlikely to validate it.

The major insight of this framework is that organizational selection biases may shape the final archival data on "wrongdoers," so that the original behavior is not the only factor in determining the sample. One can imagine two extremes. At one extreme, an allegation may have no

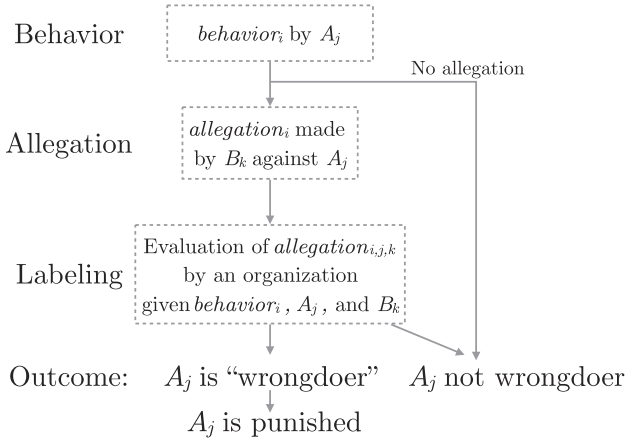

Fig. 1. The documentation of misconduct in light of a $behavior_i$ by actor A_j, a potential allegation from complainant B_k, and an organization that evaluates this $allegation_{i,j,k}$. The risk of A_j being documented as a wrongdoer in a formal and punishable sense is a direct result of this process. This outcome may be separately influenced by the characteristics of $behavior_i$, the characteristics of A_j (those enacting the behavior), and the characteristics of B_k (those free to declare whether they believe the behavior is wrongdoing). For example, the traits of A_j and B_k may influence behaviors, the evaluation of those behaviors, the decision to make allegations, and the organization's evaluation of those allegations. Attempting to separate these stages is therefore useful for understanding the organizational biases that could influence who does and does not get labeled a wrongdoer in archival data.

predictive power in determining who is labeled a wrongdoer in archival data. Individuals who function "above the law" in some fashion fall into this category. Diplomatic immunity might be a formal example of this phenomenon, where the behavior of a subset of individuals receives special evaluation no matter what actually occurs or is alleged.[3] However, diplomatic immunity is legally defined and potentially less interesting than scenarios where the difference in evaluations arises from organizational processes in spite of legal regimes. These cases will lead to under-documentation of specific populations in official datasets.

At the other extreme, an allegation may have full predictive power in labeling wrongdoers. For example, actual witch-hunts now appear clear instances of where organizations labeled wrongdoers independent of behaviors. It is now clear that archival datasets of "witches" are not useful for studying sorcery behavior. Yet it has been estimated that potentially hundreds of thousands of mostly women were executed for witchcraft in Europe during the 14th to 17th centuries (Ben-Yehuda, 1980). This provides a particularly stark example of how an allegation coming from the right person, against the right person, and evaluated by the right organizational system, can be enough to create a "wrongdoer" out of whole cloth. In the case of witch hunts, the organizations charged with detecting witchcraft appear to have been influenced by the characteristics of both those making allegations and those receiving them. A particularly extreme form of organizational selection bias can be found in the example of the Dominicans and the Inquisition, which "had a professional interest in the discovery of problems and of populations on which to exercise their specialized theological expertise in heresies and their investigative skills" (Ben-Yehuda, 1980, p. 11). Thus, organizational biases in the evaluation of allegations can even completely overwhelm the importance of the underlying behavior. This will lead to over-documentation of specific populations in official datasets.

Although anyone can make allegations, there may be practical boundaries to how closely an allegation must be tied to some version of objective fact. The avoidance of libel, slander, and defamation or more basic reputational concerns may limit the production of allegations that are complete fantasy. Further, the process of making an allegation likely entails a variety of costs that also limit their production, both real (e.g., time, transportation, money) and potential (e.g., reputation). Variance across organizations on these dimensions is another mechanism for introducing biases into archival samples.

3.1. The experimental ideal

This section outlines how a researcher could—in theory—conclusively detect organizational selection biases using allegations. This is accomplished by considering a set of "ideal experiments." These experiments are "ideal" in the sense that they would provide conclusive evidence of organizational selection biases, but they could never actually be run for ethical or practical reasons. They use the same notation introduced in the previous section and are presented in the reverse order of Fig. 1: punishment, wrongdoer label, and allegation.

Bias in punishment. The starting point for much archival data on misconduct is the sample of individuals that have been punished for a specific behavior of interest. To detect whether such samples are biased, a researcher would randomly assign sustained allegations to a set of heterogeneous actors A within the organization. There should be no correlation between the traits of specific A_j and the harshness of the punishments, unless organizational selection biases are at work.

Bias in the evaluation of allegations. The second conceptual experiment would test for bias in the evaluation of allegations on two different fronts: (1) how the traits of those being accused bias the evaluation, and (2) how the traits of those making the allegations bias the same evaluation. A researcher would ask heterogeneous actors B to make random allegations against a separate set of heterogeneous actors A. The rate at which these allegations are sustained should be uncorrelated with the traits of specific A_j unless an organization's evaluations are biased by the traits of allegation recipients. The rate at which these allegations are

[3] This has included behaviors as straightforward as parking violations (Fisman & Miguel, 2007).

sustained should also be uncorrelated with the traits of B_k unless the organization is biased by the source of the allegation.

Bias in the production of allegations. The third conceptual experiment links behavior to the receipt of an allegation. Here, a researcher would randomly assign a group of heterogeneous actors A to enact specific $behavior_i$. If the complainants, B, are not biased, then the rate at which allegations are accrued against specific A_j should be uncorrelated with the traits of A_j. However, the source of biases by B may be a function of expectations about later stages of the process itself. For example, if B believes that allegations against specific A_j will never be sustained, they will be less likely to make them in the first place—not because of their own bias but because of their beliefs about the organizational biases. This dynamic amplifies the importance of subsequent stages in the process.

Although each of the above could be conceived as a field experiment, they could never be ethically conducted given the impact they would have on participants. This means they must be approximated using other research designs. The most obvious approach is to create laboratory experiments that ask participants to imagine scenarios as if the above has occurred. For example, the punishment experiment above is roughly approximated by the lab designs found in work such as Fragale et al. (2009), who present participants with scenarios describing "guilty" individuals and ask them to interpret the severity of that guilt based on their status.

However, laboratory studies face a number of well-documented challenges related to external validity. First, lab study participants may change their behavior when they know that they are under observation (Levitt & List, 2008). This risk is increased for studies about ethics because of social desirability biases. This means that causal relationship identified in the laboratory might not be replicable in the field, even with the exact same participants. Second, there are general concerns about the ability to extrapolate from specific lab participant samples to other populations of interest (e.g., Henrich, Heine, & Norenzayan, 2010). Researchers are also often interested the biases that have resulted from specific organizational contexts or cultures (Ashforth & Anand, 2003), and creating these cultural conditions in the lab may be difficult. These potential discrepancies between laboratory and field dynamics are a major reason for the push to incorporate more behavioral field evidence into the literature (Pierce & Balasubramanian, 2015).

Randomized experiments conducted in the field have the potential to sidestep the above challenges but are ethically limited to treatments where the harm to participants is limited. For some questions and settings this is feasible. For example, Shu, Mazar, Gino, Ariely, and Bazerman (2012) randomized whether car insurance customers were asked to sign a statement regarding the veracity of their vehicle's mileage at the bottom versus the top of the actual reporting forms. This treatment resulted in the latter group reporting higher mileage, providing evidence that the salience of ethics can change behavior in practice. However, it is difficult to imagine ethical versions of the "ideal experiments" outlined above.

In cases where active manipulations are not possible, researchers sometimes employ "natural experiments" that occur in the field without the direct intervention of researchers. For example, to answer the question of how people from different cultures behave when legal enforcement is not a threat, Fisman and Miguel (2007) examined United Nation officials in New York who, because of diplomatic immunity, could accrue parking tickets without penalty until 2002; this helped them better measure the impact of cultural and legal enforcement on behavior but did not require direct intervention with the participants. Pierce and Balasubramanian (2015) review a number of other studies in this vein, and it is a powerful approach. However, in many settings of interest, natural experiments may simply not be available. Researchers must then attempt to make the optimal use of archival field data.

4. Police misconduct: an empirical example

The empirical focus of this paper is police officers that have been punished for misconduct. Police misconduct has become an important issue of public debate in the United States. Much of this discussion is concerned with race. Black Lives Matter, now considered one of the major protest movements of the millennial generation (Milkman, 2017), is closely linked to issues of police conduct, accountability, and racial bias (Chase, 2017). Black Lives Matter transformed into a national movement in 2014 after the shooting of Michael Brown in Ferguson, Missouri. By 2016, a survey by the Pew Research Center found that 45% of Americans supported the Black Lives Matter movement (Horowitz & Livingston, 2016), and another survey that year found that 63% of registered voters said the treatment of racial minorities was a very important issue in the presidential election, ranking above the environment and abortion (Fingerhut, 2016). Much of this discourse has focused on potential variance in police behavior toward minorities. Precisely estimating behavior such as the likelihood of shootings is difficult, however, because officers' exposure to situations where such shootings may occur also varies but cannot be directly observed (e.g., Johnson & Cesario, 2020; Johnson, Tress, Burkel, Taylor, & Cesario, 2019; Knox & Mummolo, 2020).

Academic criminologists have produced the bulk of existing research on police misconduct. Terrill and Ingram (2015) summarized what they considered the four general findings of past research on citizen complaints against the police: (1) allegations are not evenly distributed across officers, (2) the most frequent type of complaint is verbal discourtesy or improper use of force, (3) sustained complaints are rare, and (4) officer gender, age/experience, and education are correlated with the number of complaints against an officer. Hickman and Poore (2015) also noted from this literature that complaints from inside police departments are sustained at higher rates than citizen complaints, complaints from racial minorities are produced at disproportionately higher rates, and minority officers receive a disproportionate number of complaints.

There is also survey evidence that general perceptions about levels of police misconduct also vary by race, in part because different racial groups report having different personal experiences with police officers. For example, the conclusion by Weitzer and Tuch (2004, p. 322) that "the greater tendency for blacks and Hispanics to perceive police misconduct is largely a function of their disproportionate adverse experiences with police officers, exposure to media reports of police abuse, and residence in high-crime neighborhoods where police practices may be contentious." Racial profiling may be one reason for this, where there is evidence that minorities are searched at higher rates (e.g., Antonovics & Knight, 2009; Anwar & Fang, 2006; Knowles, Persico, & Todd, 2001).[4] These findings cast doubt on whether archival samples of officers that have been punished for wrongdoing can be used for the type of theory building of interest to academics.

4.1. Chicago Police

In 2013 the Chicago Police Department—with 12,042 full-time sworn personnel—was the second-largest local police department in the country behind New York (Reaves, 2015). Chicago is therefore an important context in its own right. At the same time, Chicago is one of the most corrupt cities, with more federal public corruption convictions than either Los Angeles or New York (during 2010–2013: 157, 114, and 70, for each of the three cities respectively; see Simpson, Gradel,

[4] Studies of racial profiling have often focused on why the difference in search rates occur (e.g., underlying prejudice versus employing race as an informational proxy), rather than whether it exists. However, Grogger and Ridgeway (2006) found limited evidence for racial profiling in traffic stops in Oakland, CA.

Mouritsen, & Johnson, 2015). Past research has indicated that Chicago police officers are not immune from such wrongdoing (Futterman, Mather, & Miles, 2007; Hagedorn et al., 2013). Hagedorn et al. (2013) presented an historical review of the Chicago Police Department, its approaches to oversight, and details of the nearly three hundred Chicago police officers criminally convicted from 1960–2012. They argued that "The 'blue code of silence,' while difficult to prove, is an integral part of the department's culture and it exacerbates the corruption problem" (p. 1). A formal investigation by the United States Department of Justice Civil Rights Division and the United States Attorney's Office for the Northern District of Illinois recently found that "Discipline is haphazard, unpredictable and does not deter misconduct" (United States Department of Justice, 2017a, p. 2). This indicates that there are systemic organizational issues that could bias the creation of archival data on misconduct in this setting.

At least three recent studies have examined data on allegations against Chicago police. A law review article by Futterman et al. (2007), whose first author was involved in the lawsuit which led to the eventual release of the data analyzed in this study, examined subsets of older data, allegations against subsets of specific officers, and officers with repeated allegations. They painted a bleak picture of the Chicago Police Department, determining that "The CPD's own data clarify that Chicago police officers can perpetrate abuse without fear of consequence" (Futterman et al., 2007, p. 265). More recently Rozema and Schanzenbach (2019) found that allegations against specific officers were predictive of civil rights litigation payouts. Finally, a working paper by Ba (2017) examined how the location of the allegation reporting center impacted who completed their allegations, a finding that I return to in the discussion section.

5. Data and methodology

The Chicago data present an opportunity to examine both the final outcomes and the inputs into the decision to document misconduct in archival data. The data analyzed for this study were the result of extensive lawsuits and have been described as the outcome of a "20-year saga" (Wildeboer, 2015) that "provides a rare look into the cloistered world of internal police discipline" (Williams, 2015). To my knowledge, the Chicago data are the most granular, complete, and timely ever publicly released. Prior studies in similar contexts have often utilized survey data at the department level (e.g., Cao, Deng, & Barton, 2000; Worrall, 2002), anonymized police departments (e.g., Griswold, 1994; Harris, 2011; Hassell & Archbold, 2010; Lersch & Kunzman, 2001; Lersch & Mieczkowski, 1996, 2000; Liederbach, Boyd, Taylor, & Kawucha, 2007), or limited subsets of officers within a department (e.g., Hickman, Piquero, & Greene, 2000). The Chicago data allow for visibility into the full allegation process.

The raw allegations data were collected by the Chicago Police Department and the City of Chicago Independent Police Review Authority[5] and released as the result of a lawsuit that deemed them public information (see legal case Kalven v. City of Chicago, 2014). The data were consolidated by the Citizens Police Data Project of the Invisible Institute, a "journalistic production company," which was run by Jamie Kalven, the plaintiff in the legal case that led to the eventual public release of these data.[6]

A number of summary accounts of these data have been publicized in the media—including by the Invisible Institute—regarding the occurrence of allegations, the low disciplinary rates, and apparent racial disparities (e.g., Wildeboer, 2015; Williams, 2015). I reproduce some of these statistics and expand them to multivariate analyses of the full data. Further, some characteristics of the data, such as the number of allegations filed without affidavits, may appear as noise but be of theoretical interest, and have been explored in other papers (Ba, 2017; Rozema & Schanzenbach, 2019).

Multiple versions of these data have now been released by the Invisible Institute, spanning different periods and with different levels of completeness.[7] I used the December 7, 2015, version of the data provided directly by the Invisible Institute, and limited my analyses to the most complete period of data, using only those allegations that contained an incident date.[8] This restricted the sample to the Freedom of Information Act release comprising the 28,588 allegations occurring from March 13, 2011, to August 19, 2015. I choose this period because it allowed for greater consistency and completeness in how the data are recorded as compared with data that spanned more years.[9] I then merged in separate data that identified which allegations were made by members of the department.[10]

Data that were likely clerical errors such as the handful of complainants documented as being born before 1900 were replaced with missing values. The police beat was used as the geographic location for each allegation. The police district was then defined as the first two digits of the four-digit beat location, per the Department's "Know Your District, Know Your Beat" webpage.[11] For the creation of maps, geographic data for the police beats were downloaded from the City of Chicago's Data Portal.[12] Crime data provided by the city at the beat level were used for crude baseline comparisons.[13] Note that there were slight shifts to the police district boundaries during the time period of the study, as three police districts were closed (Doyle & Gorner, 2012); however, the primary use of the geographic data was regression control variables, so this should have limited impact. An officer's departmental unit signified either the geographic police district where the officer worked or his or her special unit assignment if it was a city-wide unit.

5.1. Data overview and definitions

There are four main units of analysis in the data: the *incident*, the *allegation*, the *officer*, and the *complainant*. The basic unit of analysis is an *allegation* of misconduct, typically against a specific *officer*. Each allegation was the result of a specific *incident*, which could have resulted in allegations against more than one officer. Each allegation could have been filed by one or more *complainants*. These relationships are summarized with additional detail in Fig. 2. Officers are real-name identifiable and traceable across allegations along with their gender, race, age, and experience. Complainants were not uniquely identifiable across allegations, but their race, gender, and age were listed for each

[5] Created in 2007, the Independent Police Review Authority (IPRA) was "Headed by a civilian Chief Administrator and staffed entirely with civilian investigators, [and was] an independent agency of the City of Chicago, separate from the Chicago Police Department" (IPRA website). The IPRA was replaced in late 2017 by the Civilian Office of Police Accountability.

[6] For a description of the Invisible Institute, see https://invisible.institute/about/. The Invisible Institute also made the data available on an interactive public website called the Citizens Police Data Project.

[7] As of early 2019 the various data releases were available on the Invisible Institute's GitHub repository: https://github.com/invinst/chicago-police-data/.

[8] These data were the result of the Invisible Institute's FOIA request 14-5509.

[9] For example, the newer data release that includes allegations from 2000 onward does not include complainant demographics until roughly 2006. Further, the IPRA was not created until 2007, which likely also altered how data were recorded throughout the window.

[10] The Invisible Institute received these data in 2017 as a result of FOIA request P428703.

[11] "For example, Beat 521 is the 1st beat in the 2nd sector of the 5th Police District, and Beat 1913 is the 3rd beat in the 1st sector of the 19th Police District."

[12] The police beat boundary shapefiles can be found under "Boundaries - Police Beats (current)."

[13] Crime data were downloaded from the City of Chicago's data portal dataset "Crimes - 2001 to present", filtered to crimes with dates after 03/13/2011 12:00:00 AM and before 08/20/2015 12:00:00 AM

allegation, as well as whether they were a member of the Chicago Police Department themselves.

In total, there were 28,588 allegations of misconduct during this window stemming from 19,530 unique incidents of alleged misconduct. Fig. 3 plots the occurrence of allegations over time along with the general finding of the investigation into the allegation and the specific outcome of the allegation. The overall number of allegations was declining over the period.

The most frequent category of allegation was "Operation/Personnel Violations" (21.1%) followed closely by "First Amendment and Illegal Arrest" (20.1%). Allegations in subcategories potentially more related to internal department behaviors, such as "Insubordination," appeared to be sustained at higher rates (28 sustained out of 56 allegations) than potentially more external behaviors such as "Unnecessary Display Of Weapon/ On Duty" (4 of 79 sustained). These subcategories are fully enumerated in the supplement (Table A.1).

The full sample of 28,588 allegations could be subdivided into a number of important categories based on the amount of information provided during the allegation. Of the allegations, 8,070 did not list a specific officer ID, meaning that an officer was not positively identified in the allegation. The outcome of the investigation into all such allegations was listed as "Unknown." For the 20,518 allegations that did have an officer identified, the average incident resulted in allegations against 1.79 officers. There is also clear right-censoring of the outcomes. For more recent years, a sizable portion of allegations were still listed as open investigations: 63% in 2015, 22.7% in 2014, and 8.5% in 2013.

Each investigated allegation ultimately resulted in one of four findings (City of Chicago Independent Police Review Authority, 2016). 6,745 allegations (23.6% of the total sample) had both an affidavit filed and already one of these four known outcomes. 11.2% of these 6,745 allegations were "Sustained": "The allegation was supported by sufficient evidence to justify disciplinary action. Recommendations of disciplinary action may range from violation noted to separation from the Department." 48.9% were "Not sustained": "The allegation is not supported by sufficient evidence which could be used to prove or disprove the allegation." 28.7% were "Unfounded": "The complaint was not based on facts as shown by the investigation, or the reported incident did not occur." And 11.2% were "exonerated": "The incident occurred, but the action taken by the officer(s) was deemed lawful and proper." Fig. 4 summarizes this process.

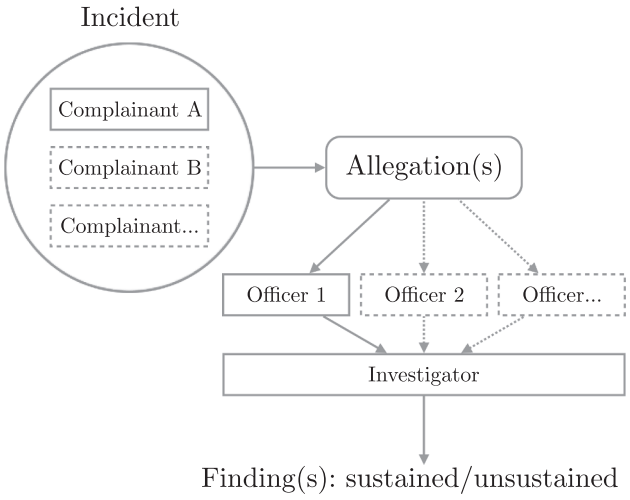

Fig. 2. The structure of the data. Each of the 19,530 incidents occurred at a specific time and location (police beat) between March 13, 2011 and August 19, 2015. These incidents lead to 28,588 allegations. Complainants are anonymous and cannot be linked across incidents; however, their demographic data are available (race, gender, and age), as is whether the complainant was a member of the Chicago Police Department. Multiple complainants per incident are possible but very rarely occurred (Table A.3). Each allegation could be filed against a specific officer(s), though a significant number (28.2%) did not positively identify an officer. Those that did alleged on average 1.8 officers per incident with a maximum of 27, resulting in at least one allegation against 7,758 unique officers. One investigator was assigned to each incident; there were 1,510 unique investigators during the time period. Each officer and investigator was uniquely and publicly (real name) identifiable, along with each officer's race, age, and gender. See Fig. 4 for how this relates to the samples used in the regressions.

The limited number of "exonerated" outcomes indicates that there was not significant disagreement about whether a specific practice was considered right or wrong in the abstract; i.e., allegations were not primarily representations of disagreements about the line demarcating right from wrong behavior. Rather, the rate of "sustained" allegations was primarily a result of whether a behavior that was agreed to be misconduct could be sufficiently demonstrated to have been enacted by a specific officer.

5.2. Methodological approach

The empirical analyses are structured into two stages. The first is simply a descriptive exploration of the allegation stages. Using summary statistics, tables, and plots, I describe where allegations came from, against whom they were directed, and their eventual outcomes. The goal is to present a foundation for understanding the role of allegations in shaping archival samples of organizationally-documented wrongdoers.

Then, building on the approach employed in previous empirical research (Terrill & Ingram, 2015), I constructed models to compare the role of individual traits in predicting versus sustaining allegations at the officer and individual allegation levels. I constructed four models, the comparison of which shed light on potential organizational selection biases that could decouple the outcome of allegations from their production. The first was a logit model that predicts the harshness of punishment for a sustained allegation. The second was a logit model that predicts whether a specific allegation was sustained. These models test the role of both officer and complainant demographics in determining these outcomes and can be viewed as the probability that an officer was deemed a wrongdoer—or received a harsher punishment—at time t based on a specific allegation, i.e., $\Pr(wrongdoer_{j,t+1}|allegation_{i,j,k,t}, A_j, B_k)$, where $allegation_{i,j,k,t}$ is the set of attributes specific to allegation i (category of the allegation, the district where the incident occurred, and the year of the incident) filed at time t by actor k against officer j; A_j is a set of officer j's traits including gender, race, age, experience, departmental unit assignment, and rank, where experience is relative to the date of the alleged incident and the total number of preceding allegations against the officer (during the window) is also included; and B_k is the set of traits of the complainant including their gender, race, and age at the time of the alleged incident, and whether they were themselves a member of the Chicago Police Department. The third model is a logit model that predicted the existence of at least one sustained allegation at the officer level using this same set of officer characteristics. This can be viewed as $\Pr(wrongdoer_j > 0|A_j)$, where $wrongdoer_j$ is the number of sustained allegations against officer j and A_j is the same as before. The fourth and final model was a zero-truncated negative binomial count model that predicted the number of allegations against a specific officer during the full time period. Using the notation introduced earlier, this can be viewed as $allegations_j = f(A_j)$, where $allegations_j$ is the number of allegations against officer j and A_j is the same as before.

Geographic variance was controlled for in two ways. For the officer-level models, controls for the officer's departmental unit assignment were included, which was often geographically defined. For the allegation-level model, controls for the district where the alleged incident occurred were included. A number of the extremely sparse categories were collapsed into catch-all categories for each variable (described in the table captions). To ensure that officers had the same risk of receiving allegations, in the first two models I limited the sample

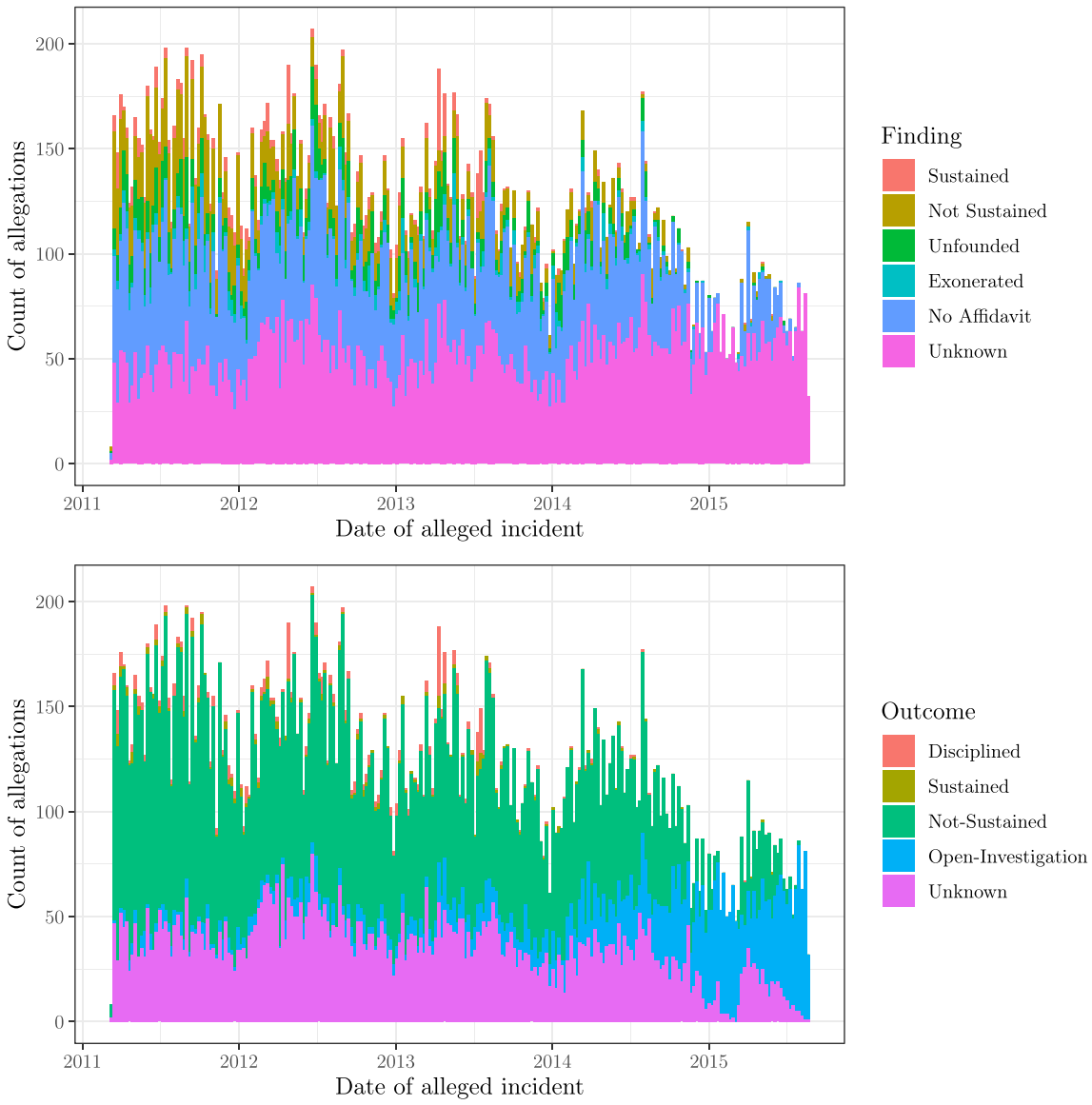

Fig. 3. Histogram of every allegation, shaded by the finding of the investigation (top) and the ultimate outcome (bottom). The number of allegations appears to be declining over time. Few allegations are sustained and even fewer result in formal discipline. A significant number (34%) are not sustained simply because the complainant never filed a required affidavit. Represents every allegation filed from March 13, 2011 to August 19, 2015. Binwidth = 7 days. Detailed tabulations of these values can be found in the supplement (Table A.11 and Table A.10).

to officers that were active throughout the time period: i.e., they were assigned before the first allegation in the sample and were confirmed to be active on June 1, 2015.

6. Results

Although there were 28,588 total records of allegations, a large number of allegations were begun by complainants but potentially not completed. No affidavit was filed by the complainant in 9,722 (34%) of the allegations, a requirement for continuing an investigation in many cases. As found by Ba (2017), this subset is theoretically interesting in its own right because it hints at the literal process required to submit a complete allegation; many individuals seemingly began but did not finish the process. Second, no officer was identified in 8,070 (28.2%) of the allegations, meaning that such allegations could not lead to findings of wrongdoing against officers even though complainants may have believed they were wronged. There were 10,796 allegations (37.8% of the total) that were both directed at a specific officer and not lacking affidavits.

6.1. The sources of allegations

Who files allegations? Nearly every incident had only one complainant (Table A.3). While it was not possible to track complainants across allegations, of the 17,027 complainants, there were at least 3,455 unique individuals based on combinations of reported age, gender, race, and whether the complainant was a member of the CPD. Of the 17,027 non-unique (minimum 3,455 unique) complainants, the average was born in 1973 (1969), 44.7% (46.1%) were women, 64.5% (39.1%) were identified as black, 22% (32.8%) as white, and 9.69% (18.7%) as white/Hispanic. For reference, census statistics indicate that in 2014 Chicago was 51.5% male with a mean age of 33.3 and was either 47.8% white and 32.3% black, or, when tabulated using the Hispanic origin census question, 32.2% white, 31.9% black, and 28.7% Hispanic. Either way, the racial composition of complainants did not appear to be strictly consistent with the general population of Chicago. Age and gender distributions also varied significantly by race. Roughly equal numbers of allegations were filed by black men and women, but many more allegations were filed by white men than white women, and black male

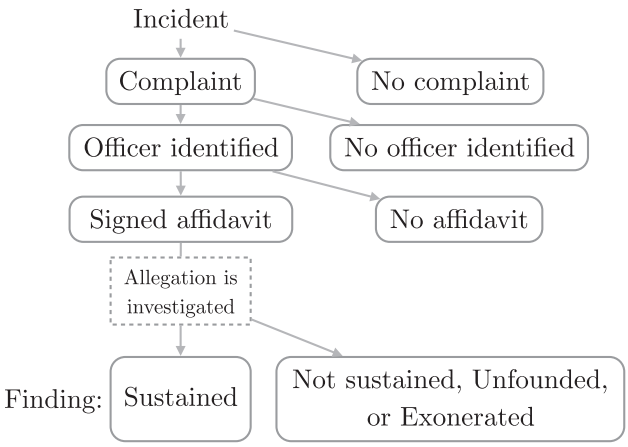

Fig. 4. To have a definitive outcome, an allegation must have identified an officer and the complainant signed an affidavit. For a more detailed flowchart of how allegations are filed and evaluated, see the flowchart distributed by the Invisible Institute (Cochrane, 2020).

complainants were younger than white male complainants (Fig. A.1). It was also possible for the complainants themselves to be members of the Chicago Police Department. Of the non-unique complainants, 14.5% were members of the CPD, and of the unique set, 22.4% were members of the CPD. More detailed complainant statistics are presented in the supplement (Table A.4), as well as trends related to timing (Fig. A.2) and geography (Figs. A.3 and A.4).

6.2. The recipients of allegations

Who are the officers? There were 7,758 unique officers during the period who had at least one allegation filed against them, with an average of 2.64 and a maximum of 30 allegations against a single officer (this distribution is plotted in Fig. A.5 and Fig. A.6). These officers were 19.3% female and had an average appointment date of 1999 and birth year of 1970. They were identified as 51% white, 24.5% black, 19.4% Hispanic, 2.75% Asian, and 2.24% white/Hispanic.[14] These demographic statistics are tabulated in the supplement (Table A.5). As of 2012 the entire Chicago Police Department consisted of 12,042 police officers, indicating that the majority received an allegation during the period (Reaves, 2015). There were 113 unique departmental units listed in the data, though the majority of officers belonged to one of the major geographic district units; this distribution is also detailed in the supplement (Table A.6; Fig. A.7 also plots allegations by unit).

Who receives allegations? As noted in prior criminology research (Terrill & Ingram, 2015), allegations were not evenly distributed across officers. In this sample, a quarter of officers attracted over half the total allegations (see Fig. A.8). The true skewness would be even greater if officers that never received even one allegation during the period were included.

Allegations may have been heavily influenced by things both within and outside of an officer's control, for example, specific job characteristics. Consistent with prior research, younger officers appear to have received more allegations than older ones (Fig. A.9). One explanation may be that less experienced officers are simply less skilled, though another might be that rookie officers are simply given more challenging assignments. Indeed, the time of day of the allegation and the officer's age and experience were correlated (Fig. A.10 and Fig. A.2). The majority of allegations were made by black complainants and were received by white officers. The interaction of officer and complainant race is tabulated in the supplement (Table A.7 and its flip-side in Table A.8).

The relationships between officers? Prior research has emphasized how organizational wrongdoing is often influenced or represented by network structures within organizations (for recent examples, see Aven, 2015; Palmer & Yenkey, 2015). Police misconduct is a prime example of where network dynamics may shape both the occurrence and the reporting of misconduct. Many of the most notorious historical reports of misconduct by Chicago police involved groups of officers, such as the torture cases from the 1970s and 1980s or the more recent "Skullcap Crew" of officers (Futterman et al., 2007). Some types of behavior also appeared more likely to result in allegations against multiple officers. The average number of officers accused together in each category of allegation ranged from 1.02 officers for something like "Alcohol Abuse" to 2.41 officers for each "First Amendment and Illegal Arrest" incident (Table A.9).

Co-accusations that resulted from the same incident can be seen as a natural type of network and are a common occurrence in the data. Using this relationship I calculated a degree centrality metric for each officer within the network of allegations: i.e., with how many other unique officers was an officer co-accused?[15] The majority of officers (81.2%) had at least one co-accused officer, meaning that at least one other officer was accused in one of the same incidents that they were. The mean number was 4.25 other officers, and the maximum number of unique co-accused officers for a single officer was 55 other officers during the period. This distribution of degree centrality appears to vary with demographics such as gender (Fig. A.11). Officers were considered one degree away from each other if they both had allegations stemming from the same incident. These co-accusation relationships were then used to construct a formal bipartite social network graph, with officers and incidents as each type of node. The network is quite dense; only one of these 10 officers was more than two degrees of separation from any of the others (Fig. A.12 plots the immediate networks of the 10 officers that received the most allegations).

6.3. The outcomes of allegations

There are two stages to the outcome of an allegation: (1) the assignment (or not) of a label (i.e., the finding), and (2) the consequence of that label (i.e., the punishment). Extant media reports of these data have largely focused on the low rates at which allegations were sustained and ultimately resulted in punishment against officers (e.g., Williams, 2015). In total there were 755 sustained allegations, representing 2.64% of the total pool of 28,588 allegations. However, as noted above, in many cases a specific allegation could by definition not be sustained; this occurred when either no specific officer was identified, a required affidavit was not filed, or the allegation was still listed as an open investigation. Accounting for this, sustained allegations represented 11.2% of the remaining 6,745 allegations. The outcome and findings of the allegations are tabulated by category and subcategory in the supplement (Table A.10 and Table A.11).

A range of consequences resulted from being labeled a wrongdoer. The absolute counts for each outcome category are plotted in Fig. 5 with additional detail. The absolute number for many categories of punishment is very low. For the subset of allegations that was sustained, the three most frequent outcomes were "reprimand," "noted," or a one-day suspension. There appeared to be a race but not a gender trend.

Outcomes also varied by the traits of the complainants. Allegations by white complainants appear to have been sustained and disciplined at higher rates than those filed by black complainants (Table A.12). This trend was largely the result of allegations by white men (Fig. A.13 and Table A.12). Finally, the category of the allegation was also important.

[14] A very small number of officers (0.412%) listed as "Italian" in the raw data were recoded as "White."

[15] Other centrality measures, such as closeness centrality and betweenness centrality, could also be calculated, though their usefulness may be limited given the number of network isolates in these data (Wasserman & Faust, 1994).

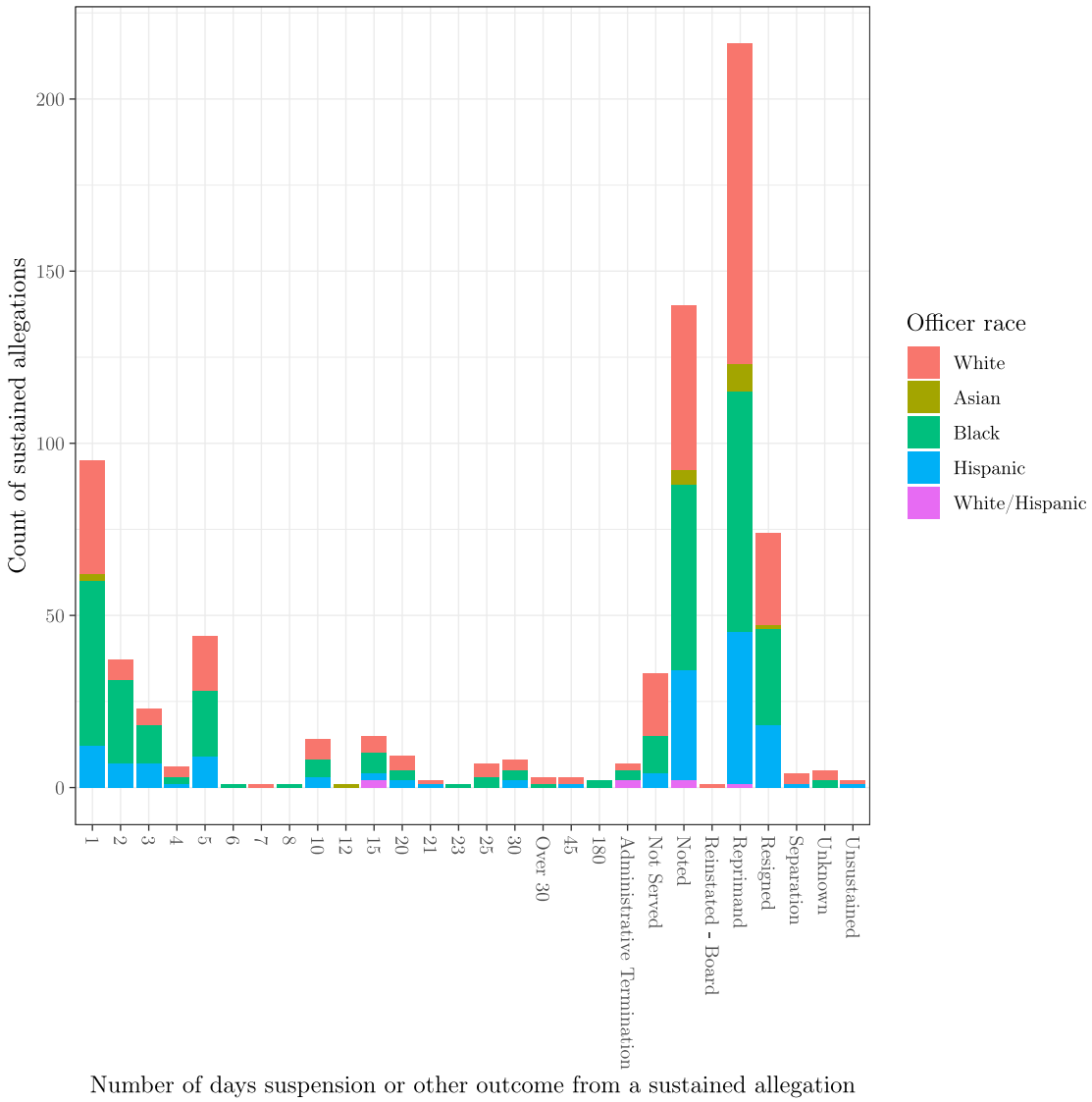

Fig. 5. The consequences of being labeled a wrongdoer. The 755 sustained allegations resulted in a range of disciplinary outcomes. The most frequent results were "Reprimand," "Noted," and a one-day suspension. The average length of a suspension when a specific number was provided was 7.5 days. 39% of these sustained allegations were for black officers, which is higher than the 24% of all allegations that were against black officers (which is roughly equal to the 25% of black officers in the sample). 21% of the sustained allegations are against women, roughly similar to the sample average of 19%. See the regression analyses presented in Table 4 for a consideration of officer race and gender effects when controlling for potentially confounding factors. See the regression presented in Table 1 for an analysis of what predicts severity of punishment.

The second most frequent category of allegation, "First Amendment and Illegal Arrest," had in total only 8 sustained allegations during the period. However, all 16 allegations of "D.U.I. - Off Duty" were sustained (incidentally, these allegations were made by CPD members).

In sum, these descriptive findings highlight the organizational process that leads to the documentation of police misconduct. From a very large "at-risk" set, only a small subset of officers are eventually punished. This is not necessarily a problem for research use as long as any false positives or false negatives are randomly generated. The next session therefore considers how one might determine whether organizational selection biases are at work.

6.4. Potential evidence of organizational selection biases

This section presents the results of four regression models that collectively address the question of organizational selection biases by exploring whether discrepancies exist between the evaluation of allegations and their production. I begin with the regression to predict the relative magnitude of punishment that resulted from sustained allegations. This is a natural starting point because—as the final stage of the process—it is often the most readily available type of archival data.

To test for biases in punishment I created two categories of punishment that roughly divided the sample of sustained allegations in half: "light" punishment was defined as the outcome being "noted," "reprimand," "not served," or "reinstated - board." All other outcomes, including any type of suspension or resignation, was then categorized as "heavy" punishment. Table 1 reports these results, controlling for the category of the complaint and district of the incident, as well as officer rank, unit, and age. Each additional previous allegation against an officer was associated with a 1.15 times higher chance of the focal sustained allegation leading to a heavier punishment. A sustained allegation against a black officer was 2.01 times as likely to result in a heavier punishment. I did not find evidence that complainant or officer gender was important, nor whether the allegation came from a member of the CPD.

I next ran the logit model that predicted whether an individual

Table 1
Allegation-level logistic regressions predicting magnitude of punishment for a sustained allegation. The sample is composed of the 707 allegations that were "sustained" and met the criteria for using complainant demographic data. The dependent variable is whether the sustained allegation lead to a "heavier" (1) or "lighter" (0) punishment (see full list of punishment categories in Fig. 5). "Lighter" is defined by an outcome of "noted," "reprimand," "not served," or "reinstated - board" and represents 53% of the sample of sustained allegations. Reference category for officer race is "White," for complainant race is "White," for allegation category is "Operation/Personnel Violations," and for the two officer and complainant age categorical variables is "missing." The inflated coefficients on complainant race "Asian" and "Native American" should not be directly interpreted because none of those seven allegations led to heavier punishment; a similar situation exists for the coefficients on the categories "Drug/Substance Abuse" and "Unknown" (all 10 led to heavier punishment). Standard errors are clustered at the officer-unit level.

	Dependent variable: Punishment was "heavier"
Officer gender: M	0.160
	(0.251)
Officer race: Asian	−0.400
	(0.582)
Officer race: Black	0.699***
	(0.205)
Officer race: Hispanic	0.341
	(0.277)
Officer race: White/Hispanic	−0.606
	(0.783)
Complainant was CPD member	0.289
	(0.312)
Year of incident (mean centered)	−0.411***
	(0.134)
Officer tenure at incident (mean centered)	0.0004
	(0.027)
Officer age at incident: <30	−0.372
	(0.720)
Officer age at incident: [30, 45)	0.100
	(0.309)
Officer age at incident: [45, 60)	0.159
	(0.334)
Officer age at incident: >=60	0.435
	(0.565)
Count of previous allegations against officer	0.141**
	(0.068)
Complainant race: Asian	−16.883***
	(0.814)
Complainant race: Black	0.157
	(0.337)
Complainant race: Hispanic	−0.338
	(0.399)
Complainant race: Native American	−17.129***
	(1.037)
Complainant race: Unknown	−1.936
	(1.701)
Complainant race: White/Hispanic	−0.468
	(0.862)
Complainant gender: M	−0.049
	(0.207)
Complainant age at incident: <30	−0.511
	(0.777)
Complainant age at incident: [30, 45)	−0.690
	(0.551)
Complainant age at incident: [45, 60)	−0.845
	(0.558)
Complainant age at incident: >=60	−1.912**
	(0.762)
Cat: Alcohol Abuse	1.340
	(0.862)
Cat: Arrest/Lock-up Procedures	−0.056
	(0.342)
Cat: Bribery/ Official Corruption	1.552
	(1.542)
Cat: Conduct Unbecoming (Off-duty)	0.004
	(0.402)
Cat: Criminal Misconduct	0.939*
	(0.534)
Cat: Drug/Substance Abuse	17.690***
	(1.334)
Cat: First Amendment and Illegal Arrest	−1.800
	(1.349)
Cat: Search-Related	0.085
	(0.341)
Cat: Supervisory Responsibilities	−0.858
	(1.267)
Cat: Traffic	−0.594
	(0.721)
Cat: Unknown	17.581***
	(0.970)
Cat: Verbal Abuse	−0.046
	(1.783)
Constant	−17.713***
	(1.202)
District controls	Yes
Officer rank controls	Yes
Officer unit controls	Yes
Observations	707
Log Likelihood	−369.266
Akaike Inf. Crit.	988.531

Note: *p < 0.1; **p < 0.05; ***p < 0.01.

allegation was "sustained." The sample of allegations was limited to those at risk of being sustained: i.e., the allegations that identified a specific officer, were not missing an affidavit, and had a known finding. Further, to allow for the inclusion of accurate complainant traits in the model, I removed allegations that came from multiple complainants. The result of this regression is reported in Table 2. The first model includes only officer characteristics, the second only complainant characteristics, and the third both officer and complainant characteristics. I discuss the results from the last of these specifications, as the results are similar across the models.

Individual allegations directed at black officers were sustained at a rate 1.78 times higher than those directed at white officers; a similar relationship existed for allegations against Hispanic officers (1.46 times higher). Consistent with the previous model, I did not find evidence that male officers were more or less likely than female officers to have an individual allegation against them sustained. The number of previous allegations against an officer did not predict whether an allegation was sustained (i.e., an accumulation of allegations did not increase the probability that a subsequent one was sustained), indicating that past allegations may not meaningfully influence current decisions at this stage.

Multiple complainant characteristics were predictive of whether an allegation was sustained. Compared with allegations filed by white complainants, allegations from black complainants were 0.499 times less likely to be sustained. I did not find evidence that allegations made by men were more or less likely to be sustained than allegations made by women. However, by far the strongest predictor was whether the allegation was made by a member of the department itself. Allegations made by CPD members (17.4% of the sample) were sustained at a rate 21.6 times higher than allegations made by outsiders.

I next ran the logit model to predict whether an officer ever received a "sustained" allegation during the period. These results are presented in Table 3. Black and Hispanic officers were both at greater risk (by 1.71 and 1.38 times, respectively) of receiving a sustained allegation than white officers. However, I did not find evidence that men received sustained allegations at different rates than women. I also expanded the sample of officers to include those that were not confirmed to be active

Table 2
Allegation-level logistic regressions predicting whether a specific allegation was sustained. Only includes allegations that met the following criteria: identified a specific officer, included an affidavit, were not "open-investigation," and came from a single complainant with gender listed as "M" or "F." In addition, to avoid quasi/complete separation of the logit model, a number of very low-frequency-incident-district categories (the three least frequent) and departmental unit categories (those with fewer than 10 officers) were collapsed into new categories. This left a potential 6,370 allegations, of which 707 were sustained. Reference categories for officer race is "White," for complainant race is "White," for allegation category is "Operation/Personnel Violations," and for the two officer and complainant age categorical variables is "missing." Standard errors are clustered at the officer-unit level.

	Dependent variable: Allegation was sustained (0/1)		
	(1)	(2)	(3)
Officer gender: M	0.070		0.021
	(0.135)		(0.168)
Officer race: Asian	0.134		0.125
	(0.272)		(0.364)
Officer race: Black	0.526***		0.578***
	(0.150)		(0.179)
Officer race: Hispanic	0.303**		0.375***
	(0.129)		(0.131)
Officer race: White/Hispanic	0.034		0.071
	(0.459)		(0.419)
Complainant race: Asian		−0.204	−0.036
		(0.554)	(0.502)
Complainant race: Black		−0.742***	−0.696***
		(0.187)	(0.180)
Complainant race: Black/Hispanic		−12.552***	−13.293***
		(0.532)	(0.529)
Complainant race: Hispanic		0.117	0.306
		(0.198)	(0.204)
Complainant race: Native American		2.014**	2.004**
		(0.865)	(0.843)
Complainant race: Unknown		−0.732	−0.541
		(0.483)	(0.493)
Complainant race: White/Hispanic		−0.103	−0.021
		(0.347)	(0.339)
Complainant gender: M		0.016	0.040
		(0.129)	(0.135)
Year of incident (mean centered)	−0.034	−0.107*	−0.105
	(0.064)	(0.059)	(0.076)
Officer tenure at incident (mean centered)	−0.006		0.001
	(0.012)		(0.014)
Officer age at incident: <30	−0.520**		−0.637*
	(0.265)		(0.327)
Officer age at incident: [30, 45)	−0.078		−0.106
	(0.149)		(0.168)
Officer age at incident: [45, 60)	0.144		0.034
	(0.149)		(0.216)
Officer age at incident: >=60	−0.278		−0.183
	(0.302)		(0.395)
Count of previous allegations against officer	−0.066**		−0.019
	(0.032)		(0.045)
Complainant was CPD member		2.902***	3.071***
		(0.225)	(0.236)
Complainant age at incident: <30		−0.881***	−0.818***
		(0.307)	(0.316)
Complainant age at incident: [30, 45)		−1.068***	−1.079***
		(0.233)	(0.246)
Complainant age at incident: [45, 60)		−1.029***	−0.997***
		(0.268)	(0.297)
Complainant age at incident: >=60		−1.310***	−1.412***
		(0.387)	(0.432)
Cat: Alcohol Abuse	4.012***	3.481***	3.951***
	(0.812)	(0.511)	(0.735)
Cat: Arrest/Lock-up Procedures	−2.169***	−0.937***	−1.018***
	(0.203)	(0.241)	(0.249)
Cat: Bribery/ Official Corruption	−0.022	0.847	1.028
	(0.563)	(0.627)	(0.653)
Cat: Conduct Unbecoming (Off-duty)	0.900***	0.582***	0.445*
	(0.205)	(0.220)	(0.249)
Cat: Criminal Misconduct	0.764*	0.421	0.038
	(0.397)	(0.593)	(0.532)
Cat: Drug/Substance Abuse	−0.534	−0.776	−1.798***
	(0.648)	(0.695)	(0.684)
Cat: First Amendment and Illegal Arrest	−3.921***	−2.259***	−2.285***
	(0.409)	(0.384)	(0.425)
Cat: Search-Related	−0.206	−0.176	−0.237
	(0.185)	(0.255)	(0.276)
Cat: Supervisory Responsibilities	−0.869	−0.328	−0.045
	(0.568)	(0.766)	(0.883)
Cat: Traffic	−0.976***	0.062	−0.041
	(0.336)	(0.413)	(0.449)
Cat: Unknown	−0.476	−0.641	−1.578
	(0.740)	(0.951)	(0.974)
Cat: Verbal Abuse	−3.076***	−2.023***	−2.067***
	(0.580)	(0.577)	(0.632)
Constant	−17.550***	−1.340***	−19.514***
	(0.970)	(0.336)	(1.133)
District controls	Yes	Yes	Yes
Officer rank controls	Yes	No	Yes
Officer unit controls	Yes	No	Yes
Observations	6,370	6,370	6,370
Log Likelihood	−1,567.102	−1,243.436	−1,155.935
Akaike Inf. Crit.	3,366.203	2,590.872	2,569.869

Note: *p < 0.1; **p < 0.05; ***p < 0.01.

at the end of the period (second model in Table 3); the results were very similar.[16]

Finally, Table 4 reports the results of the zero-truncated negative binomial count model predicting the total number of allegations against a given officer, the very earliest stage of this process (alternate specifications using a standard Poisson and zero-truncated Poisson are also reported in Table A.13 and Table A.14, respectively). Male officers received allegations at a rate 1.37 times that of female officers. I did not find strong evidence that black officers received allegations at a rate different from white officers, but Hispanic officers received them at a rate 0.927 times lower than white officers.

6.5. Comparison of results

When an analysis of organizational data (i.e., punishment) is complemented with external input into that process (i.e., allegations), the scope of potential organizational selection biases becomes clearer. Comparing the results from the four regressions highlights that demographic traits did not equally predict punishment, the evaluation of allegations, and the accrual of allegations. Contingent on the existence of allegations, allegations were sustained at different rates according to demographic traits of officers and complainants. These relationships existed even with a wide range of control variables.

For example, black officers received heavier punishments when they received sustained allegations. Therefore, if one were to analyze only the sample of officers that were punished, the data may be unrepresentative of the actual behavior of interest. This type of organizational selection bias would obviously complicate the use of archival data. It is consistent

[16] This test was run because limiting the sample to officers active for the entire time period non-randomly removes a subset; i.e., sustained allegations against 68 unique officers resulted in an outcome of "Resigned," "Administrative Termination," or "Separation."

Table 3
Officer-level logistic regressions predicting whether an officer ever received a sustained allegation. Limited to the sample of 6,059 officers that were active during the entire period. 503 of these officers had at least one sustained allegation. The second model expands this to include all officers (7,758) regardless of when they were active in the department. The reference category for race for officer race is "White" and for birth year is "missing." Standard errors are clustered at the unit level.

	Dependent variable: Officer received a sustained allegation (0/1)	
	(1)	(2)
Officer race: Asian	−0.056	0.215
	(0.413)	(0.311)
Officer race: Black	0.539***	0.464***
	(0.138)	(0.106)
Officer race: Hispanic	0.319**	0.351***
	(0.148)	(0.126)
Officer race: White/Hispanic	−13.881***	−0.707*
	(0.571)	(0.402)
Officer gender: M	0.053	0.056
	(0.160)	(0.153)
Total number of allegations	0.037	0.027
	(0.027)	(0.026)
Total allegations from CPD members	1.812***	1.788***
	(0.163)	(0.149)
Officer birth year category: <1960	0.383*	0.128
	(0.231)	(0.156)
Officer birth year category: [1960, 1970)	−0.038	−0.187
	(0.201)	(0.177)
Officer birth year category: [1970, 1980)	−0.190	−0.238
	(0.175)	(0.198)
Officer birth year category: >=1980	−0.340	−0.418*
	(0.224)	(0.217)
Officer appointment year (mean centered)	0.019	−0.009
	(0.016)	(0.010)
Constant	−17.741***	−16.909***
	(0.393)	(0.919)
Officer rank controls	Yes	Yes
Departmental unit controls	Yes	Yes
Observations	6,059	7,758
Log Likelihood	−1,274.090	−1,711.130
Akaike Inf. Crit.	2,704.180	3,598.259

Note: *p < 0.1; **p < 0.05; ***p < 0.01.

with concerns noted in the United States Department of Justice (2017b, p. 9) report that, "In the rare instances when complaints of misconduct are sustained, we found that discipline is haphazard and unpredictable, and is meted out in a way that does little to deter misconduct."

The scope of this specific problem is made clearer when stages prior to punishment are also considered. Allegations directed against black officers were also more likely to be sustained, and black officers were more likely to receive at least one sustained allegation. This is consistent with the trend at the punishment stage, although if one believes the strength of the controls, then there should still be cause for concern. However, black and white officers received allegations at similar rates. This is inconsistent with the subsequent evaluation and punishment stages, which calls into question whether organizational biases may be at work.

The gender analyses present a more subtle picture, because if one were to look only at the archival data on punishment, there is no evidence that gender is related to the harshness of punishment. Similarly, there was no evidence that gender influenced whether a specific allegation was sustained, nor whether it influenced the risk of receiving a sustained allegation. However, in the earliest stage of allegation accrual, women received allegations at a lower rate than men. This helps interpret the later stages, where one might expect that if women are receiving fewer allegations than men, then they should also be less likely to receive a sustained allegation. This highlights that the lack of a bias can also be evidence for the existence of a bias if other stages are incorporated. Thus, the multi-stage analysis highlights that interpreting a finding as a lack of bias still requires assumptions about relationships at prior stages. Examining the allegation process allows for this type of inference.

Table 4
Zero-truncated negative binomial regression predicting the total number of allegations accrued at the officer level. Only includes officers appointed before (and allegations filed against them during) the data window to avoid selection biases (i.e., 6,059 officers with 16,986 total allegations). The reference category for officer race is "White," the reference category for officer rank is "PO," and the reference category for birth year is "missing." Standard errors are clustered at the unit level. Alternative models are reported in the supplement (standard Poisson in Table A.13 and zero-truncated Poisson in Table A.14).

	Total number of allegations
Officer gender: M	0.316***
	(0.0456)
Officer race: Asian	−0.0811
	(0.0730)
Officer race: Black	−0.0221
	(0.0572)
Officer race: Hispanic	−0.0755**
	(0.0373)
Officer race: White/Hispanic	−0.554**
	(0.220)
Officer birth year category: <1960	−0.369***
	(0.0651)
Officer birth year category: [1960, 1970)	−0.320***
	(0.0582)
Officer birth year category: [1970, 1980)	−0.198***
	(0.0512)
Officer birth year category: ⩾1980	−0.0324
	(0.0585)
Officer appointment year (mean centered)	0.0236***
	(0.00500)
Officer rank: CMDR	−1.178**
	(0.480)
Officer rank: Cpt	−1.520***
	(0.404)
Officer rank: DET	−0.569***
	(0.194)
Officer rank: ET	−0.133**
	(0.0518)
Officer rank: FTO	0.132
	(0.119)
Officer rank: LT	−0.113
	(0.200)
Officer rank: SGT	0.0537
	(0.0580)
Constant	0.267***
	(0.0580)
lnalpha	−0.334***
	(0.0968)
Officer departmental unit controls	Yes
Observations	6,059

Note: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

7. Discussion

Sample selection biases are a fundamental challenge to successfully building and testing theory about ethics and misconduct. This can result from the standard individual-level selection biases frequently acknowledged in archival research, but also from the organizational processes that create archival samples of wrongdoers. The findings in this study illustrate the specific challenge of intra-organizational selection biases and should signal caution in the ability to use "official" samples of wrongdoers to accurately represent underlying behavior. Greve et al. (2010) noted that "A frequent dilemma in research on misconduct is that data become available when a social-control agent detects misconduct and decides to act against it." This paper presented allegations as a promising approach for addressing this challenge. In the

context of the Chicago Police Department, I highlighted how systematic variance in who makes allegations, who receives allegations, and variance in how allegations are evaluated can be used to better understand how misconduct is ultimately documented.

The approach is useful for understanding the scope of organizational selection biases, even if allegations are themselves an imperfect measure of perceptions of misconduct. For example, Futterman et al. (2007, p. 267) noted from field research that "Only a small fraction of people who believe that they have been abused by the police actually file a complaint with the Chicago Police Department" and cited a national statistic that while 75% of people experiencing force from the police believed it to be excessive, only 10% filed a complaint (Hickman, 2006). I found similar evidence using more recent data from the 2011 Police-Public Contact Survey by the Department of Justice (United States Department of Justice, 2014). Of the 10,196 respondents to the question "Looking back on this contact, do you feel the police behaved properly?" approximately 8% answered "no." Of those that answered "no," however, 95% did not file a complaint. It is likely that many individuals are precluded from being labeled wrongdoers simply because allegations are never filed against them, even if an allegation would have been sustained at some unknown rate if it had been filed. However, as alluded to in earlier sections, the propensity to file an allegation is itself likely a function of beliefs about how an allegation will be evaluated. If people believe there will be a bias in how their allegations are evaluated, they may simply not make them in the first place. This makes intra-organizational selection biases even more important, because they may not only directly introduce biases during documentation but also influence the upstream process and alter what gets alleged in the first place. In this case, the magnitude of the actual importance of organizational selection biases would be even greater.

A second feature of this approach—common to most approaches to studying misconduct—is that it cannot "prove" that organizational selection biases exist; it can only highlight that some type of bias exists. The United States Department of Justice (2017b) wrote that "Our investigation also found that CPD has tolerated racially discriminatory conduct that not only undermines police legitimacy, but also contributes to the pattern or practice of unreasonable force" (p. 143). However, on the specific disparities between complainant race and the outcome of allegations, it noted, "This does not necessarily indicate that the complaint process is biased, as these numbers do not say anything about the quality of the complaint" (p. 69). It is not clear whether they attempted to account for the complaint category and other control variables. However, it is possible that if the generation of allegations is biased by something other than expectations about the evaluation process, then what appear to be organizational selection biases may simply be a correction for this.[17] Regardless, detecting and addressing these disparities is a clear policy concern.

The single empirical context is both a strength and a limitation of this study. On the one hand, the particulars of Chicago make it an ideal setting to explore these relationships, and the focus on a single context potentially allows for greater consistency in how data are recorded and ultimately interpreted. On the other hand, Chicago represents just one setting in one particular context. There is likely limited theoretical reason to believe that organizational biases should be more influenced by race, gender, and occupation than other status characteristics. On the wide range of status characteristics, Berger et al. (1980, p. 479) noted,"[e]xamples include age, sex, race, ethnicity, education, occupation, physical attractiveness, intelligence quotients, reading ability—but there are many others." This context happened to be primary male and white, which may explain these particular results. Researchers should be cognizant of the most salient status characteristics in their particular archival data. The importance of the finding about whether the complainant was a CPD member is likely mirrored in many organizational settings, where the occupational or positional status of individuals may result in organizational selection biases simply because such individuals have more direct or indirect influence within the organization. This highlights a promising direction for future research, expanded on next, related to better understanding the underlying sources of intra-organizational selection biases.

7.1. What causes intra-organizational selection biases?

The empirical focus of the paper was explicating a methodological approach—using allegations—to detect and account for potential organizational biases in archival data. However, this approach cannot explain *why* selection biases exist in the first place. Theory and additional context are still needed to answer that question. In this paper, the dominant explanation drawn from prior academic work (Hagedorn et al., 2013) and an investigation by the Department of Justice (United States Department of Justice, 2017b) was that organizational culture was a plausible explanation. However, if a researcher were able to run the "ideal experiments" outlined earlier in this paper across multiple settings, what would potentially lead to different results in different organizations? Organizational culture and organizational policies are two potential explanations that also have important implications for organization-level research.

Organizational culture. Culture is perhaps the most obvious explanation and directly related to the empirical context of the paper. In the policing context, a "code of silence" is one explanation for why police misconduct may go unpunished (Hagedorn et al., 2013). The United States Department of Justice (2017b, p. 75) noted of Chicago: "We cannot determine the exact contours of this culture of covering up misconduct, nor do we know its precise impact on specific cases. What is clear from our investigation, however, is that a code of silence exists, and officers and community members know it." However, a "blue code of silence" may exist in one police department but not another. Similar variance in cultural factors likely exists in different industries where power dynamics vary (e.g., gender in Hollywood). This creates the problem that a corrupt culture could plausibly simultaneously lead to more bad behavior but less documentation of that behavior. This would result in underestimating the magnitude of culture's importance from archival data, as a corrupt culture may both increase unethical behavior and decrease documentation of that behavior.

Culture may also influence earlier stages of the allegation process. For example, a working paper by Ody-Brasier and Mohliver (2020) studies health violations at nursing homes, where variations in religious affiliation appear related to the amount and nature of officially recorded violations by these organizations. The proposed mechanism is that religious homogeneity influences how organizational members make complaints in the first place. Thus, there are likely multiple channels by which organizational culture can bias archival data.

Organizational policy. Organizations also employ a range of explicit or implicit policies that should influence how data are recorded. Many of these are directly related to reporting and monitoring of behavior. For example, variance in human resource policies defining good and bad behavior determine what can be documented as misconduct in the first place and will clearly influence what gets recorded. A more explicit example is the use of mandatory arbitration agreements in firm-level employment contracts that dictate whether employees are able to take complaints to the courts or must use a private arbitrator for grievances such as sexual harassment. Recent estimates indicate that the majority of non-union private-sector employees in America must sign mandatory arbitration agreements (Colvin, 2018). Yet such agreements "often draw a heavy veil of secrecy around allegations of misconduct and their resolution" (Estlund, 2018, p. 681) by removing disputes from the public courts. Even within the group of firms that use mandatory arbitration, variance in arbitrators may lead to additional biases. Estlund (2018)

[17] For example, it has been reported that the Chicago Police Department believes gang members use false allegations as a tool against effective officers (Williams, 2015).

notes that given the ratio of employees covered by mandatory arbitration agreements, one might expect a greater number of claims to be filed through arbitration than the federal courts. However, in practice it appears that a minuscule number of arbitration claims are actually made. This emphasizes the importance of expectations about bias influencing earlier stages of the process. This lack of transparency is one of the reasons there has been pressure related to the #MeToo movement to place limits on the use of mandatory arbitration agreements (McCullough, 2019). If the use of such agreements is correlated with the existence of misconduct, then archival data will be biased.

More mundane organizational policies also have the potential to bias archival data. In the Chicago police context, Ba (2017) found that the location of a physical office altered who was willing and able to fully submit allegations of misconduct. Allegations with an affidavit were more likely for incidents that occurred closer to the office because of the costs of travel. If geography is correlated with demographics (likely), then the final sample may also be biased simply by where the intake of complaints occurs.

Beyond general organizational policies, the adoption of specific technology may also alter how data are recorded, particularly in cases where technology itself is related to the monitoring of behavior. For example, in the policing context, a recent stream of research has focused on the effects of officer-worn body cameras that record interaction with the public (Ariel et al., 2017; Ariel, Farrar, & Sutherland, 2015). Such technology appears to lower the number of complaints against officers that use it. In many settings, information technology provides greater visibility into how people inside the organization behave but also changes behavior (Pierce, Snow, & McAfee, 2015). Problems will arise if such technology itself is biased. For example, there are concerns related to the impartiality of artificial intelligence and other technology, where "AI carries the serious risk of perpetuating, amplifying, and ultimately ossifying existing social biases and prejudices" (Raso, Hilligoss, Krishnamurthy, Bavitz, & Kim, 2018, p. 18). In sum, researchers interested in the effects of organizational policy on behavior need to account for the possibility that those policies also change how behavior is recorded in non-random ways.

7.2. Implications for organization-level research

This paper examined a single context: misconduct by Chicago police. Despite this, there are clear implications for multi-organization studies that attempt to build and test theory at the organization level. This is because the organization-level traits discussed in the previous section are often related to the organizational traits that have been proposed as direct causes of organizational misconduct itself (Palmer, 2012). These include organizational culture and environmental strain (Greve et al., 2010), as well as organizational structure, processes, and tasks (Vaughan, 1999). If such traits both cause behavior and influence how that behavior is documented, then directly testing theory with archival data will be challenging.

A single third-party who records misconduct may be biased by organizational traits in the same way that individual-level traits may bias evaluation. This may be addressable using a direct extension of the empirical approach of this paper where a third-party organization becomes the focal organization at risk of introducing selection biases. However, a researcher may also aggregate data from multiple organizations into a single dataset. If organizational traits influence the process by which behavior is independently documented by different organizations, the direct comparability across organizations will be difficult without separately testing for biases in each organization.

7.3. Conclusion

The appeal of using archival behavioral field data for research is that it is often readily available and represents "real" outcomes of clear importance. However, there is a risk that such data are biased by the organizational processes that create them. In this paper, I outlined the nature of these intra-organizational selection biases and proposed a way to detect them using data on allegations. In the context of police misconduct in Chicago, I used this approach to present evidence that archival data on officers punished for misconduct are unlikely to perfectly represent the behavior of most interest to researchers. By applying this general approach to other settings, researchers may be able to better understand the extent of intra-organizational selection biases and thus the optimal use of archival field data.

Acknowledgments

I would like to thank the Invisible Institute for fielding questions about the data, Lamar Pierce and the anonymous reviewers for guidance during the revision, as well as audiences at the 2017 Junior Faculty Organizational Theory Conference at Yale University, the 2017 Academy of Management Conference, Imperial College London, and London Business School for feedback on earlier versions of the paper.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.obhdp.2020.03.003.

References

Adut, A. (2005). A theory of scandal: Victorians, homosexuality, and the fall of Oscar Wilde. *American Journal of Sociology, 111*(1), 213–248.
Antonovics, K., & Knight, B. G. (2009). A new look at racial profiling: Evidence from the Boston Police Department. *Review of Economics and Statistics, 91*(1), 163–177.
Anwar, S., & Fang, H. (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review, 96*(1), 127–151.
Ariel, B., Farrar, W. A., & Sutherland, A. (2015). The effect of police body-worn cameras on use of force and citizens' complaints against the police: A randomized controlled trial. *Journal of Quantitative Criminology, 31*(3), 509–535.
Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., & Henderson, R. (2017). "Contagious Accountability": A global multisite randomized controlled trial on the effect of police body-worn cameras on Citizens' complaints against the police. *Criminal Justice and Behavior, 44*(2), 293–316.
Ashforth, B. E., & Anand, V. (2003). The normalization of corruption in organizations. *Research in Organizational Behavior, 25*, 1–52.
Aven, B. L. (2015). The paradox of corrupt networks: An analysis of organizational crime at Enron. *Organization Science, 26*(4), 980–996.
Ba, B. A. (2017). Going the Extra Mile: the Cost of Complaint Filing, Accountability, and Law Enforcement Outcomes in Chicago. Working Paper.
Ben-Yehuda, N. (1980). The European witch craze of the 14th to 17th centuries: A sociologist's perspective. *American Journal of Sociology, 86*(1), 1–31.
Berger, J., Rosenholtz, S. J., & Zelditch, M. (1980). Status organizing processes. *Annual Review of Sociology, 6*, 479–508.
Bowles, H. R., & Gelfand, M. (2010). Status and the evaluation of workplace deviance. *Psychological Science, 21*(1), 49–54.
Cao, L., Deng, X., & Barton, S. (2000). A test of Lundman's organizational product thesis with data on citizen complaints. *Policing: An International Journal of Police Strategies & Management, 23*(3), 356–373.
Chase, G. (2017). The Early History of the Black Lives Matter Movement, and the Implications Thereof. Nev. LJ, 18:1091.
City of Chicago Independent Police Review Authority (2016). Investigative results. Retrieved from http://www.iprachicago.org/results.html.
Cochrane, J. (2020). CPD Complaints; Where do they go? Retrieved from https://data.cpdp.co/static/allegation/img/complaint-flowchart-1.png.
Colvin, A. J. (2018). The growing use of mandatory arbitration: Access to the courts is now barred for more than 60 million American workers. Technical report. Economic Policy Institute. Retrieved from https://www.epi.org/publication/the-growing-use-of-mandatory-arbitration-access-to-the-courts-is-now-barred-for-more-than-60-million-american-workers/.
Davey, M., & Smith, M. (2015). *Mayor Rahm Emanuel Fires Chicago Police Superintendent*. The New York Times. Retrieved from https://www.nytimes.com/2015/12/02/us/chicago-police-rahm-emanuel-laquan-mcdonald.html (Dec. 1, 2015).
Doyle, B., & Gorner, J. (2012). *Chicago police districts close in cost-cutting plan*. Chicago Tribune. Retrieved from http://articles.chicagotribune.com/2012-03-03/news/chi-chicago-police-districts-close-in-costcutting-plan-20120303_1_third-station-detectives-chicago-police (March 3, 2012).
Estlund, C. (2018). The black hole of mandatory arbitration. *North Carolina Law Review, 96*(3), 679.
Faulkner, R. R. (2011). *Corporate wrongdoing and the art of the accusation*. Anthem Press. ISBN 978-0-85728-794-6.

Fielding, L. (2016). U.S. Attorney: Fed Probe Into Chicago Police Department Is Biggest Ever, Ongoing. CBS Chicago. Retrieved from http://chicago.cbslocal.com/2016/09/26/u-s-attorney-fed-probe-into-chicago-police-department-is-biggest-ever-ongoing/.
Fingerhut, H. (2016). Is treatment of minorities a key election issue? Views differ by race, party. Pew Research Center. Retrieved from https://www.pewresearch.org/fact-tank/2016/07/13/partisan-racial-divides-exist-over-how-important-treatment-of-minorities-is-as-a-voting-issue/.
Fisman, R., & Miguel, E. (2007). Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets. *Journal of Political Economy, 115*(6), 1020–1048.
Fragale, A. R., Rosen, B., Xu, C., & Merideth, I. (2009). The higher they are, the harder they fall: The effects of wrongdoer status on observer punishment recommendations and intentionality attributions. *Organizational Behavior and Human Decision Processes, 108*(1), 53–65.
Futterman, C. B., Mather, H. M., & Miles, M. (2007). Use of statistical evidence to address police supervisory and disciplinary practices: The Chicago Police Department's Broken System. *DePaul Journal for Social Justice, 1*, 251.
Greve, H. R., Palmer, D., & Pozner, J.-E. (2010). Organizations gone wild: The causes, processes, and consequences of organizational misconduct. *The Academy of Management Annals, 4*(1), 53–107.
Griswold, D. B. (1994). Complaints against the police: Predicting dispositions. *Journal of Criminal Justice, 22*(3), 215–221.
Grogger, J., & Ridgeway, G. (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association, 101*(475), 878–887.
Hagedorn, J., Kmiecik, B., Simpson, D., Gradel, T. J., Zmuda, M. M., & Sterrett, D. (2013). Crime, Corruption and Cover-ups in the Chicago Police Department. Anti-Corruption Report 7. University of Illinois at Chicago Department of Political Science. Retrieved from https://pols.uic.edu/chicago-politics/anti-corruption-reports/.
Harris, C. J. (2011). The relationship between career pathways of internal and citizen complaints. *Police Quarterly, 14*(2), 142–165.
Hassell, K. D., & Archbold, C. A. (2010). Widening the scope on complaints of police misconduct. *Policing: An International Journal of Police Strategies & Management, 33*(3), 473–489.
Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83.
Hickman, M. J. (2006). Citizen complaints about police use of force. Technical report. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. Retrieved from https://www.ncjrs.gov/App/AbstractDB/AbstractDBDetails.aspx?id=210296.
Hickman, M. J., Piquero, A. R., & Greene, J. R. (2000). Does community policing generate greater numbers and different types of citizen complaints than traditional policing? *Police Quarterly, 3*(1), 70–84.
Hickman, M. J., & Poore, J. E. (2015). National data on citizen complaints about police use of force data quality concerns and the potential (Mis) use of statistical evidence to address police agency conduct. *Criminal Justice Policy Review*, 1–25.
Horowitz, J. M., & Livingston, G. (2016). How Americans view the Black Lives Matter movement. Pew Research Center. Retrieved from https://www.pewresearch.org/fact-tank/2016/07/08/how-americans-view-the-black-lives-matter-movement/.
Johnson, D. J., & Cesario, J. (2020). Reply to Knox and Mummolo and Schimmack and Carlsson: Controlling for crime and population rates. *Proceedings of the National Academy of Sciences, 117*(3), 1264–1265.
Johnson, D. J., Tress, T., Burkel, N., Taylor, C., & Cesario, J. (2019). Officer characteristics and racial disparities in fatal officer-involved shootings. *Proceedings of the National Academy of Sciences, 116*(32), 15877–15882.
Kakkar, H., Sivanathan, N., & Gobel, M. (2020). Fall from grace: The role of dominance and prestige in the punishment of high-status actors. *Academy of Management Journal, 63*(2), 530–553.
Kalven v. City of Chicago (2014). Kalven v. City of Chicago. Technical Report 1–12-1846, 1–12-1917 cons., Ill: Appellate Court. Retrieved from http://www.illinoiscourts.gov/opinions/appellatecourt/2014/1stdistrict/1121846.pdf.
Knowles, J., Persico, N., & Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy, 109*(1), 203–229.
Knox, D., & Mummolo, J. (2020). Making inferences about racial disparities in police violence. *Proceedings of the National Academy of Sciences, 117*(3), 1261–1262.
Lersch, K. M., & Kunzman, L. L. (2001). Misconduct allegations and higher education in a southern sheriff's department. *American Journal of Criminal Justice: AJCJ, 25*(2), 161–172.
Lersch, K. M., & Mieczkowski, T. (1996). Who are the problem-prone officers? An analysis of citizen complaints. *American Journal of Police, 15*(3), 23–44.
Lersch, K. M., & Mieczkowski, T. (2000). An examination of the convergence and divergence of internal and external allegations of misconduct filed against police officers. *Policing: An International Journal of Police Strategies & Management, 23*(1), 54–68.
Levitt, S. D., & List, J. A. (2008). Homo economicus Evolves. *Science, 319*(5865), 909–910.
Liederbach, J., Boyd, L. M., Taylor, R. W., & Kawucha, S. K. (2007). Is it an inside job? An examination of internal affairs complaint investigation files and the production of nonsustained findings. *Criminal Justice Policy Review, 18*(4), 353–377.
McCullough, K. (2019). Mandatory arbitration and sexual harassment claims: #MeToo-and time's up-inspired action against the federal arbitration act. *Fordham Law Review, 87*(6), 2653.
McDonnell, M.-H., & King, B. G. (2018). Order in the Court: How firm status and reputation shape the outcomes of employment discrimination suits. *American Sociological Review, 83*(1), 61–87.
Milkman, R. (2017). A new political generation: Millennials and the post-2008 wave of protest. *American Sociological Review, 82*(1), 1–31.
Mishina, Y., Dykes, B. J., Block, E. S., & Pollock, T. G. (2010). Why "Good" firms do bad things: The effects of high aspirations, high expectations, and prominence on the incidence of corporate illegality. *Academy of Management Journal, 53*(4), 701–722.
Ody-Brasier, A., & Mohliver, A. (2020). How Religious Affiliation Affects Organizational Wrongdoing. Working Paper.
Palmer, D. A. (2012). *Normal organizational wrongdoing: A critical analysis of theories of misconduct in and by organizations* (1st ed.). Oxford University Press. ISBN 978-0-19-957359-2.
Palmer, D., & Yenkey, C. B. (2015). Drugs, sweat, and gears: An organizational analysis of performance-enhancing drug use in the 2010 Tour de France. *Social Forces, 94*(2), 891–922.
Peirce, E., Smolinski, C. A., & Rosen, B. (1998). Why sexual harassment complaints fall on deaf ears. *Academy of Management Executive, 12*(3), 41–54.
Phillips, D. J., & Zuckerman, E. W. (2001). Middle-status conformity: Theoretical restatement and empirical demonstration in two markets. *The American Journal of Sociology, 107*(2), 379–429.
Pierce, L., & Balasubramanian, P. (2015). Behavioral field evidence on psychological and social factors in dishonesty and misconduct. *Current Opinion in Psychology, 6*, 70–76.
Pierce, L., Snow, D. C., & McAfee, A. (2015). Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Management Science, 61*(10), 2299–2319.
Polman, E., Pettit, N. C., & Wiesenfeld, B. M. (2013). Effects of wrongdoer status on moral licensing. *Journal of Experimental Social Psychology, 49*(4), 614–623.
Raso, F. A., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks. SSRN Scholarly Paper ID 3259344, Social Science Research Network, Rochester, NY. doi: 10.2139/ssrn.3259344.
Reaves, B. A., (2015). Local Police Departments, 2013: Personnel, Policies, and Practices. Technical report. Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice. Retrieved from http://www.bjs.gov/content/pub/pdf/lpd13ppp.pdf.
Rosnow, R. L., & Foster, E. K. (2005). Rumor and gossip research. *Psychological Science Agenda, 19*(4).
Rozema, K., & Schanzenbach, M. (2019). Good cop, bad cop: Using civilian allegations to predict police misconduct. *American Economic Journal: Economic Policy, 11*(2), 225–268.
Sharkey, A. J. (2014). Categories and organizational status: The role of industry status in the response to organizational deviance. *American Journal of Sociology, 119*(5), 1380–1433.
Shaver, J. M. (1998). Accounting for endogeneity when assessing strategy performance: does entry mode choice affect FDI survival? *Management Science, 44*(4), 571–585.
Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences, 109*(38), 15197–15200.
Simcoe, T. S., & Waguespack, D. M. (2010). Status, quality, and attention: What's in a (missing) name? *Management Science, 57*(2), 274–290.
Simpson, D., Gradel, T. J., Mouritsen, M., & Johnson, J. (2015). Chicago: Still the Capital of Corruption. Anti-Corruption Report 8. University of Illinois at Chicago Department of Political Science. Retrieved from https://pols.uic.edu/chicago-politics/anti-corruption-reports/.
Smith, M., Williams, T., & Davey, M. (2018). *'Justice for Laquan!' Demonstrators Chant, as Chicago Officer Is Convicted of Murder*. The New York Times. Retrieved from https://www.nytimes.com/2018/10/05/us/van-dyke-guilty-laquan-mcdonald.html (October 5, 2018).
Terrill, W., & Ingram, J. R. (2015). Citizen complaints against the police: An eight city examination. *Police Quarterly*, 1–30.
United States Department of Justice (2014). Police-Public Contact Survey, 2011. Office of Justice Programs. Bureau of Justice Statistics, ICPSR - Interuniversity Consortium for Political and Social Research. doi: 10.3886/ICPSR34276.v1.
United States Department of Justice (2017a). Fact Sheet: The Department of Justice Pattern or Practice Investigation of the Chicago Police Department. Technical report, Civil Rights Division and United States Attorney's Office Northern District of Illinois. Retrieved from https://www.justice.gov/opa/pr/justice-department-announces-findings-investigation-chicago-police-department.
United States Department of Justice (2017b). Investigation of the Chicago Police Department. Technical report, Civil Rights Division and United States Attorney's Office Northern District of Illinois. Retrieved from https://www.justice.gov/opa/pr/justice-department-announces-findings-investigation-chicago-police-department.
Vaughan, D. (1999). The dark side of organizations: Mistake, misconduct, and disaster. *Annual Review of Sociology, 25*, 271–305.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge Univ Pr.

Weitzer, R., & Tuch, S. A. (2004). Race and perceptions of police misconduct. *Social Problems, 51*(3), 305–325.

Wildeboer, R. (2015). *Complaints against Chicago cops published after 20-year saga*. Chicago: WBEZ News. Retrieved from http://www.wbez.org/news/complaints-against-chicago-cops-published-after-20-year-saga-113715 (November 9, 2015).

Williams, T. (2015). *Chicago Rarely Penalizes Officers for Complaints, Data Shows*. The New York Times. Retrieved from http://www.nytimes.com/2015/11/19/us/few-complaints-against-chicago-police-result-in-discipline-data-shows.html (November 18, 2015).

Worrall, J. L. (2002). If You Build It, They Will Come: Consequences of Improved Citizen Complaint Review Procedures. *Crime & Delinquency, 48*(3), 355–379.

Zitzewitz, E. (2012). Forensic economics. *Journal of Economic Literature, 50*(3), 731–769.